PS C236A / Stat C239A
Problem Set 1 - Solutions

1: For PID to elicit potential outcomes, we must assume (a) that party identity is able to be manipulated, and can be done so in a way that does not inherently affect other attributes. Good prediction here is not the same as good prediction in a scientific lab, since we don't know what we are manipulating that changes the amount of donations we observe. For Hooke's Law we still must make some assumptions to elicit potential outcomes in F from manipulating x weight on a spring. We must at least assume (b) temporal stability between the result of an experiment at t and at $t+1$. We also typically invoke a (c) homogeneity assumption about different springs and weights being tested that are identical in every relevant way but in x. Also, there is an implicit (d) measurement assumption, which is that the scale of measurement error is small relative to the size of the intervention examined or the outcome expected, under scientific control; (d) is actually a special case of (c).

2: a) Without making a SUTVA assumption, there are 2^n possible ways that treatment can be assigned to the n units. Each unit i can be assigned either treatment $T_i = 1$ or $T_i = 0$. There is no restriction in the problem description on the number of units that can be assigned treatment. Since there are n units in total, there are $n2^n$ total potential outcomes in this experiment.

b) Now for the case where there is interference if and only if $\sum_{i=1}^{n} T_i \geq n/2$

We break this down into two cases.

Case 1: $\sum_{i=1}^{n} T_i < n/2$. The outcome when unit i is treated is the same for all treatment assignments satisfying this case with $T_i = 1$. The outcome when unit i is not treated is the same for all treatment assignments satisfying this case with $T_i = 0$. Thus, there are two outcomes for each unit when treatment assignment satisfies the above condition, the one when the unit is assigned to treatment and the one when assigned to control.

Case 2: $\sum_{i=1}^{n} T_i \geq n/2$. When this happens, each treatment assignment has its own potential outcome. It follows that the number of potential outcomes is

$$\begin{aligned}
\#\text{outcomes} &= \#\text{outcomes when } \sum_{i=1}^{n} T_i < n/2 + \#\text{outcomes when } \sum_{i=1}^{n} T_i \geq n/2 \\
&= 2 + \sum_{i=\lceil n/2 \rceil} \#\text{ways exactly } i \text{ treatments can be assigned} \\
&= 2 + \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} \\
&= \begin{cases} 2 + 2^{n-1} & n \text{ is odd} \\ 2 + 2^{n-1} + \frac{1}{2}\binom{n}{n/2} & n \text{ is even} \end{cases}
\end{aligned}$$

where $\lceil x \rceil$ denotes "round x up to the nearest whole number."

These formulas can be derived by the properties:

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n$$

and

$$\sum_{i=0}^{k} \binom{n}{i} = \sum_{i=n-k}^{n} \binom{n}{i}.$$

c) Now suppose there is interference only if both adjacent units receive treatment.

There are two possible interpretations of this problem. For the first interpretation, only neighbors of a unit can affect which potential outcome is observed by the unit.

In this case, units at the end of the street ($i = 1$ and $i = n$) only have one neighbor, and thus, have two potential outcomes: one for receiving treatment and one for receiving control. Those units with two neighbors have four potential outcomes: Receiving treatment while both neighbors receive treatment; receiving control while both neighbors receive treatment; receiving treatment, and neighbors do not both receive treatment, receiving control, and neighbors do not both receive treatment.

It follows that the the total number of potential outcomes (for $n \geq 2$) is:

$$2 + \sum_{i=2}^{n-1} 4 + 2 = 4 + 4(n-2) = 4(n-1)$$

Another interpretation suggests that the neighboring condition continues inductively. For example: Unit i's response depends on whether $i+1$ and $i-1$ both get treatment. However, unit $i+1$'s response is affected by whether units i and $i+2$ get treatment; the outcome of i when units $i-1$, i, $i+1$ and $i+2$ all get treatment may be different from the outcome where units $i-1$, i, $i+1$ get treatment, but $i+2$ gets control. Thus, we obtain the following cases:

Again, units at the end of the street have two potential outcomes. We break up the cases for units $2 \leq i \leq n-1$ as follows:

Case 1: Units $i-1$ and units $i+1$ are not both treated.

In this case, unit i has two potential outcomes, response under treated and response under control.

Case 2: Units $i-1$ and units $i+1$ are both treated, but unit i is assigned to control.

