University of Illinois
Fall 2016 | Department of Economics
Roger Koenker

University of Illinois
Fall 2016

Department of Economics
Roger Koenker

Economics 536
Lecture 24
Treatment Effects, Matching and Propensity Scores

In randomized experiments we are assured that the treatment indicator, say D_i, is assigned independently of both the potential responses, $D_i \perp\!\!\!\perp (Y_{0i}, Y_{1i})$ and any observed covariates, $D_i \perp\!\!\!\perp X_i$. This is obviously advantageous since it permits us to estimate the causal effect of the treatment,

$$\mathbb{E}(Y_{1i} - Y_{0i}) = \alpha + \beta D_i.$$

Here, we adopt the now standard Rubin potential outcomes formulation with Y_{1i} denoting the response of subject i under the treatment, $D_i = 1$, while Y_{0i} denotes the possibly counter factual response of subject i under the control regime, $D_i = 0$. Inevitably, we cannot observe both Y_{1i} and Y_{0i}, instead we observe

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

However, in the idealized framework of the randomized experiment we can obtain consistent estimates of the causal effect of treatment, β, by simply regressing Y_i on D_i, i.e. by simply computing the difference in mean response for those in the treatment and control groups, $\hat{\beta} = \bar{Y}_1 - \bar{Y}_0$. Typically, the observed responses are a change in the level of something, wages, blood pressure, etc. in response to the treatment so $\hat{\beta}$ is really $\Delta\bar{Y}_1 - \Delta\bar{Y}_0$. Thus the treatment effect is frequently referred to as a difference in differences, "diff-in-diff". Recall, for example. the Lanarkshire milk experiment, Student (1931).

1. Matching

We are, however, rarely in the ideal world of randomized experiments with perfect compliance. The next best setting seems to be one in which treatment assignment is *conditioned on observables*, so $D_i \perp\!\!\!\perp (Y_{0i}, Y_{1i}) | X_i$. This situation is illustrated in PS 5 where assignment to the treatment depends upon the observed covariates, sex, dex, lex, and we can see that employees who scored poorly on the preliminary dexterity exam were more likely to be assigned to the subsequent training program. It is important to emphasize that this need not be a deterministic assignment scheme, but the random component of treatment assignment must be independent of the experimental response.

When treatment is assigned conditional on observables an attractive estmation strategy can be formulated by simply applying the fully randomized strategy described above for each distinct setting of the covariates, so we can estimate a covariate specific treatment effect,

$$\hat{\beta}(x) = \bar{Y}_1(x) - \bar{Y}_0(x),$$

where, now, we have a mean treatment effect for each $X = x$. Of course, for continuously distributed covariates we would have to make some provision to "bin" neighboring x's resulting in some degree of ambiguity in the resulting estimator. More seriously, we may encounter covariate settings for which either $\bar{Y}_1(x)$ or $\bar{Y}_0(x)$ cannot be computed because all the subjects with $x_i = 1$ are either in the treatment, or control, group. This is generally referred to as a failure of "overlap" in the support of experimental design. Unless we are able to estimate a treatment effect at a given x, there is no way to know how this region of the design space $\mathcal{X}$ contributes to the overall treatment effect. In the fortunate circumstance that we are able to compute $\hat{\beta}(x)$ over all of $\mathcal{X}$, we can then integrate,

$$\tilde{\beta} = \int_{\mathcal{X}} \hat{\beta}(x) d\mu(x),$$

to obtain an average treatment effect unconditioned on covariates. Here, μ denotes some form of weighting that accounts for the relative precision of the $\hat{\beta}(x)$ estimates. In the simplest case this is just the relative sample sizes in the x-cells. Restricting the domain $\mathcal{X}$ to the region of effective overlap is fine, and a major virtue of the "matching on observables" procedures is that it forces us to be aware of these limitations.

A more automatic, and therefore potentially more dangerous approach, is to simply regress our response on observed treatment and the conditioning covariates, X. Under our treatment assignment assumption, this will yield consistent estimates of the treatment effect. However, as usual, it depends crucially on functional form assumptions and in particular on the assumption that $\beta(x)$ is constant over the design space $\mathcal{X}$. Matching provides clear intermediate evidence on this.

2. Propensity Scores

The downside of matching on the full vector of covariates is often referred to as the "curse of dimensionality." We are essentially required to solve a high dimensional nonparametric regression problem. The literature abounds in claims to annul this curse. Such claims should be viewed with scepticism, as in Verdi – once cursed, cursed forever. Nevertheless, it is worthwhile to consider some fashionable dimension reduction devices and the assumptions under which they offer some respite from this curse. Chief among these devices is the propensity score.

