Causal Inference in Observational Studies

Contents

1 Causal Inference and Predictive Comparison

Causal Inference and Predictive Comparison

- We have been using regression in the *predictive* sense, to determine what values of Y tend to be associated with particular values of X in a given hypothetical "superpopulation" modeled with random variables and probability distributions.
- In causal inference, we attempt to answer a fundamentally different question, namely, what would happen if *different treatments* had been applied to *the same units.*

1.1 How Predictive Comparison Can Mislead

Examples

Example 1 (Example 1). • Suppose a medical treatment is of no value. It has no effect on any individual.

- However, in our society, healthier people are more likely to receive the treatment.
- What would/could happen? (C.P.)

Example 2 (Example 2). • Suppose a medical treatment has positive value. It increases IQ on any individual.

- However, in our society, lower IQ people are more likely to receive the treatment.
- What would/could happen? (C.P.)

1.2 Adding Predictors as a Solution

Adding Predictors as a Solution

- In the preceding two examples, there was a solution, i.e., to compare treatments and controls *conditional on previous health status.* Intuitively, we compare current health status across treatment and control groups only within each previous health strategy.
- Another alternative is to include treatment status *and* previous health status as predictors in a regression equation.

- Gelman and Hill assert that “in general, causal effects can be estimated using regression if the model includes all confounding covariates and if the model is correct.”

1.3 Omitted Variable Bias

Omitted Variable Bias

Suppose the “correct” specification for confounding covariate x_i is

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i \tag{1}$$

Moreover, suppose that the regression for predicting x_i from the treatment is

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i$$

Omitted Variable Bias – 2

Substituting, we get

$$\begin{aligned} y_i &= \beta_0 + \beta_1 T_i + \beta_2(\gamma_0 + \gamma_1 T_i + \nu_i) + \epsilon_i \\ &= \beta_0 + \beta_2\gamma_0 + \beta_1 T_i + \beta_2\gamma_1 T_i + (\epsilon_i + \beta_2 \nu_i) \\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1) T_i + (\beta_2\gamma_1 T_i) \end{aligned} \tag{2}$$

Note that this can be written as

$$y_i = \beta_0^* + \beta_1^* T_i + \epsilon_i^*$$

where

$$\beta_1^* = \beta_1 + \beta_2\gamma_1$$

2 Causal Inference – Problems and Solutions

2.1 The Fundamental Problem

The Fundamental Problem

- The *potential outcomes* of y_i^1 and y_i^0 under T are the values that the ith unit would have demonstrated had level 1 or level 0 of the treatment actually been received by that unit.
- In general, of course, the ith unit (or, for simplicity, individual i) will not receive both treatments so either y_i^1 or y_i^2 is a *counterfactual* and will not be observed. We can think of the counterfactuals as "missing data."

2.2 Ways of Getting Around the Problem

Possible Solutions

We can think of causal inference as a prediction of what would happen to unit i if $T_i = 1$ or $T_i = 0$.

There are 3 basic strategies:

1. Obtain close substitutes for the potential outcomes. Examples:
   (a) T=1 one day, T=0 another
   (b) Break plastic into two pieces and test simultaneously
   (c) Measure new diet using previous weight as proxy for y_i^0.
2. Randomize. Since we cannot compare on identical units, compare on similar units. In the long run, randomization confers similarity.
3. Do a statistical adjustment. Predict with a more complex model, or block to achieve similarity.

2.3 Randomization

Randomization

In a completely randomized experiment, we can estimate the average treatment effect easily as

$$\text{average treatment effect} = \text{avg } (y_i^1 - y_i^0)$$

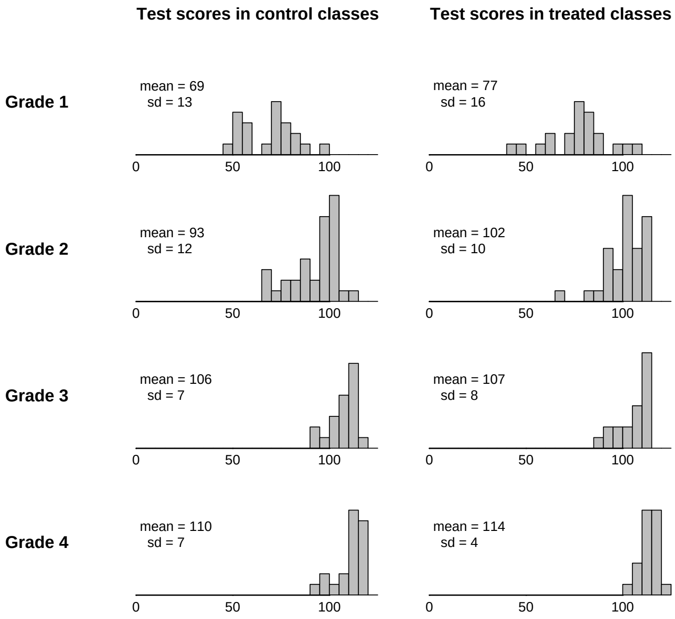
The standard test on means can be applied. Of course, issues of *external validity* apply too. The results are relevant only for the population from which the sample was taken.

