

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?

First of all, let's let model (1) refer to the model with covariates added (the researcher's model), and model (2) refer to the model with no covariates (the reviewer's model). Since this is a perfectly randomized experiment, we know that the assignment of vouchers is uncorrelated with all observable and unobservable characteristics of students. It follows that there can be no omitted variable bias in model (2), and presumably this is the rationale behind the reviewer's model, model (2). Since all observable and unobservable characteristics of students are orthogonal to treatment assignment, the reviewer knows that the estimate of β_1 from model (2) is guaranteed to be unbiased and he sees no point in adding covariates to the model. We know that race, parents education and parents income did not have an impact on the treatment assignment decision since assignment was performed randomly.

We should note, however, that we can turn the reviewer's rationale against him. Since all covariates are orthogonal to treatment assignment, adding these covariates to the model will never bias the estimation of β_1 . To make this claim, though, we do have to assume that all covariates added to the model are pre-treatment, so that the problem of post-treatment bias does not arise (recall that even under random assignment, if we condition on covariates that were affected by the treatment we will obtain a biased estimate of the treatment effect). Since under these assumptions, adding covariates to model (2) will still yield an unbiased estimate of β_1 , the possibility of introducing bias cannot be used as an argument against model (1).

- 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.]

There is an argument in favor of using model (1), and that is that adding (pre-treatment) covariates to model (2) increases the precision in the estimation of β_1 (i.e., it reduces the standard error estimate of β_1). This is the reason why the researcher finds that this estimate is no longer significant, although still has the same point estimate, when the covariates are no longer included in the model.

Now, on to the proof of why the results differ between the two models. Let $\tilde{x} = x - \bar{x}$ refer to a mean deviated variable for any variable x . The following proof will rely on mean deviated variables in order to remove the constant term and make the proof easier to follow. For a discussion on how mean deviated regression works, see any econometrics book that discusses partitioned regression (for example, Greene's "Econometric Analysis"). Now, let's rewrite the models. Let the researcher's model, model (1), be written more generally as:

$$\tilde{Y} = \beta_1 \tilde{T} + \gamma \tilde{X} + \epsilon \quad (1^*)$$

where \tilde{T} is an $n \times 1$ treatment indicator (mean deviated), \tilde{Y} is an $n \times 1$ vector of mean deviated outcomes, and \tilde{X} is an $n \times k$ matrix of mean deviated pre-treatment covariates, β_1 is a scalar and γ is a $k \times 1$ vector of coefficients for the pre-treatment covariates. Let the generalized version of model (2), the reviewer's model be:

$$\tilde{Y} = \beta_2 \tilde{T} + \nu \quad (2^*)$$

We know that the OLS estimator of β_2 is:

$$\hat{\beta}_2 = (\tilde{T}^T \tilde{T})^{-1} \tilde{T}^T \tilde{Y}$$

and that the variance of this estimator is:

$$Var(\hat{\beta}_2) = \frac{\hat{\nu}^T \hat{\nu}}{n-1} (\tilde{T}^T \tilde{T})^{-1}$$

with $\hat{\nu}$ defined as the vector of residuals defined by:

$$\hat{\nu} = \tilde{Y} - \hat{\beta}_2 \tilde{T}$$

Now, let's derive $\hat{\beta}_1$ and its variance. Turning to model (1*) and using partitioned regression, we can write the OLS estimators as:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \tilde{T}^T \tilde{T} & \tilde{T}^T \tilde{X} \\ \tilde{X}^T \tilde{T} & \tilde{X}^T \tilde{X} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{T}^T \tilde{Y} \\ \tilde{X}^T \tilde{Y} \end{bmatrix}$$

where the $(k+1) \times (k+1)$ matrix $\begin{bmatrix} \tilde{T}^T \tilde{T} & \tilde{T}^T \tilde{X} \\ \tilde{X}^T \tilde{T} & \tilde{X}^T \tilde{X} \end{bmatrix}^{-1}$ must be calculated using the partitioned inverse.

However, because we know we have orthogonality of the treatment assignment, \tilde{T} and the covariates \tilde{X} , due to the randomization, then this matrix is block-diagonal.

$$\begin{bmatrix} \tilde{T}^T \tilde{T} & \tilde{T}^T \tilde{X} \\ \tilde{X}^T \tilde{T} & \tilde{X}^T \tilde{X} \end{bmatrix}^{-1} = \begin{bmatrix} (\tilde{T}^T \tilde{T})^{-1} & \mathbf{0}_{1 \times k} \\ \mathbf{0}_{k \times 1} & (\tilde{X}^T \tilde{X})^{-1} \end{bmatrix}^{-1}$$

Therefore, it follows that:

