

PS C236A/ Stat C239A
Problem Set 2
Due: Sept. 17, 2009

Problem 1: A researcher wants to figure out once and for all if giving students high school vouchers for private schools would significantly improve the educational outcomes, in this case measured by standardized test scores of the students. The researcher receives a government sponsored grant to conduct a randomized controlled trial. Assume that the researcher's experiment does not suffer from any of the standard problems we typically encounter in implementing randomized trials, namely there are no compliance issues. This means that everyone who received a private school voucher goes to private school and everyone who did not receive a voucher (those people in the control group) does not go to private school. After running the experiment and collecting all of her data, the researcher estimates the following model to determine the treatment effect:

$$Y_i = \alpha + \beta_1 \text{voucher}_i + \beta_2 \text{race} + \beta_3 \text{parent's education}_i + \beta_4 \text{parent's income}_i + \epsilon_i$$

where Y is the student's test score at the end of the experiment, "voucher" is a dummy variable for whether a student received a voucher or not, and the other variables are "controls" which predict test scores in particular and other educational outcomes in general. The researcher finds that her estimate for β_1 is significant and sends in her work for publication.

Many months later she receives her reviewers comments. One reviewer really doesn't like the researcher's model. The reviewer argues, "Since the experiment has no compliance problems, and because it is a randomized controlled trial, the researcher should only estimate the following model:

$$Y_i = \alpha + \beta_1 \text{voucher} + \epsilon_i$$

The researcher estimates the recommended model, but finds that while her point estimate for β_1 has not changed, her standard errors have grown and her estimate is no longer significant.

- a. What is the reviewer's rationale?
- b. Why do the results differ between the two models? Prove your answer using the OLS framework. [Hint: This is an experiment, so some of the problems that plague observational studies are not relevant.]
- c. Which model should we prefer? Why? Is this generally true, or are there exceptions?
- d. If the researcher hadn't had been so fortunate, and had experienced the common problems that plague experiments such as compliance problems, how might this have affected her estimates.
- e. *Bonus:* As an attentive student, you say "but if this is a randomized trial, why don't we use the potential outcomes framework?". Therefore, you estimate the treatment effect under this framework. Should your estimate and the reviewer's model yield the same answer? Prove it.

Problem 2 Consider a field experiment that compares treatments A and B. Suppose there are N subjects, indexed by $i = 1, \dots, N$. Let x_i be the response of subject i to treatment A; likewise, y_i is the response to B. For each i , either x_i or y_i can be observed, but not both. Let S be a random subset of $\{1, \dots, N\}$, with n elements; this group gets treatment

A, so x_i is observed for i in S . Let T be a random subset of $\{1, \dots, N\}$, with m elements, disjoint from S . This group gets treatment B , so y_i is observed for i in T .

We estimate population means \bar{x} and \bar{y} by the sample means:

$$\bar{X} = \frac{1}{n} \sum_i^n x_i \qquad \bar{Y} = \frac{1}{m} \sum_i^m y_i$$

Using simple sampling without replacement formulas:

$$\text{var}(\bar{X}) = \frac{N-n}{n-1} \frac{\sigma^2}{n} \qquad \text{var}(\bar{Y}) = \frac{N-m}{N-1} \frac{\tau^2}{m}$$

$$\text{cov}(\bar{X}, \bar{Y}) = -\frac{1}{N-1} \text{cov}(x, y)$$

- What is the average treatment effect parameter? Write it using the above notation and also explain what it is in words.
- What is the variance of the average treatment effect, i.e. $\text{var}(\bar{X} - \bar{Y})$, using the above notation?
- Is this variance identified, i.e. estimable from data?
- The usual two sample difference-in-means variance (without replacement) found in sampling textbooks is:

$$\frac{N}{N-1} \left(\frac{\sigma^2}{n} + \frac{\tau^2}{m} \right)$$

What is the difference, if any, between the usual two sample difference-in-means variance and the variance expression you derived in part b?

Problem 3 Table 1 contains the potential outcomes from a hypothetical experiment with 6 units. Complete the following calculations using R.

Table 1: Potential Outcomes

Unit	Y_T	Y_C
1	2	1
2	6	2
3	33	13
4	17	14
5	2	10
6	54	3

- What are the unit-level treatment effects? What is the “true” average treatment effect? Is the average treatment effect a reasonable way of summarizing causal effects in this case?
- What is the true variance of the average treatment effect, using the formula you derived in part b of the previous question? What is the variance using the “usual” formula written in part d of the previous question?
- Write a function that randomly assigns treatment to three out of the six units and then produces the observed values of the dependent variable. The function should also calculate the estimated average treatment effect from the observed values, as well as its standard errors. [Hint: You may want to look at the help file for the function `rbinom(n, size, prob)` with `size = 1` and `prob = 0.5`.]

- d. Calculate the estimated treatment effect for every possible combination of treatment assignment. Summarize this distribution of estimates using a plot. [Hint: You may want to look at the help file for the function `combn(x, m)`.]