Counterfactuals and the Potential Outcome Model Summary

Siddhartha Basu

April 23, 2017

1 Counterfactuals and the Potential Outcome Model

1.1 Potential Outcomes

In causal inference, we are generally interested in individual-level causal impacts of the treatment:

$$\delta_i = y_i^1 - y_i^0$$

Let D be the causal exposure variable, which is 1 if you are exposed to the treatment state and 0 if you are exposed to the control state. The observed outcome variable Y is therefore:

$$Y = Y^1 \text{ if } D = 1$$
$$Y = Y^0 \text{ if } D = 0$$
$$Y = DY^1 + (1 - D)Y^0$$

The last equation shows the **fundamental problem of causal inference**, that we cannot observe the potential outcome under the control state for those in the treatment state, and that we cannot observe the potential outcome under the treatment state for those in the control state. This means that it is **impossible to calculate individual-level causal effects**.

1.2 The Average Treatment Effect

Since we cannot calculate individual-level causal effects, we focus our attention on carefully defined aggregate causal effects. The broadest possible average effect is the average treatment effect (ATE) for the population as a whole:

$$E[\delta] = E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$$

1.3 The Stable Unit Treatment Value Assumption

1.3.1 Definition

In most applications, the potential outcome model retains its tractability through the presence of the stable unit treatment value assumption (SUTVA). SUTVA requires that the potential outcomes of individuals be unaffected by changes in the treatment exposures of all other individuals. Mathematically, it requires that, if $\mathbf{d}$ is the vector of treatment assignments for all N individuals, the treatment effect for each individual i:

$$\delta_i(\mathbf{d}) = y_i^1(\mathbf{d}) - y_i^0(\mathbf{d})$$

depends on $\mathbf{d}$, only through d_i, individual i's assignment. SUTVA is what allows us to declare $y_i^1(\mathbf{d}) = y_i^1$ and $y_i^0(\mathbf{d}) = y_i^0$, and as a result, individual-level causal results δ_i exist that are independent of the overall configuration of causal exposure.

1.3.2 Violations

Typical SUTVA violations share two interrelated features:

- **Influence patterns** that result from contact across individuals in social or physical space.
- **Dilution/concentration patterns** that one can assume would result from changes in the prevalence of treatment.

In the worker training example, SUTVA is violated for large programs in small markets but not small programs in large markets. For Catholic schooling, there are both influence patterns between students and potential dilution patterns.
If the violation can be interpreted as a dilution/concentration problem, even when generated in part by an underlying influence pattern, the analyst can proceed by scaling back the asserted relevance of any estimates where the prevalence of treatment is not substantially different. The idea is to state that estimates of average causal effects hold only for what-if movements of relatively small numbers of individuals. This doesn't work when influence patterns are inherent to the causal process of interest, eg. vaccinations. In situations like this you have:

- The indirect effect: difference in outcome for a non-vaccinated person in a community with a vaccination program vs. his outcome in a similar community without a vaccination program.
- Mathematically, Y_i^0(community vaccination program) $- Y_i^0$(no community vaccination program)
- The total effect: Y_i^1(community vaccination program) $- Y_i^0$(no community vaccination program)

Effectively estimating these types of effects requires a **nested randomization structure**, where (1) vaccine programs are randomly assigned to a subset of communities and (2) people within these communities are randomly given vaccinations.

1.4 Treatment Assignment and Observational Studies

For a properly-run randomized experiment, treatment status gives no information about subject i's what-if outcome under the treatment and control states: y_i^1 and y_i^0. Mathematically we have:

$$(Y^1, Y^0) \perp\!\!\!\perp D$$

This implies that, in the full population, ex ante to any pattern of treatment assignment, D is independent of Y^0, Y^1. Then treatment is assigned, and values of Y emerge, at which point Y and D are dependent.
The main difference between an experiment and an **observational study** is that the analyst can control treatment assignment in an experiment, whereas they only observe them in an observational study. An observational researcher can only observe pairs $\{y_i, d_i\}_i^N$. Then they have to understand the treatment selection mechanism and account for it (this is often impossible).