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\gamma} \end{bmatrix} &= \begin{bmatrix} \tilde{T}^T \tilde{T} & \tilde{T}^T \tilde{X} \\ \tilde{X}^T \tilde{T} & \tilde{X}^T \tilde{X} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{T}^T \tilde{Y} \\ \tilde{X}^T \tilde{Y} \end{bmatrix} \\ &= \begin{bmatrix} (\tilde{T}^T \tilde{T})^{-1} & \mathbf{0}_{1 \times k} \\ \mathbf{0}_{k \times 1} & (\tilde{X}^T \tilde{X})^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{T}^T \tilde{Y} \\ \tilde{X}^T \tilde{Y} \end{bmatrix} \\ &= \begin{bmatrix} (\tilde{T}^T \tilde{T})^{-1} \tilde{T}^T \tilde{Y} \\ (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y} \end{bmatrix} \end{aligned}$$

From this we see that our estimate of $\hat{\beta}_1 = \hat{\beta}_2$, and we can see that the variances of our estimates for model (2*) are:

$$Var \begin{bmatrix} \hat{\beta}_1 \\ \hat{\gamma} \end{bmatrix} = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - (k + 1)} \begin{bmatrix} (\tilde{T}^T \tilde{T})^{-1} & \mathbf{0}_{1 \times k} \\ \mathbf{0}_{k \times 1} & (\tilde{X}^T \tilde{X})^{-1} \end{bmatrix}$$

where $\hat{\epsilon}$ is the vector of residuals defined by

$$\hat{\epsilon} = \tilde{Y} - \hat{\beta}_2 \tilde{T} - \hat{\gamma} \tilde{X}$$

Therefore, we see that

$$Var[\hat{\beta}_1] = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - k - 1} (\tilde{T}^T \tilde{T})^{-1}$$

Now that we know the variance for each of the estimates of the treatment effect (note, we also know that the two estimates are equal), we can compare the variances.

$$Var[\hat{\beta}_2] - Var[\hat{\beta}_1] = \frac{\hat{\nu}^T \hat{\nu}}{n - 1} (\tilde{T}^T \tilde{T})^{-1} - \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - k - 1} (\tilde{T}^T \tilde{T})^{-1} \quad (1)$$

$$= \left(\frac{\hat{\nu}^T \hat{\nu}}{n - 1} - \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - k - 1} \right) (\tilde{T}^T \tilde{T})^{-1} \quad (2)$$

However, we want to prove that $Var[\hat{\beta}_2] - Var[\hat{\beta}_1] > 0$. We know the matrix $(\tilde{T}^T \tilde{T})^{-1}$ is positive definite, so the sign of $Var[\hat{\beta}_2] - Var[\hat{\beta}_1]$ will be equal to the sign of $\left(\frac{\hat{\nu}^T \hat{\nu}}{n - 1} - \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - k - 1} \right)$. We can prove that the sum of squared residuals always falls when more covariates are added to the regression (not shown here, but this proof can be found in many econometric textbooks, including the proof in Chapter 3 of Green's "Econometric Analysis"). Note that this is the same proof that says that R^2 increases when more covariates are added to the regression, which is actually true because the sum of squared residuals falls.

So, we know that $\hat{\nu}^T \hat{\nu} \geq \hat{\epsilon}^T \hat{\epsilon}$ due to the reduction in sum of squared residuals from adding covariates, but this is not enough to show that $Var[\hat{\beta}_2] - Var[\hat{\beta}_1] > 0$. This is because of the fact that we divide the sum of squared residuals by the degrees of freedom, and $n - 1 > n - (k + 1)$, so while $\hat{\nu}^T \hat{\nu}$ is greater than $\hat{\epsilon}^T \hat{\epsilon}$, it is also divided by a bigger number. So, the result is surprisingly ambiguous. Whether or not the variance decreases will depend on how the degrees of freedom and reduction in sum of squared residuals relate. The precise answer, then, is that whether $Var[\hat{\beta}_1]$ rises or falls when we add k covariates depends on whether the fall in the sum of squared residuals is bigger than the increase caused by the loss in the degrees of freedom (noting that we lose exactly k degrees of freedom).

We can, in this example, determine that the effect of the sum of squared residuals dominates the effect of the loss in the degrees of freedom because we are told that when we add the covariates, the standard error of the treatment effect decreases (while the point estimate remains the same). In this case, the ambiguity in the proof does not cause problems in answering the question, we know that adding covariates reduced the standard error.

- c. Which model should we prefer? Why? Is this generally true, or are there exceptions?

