PS C236A / Stat C239A
Problem Set 1
Due: Sept. 21, 2012

Instructions

This assignment is due **4 pm Friday, Sept. 21.** You may submit your analytical work either electronically or in paper form. Electronic versions must be sent as a .pdf to <`jahenderson[at]berkeley.edu`>. Paper copies should be placed in my mailbox in 210 Barrows. For the computing portion of the assignment, you must submit a fully executable version of all .R code, along with any data used in the code (excepting that provided through the course webpage) to the email above. All files for each assignment sent electronically should be included in one omnibus email, with the subject line containing the course and homework number, and your last name (e.g., PS239A/STAT236A: HW1 - Obama).

You are encouraged to work together in groups to complete the assignments. However, you must hand in your own individual answers. Photocopies and other reproductions of someone else's answers are not acceptable. Please also list the names of everyone with whom you have collaborated on this assignment.

Potential Outcomes

Problem 1: Hooke's Law of elasticity for the restoring force of a spring out of equilibrium is $F = -kx$, with x being a measure of displacement, and k being a spring constant rate. It is an approximation. Compare Hooke's Law to a model of 'force' in political science, where a person's party identification, PID_i (e.g., an ordinal scale of seven points which ranges from strong Democrat to strong Republican), influences the rate of i's campaign giving to candidate j, denoted as d_{ij}. Frequently this is modeled as: $d_{ij} = \alpha + \gamma PID_i + \beta(v_i - m_j)^2 + \epsilon_{ij}$, where v_i and m_j control for the ideal policies i and j prefer. In this model, α, β and γ are parameters that are estimated, ϵ_{ij} is a stochastic term, and PID is measured through a random survey. Assume v_i and m_j are fixed covariates measured through a survey asking respondents to place themselves on an ideological scale. In real data of contributors, this model predicts a person's political donations extremely well. Is this sufficient for the donations equations to provide potential outcomes for PID? Why or why not? What about Hooke's Law for x? What (if any) additional assumptions are required in either case?

Problem 2: Imagine n people who live on the same street are randomly assigned to some treatment $T_i = \{0, 1\}$.

a. How many potential outcomes in total are there in this experiment *without* making the SUTVA assumption?

b. Now, assume there is interference only if $\sum_{i=1}^{n} T_i \geq \frac{n}{2}$. When this condition is met, how many potential outcomes does every i unit have?

c. Define *adjacent units* on this street to be each i's nearest neighbors $\{i-1, i+1\}$, where location on the street defines i's ordering. Assume interference for i occurs *only* if $T_{i-1} = T_{i+1} = 1$, that is both neighbors $i-1$ and $i+1$ are assigned treatment. Now how many potential outcomes are there for each i?

Problem 3: Consider a field experiment that compares treatments A and B. Suppose there are N subjects, indexed by $i = 1, ..., N$. 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 $\{1, ..., N\}$, with n elements; this group gets treatment A, so x_i is observed for i in S. Let T be a random subset of $\{1, ..., N\}$, 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.

b. What is the variance of the average treatment effect (ATE), i.e. var$(\bar{X} - \bar{Y})$, using the above notation?

c. 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)?

d. The variance calculated using the "usual" formula can be biased, but only in one direction. What is the direction of the bias in the "usual" variance estimate? Prove it. Under what conditions will this bias be 0?

Linear Regression

Problem 4: 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. Assume the correct model is:

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

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:

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

a. Write down or describe the design matrix for the correct model of the world (model 1) as well as the naive model (model 2).

b. Show that $\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i$

c. Which, if any, assumptions and conditions are necessary for part (b) to be true?

d. 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.

e. Is this estimate of β_1 from (d) BLUE? Why or why not?

f. What is cov$(\hat{\beta}_1|X)$?

Problem 5: Researchers run a randomized experiment to measure the effect of school vouchers T_t in the 8th grade on student test scores S_{t+2} by grade 10.

a. Researchers first estimate: $S_{t+2} = \alpha + \beta_1 T_t + \beta_2 S_{t-1} + \epsilon$. Assume there is successful randomization, no compliance problem, and the data are full rank. Is $\hat{\beta}_1$ unbiased? Prove it.

b. Subsequently, researchers estimate: $S_{t+2} = \alpha + \beta_1 T_t + \beta_2 S_{t-1} + \beta_3 S_{t+1} + \epsilon$. Again, assume successful randomization, full compliance, and full rank data. Is $\hat{\beta}_1$ unbiased? Prove it.

c. *Bonus:* In part (b), is $\hat{\beta}_1$ unbiased asymptotically? Prove it.

Applications In R

Problem 6: 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?

b. What is the variance of the average treatment effect, using the formula you derived in part 3(b) from the above question? What is the variance using the "usual" formula written in 3(d) from the above question?

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.

d. Calculate the estimated treatment effect for every possible combination of treatment assignment. Summarize this distribution of estimates using a plot.

e. What is the "true" variance of the treatment effect estimate? Calculate this using your treatment effect estimates from part (d).

Olken Data

For Problems 7 and 8, you will use R to calculate 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://sekhon.berkeley.edu/causalinf/data/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 2: Variables

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

Other variables in the dataset are listed in Table 2.

Problem 7: Complete the following computations, using your own custom-written function(s). Where appropriate, your function(s) should be flexible to input data up to $n \times k$ dimensions, and produce well-formatted results. (Note: You should not use any existing statistical functions, i.e., `lm(),t.test()`; mathematic functions are okay, i.e. `mean(),var()`. When in doubt, if it produces a p-value, don't use it.)

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. Is there a reason to prefer one of these methods of estimating treatment effects over the others? What can you conclude about the effectiveness of this intervention?

Problem 8: Making a selection on X observables assumption, the average treatment effect for the treated (ATT) is defined as:

$$\tilde{\tau}|(T_i = 1) = E\left\{E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0) \mid T_i = 1\right\}$$

Estimate this quantity by OLS in R, for the following two models from the Olken experimental data. Unlike the previous question, feel free to use any function in R. In the models below:

$$Y = \texttt{pct.missing}$$

$$T = \texttt{treat.invite}$$

$$X_1 = \texttt{unskilled.transformed}$$

$$X_2 = \texttt{mosques}$$

a. $Y_i = \alpha + \tau T_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$

b. $Y_i = \alpha + \tau T_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma X_{1i} * T_i + \epsilon_i$