In this case, units $i-1$ and units $i+1$'s response are not affected by neighboring units, and so, there is only one outcome.

Case 3: Units $i-1$, i and $i+1$ are all treated.

In this case, the response of i is also going to depend on whether $i-2$ is treated (as if $i-2$ is treated, then $i-1$ will have interference, otherwise it will not). Likewise, it will also depend on whether $i+2$ is treated. Moreover, if $i-2$ is also treated, the response of i will also depend on whether or not $i-3$ is treated, and so on.

It follows that, when counting the number of cases to the left of unit i, it suffices to count the number of possible locations of the first control unit left of unit $i-1$ (including the possibility of no control unit to the left of unit i). There will be $i-1$ cases in total. Similarly, there will be $n-i$ cases to consider to the right of unit i. To get the total number of potential outcomes under this case, we multiply these numbers together: there are $(i-1)(n-i)$ different potential outcomes for unit i

Combining cases 1, 2, 3, and the case where units are on the end of the street, the total number of potential outcomes is:

$$2 + \sum_{i=2}^{n-1} [2 + 1 + (i-1)(n-i)] + 2$$

Using the identities

$$\sum_{i=1}^{n} i = n(n+1)/2$$

and

$$\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$$

we simplify this sum.

$$\begin{aligned}
&\sum_{i=2}^{n-1} 2 + 1 + (i-1)(n-i) = \sum_{i=2}^{n-1}[3 - n + ni + i - i^2] \\
= \; & -\sum_{i=2}^{n-1}[n-3] + \sum_{i=2}^{n-1}(n+1)i - \sum_{i=2}^{n-1} i^2 \\
= \; & -(n-2)(n-3) + (n+1)(n(n+1))/2 - (n+1)(n+1) - n(n+1)(2n+1)/6 + n^2 + 1 \\
= \; & -n^2 + 5n - 6 - n^2 - 2n - 1 + n^2 + 1 + (n+1)(n(n+1))/2 - n(n+1)(2n+1)/6 \\
= \; & -n^2 + 3n - 6 + (n+1)(n(n+1))/2 - n(n+1)(2n+1)/6
\end{aligned}$$

And so, the total number of potential outcomes simplifies to:

$$\begin{aligned}
& 4 - n^2 + 3n - 6 + (n+1)(n(n+1))/2 - n(n+1)(2n+1)/6 \\
= \; & -n^2 + 3n - 2 + n(n+1)((n+1)/2 - (2n+1)/6) \\
= \; & n(n+1)((n+1)/2 - (2n+1)/6) - (n-2)(n-1)
\end{aligned}$$

3) a) The average treatment effect parameter is

$$\bar{\tau} = \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}\sum_{i=1}^{N} y_i.$$

That is, $\bar{\tau}$ is the difference of the average effect of treatment A over all N subjects and the average effect of treatment B over all N subjects.

b)

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

c) From part b),

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

The right-hand side is the difference between our formula and the usual formula. This quantity is not identifiable as the covariance of x and y is never observed.

d) Note that,

$$\sigma^2 + \tau^2 - 2\text{cov}(x,y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x,y) = \text{var}(x-y) \geq 0.$$

Thus, the "usual" estimate greater than or equal to the truth asymptotically. The bias will be 0 when $\text{var}(x-y) = 0$. This is only true if $x_i = y_i$ for $1 \leq i \leq N$. That is, this is only true if the sharp null of no treatment effect of B relative to A for all subjects i in the population is true.

4) a) The design matrix for the correct model is

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

and for the incorrect model is

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

b) It is equivalent to show

$$\sum_{i=1}^{N}(y_i - \hat{y}_i) = 0.$$

Consider

$$Y - \hat{Y} = Y - X(X'X)^{-1}X'Y = (1 - X(X'X)^{-1}X)Y.$$

Note that

$$\begin{aligned} X'(Y - \hat{Y}) =& X'(1 - X(X'X)^{-1}X')Y = (X' - X'X(X'X)^{-1}X')Y \\ =& (X' - X')Y = \mathbf{0}. \end{aligned}$$

Since X has a column of 1's, then we must have that

$$(1, 1, \ldots, 1)(Y - \hat{Y}) = \sum_{i=1}^{N}(y_i - \hat{y}_i) = 0.$$

c) Part b) is guaranteed to be true by the properties of OLS. All that is necessary is that the design matrix is full rank and that the model has an intercept term (hence, the design matrix has a column of ones).