There is an argument to be made for the researcher's model. We know that adding the covariates didn't change the point estimate, so we know that correlation among T and X wasn't causing bias in the estimate of the treatment effect. We also know that adding the covariates reduced the standard error and thus increased the precision of our estimate, so it would appear that the researcher wasn't erring, at least not in terms of logic based in the OLS framework. However, we can't always say that it will be the case that adding covariates will help. If the point estimate had changed, then there would be a cause for concern since it would indicate that there was some correlation in T and X , and we know that the simple regression is an unbiased estimate of the ITT estimate. We also know, from the proof in part b, that is isn't a guarantee that adding covariates will necessarily reduce the standard error of your estimate. Depending on which force dominates, the reduction in sum of squares or the loss of degrees of freedom, it may help or hurt. Also, given what we've learned from the Freedman discussion of

using OLS on experimental data, we can be weary of the researcher’s model since randomization doesn’t justify regression. The key here is that the point estimate didn’t change, so we can probably allow the researcher the argument that she was improving the precision of her estimate.

- 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.

Many of the common problems that we see in experiments could cause bias. While the “intention to treat” estimate will be unbiased, compliance problems in which “treated” units don’t take treatment, or “control” units manage to get treatment, will cause the “ITT” estimate to be deflated compared to the “treatment of treated” estimate. Only measuring the effect of those who were treated and characterizing all others as control will be a biased estimate of the treatment effect.

- 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.

The point estimate will be the same as the simple regression model, but the nominal standard errors may be different. See the proof in Section 3 notes on Freedman’s “On regression adjustment to experimental data”.

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:

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

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

- a. What is the average treatment effect parameter? Write it using the above notation and also explain what it is in words.

The average treatment effect parameter is $\bar{x} - \bar{y} = \frac{1}{N} \sum_{i=1}^N x_i - \frac{1}{N} \sum_{i=1}^N y_i$. The average treatment effect measures the difference between putting all the subjects into regime A and putting all the subjects into regime B.

- b. What is the variance of the average treatment effect, i.e. $\text{var}(\bar{X} - \bar{Y})$, using the above notation? The variance of the difference of two random variables X and Y is $\text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)$. Apply this formula to

the random variable $X - Y$:

$$\begin{aligned}\text{var}(\bar{X} - \bar{Y}) &= \frac{N-n}{N-1} \frac{\sigma^2}{n} + \frac{N-m}{N-1} \frac{\tau^2}{m} - 2\text{cov}(\bar{X}, \bar{Y}) \\ \text{var}(\bar{X} - \bar{Y}) &= \frac{N-n}{N-1} \frac{\sigma^2}{n} + \frac{N-m}{N-1} \frac{\tau^2}{m} + \frac{2}{N-1} \text{cov}(x, y) \\ &= \frac{1}{N-1} \left(\frac{N\sigma^2}{n} - \frac{n}{n} \sigma^2 + \frac{N\tau^2}{m} - \frac{m}{m} \tau^2 \right) + \frac{2}{N-1} \text{cov}(x, y) \\ &= \frac{N}{N-1} \left(\frac{\sigma^2}{n} + \frac{\tau^2}{m} \right) + \frac{1}{N-1} (2\text{cov}(x, y) - \sigma^2 - \tau^2)\end{aligned}$$

c. Is this variance identified, i.e. estimable from data?

No, since we can never observe $\text{cov}(x, y)$.

d. 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?

The difference is: $\frac{1}{N-1}(2\text{cov}(x, y) - \sigma^2 - \tau^2)$. It's not necessary to prove the direction of this bias, but the bias will always be negative (or zero). As a result, the true variance will always be less than or equal to the usual variance. Hence, the standard it variance estimator is conservative.

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

a. 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?

The unit level treatment effects are as follows: $\{1, 4, 19, 3, -8, 51\}$. The average treatment effect is 11.8. The problem with ATE in this particular case, is that it masks a great deal of heterogeneity in response to treatment. So while ATE may be accurate, it also may not shed much light on the scientific question of interest.

b. 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?

The “true” variance is 82.5. The usual variance is 157.

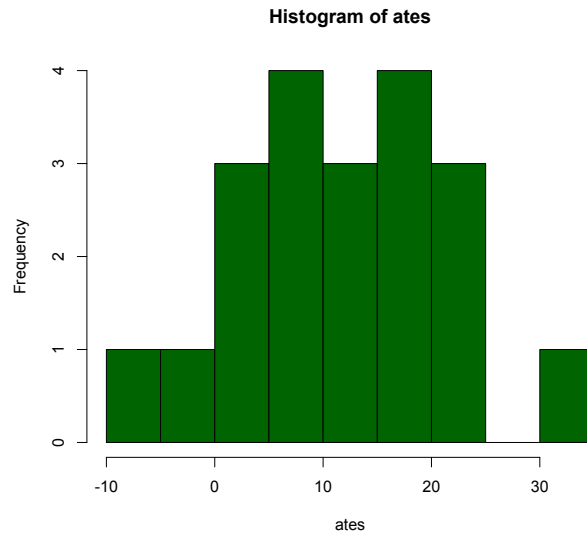


Figure 1: A histogram showing the distribution of all possible treatment effects.

- c. 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.

See the R code.

[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.

See figure 1.

[Hint: You may want to look at the help file for the function `combn(x, m)`.]