Rosenbaum and Rubin (1983) introduced propensity score methods for binary treatment models. Suppose we have a binary treatment that is assigned conditional on observable covariates, X, so as above, $D_i \perp\!\!\!\perp (Y_{0i}, Y_{1i}) | X_i$. Let $p(X) = \mathbb{P}(D = 1|X)$, denote the propensity score, the probability of being assigned to treatment given X, then we have the following result from Rosenbaum and Rubin.

Theorem 1. $D_i \perp\!\!\!\perp (Y_{0i}, Y_{1i})|X_i$ *implies* $D_i \perp\!\!\!\perp (Y_{0i}, Y_{1i})|p(X_i)$.

Proof. It suffices to show that $\mathbb{P}(D_i = 1|(Y_{0i}, Y_{1i}), p(X_i)) = \mathbb{P}(D_i = 1|p(X_i))$, which follows by the following nested conditioning argument:

$$\begin{aligned}
\mathbb{P}(D_i = 1|(Y_{0i}, Y_{1i}), p(X_i)) &= \mathbb{E}(D_i = 1|(Y_{0i}, Y_{1i}), p(X_i)) \\
&= \mathbb{E}(\mathbb{E}(D_i|(Y_{0i}, Y_{1i}), p(X_i), X_i)|(Y_{0i}, Y_{1i}), p(X_i)) \\
&= \mathbb{E}(\mathbb{E}(D_i|(Y_{0i}, Y_{1i}), X_i)|(Y_{0i}, Y_{1i}), p(X_i)) \\
&= \mathbb{E}(\mathbb{E}(D_i|X_i)|(Y_{0i}, Y_{1i}), p(X_i)) \\
&= \mathbb{E}(p(X_i)|(Y_{0i}, Y_{1i}), p(X_i)) \\
&= p(X_i)
\end{aligned}$$

□

Recall that in ordinary regression if we wanted $\mathbb{E}(Y|D, X)$ and $D \perp X$ we could ignore X and simply estimate $\mathbb{E}(Y|D)$. Here we need simply to "control for" $p(X_i)$. We are interested in estimating the mean treatment effect,

$$\mathbb{E}((Y_{0i}, Y_{1i}), X_i|D_i) = \mathbb{E}(\mathbb{E}(Y_i|p(X_i)), D_i = 1) - (\mathbb{E}((Y_i|p(X_i)), D_i = 0)|D_i = 1)$$

This can be reformulated somewhat by noting that,

$$\mathbb{E}\frac{Y_i D_i}{p(X_i)} = \mathbb{E}Y_{1i} \quad \text{and} \quad \mathbb{E}\frac{Y_i(1 - D_i)}{1 - p(X_i)} = \mathbb{E}Y_{0i}$$

hence

$$\mathbb{E}(Y_{0i} - Y_{1i}) = \mathbb{E}\frac{Y_i(D_i - p(X_i))}{p(X_i)(1 - p(X_i))} \tag{1}$$

In this form we can now easily imagine constructing an estimation procedure that first estimates $p(X_i)$ and then computes sample averages based on (1). This approach is closely related to the Horvitz and Thompson (1952) estimator intended to estimate means based on sample survey data.

2.1. Digression on the Horvitz Thompson estimator.

In sample surveys a common problem is estimating a mean or a total from a sample on a finite population. Suppose we denote Y_i as the response of the ith observation and D_i as the indicator of whether the ith observation was sampled. Typically, we have an explicit sampling plan so we know $p_i = P(D_i = 1)$, so the HT estimator of the total is simply

$$T_n = \sum_{i=1}^{n} D_i Y_i / p_i.$$

Taking expectations conditionally shows that this gives an unbiased estimate of the total. In fact it can be shown to be the UMVUE. To see the connection with $\hat{\beta}$ above suppose we want to estimate averages now, and we think of

D_i as a treatment rather than a sampling decision, then the same argument shows that $ED_iY_i/p_i = EY_{1i}$ and also that $E(1 - D_i)Y_i/(1 - p_i) = EY_{0i}$, and combining these facts we have our expression for $\hat{\beta}$.

There is a nice cautionary tale due to Basu (1971) about the HT estimator: A circus impresario has to ship his 50 elephants and needs a rough estimate of their total weight. He speaks with his trainer who suggests weighing Sambo the elephant who is roughly middle size and then multiplying Sambo's weight by 50. But the circus statistician intervenes and says "No, we need a sampling plan." So he proposes chosing Sambo with probability 99/100 and assigning the rest of the probability uniformly to the rest of the elephants. So the impresario agrees and they randomly draw a $U[0, 1]$ which turns out to be less than .99 and so they weigh Sambo. The impresario is about to multiply by 50 when the statistician says "No, stop, the Horvitz Thompson estimator is known to be UMVUE and it tells us to divide Sambo's weight by $p_i = 99/100$, so we want to multiply by 100/99 not 50. Moral: Unbiasedness isn't always such a great property.