d,e) Although it's not explicitly stated, we assume

$$E(\epsilon_{i1}|\text{correct design matrix}) = 0.$$

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:

Let X denote the incorrect model, let $Z = (\text{intelligence}_1, \text{intelligence}_2, \ldots, \text{intelligence}_N)'$, let $\beta_0 = (\alpha_2, \beta)'$, and let $\epsilon_1 = (\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{1n})$. The expectation of the OLS estimates for the coefficients in the second model (assuming that the values of the covariates is fixed) is

$$\begin{aligned} E(\hat{\beta}_0|X, Z) &= E((X'X)^{-1}X'Y) = E((X'X)^{-1}X'(X\beta_0 + Z\gamma_2 + \epsilon_1)|X, Z) \\ &= E((X'X)^{-1}(X'X)\beta_0 + (X'X)^{-1}(X'Z)\gamma_2 + (X'X)^{-1}X'\epsilon_1|X, Z) \\ &= E(\beta_0 + (X'X)^{-1}(X'Z)\gamma_2 + (X'X)^{-1}X'\epsilon_1|X, Z) \\ &= \beta_0 + (X'X)^{-1}(X'Z)\gamma_2 + 0 \neq \beta_0. \end{aligned}$$

In particular, our parameter β is not estimated unbiasedly unless the second row of the matrix $(X'X)^{-1}(X'Z) = 0$. We now try to identify exactly when this happens.
Let ℓ_i denote education level$_i$ and let g_i denote intelligence$_i$. We find that

$$(X'X) = \begin{pmatrix} n & \sum \ell_i \\ \sum \ell_i & \sum \ell_i^2 \end{pmatrix}$$

and so

$$(X'X)^{-1} = \frac{1}{n\sum \ell_i^2 - (\sum \ell_i)^2} \begin{pmatrix} \sum \ell_i^2 & -\sum \ell_i \\ -\sum \ell_i & n \end{pmatrix}.$$

It follows that the bottom row of the matrix $(X'X)^{-1}(X'Z)$ is proportional to

$$(-\sum \ell_i, n)(\sum g_i, \sum \ell_i g_i)' = n\sum \ell_i g_i - \sum \ell_i \sum g_i.$$

It is easy to show that this is equation zero if and only if the correlation between $(\ell_i)_{i=1}^n$ and $(g_i)_{i=1}^n$ is zero.

Since intelligence and education are positively correlated, and since the estimate of β is unbiased if and only if intelligence and education are uncorrelated it follows that our estimate of β is biased. Thus, our estimate of β is not BLUE, since it is not unbiased.

f) Note that:

$$\begin{aligned} \hat{\beta}_0 - E(\hat{\beta}_0) &= (X'X)^{-1}X'Y - (\beta_0 + (X'X)^{-1}(X'Z)\gamma_2) \\ &= (X'X)^{-1}X'(X\beta_0 + Z\gamma_2 + \epsilon_1) - \beta_0 - (X'X)^{-1}(X'Z)\gamma_2 \\ &= \beta_0 + (X'X)^{-1}X'Z\gamma_2 + (X'X)^{-1}X'\epsilon_1 - \beta_0 - (X'X)^{-1}(X'Z)\gamma_2 \\ &= (X'X)^{-1}X'\epsilon_1. \end{aligned}$$

It follows that:

$$\begin{aligned} \text{cov}(\hat{\beta}_0|X,Z) &= E[(\hat{\beta}_0 - E(\hat{\beta}_0))(\hat{\beta}_0 - E(\hat{\beta}_0))'|X,Z] \\ &= E[((X'X)^{-1}X'\epsilon_1)((X'X)^{-1}X'\epsilon_1)'|X,Z] \\ &= E[((X'X)^{-1}X'\epsilon_1\epsilon_1'X((X'X)^{-1})')|X,Z] \\ &= (X'X)^{-1}X'E[\epsilon_1\epsilon_1'|X,Z]'X((X'X)^{-1})' \\ &= (X'X)^{-1}X'I_{n\ x\ n}\sigma_1^2X((X'X)^{-1})' \\ &= \sigma_1^2(X'X)^{-1}X'I_{n\ x\ n}X((X'X)^{-1})' \\ &= \sigma_1^2(X'X)^{-1}X'X((X'X)^{-1})' \\ &= \sigma_1^2(X'X)^{-1}. \end{aligned}$$