An Electric Example

Example 3 (Electric Company Study).

- 4 grades, 2 cities
- For each city and grade, approximately 10-20 schools were chosen
- 2 weakest classes randomly assigned to either treatment or control
- $T = 1$ classes given opportunity to watch The Electric Company, and educational show
- At the end of the year, students in all classes were given a reading test

Post-Test Results



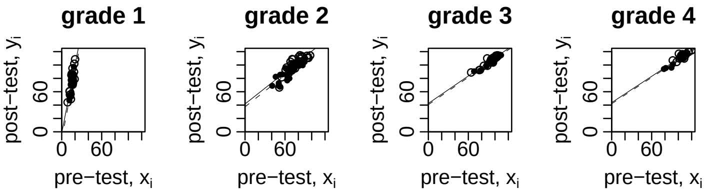
2.4 Controlling for a Pre-Treatment Predictor

Controlling for Pre-Treatment Score

- The preceding results are suggestive.

- However, in this study, a pre-test was also given. In this case, the treatment effect can also be estimated using a regression model: $y_i = \alpha + \theta T_i + \beta x_i + \text{error}_i$.
- First, we fit a model where post-test score is predicted from pre-test score, with constant slopes treatment and control groups.
- Treatment group is represented by a solid regression line and circles, control by dotted regression line and filled dots.
- In this case,
  - The treatment effect is estimated as a constant across individuals within treatment group
  - The regression lines are parallel, and
  - The treatment effect is the difference between the lines.

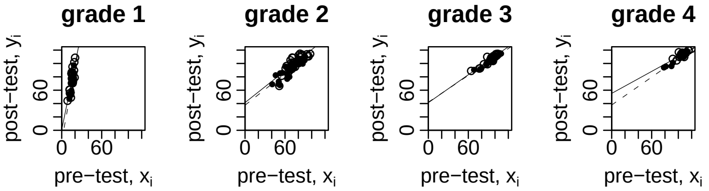
Results with No Interaction



Including an Interaction Term

- The preceding model failed to take into account the fact that the relationship between pre-test and post-test scores might differ between treatment and control groups.
- We can add an interaction term to the model, thus allowing treatment and control groups to have regression lines with differing slopes.
- In this model, $y_i = \alpha + \theta T_i + \beta_1 x_i + \beta_2 T_i x_i + \text{error}_i$.
- Note that in this mode, the treatment effect can be written as $\theta + \beta_2 x_i$. In other words, *the treatment effect changes as a function of pre-test status.*

Interaction Model Results

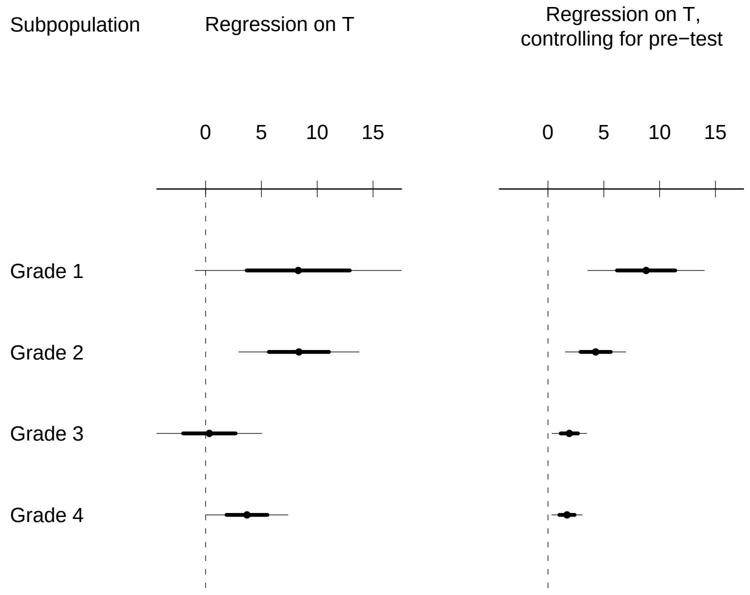


A Combined Picture

This next plot shows T regression coefficient estimates, 50% and 90% confidence intervals by grade. You can see clearly how "controlling" for pre-test score reduces variability in the estimators and smooths them out.