Commentary: Note that if, by chance, we would have chosen Jumbo the biggest elephant whose p_i was only 1/5000, we would have multiplied his weight by 5000, which would have compensated for the drastic underestimate obtained with Sambo. But of course we are only doing this procedure once, so the fact that repeatedly doing it gives us something unbiased is not much comfort. Each time we would be getting something quite stupid for an answer.

2.2. **Propensity Score Matching.** The propensity score appears to be an attractive way to reduce dependence of the treatment assignment on covariates to a convenient scalar quantity. This is a bit misleading since we still need to estimate $p(x)$ and there is usually little guidance as to how to do this. In practice, one is usually left with the familiar binary response models. In Yoon and Koenker (2009) we argue that it is worthwhile to consider alternatives to the usual logit and probit link functions. Given an estimated propensity score, $\hat{p}(x)$, there are a variety of treatment effect estimation options, a general class considered by Hirano, Imbens and Ridder (2003) takes the form

$$\mathbb{E}\left\{ g(X_i) \left[\frac{Y_i(D_i - p(X_i))}{p(X_i)(1 - p(X_i))} \right] \right\},$$

where $g(X_i)$ is a weight function designed to focus on particular groups. For example, $g(X_i) \equiv 1$ gets us back to the HT estimate of the mean treatment effect, and $g(X_i) = p(X_i)/\mathbb{P}(D_i = 1)$ yields the average treatment effect on the treated. For further discussion see Angrist and Pischke (2009).

2.3. **Imperfect Compliance.** In many randomized experiments it is impossible to compel subject to undergo the treatment. Thus, subjects are randomized into a group that is offered treatment and a group that is not. Generally, we can verify that those not offered treatment do not find a way to participate in the treatment, but the worry is that the decision to undergo treatment creates an endogeneity bias that requires attention. Fortunately, in such cases we have a natural instrument for the observed treatment variable, D_i, and that is the intent-to-treat variable, say Z_i. Since Z_i can be randomized, perhaps conditioning on other observable covariates as in the foregoing discussion, it serves a plausible instrument as long as compliance with the treatment offer isn't negligible.

In the simplest settings without other covariates this brings us back to the Ur-IV estimator proposed by Wald, see L11. With other covariates we have a binary endogenous variable that requires some care. There is a temptation to try to simply apply 2SLS ideas: estimate a probit specification for D_i as a function of Z_i and other covariates, X_i, construct $\hat{p}(Z_i, X_i$ for each observation, and now estimate the structural model replacing D_i by $\hat{p}(X_i)$. You could do this, *but it would be wrong.*[1] The difficulty with this approach is that the nonlinearity of the $\hat{p}(x)$ function violates the underlying orthogonal projection objective of 2SLS. Instead, what can and should be done is simply to estimate the structural model by instrumental variables using $\hat{p}$ as an instrument for D_i. For more details see the extensive discussion in Angrist and Pischke (2009).

3. Ecological Regression

The phrase "ecological regression" refers to the common desire to use standard regression methods to infer individual behavior from spatial aggregated data. The most typical situation would seem to be the attempt to infer ethnic voting behavior from precinct level data in political science. Thus, for example we might like to know what proportion of Hispanics voted for Obama in the 2012 election.

3.1. **Goodman's Regression.**

Given data on the proportion of Hispanics, x_i and the proportion of the vote going to Obama, y_i, in a large number of precincts, we are tempted to estimate the ecological regression,

$$y_i = a + bx_i + u_i.$$

If it were reasonable to assume that the coefficients a and b were constant over precincts and u_i was well-behaved, then we could interpret $\hat{a}$ as an estimate of the proportion voting for Obama among non-Hispanics, and $a + b$ as the proportion of Hispanics voting for Obama, corresponding to the extreme cases of $x_i = 0$ and $x_i = 1$, respectively. This is sometimes described as Goodman's (1953) regression. Note that a potentially embarrassing drawback of this approach is the possibility that we could end up with estimates of the two parameters a and b that fall outside the interval $[0, 1]$. Various

[1] During the Watergate scandal, Nixon was quoted by H.R. Haldeman his chief of staff as saying: "There is no problem in raising a million dollars [to keep the Watergate burglers quiet] – we can do that – but it would be wrong." This phrase has entered into the econometrics folklore with some encouragement from Joel Horowitz.

refinements are possible, most obviously a weighting by the size of the regions, but there is nothing to ensure that such refinements are going to help make the estimates more accurate.

