

PS C236A/ Stat C239A

Problem Set 1 Solutions

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:

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

- 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).

For the true model:

$$\begin{pmatrix} 1 & \text{education level}_1 & \text{intelligence}_1 \\ 1 & \text{education level}_2 & \text{intelligence}_2 \\ \vdots & \vdots & \vdots \\ 1 & \text{education level}_n & \text{intelligence}_n \end{pmatrix}$$

For the naive model:

$$\begin{pmatrix} 1 & \text{education level}_1 \\ 1 & \text{education level}_2 \\ \vdots & \vdots \\ 1 & \text{education level}_n \end{pmatrix}$$

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

First order conditions gives us that:

$$(X^T X) \hat{\beta} = X^T Y$$

$$\Rightarrow X^T (Y - X \hat{\beta}) = 0$$

However, using the fact that:

$$e = Y - X \hat{\beta}$$

$$\Rightarrow X^T e = 0$$

Now, if we write out the matrix $X^T e$ then we see:

$$X^T e = \begin{pmatrix} \sum_{i=1}^N e_i \\ \sum_{i=1}^N x_i * e_i \\ \vdots \\ \sum_{i=1}^N x_n * e_i \end{pmatrix} = 0$$

$$\Rightarrow \sum_{i=1}^N e_i = 0$$

$$\begin{aligned}
&\text{where } e_i = y_i - \hat{\beta}x_i = y_i - \hat{y}_i \\
&\Rightarrow \sum_{i=1}^N y_i = \sum_{i=1}^N (\hat{y}_i + e_i) \\
&\Rightarrow \sum_{i=1}^N y_i = \sum_{i=1}^N \hat{y}_i + \sum_{i=1}^N e_i \\
&\Rightarrow \sum_{i=1}^N y_i = \sum_{i=1}^N \hat{y}_i \\
&\Rightarrow \frac{1}{N} \sum_{i=1}^N y_i = \frac{1}{N} \sum_{i=1}^N \hat{y}_i \\
&\Rightarrow \bar{\hat{y}} = \bar{y}
\end{aligned}$$

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

The fact that $\bar{\hat{y}} = \bar{y}$ is true by construction. So long as we have full rank and can calculate an estimator, and so long as we include an intercept, it will follow that $\bar{\hat{y}} = \bar{y}$.

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.

By running the naive model instead of the true model, you are introducing “omitted variable bias” into your estimate of the effect of education. The proof is as follows:

First, for simplicity, we will still assume that $E[\epsilon_{2i}|X] = 0$ where X is our design matrix for the naive model. We will also rewrite the true model as $X\beta + X\gamma_2 + \epsilon_{1i}$ where X is as defined above and Z is our intelligence vector.

Now we want to see if it is true that $E[\hat{\beta}|X] = \beta$

$$\begin{aligned}
E[\hat{\beta}|X] &= E[(X^T X)^{-1} X^T Y | X] \\
&= E[(X^T X)^{-1} X^T (X\beta + Z\gamma_2 + \epsilon_{1i}) | X] \\
&= E[(X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T Z\gamma_2 + (X^T X)^{-1} X^T \epsilon_{1i} | X] \\
&= E[\beta | X] + (X^T X)^{-1} E[X^T Z\gamma_2 | X] + (X^T X)^{-1} E[X^T \epsilon_{1i} | X] \\
&= \beta + (X^T X)^{-1} E[X^T Z\gamma_2 | X] \\
\Rightarrow E[\hat{\beta}|X] &\neq \beta
\end{aligned}$$

So, from above we see that we introduce bias. In this case, the bias will be however large $(X^T X)^{-1} E[X^T Z\gamma_2 | X]$ is, but if we assume that education level and intelligence are positively correlated, then we’d have a positive bias. Therefore, our estimate of $\hat{\beta}$ would be too large. If the $E[\epsilon_{2i}|X] \neq 0$, then the problem will only be worse.

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.

True. So long as the design matrix has full rank, the mechanics of OLS will go through and the computer will be able to calculate the projection matrix and determine the fitted values and the residuals. The assumptions are what allow us to say that our estimate is a “BLUE” estimator. Without the assumptions, we will get an estimate of $\hat{\beta}$, however it is unclear *what* that estimate means.

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 —

a. $\text{cov}(x, y) = r s_x s_y$

The correlation coefficient is $r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y} \right)$. Multiplying both sides of the expression by $s_x \cdot s_y$, produces $r \cdot s_x \cdot s_y = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$, with the second expression being the formula for covariance.

b. The slope of the regression line for predicting y from x is

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

In the bivariate case, the regression slope is $\frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$. This can be re-expressed as $\frac{\frac{1}{n} \cdot \sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \cdot \sum_i^n (x_i - \bar{x})^2}$. Notice that the numerator now equals the formula for $\text{cov}(x, y)$ and the denominator equals the formula for $\text{var}(x)$.

c. $\text{var}(x) = \text{cov}(x, x)$ This is true by the definition of variance.

d. $\text{cov}(x, y) = \overline{xy} - \bar{x}\bar{y}$

Note that $(x_i - \bar{x})(y_i - \bar{y}) = x_i y_i + x_i \bar{y} - \bar{x} y_i - \bar{x} \bar{y} = x_i y_i - \bar{x} \bar{y}$. Plug into the covariance formula: $\frac{1}{n} \sum_i^n (x_i y_i - \bar{x} \bar{y}) = \frac{1}{n} \sum_i^n x_i y_i - \frac{1}{n} \sum_i^n \bar{x} \bar{y} = \overline{xy} - \bar{x} \bar{y}$

e. $\text{var}(x) = \overline{x^2} - \bar{x}^2$

Just use the result from *d* and replace y with x .