Note that all classes improved whether treated or not, so it is hard to see what is going on. (The pre-was identical to the post-test except in grade 1, so the improvement is hardly a shock.)

Combined Plot of Treatment Effects



2.5 The assumption of no interference between units

No Interference between Units

- An important assumption in these modeling efforts is that treatment assignment for individual i does not effect treatment effect for individual j.
- Without this assumption, we'd need to define a different potential outcome for the ith person not only for each treatment, but also for every other treatment received by every other relevant individual!

3 Treatment interactions and post-stratification

Treatment Interactions and Post-Stratification

- Once we include pre-test information in the model, it is natural to allow an interaction between the pre-test (x) and the treatment effect (T).
- As mentioned above, once the interaction term is included, the effect of the treatment varies for each individual as a function of the pre-test score.

Simple Model

```
> display(lm(post.test˜treatment,
+ subset=(grade==4)))
```

```
lm(formula = post.test ˜ treatment, subset = (grade == 4))
            coef.est coef.se
(Intercept) 110.36     1.30
treatment     3.71     1.84
---
n = 42, k = 2
residual sd = 5.95, R-Squared = 0.09
```

The estimated effect is 3.7 with a standard error of 1.8.

Including the Pre-Test

We can get a more efficient (lower s.e.) error by including the pre-test as a predictor.

```
> display(lm(post.test˜treatment+pre.test,
+ subset=(grade==4)))
```

```
lm(formula = post.test ˜ treatment + pre.test, subset = (grade ==
    4))
            coef.est coef.se
(Intercept) 41.99     4.28
treatment    1.70     0.69
```

```
pre.test     0.66     0.04
---
n = 42, k = 3
residual sd = 2.18, R-Squared = 0.88
```

The new estimated treatment is only 1.7 with a standard error of 0.7. Next we add the interaction.

Adding a $T \times x$ Interaction

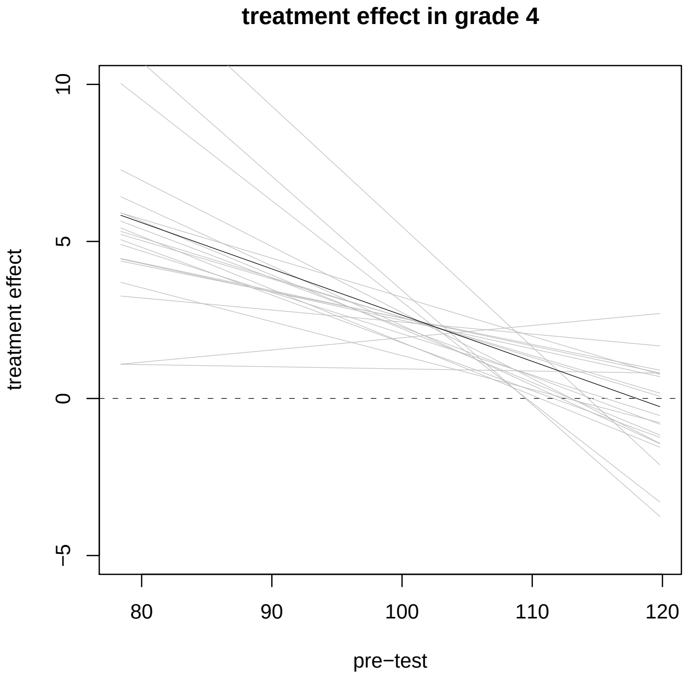
```
> display(lm(post.test~treatment+pre.test +
+ treatment:pre.test, subset=(grade==4)))
```

```
lm(formula = post.test ~ treatment + pre.test + treatment:pre.test,
    subset = (grade == 4))
                   coef.est coef.se
(Intercept)        37.84     4.90
treatment          17.37     9.60
pre.test            0.70     0.05
treatment:pre.test -0.15     0.09
---
n = 42, k = 4
residual sd = 2.14, R-Squared = 0.89
```

The effect is now $17.37 - 0.15x$. Looking at a previous plot, we can see that pretest scores range from about 80 to 120, and plugging into the formula, we see that the treatment effect varies from about 5.37 to $-.63$. This is an estimate of the *range* of the effect, and does *not* include statistical uncertainty indications.