1.5 Average Causal Effects and Naive Estimation

The fundamental problem of causal inference means that we have to zoom out to non-individual level causal effects.

1.5.1 Conditional Average Treatment Effects

The unconditional average treatment effect $E[\delta] = E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$ can't be estimated due to the fundamental problem of causal inference. Some other average treatment effects that we can't directly compute are the average treatment effect for those who typically take the treatment (**ATT**):

$$E[\delta|D = 1] = E[Y^1|D = 1] - E[Y^0|D = 1]$$

And the average treatment effect for those who typically do not take the treatment (**ATC**):

$$E[\delta|D = 0] = E[Y^1|D = 0] - E[Y^0|D = 0]$$

Sometimes these are as or more interesting than the ATE. For example, in the worker training problem, the ATT is the average wage gain for those who enrolled in the program, which is quite important.

1.5.2 Naive Estimation of Average Treatment Effects

Let π be the fraction of the population in treatment. Then the ATE can be computed:

$$\begin{aligned} E[\delta] = &\{\pi E[Y^1|D = 1] + (1 - \pi)E[Y^1|D = 0]\} \\ &- \{\pi E[Y^0|D = 1] + (1 - \pi)E[Y^0|D = 0]\} \end{aligned}$$

Unfortunately, we are only able to compute the following three of the five unknowns in the above equation:

$$\begin{aligned} E_N[d_i] &\to \pi \\ E_N[y_i|d_i = 1] &\to E[Y^1|D = 1] \\ E_N[y_i|d_i = 0] &\to E[Y^0|D = 0] \end{aligned}$$

Without assumptions, there is no way to compute unbiased estimates of $E[Y^1|D = 0]$ and $E[Y^0|D = 1]$

1.5.3 The Typical Inconsistency and Bias of the Naive Estimator

Note the following decomposition of the naive estimator:

$$\begin{aligned} E[Y^1|D = 1] - E[Y^0|D = 0] = &E[\delta] + \{E[Y^0|D = 1] - E[Y^0|D = 0]\} \\ &+ (1 - \pi)\{E[\delta|D = 1] - E[\delta|D = 0]\} \end{aligned}$$

This identifies the two major sources of bias:

- **Baseline bias:** The difference in expected outcome in the absence of the treatment between those in the treatment group and those in the control group.
- **Differential treatment effect bias:** The expected difference in the treatment effect between those in the treatment group and those in the control group.

For example if we looked at the effect of college attendance on earnings, there could be three sources of higher incomes by college educated individuals. Firstly, attending college may give you the skills to earn more ($E[\delta]$). Secondly, those that attend college may have higher baseline expected earnings, eg. because their parents are richer. Thirdly, college may increase earnings for those that select to go to college more than for those who choose not to go to college.

1.5.4 Estimating Causal Effects Under Maintained Assumptions About Potential Outcomes

There are two basic classes of assumptions that are variants of each other: (1) assumptions about potential outcomes for the treatment and control groups, (2) assumptions about the treatment assignment/selection process. We discuss (1) in this section. Consider the following two assumptions:

$$E[Y^1|D=1] = E[Y^1|D=0]$$
$$E[Y^0|D=1] = E[Y^0|D=0]$$

If both of these are maintained, then ATT, ATC and ATE are all equal, and the naive estimator works. Random assignment makes these assumptions reasonable. They are not reasonable for observational studies.

If only the first assumption holds, then we can still estimate ATC, since $E[\delta|D=0] = E[Y^1|D=0] - E[Y^0|D=0] = E[Y^1|D=1] - E[Y^0|D=0]$.

If only the second assumption holds, then we can still estimate ATT, since $E[\delta|D=1] = E[Y^1|D=1] - E[Y^0|D=1] = E[Y^1|D=1] = E[Y^0|D=0]$.

2 Extra Class Notes