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.

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

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.

Other variables in the dataset are listed in Table 1.

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

Variable	Missing.Values
id	0
treat.invite	0
pct.missing	90
share.total.unskilled	73
head.edu	5
mosques	2
pct.poor	7
total.budget	2

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.

	Treated.Mean	Control.Mean	Treated.SD	Control.SD	Treated.Max	Control.Max	Treated.Min	Control.Min
head.edu	11.43	11.50	2.71	2.70	20.00	20.00	6.00	6.00
mosques	1.41	1.47	0.83	0.83	6.89	4.52	0.00	0.12
pct.poor	0.41	0.41	0.21	0.21	0.95	0.94	0.02	0.03
total.budget	80.22	81.98	56.65	41.21	890.24	273.48	8.76	19.07

Note that the “`share.total.unskilled`” variable is not pre-treatment, since decisions about how the funds would be allocated were decided after the intervention. Controlling for a post-treatment variable will introduce bias

into the treatment effect estimate, even in a randomized experiment¹. The treatment and control units appear to be broadly similar on pre-treatment covariates, i.e. *balanced*. While there are some dissimilarities, we can't know if these differences are large relative to what we would observe by chance without more formal statistical tests.

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?

The estimate of the treatment effect is -0.023. The standard error of this estimate (using the formula for the standard error for the difference of two independent samples) is about 0.03.

- 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?

The simple regression estimate is -0.023 with a standard error of about 0.033. The simple regression estimate is the same as the difference in means estimate. The standard errors are not equal, but the difference is very small.

- 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?

The estimate is -0.026 with a standard error of about 0.033. The multivariate regression estimate is slightly larger than the difference-in-means estimate and the standard error is almost identical.

- d. Bonus question: Is there a reason to prefer one of these methods of estimating treatment effects over the others?

In this example, each method produces similar results. In general, however, the estimator most consistent with the Neyman-Rubin potential outcomes model is the simple difference-in-means estimator. We will cover this further in future sections.

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

The treatment effect estimate is small and negative, perhaps indicating that the intervention suppressed some corruption. The standard error is large relative to the point estimate, however, so this negative effect could easily be a result of chance. We cannot reject the null hypothesis of no effect of the intervention at conventional levels of statistical significance.

¹For more on this issue, see Paul Rosenbaum. 1984. “The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by Treatment.” *Journal of the Royal Statistical Association* 147: 656-666.

Example R Code:

```
#load library for making tables
library(xtable)
rm(list=ls(all=TRUE))

#load data
load("/Users/ElGuapo/Documents/projects/ps236/hw/hw1/hw1data.RData")

#Problem 4
##Check the number of missing values using "apply"
missing.num <- data.frame(Missing.Values = apply(data, 2, function(x) sum(is.na(x))))
xtable(missing.num)

#Problem 5
##create 2 datasets for pre-treatment covariates
treated.covar <- data[data$treat.invite == 1, c(5, 6, 7, 8)]
control.covar <- data[data$treat.invite == 0, c(5, 6, 7, 8)]

#calculate means
treated.covar.mean <- apply(treated.covar, 2, mean, na.rm=T)
control.covar.mean <- apply(control.covar, 2, mean, na.rm=T)
#calculate standard deviations
treated.covar.sd <- apply(treated.covar, 2, sd, na.rm=T)
control.covar.sd <- apply(control.covar, 2, sd, na.rm=T)
#calculate max
treated.covar.max <- apply(treated.covar, 2, max, na.rm=T)
control.covar.max <- apply(control.covar, 2, max, na.rm=T)
#calculate min
treated.covar.min <- apply(treated.covar, 2, min, na.rm=T)
control.covar.min <- apply(control.covar, 2, min, na.rm=T)

#make table of data
summary.stat <- data.frame( Treated.Mean = treated.covar.mean,
  Control.Mean = control.covar.mean,  Treated.SD = treated.covar.sd,
  Control.SD = control.covar.sd,  Treated.Max = treated.covar.max,
  Control.Max = control.covar.max,  Treated.Min = treated.covar.min,
  Control.Min = control.covar.min )
xtable(summary.stat)

#Problem 6
##Part a
##Calculate the treatment effect estimate using a simple difference of means
itt.estimate <- mean(data$pct.missing[data$treat.invite == 1], na.rm = TRUE)
  - mean(data$pct.missing[data$treat.invite==0], na.rm = TRUE)
#for Standard error, you need number of units in control, number in treatment,
#sample variance of treated units, sample variance of control units
n.treat <- sum(data$treat.invite)
n.control <- sum(1 - data$treat.invite)
var.treated <- var(data$pct.missing[data$treat.invite == 1], na.rm = TRUE) / n.treat
var.control <- var(data$pct.missing[data$treat.invite == 0], na.rm = TRUE) / n.control
itt.se <- sqrt(var.treated + var.control)
```

```
##Part b
reg.simple <- lm(pct.missing ~ treat.invite, data = data)
summary(reg.simple)

##Part c
reg.covar <- lm(pct.missing ~ treat.invite + head.edu + mosques + pct.poor + total.budget,
  data = data)
summary(reg.covar)
```