

PS C236A/ Stat C239A

Problem Set 1

Due: Sept. 10, 2009

Problem 1: Suppose you are in a simplified world, and you wish to determine the returns to education for a group of N workers you have data for. In this simplified version of the world, there are two factors that influence a worker's income, level of education and intelligence. The correct model would, therefore, be:

$$y_i = \alpha_1 + \gamma_1 * \text{education level}_i + \gamma_2 * \text{intelligence}_i + \epsilon_{1i} \quad (1)$$

Where Y_i is individual i 's income. However, you naively assume that the only factor that influences income is education level, and you run a regression using the following model:

$$y_i = \alpha_2 + \beta_1 * \text{education level}_i + \epsilon_{2i} \quad (2)$$

- Write down or describe the design matrix for the correct model of the world (model 1) as well as the naive model (model 2).
- Show that $\frac{1}{N} \sum_{i=1}^N y_i = \frac{1}{N} \sum_{i=1}^N \hat{y}_i$
- Which, if any, assumptions and conditions are necessary for part b to be true?
- Assume that education level and intelligence are positively correlated. By using the naive model instead of the true model, what happens to your estimate of β_1 ? How would it relate to your estimate of γ_1 if you ran a regression using the true model? Prove it.

Problem 2: True or False, and explain: as long as the design matrix has full rank, the computer can find the OLS estimator $\hat{\beta}$. If so, what are the assumptions good for? Discuss briefly.

Problem 3: Suppose x_1, \dots, x_n and y_1, \dots, y_n have means \bar{x}, \bar{y} , the standard deviations are $s_x > 0, s_y > 0$; and the correlation is r . Let

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

“cov” is shorthand for covariance. Show that —

- $\text{cov}(x, y) = r s_x s_y$
- The slope of the regression line for predicting y from x is

$$\frac{\text{cov}(x, y)}{\text{var}(x)}$$

- $\text{var}(x) = \text{cov}(x, x)$
- $\text{cov}(x, y) = \overline{xy} - \bar{x}\bar{y}$
- $\text{var}(x) = \overline{x^2} - \bar{x}^2$

Application

In this section, you will use R to calculate descriptive statistics and treatment effect estimates from a dataset used in:

Benjamin A. Olken. 2007. “Monitoring Corruption: Evidence from a Field Experiment in Indonesia.”
Journal of Political Economy 115: 300-249

Note: You can download the data file on the class website at:

<http://www.sekhon.berkeley.edu/causalinf/fa2009/hw1data.RData>
The data are contained in an object called `data`.

This objective of this experiment was to evaluate two interventions thought to reduce corruption in road building projects in Indonesian villages. The two treatments were audits by engineers and efforts to encourage communities to monitor the projects themselves. i.e. “grassroots participation”. While the actual experimental design is somewhat involved, in this exercise we will focus on the intervention designed to increase community monitoring. The full paper can be found here:

<http://econ-www.mit.edu/files/2913>

Olken describes the intervention to be analyzed as follows:

...[T]he experiments sought to enhance participation at “accountability meetings”, the village-level meetings in which project officials account for how they spent project funds. ...[H]undreds of invitations to these meetings were distributed throughout the village, to encourage direct participation in the monitoring process and to reduce elite dominance of the process.

Note that residents in treatment villages were notified about these meetings *before* construction began, but after the total budget was decided. While the total budget was allocated before assignment to treatment, decisions about how the budget was to be spent was decided after the intervention.

The main dependent variable is `pct.missing`, which is a measure of the difference between what the villages claimed they spent on road construction and an independent estimate of what the villages actually spent. Treatment status is indicated by the dummy variable `treat.invite`, which takes a value of 1 if the village received the intervention and 0 if it did not.

Table 1: Variables

Variable	Definition
<code>pct.missing</code>	Percent expenditures missing
<code>treat.invite</code>	Treatment assignment
<code>head.edu</code>	Village head education
<code>mosques</code>	Mosques per 1,000
<code>pct.poor</code>	Percent of households below the poverty line
<code>total.budget</code>	Total budget (Rp. million)
<code>share.total.unskilled</code>	Share of road construction expenses spent on unskilled labor

Other variables in the dataset are listed in Table ??.

Problem 4: Check whether the variables in the dataset have missing values, and report the number of missing values by variable.

Problem 5: Report the minimum, maximum, mean, and standard deviation of the *pre*-treatment covariates in the data set, separately for treatment and controls. Are treatment and control units similar in terms of these characteristics? Be sure that you only include variables that were measured before the treatment was applied.

If you can, use a “for loop” or the `apply` function to calculate these summary statistics.

Problem 6:

- a. Report the average difference in the outcome variable by treatment assignment status (the “treatment effect”). What is the standard error of this estimate?
- b. Now estimate the treatment effect using a regression model with no covariates. Is this estimate different from the difference-in-means estimate? Are the standard errors of the two estimates different?
- c. Finally, estimate the treatment effect using a regression model, but this time include all pre-treatment covariates as additional independent variables. What is your estimated treatment effect? What is the standard error of this estimate? Is this estimate substantively different from the difference-in-means estimate?
- d. Bonus question: Is there a reason to prefer one of these methods of estimating treatment effects over the others?

Problem 7: In a couple of sentences, what can you conclude about the effectiveness of this intervention?