Using Differences-in-Differences Methods for Causal Analysis

Jodi Beggs

Akamai Technologies

July 2015

One way to investigate the effect of a particular treatment, feature, etc. is to compare an outcome of interest (some measure of performance in this case) before and after the treatment was initiated. Unfortunately, this approach is vulnerable to the *post hoc ergo propter hoc* logical fallacy[1]- namely, the flawed notion that because one thing happened after another thing, one thing happened because of the other thing. In order to more convincingly establish a causal relationship, one would like to compare the before-after trajectory of a group that received treatment to the before-after trajectory of a group that didn't receive the treatment in order to control for general environmental factors that affect the outcome trajectory. This is where differences-in-differences analysis comes in.

The logic of differences-in-differences is simple enough- consider the following matrix of outcomes:

	Before	After	Difference
Treated	A	B	B-A
Untreated	C	D	D-C

We can then analyze whether the treatment had an effect by investigating whether the incremental difference of the treatment group (B-A) – (D-C) is different from zero. (If the difference is positive, the treatment had a positive effect and vice versa, assuming that larger outcomes are better.) In order to be more rigorous in the analysis, the next logical step would be to test whether this difference in differences is statistically significantly different from zero. So how is this done?

In our context, our outcome of interest is some measure of performance, and let's assume for initial simplicity that there is just one feature under consideration. Our independent variables are binary indicators of whether a page/site/etc. is in the group that had a feature turned on (or actually applied) between the before and after period (i.e. is in the treatment group) along with some set of relevant controls. Since we are using data for two time periods, we can also create a binary indicator after that is equal to 1 if an observation is from the later time period and zero otherwise. An appropriate regression specification would look something like the following:

$$performance_{it} = \beta_0 + \beta_1 treatment_i + \beta_2 after_t + \beta_3 after_t * treatment_i + \gamma controls_{it} + \varepsilon_{it}$$

[1] Not just an episode of *The West Wing*!

where subscript i references a particular page/site.etc. and subscript t indicates time period. In differences form (where differences are over time), this reduces to

$$\Delta performance_{it} = \beta_2 + \beta_3 treatment_i + \gamma \Delta controls_{it}$$

Therefore, B_3 is our variable of interest, since it shows the incremental impact on the change in performance for being in the treatment group, and we can very easily test its significance.

Alternatively, one may consider a slightly different specification that directly uses a binary indicator for whether a particular feature is enabled (or actually applied) at a point in time:

$$performance_{it} = \beta_0 + \beta_1 enabled_{it} + \beta_2 after_t + \gamma controls_{it} + \varepsilon_{it}$$

Or, in differences form:

$$\Delta performance_{it} = \beta_1 \Delta enabled_{it} + \beta_2 + \ \gamma \Delta controls_{it}$$

In this case, B_1 is the variable of interest because it represents the effect of a feature being activated between the before and after time periods. This latter approach has its pros and cons:

- Pro: Can be easily extended to more than two time periods by adding in more time dummy variables. (Note that a single time trend would likely not be appropriate since there's no reason to think that the relationship between time and performance is linear.)
- Con: Doesn't address control for potential differences in the group of pages/sites/etc. that have features turned on versus those that don't.

Note that, regardless of specification, the analysis must account for the fact that there is potential correlation across the error terms due to the fact that multiple observations represent a single page/site/etc. (This for the most part involves a tweak to how standard errors are calculated that statistical software can do automatically.)

What do we do if we can't observe feature enablement/application directly?

In this case, we can do a couple of things, but we will need to gather some additional data first from a (random) sample of pages/sites/etc.:

1. Data on how often a feature is actually used/applied once it is enabled. (This would of course become irrelevant if we collect data directly on feature application.)
2. Data on what characteristics of a site/page/etc. predict whether a feature will be applied (we would probably run a regression to calculate this).

With this data, we can try to back out the effect of application from the effect of having the feature enabled and the likelihood that the feature is applied once it is enabled. For example, if the feature being enabled has a .01 impact on performance but the feature is only applied 20% of the time it is enabled, we can approximate that the application has a .05 impact on performance. Alternatively, we can use a propensity score approach to estimate the effect of application- to do this, we calculate a summary statistic for the likelihood that a feature gets applied given the characteristics of the site/page/etc. (called a propensity score) and then we use the propensity score as a regressor since we can't see application directly. The coefficient on this propensity score should then approximate the effect of the feature being applied.