Summary of Chapter 12 of Imbens and Rubin, "Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction"

James McInerney

August 20, 2015

1 Running Examples

- From the book: assessing impact of a jobs program on income
- My example: does sun screen cause skin cancer[1]? (Based on an article I read today `http://www.telegraph.co.uk/beauty/skin/is-sunscreen-toxic/`)

2 Summary

12.1 Recap: consider N units, with assignments W for all units to treatments (assumed binary for simplicity), covariates X, and outcomes Y. There are three key assumptions we typically make in causal analysis,

- *Individualistic*, meaning $p(w_i \mid X, Y) = p(w_i \mid x_i, y_i), \quad \text{for } i = 1, \ldots, N$
- *Probabilistic*, meaning $0 < p(w_i = k \mid X, Y) < 1, \quad \text{for } i = 1, \ldots, N, \text{ for } k = 0, 1$
- *Unconfounded*, meaning $p(W \mid X, Y) = p(W \mid X)$.

Under these assumptions, we know that units with identical covariates have the same probability of being assigned treatment. Therefore, subpopulations of units with the same covariate values can be interpreted as separate completely randomized experiments. This has limited use in practice because there are often too many possible covariate values (resulting in too few units per value). However, the intuition carries over to propensity and balancing scores.

[1] That was not a typo; increased use of sunscreen has been associated with increased risk of skin cancer.

12.2.1 Using the three assumptions from Section 12.1, we can write the assignment mechanism as,

$$p(W \mid X, Y) = c \prod_{i}^{N} e(x_i)^{w_i} (1 - e(x_i))^{1-w_i} \tag{1}$$

where we introduced $e(x)$, the *propensity score*.

Example on p.259: when we use the three key assumptions in a study of men and women, we can treat the treatment effects on men and women as two separate completely randomized experiments. But stratification is rarely useful in practice.

12.2.2 Super-population perspective: it is helpful to view covariates as being drawn from an approximately continuous distribution.

Another way of stating unconfoundedness is $w_i \perp\!\!\!\perp y_i \mid x_i$.

12.2.3 There is nothing in the data that can tell us whether or not unconfoundedness holds. We must use outside information (e.g., scientific insight).

Super-population unconfoundedness: $p(y_i^{\text{missing}} \mid w_i = w, x_i) = p(y_i^{\text{missing}} \mid w_i = 1 - w, x_i)$.

This makes it clear why unconfoundedness can't be tested from the data: no dataset will allow us to infer the distribution of y^{missing} given w_i and x_i (without begging the question of unconfoundedness).

12.2.4 Out of the three key assumptions, why do we worry about unconfoundedness the most?

1. If any assignments are non-probabilistic then we could/would have to ignore the strata that do not follow it.
2. Non-individualistic assignment rarely occurs in practice, outside of sequential assignment mechanisms. (And in any case, for non-sequential data, we could often change the definition of a unit such that individualistic assignment does hold).
3. Most causal studies rely on unconfoundedness. And it's not always made explicit, leading to stronger assumptions than may be necessary. For example, methods that use linear regression to estimate causal effects talk about the *exogeneity* of the noise on y_i^{observed}, which specifies a functional form of unconfoundedness and constant treatment effect.
4. Because unconfoundedness encodes the intuition that we should compare "like with like" in causal studies. (In other words, it's implausible that "individuals who differ in terms of pre-treatment characteristics would be more suitable comparisons").
5. Because ensuring that the assignments are unconfounded forces us to think more clearly about the assignment mechanism (e.g., patients with differing health insurance plans who are otherwise identical).

12.2.5 One should avoid including covariates that might be affected by the treatment (e.g., applying sun screen might give you a false sense of security leading you to spend more time in the sun). Otherwise, in general, we want to control for all pre-treatment variables. One exception is when we have instrumental variables, we do not want to include all pre-treatment variables.

12.3.1 Balancing Scores and the Propensity Score Without further ado, we move on to using the three key assumptions to remove biases between treated and control units. As previously mentioned, a large number of covariates makes this hard, so we next try to remove bias by finding a low dimensional *balancing score* $b(x)$ with the following property,

$$w_i \perp\!\!\!\perp x_i \,|\, b(x_i). \tag{2}$$

x_i is an example of a balancing score. But we prefer low dimensional balancing scores, e.g., the propensity score $e(x_i)$.

Why is $e(x_i)$ a balancing score?

Partly because $p(w_i \,|\, e(x_i)) = \mathbb{E}[w_i \,|\, e(x_i)] = \mathbb{E}[\mathbb{E}[w_i \,|\, x_i, e(x_i)] \,|\, e(x_i)] = \mathbb{E}[e(x_i) \,|\, e(x_i)] = e(x_i)$. (There is also another part to proof).