Picturing the Uncertainty

To get a sense of the uncertainty, we can plot the estimated treatment effect as a function of x, including random simulation draws to create a picture of the



uncertainty involved.

Computing Uncertainty

- We can also estimate a mean treatment effect across classrooms by averaging. Across classrooms, we calculate the treatment effect as $\theta_1 + \theta_2 x_i$ and simply average.
- We can also compute the mean and standard deviation of these estimated average effects across the simulations depicted in the preceding graph. In this case, we get

```
> mean(avg.effect)

[1] 1.760238

> sd(avg.effect)

[1] 0.6851486
```

- The result is 1.8 with a standard deviation of 0.7, quite similar to the estimate obtained by fitting the model with no interactions. The main virtue of fitting the interaction is to get an estimate of how the treatment effect varies as a function of the pretest.

3.1 Post-stratification

Computing Average Treatment Effects

Treatment effects may vary as a function of pre-treatment indicators. To estimate an average treatment effect, we average over the population.

For example, if we have the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 T + \beta_4 x_1 T + \beta_5 x_2 T + \epsilon$$

the estimated treatment effect is then

$$\beta_3 + \beta_4 x_1 + \beta_5 x_2$$

The mean treatment effect is then

$$\beta_3 + \beta_4 \mu_1 + \beta_5 \mu_2$$

where μ_1, μ_2 are the means of x_1, x_2.

Standard errors can be computed via simulation or by analytic derivation.

4 Observational Studies

4.1 Electric Company example

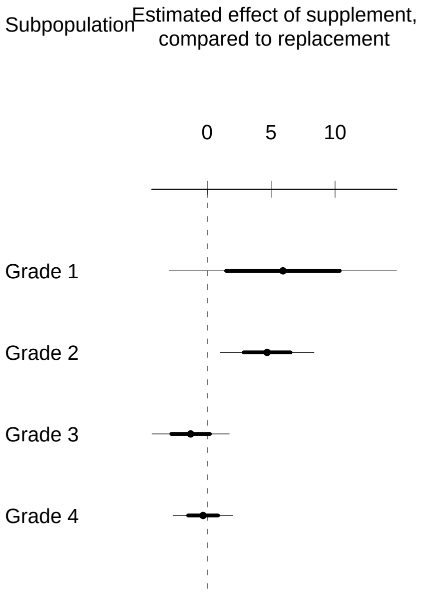
The Electric Company Revisited

According to Gelman & Hill , this is an observational study “for which a simple regression analysis, controlling for pre-treatment information, may yield reasonable causal inferences.”

It turns out that, in the study, once $T = 1$, the teacher decided whether to *replace* or *supplement* the regular reading program with the Electric Company show.

The results are on the next slide. Supplementing seems to more effective than replacing, at least in the lower grades, although the low precision compromises our ability to judge.

The Electric Company Revisited



4.2 Assumption of ignorable treatment assignment

Ignorability

Formally, ignorability states that

$$y^0, y^1 \perp T \mid X$$

This says that the distribution of *potential* outcomes is the same across levels of the treatment variable T, once we condition on the confounding covariates X.

Note:

- We would not necessarily expect any two classes to have the same probability of receiving the supplemental version of the treatment;
- However, we do expect any two classes at the same level of the confounding variable (in this case pre-test score) to have had the same probability of receiving the treatment.

Ignorability

A non-ignorable assignment mechanism might occur if, for example, brighter more motivated teachers assigned students to a treatment based on their knowledge of the characteristics students, and that motivation also led to higher scores.

It is always possible that ignorability does not hold. If it seems likely that treatment assignments depended on information not included in the model, then we need to choose a different analysis strategy.

4.3 Judging the reasonableness of regression as a modeling approach, assuming ignorability

Lack of Overlap

- Even if ignorability is satisfied, regression on the covariates and treatment may not be the best approach, especially if there is lack of overlap and balance.
- For example, suppose in the Electric Company experiment, students in the supplementary condition tended to have higher pre-test scores. This can lead to misleading results, because the data are being plotted in different regions of the range of pretest scores.