4. Method of Bounds

An alternative is the "method of bounds" introduced by Duncan and Davis (1953). Freedman (1999) illustrates this approach with an example using CPI data from Washington state. Suppose we know that 0.079 of the population is foriegn born and 0.344 of the population have "high income." We are interested in the proportion, p of the foreign born who have "high income." We know that

$$0.344 = 0.079p + (1 - 0.079)q$$

where q denotes the proportion of the native born with high income. This reveals the essential problem: we have only one equation to determine two unknowns. But all is not lost, suppose we solve for q in terms of p and then observe that p must be between zero and one. This implies that $q \in [0.288, 0.374]$. Try it! Manski (2007) elaborates this idea in many other contexts. Manski argues that many problems in econometrics have this form, that models are inherently underidentified, but some bounds can be placed on parameter estimates based on careful analysis of the probability structure of the problem. Another example of this sort of analysis is the case of regression data in which we observe intervals $y_i \in [\underline{y_i}, \overline{y_i}]$ and we would like to make inferences about the standard regression model. The challenge in all such models is to carefully specify the probability structure of the model, and when the identified set in non-unique to find a practical way to make inferences about these sets.

5. Random Coefficients

Now, suppose that we have many observations on (x_i, y_i) as above and we would like to consider the random coefficient model,

$$y_i = p_i x_i + q_i(1 - x_i).$$

This is obviously a generalization of the Goodman model. King (1997) in an influential (and controversial) book on the subject assumes that (p_i, q_i) are drawn iid-ly from a bivariate normal distribution truncated to respect the requirement that they should lie in $[0, 1]^2$. This model has a reasonably tractable likelihood and can be therefore estimated by maximum likelihood. Freedman compares three methods of estimating (p_i, q_i): the Goodman regression, the King regression, and a simple model that he calls the neighborhood model that assumes that outcomes are determined by geography not demography. In his formulation, the neighborhood model assumes that $p_i = q_i = y_i$ in each region. This is obviously quite extreme, but in Freedman's example, where we know the correct answer thanks to the crosstabs

Method	p	q
Truth	0.35	0.28
Nbd	0.34	0.36
Goodman	0.29	0.85
King	0.30	0.72

provided by the CPI, the neighborhood approach is better than the others in estimating the mean of the p's and q's. I reproduce his table here.

There is an interesting connection of the King method to medical imaging called tomography. In tomography a 3d image is reconstructed from many 2d slices. In King's tomography plot, each point (x_i, y_i) appears as a line in the parameter space of (p, q)'s. Pairs of these lines have intersections that can be taken as meta-observations to which we try to fit the normal model. Of course, there doesn't seem to be any compelling reason to think that the normal model is appropriate. So it would seem prudent to explore other less parametric approaches. One such approach is the Kiefer-Wolfowitz (1956) nonparametric MLE, which would replace the normal model with a discrete mixing distribution with a relatively small number of mass points. If you were very lucky these mass points might yield an interpretable clustering of the regions.

References

Angrist, J. and J.-S Pischke, (2009) Mostly Harmless Econometrics, Princeton.

Basu, D. (1971) An Essay on the Logical Foundations of Survey Sampling, in Foundations of Statistical Inference, eds V.P. Godambe and D.A. Sprott, Holt Rinehart and Winston.

Freedman, D.A. (1999) Ecological Inference and the Ecological Fallacy, in *International Encyclopedia of the Social & Behavioral Sciences*, available as `http://statistics.berkeley.edu/tech-reports/549.pdf`

Goodman, L. 1953 Ecological regression and the behavior of individuals. *American Sociological Review* 18: 66364

Duncan, O. D, Davis B 1953 An alternative to ecological correlation. *American Sociological Review* 18: 66566

Hirano, K., G. Imbens, and G. Ridder (2003) Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, Econometrica, 71, 1161-89.

Kiefer, J. and J. Wolfowitz (1956) Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters, The Annals of Mathematical Statistics, 27, 887906.

Horvitz, D.G. and D.J. Thompson, (1952) A Generalization of of Sampling without Replacement from a Finite Population, JASA, 47, 663-685.

King, G. 1997 *A Solution to the Ecological Inference Problem.* Princeton University Press

Rosenbaum, P. and D. Rubin (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika, 70, 451-55.

Student, (1931) The Lanarckshire Milk Experiment, Biometrika, 23, 398-406.

Yoon, J. and R. Koenker (2009) Parametric Links for Binary Response Models, J. of Econometrics, 152, 120-130 .

Manski, C. (2007) *Identification for Prediction and Decision*, Harvard U. Press