The last equality uses the fact that $(X'X)^{-1}$ is symmetric. The covariance of our estimate for β is just simply the entry in the second row and second column of $(X'X)^{-1}\sigma_1^2$. From part d), it follows that the covariance of our estimate is

$$\sigma^2 \frac{n}{n\sum \ell_i^2 - (\sum \ell_i)^2} = \sigma^2 \frac{1}{\sum \ell_i^2 - n(\bar{\ell})^2} = \sigma^2 \frac{1}{\sum (\ell_i - \bar{\ell})^2}.$$

5) We will prove these results in a similar way as in problem 4. Let X denote the design matrix:

$$X = \begin{pmatrix} 1 & \mathrm{T}_1 \\ 1 & \mathrm{T}_2 \\ \vdots & \vdots \\ 1 & \mathrm{T}_N \end{pmatrix}$$

Let Z_a denote the design matrix for the model in part a).

$$Z_a = \begin{pmatrix} 1 & \mathrm{T}_1 & S_{t-1,1} \\ 1 & \mathrm{T}_2 & S_{t-1,2} \\ \vdots & \vdots & \vdots \\ 1 & \mathrm{T}_N & S_{t-1,N} \end{pmatrix}$$

Let Z_b denote the design matrix for the model in part b).

$$Z_b = \begin{pmatrix} 1 & \mathrm{T}_1 & S_{t-1,1} & S_{t+1,1} \\ 1 & \mathrm{T}_2 & S_{t-1,2} & S_{t+1,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathrm{T}_N & S_{t-1,N} & S_{t+1,N} \end{pmatrix}$$

Note, if we estimate the model

$$S_{t+2,i} = \alpha + T_i\beta_1 + \epsilon_i \tag{1}$$

using OLS, we find that our estimate for β_1 is

$$\frac{\sum(T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}})}{\sum(T_i - \overline{T})^2} = \left(\frac{1}{\sum T_i}\sum_{i:T_i=1} S_{t+2,i}\right) - \left(\frac{1}{N - \sum T_i}\sum_{i:T_i=0} S_{t+2,i}\right)$$

which is an unbiased estimate (asymptotically) for the ATE.

a) To show that the estimate for β_1 is unbiased, we will show that the estimate of this model is the same (asymptotically) as the estimate for β_1 in **??**. Are argument will be entirely brute-forced; there are far more elegant ways to prove this result.

Let $\beta_0 = (\alpha, \beta_1, \beta_2)'$. The OLS estimate for β_0 is

$$(Z_a'Z_a)^{-1}Z_a'S_{t+2}$$

We are interested in the second row of this estimate. Now (thanks to the magic of Wolfram Alpha), the second row of $(Z_a'Z_a)^{-1}$ is

$$C\left(\left(\sum T_i \sum S_{t-1,i}^2 - \sum S_{t-1,i}\sum S_{t-1,i}T_i\right), \left((\sum S_{t-1,i})^2 - n\sum S_{t-1,i}^2\right), \left(n\sum S_{t-1,i}T_i - \sum T_i \sum S_{t-1,i}\right)\right)$$

where

$$C = \frac{1}{-n\sum T_i^2 \sum S_{t-1,i}^2 + n(\sum T_i S_{t-1,i})^2 + (\sum T_i)^2 \sum S_{t-1,i}^2 - 2\sum T_i \sum S_{t-1,i} \sum T_i S_{t-1,i} + (\sum S_{t-1,i})^2 \sum T_i^2}$$

Also,

$$Z_a'S_{t+2} = \left(\sum S_{t+2,i}, \sum T_i S_{t+2,i}, \sum S_{t-1,i}S_{t+2,i}\right)'$$

It follows that the second row of the OLS estimate is proportional to

$$\begin{aligned} & \sum S_{t+2,i} \sum T_i \sum S_{t-1,i}^2 - \sum S_{t+2,i} \sum S_{t-1,i} \sum S_{t-1,i}T_i \\ + \quad & \sum T_i S_{t+2,i} (\sum S_{t-1,i})^2 - n\sum T_i S_{t+2,i} \sum S_{t-1,i}^2 \\ + \quad & n\sum S_{t-1,i}S_{t+2,i} \sum S_{t-1,i}T_i - \sum S_{t-1,i}S_{t+2,i} \sum T_i \sum S_{t-1,i} \end{aligned}$$