Two properties of balancing scores are,

1. $w_i \perp\!\!\!\perp y_i \,|\, b(x_i)$. See p. 267 for proof (it uses iterated expectations as well).
2. $e(x_i)$ is the *coarsest* balancing score, i.e., it is a function of every balancing score. Proof by contradiction on p. 268: if this does not hold, then we violate the definition of a balancing score.

12.4.1 Estimation and Inference Let the super-population average treatment effect be,

$$\begin{aligned} \tau_{\text{sp}} &= \mathbb{E}_{\text{sp}}[y_i(1) - y_i(0)] = \mathbb{E}_{\text{sp}}[\tau_{\text{sp}}(x_i)] \\ &\quad \text{where } \tau_{\text{sp}}(x) = \mathbb{E}_{\text{sp}}[y_i(1) - y_i(0) \,|\, x_i = x] \end{aligned} \tag{3}$$

This is distinct from τ_{fs}, the finite-sample average treatment effect, and τ_{cond}, the conditional average treatment effect.

Using unconfoundedness and probabilistic assignment (and smoothness [in what?]), a bound on the variance of the estimator τ_{fs} can be calculated using the *semiparametric efficiency bound*,

$$\text{Var}_{\text{sp}}^{\text{eff}} = \mathbb{E}_{\text{sp}}[\frac{\sigma_{\text{c}}^2(x_i)}{1 - e(x_i)} + \frac{\sigma_{\text{t}}^2(x_i)}{e(x_i)} + (\tau_{\text{sp}}(x_i) - \tau_{\text{sp}})^2], \tag{4}$$

where $\sigma_{\text{c}}^2(x_i)$ and $\sigma_{\text{t}}^2(x_i)$ are the super-population control and treatment variances conditioned on x_i, respectively.

The intuition for Eq. 4 is as follows. The first two terms say extreme assignment probabilities (i.e., close to 0 or 1) for the covariates present in the data result in higher variance estimates of the treatment effect. The third term is the "variance of the treatment effect conditional on the pretreatment variables".

The bound on the variance of the conditional treatment effect is,

$$\mathrm{Var}_{\mathrm{cond}}^{\mathrm{eff}} = \mathbb{E}_{\mathrm{sp}}[\frac{\sigma_{\mathrm{c}}^2(x_i)}{1-e(x_i)} + \frac{\sigma_{\mathrm{t}}^2(x_i)}{e(x_i)}], \tag{5}$$

Thus, we can, "in principle, estimate τ_{cond} more accurately than τ_{sp}". This is because Eq. 5 does not have to worry about the difference in covariate distribution between the sample and the population.

12.4.2 There are five kinds of methods for estimating treatment effects, all of which use the three key assumptions.

1. Model-based imputation, in which we specify the distributions $p(y_i \mid x_i, \theta)$ and $p(W \mid X, \phi)$ for some unknown parameters θ and ϕ, then calculate the treatment effect from the inferred missing potential outcomes.
2. Weighting, in which observed outcomes are weighted by propensity scores.
3. Subclassification, in which units are grouped by similar propensity scores.
4. Matching, in which each treated unit is matched with a control unit with a similar set of covariates.
5. Mixed esimators, any combination of the above.

12.5 Design Phase In the design phase, we "investigate the extent of overlap in the covariate distributions" that will allow us to construct better subsamples. Phase does not involve outcome data (all the better to avoid "contamination" of the treatment effect estimate with prejudices of the experimenter – e.g., choosing control villages in Millennium Villages Project).

The authors highlight three types of design:

1. Assess balance of covariates.
2. Subsample selection using matching.
3. Subsample selection using trimming, where we remove units with extreme propensity scores (i.e., $e(x_i)$ close to 0 or 1).

12.6 Assessing Unconfoundedness We cannot test unconfoundedness, only assess it. The authors highlight three ways:

1. *Estimate effect of treatment on unaffected outcome*. Choose a variable that we know for certain is not affected by the treatment (e.g., the country in which someone is using sunscreen). They highlight lagged outcomes. Next, use the rest of the covariates to estimate the treatment effect on the variable chosen. The size of the treatment effect is small, unconfoundedness is more plausible.

2. *Estimate effect of pseudo-treatment on outcome*. Choose an alternative treatment that is known *a priori* not to have an effect on the outcome (a "psuedo-treatment"). One would expect the estimated treatment effect within the control group for this to be zero. Significant treatment effects within the control group suggests that unconfoundedness is violated.

3. *Assess sensitivity of treatment estimates to choice of pre-treatment variables*. Partition the covariates into two parts. Compare treatment effects of each part to each other. A large differential suggests that either unconfoundedness does not hold, or that it "relies critically on all covariates".