Now, treatment should be (asymptotically) uncorrelated with $S_{t-1,i}$, since treatment is assigned independently of previous test scores. Thus, the constant C simplifies to

$$\begin{aligned} C &= \frac{1}{-n\sum T_i^2 \sum S_{t-1,i}^2 + n(\sum T_i S_{t-1,i})^2 + (\sum T_i)^2 \sum S_{t-1,i}^2 - 2\sum T_i \sum S_{t-1,i} \sum T_i S_{t-1,i} + (\sum S_{t-1,i})^2 \sum T_i^2} \\ &= \frac{1}{\begin{array}{l} \sum S_{t-1,i}^2((\sum T_i)^2 - n\sum T_i^2) + \sum T_i S_{t-1,i}(n(\sum T_i S_{t-1,i}) - \sum T_i \sum S_{t-1,i}) \\ - \quad \sum T_i \sum S_{t-1,i} \sum T_i S_{t-1,i} + (\sum S_{t-1,i})^2 \sum T_i^2 \end{array}} \\ &= \frac{1}{-n\sum S_{t-1,i}^2 \sum (T_i - \overline{T})^2 + \sum T_i S_{t-1,i}(0) - \sum S_{t-1,i}(\sum T_i \sum T_i S_{t-1,i} - \sum S_{t-1,i} \sum T_i^2)} \\ &= \frac{1}{-n\sum S_{t-1,i}^2 \sum (T_i - \overline{T})^2 - \sum S_{t-1,i}(\sum T_i \sum T_i S_{t-1,i} - \sum S_{t-1,i} \sum T_i^2)} \end{aligned}$$

and we can simplify

$$\begin{aligned}
& \sum S_{t+2,i} \sum T_i \sum S_{t-1,i}^2 - \sum S_{t+2,i} \sum S_{t-1,i} \sum S_{t-1,i} T_i \\
& + \sum T_i S_{t+2,i} (\sum S_{t-1,i})^2 - n \sum T_i S_{t+2,i} \sum S_{t-1,i}^2 \\
& + n \sum S_{t-1,i} S_{t+2,i} \sum S_{t-1,i} T_i - \sum S_{t-1,i} S_{t+2,i} \sum T_i \sum S_{t-1,i} \\
= \; & - \sum S_{t-1,i}^2 (n \sum T_i S_{t+2,i} - \sum S_{t+2,i} \sum T_i) \\
& - \sum S_{t-1,i} (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i}) \\
& + \sum S_{t-1,i} S_{t+2,i} (n \sum S_{t-1,i} T_i - \sum T_i \sum S_{t-1,i}) \\
= \; & - n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}}) \\
& - \sum S_{t-1,i} (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i})
\end{aligned}$$

Moreover,

$$\begin{aligned}
& n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})^2 \sum S_{t-1,i} (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i}) \\
= \; & - \sum S_{t-1,i}^2 ((\sum T_i)^2 - n \sum T_i^2) \sum S_{t-1,i} (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i}) \\
= \; & - \sum S_{t-1,i}^2 \sum S_{t-1,i} ((\sum T_i)^2 - n \sum T_i^2) (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i}) \\
= \; & \sum S_{t-1,i}^2 \sum S_{t-1,i} (n \sum T_i^2 - (\sum T_i)^2) (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i}) \\
= \; & \sum S_{t-1,i}^2 \sum S_{t-1,i} (n \sum T_i^2 \sum S_{t+2,i} \sum S_{t-1,i} T_i + (\sum T_i)^2 \sum T_i S_{t+2,i} \sum S_{t-1,i} \\
& - (\sum T_i)^2 \sum S_{t+2,i} \sum S_{t-1,i} T_i - n \sum T_i^2 \sum T_i S_{t+2,i} \sum S_{t-1,i}) \\
= \; & \sum S_{t-1,i}^2 \sum S_{t-1,i} (\sum T_i^2 \sum S_{t+2,i} \sum S_{t-1,i} \sum T_i + n (\sum T_i) \sum T_i S_{t+2,i} \sum T_i S_{t-1,i} \\
& - (\sum T_i)^2 \sum S_{t+2,i} \sum S_{t-1,i} T_i - n \sum T_i^2 \sum T_i S_{t+2,i} \sum S_{t-1,i}) \\
= \; & \sum S_{t-1,i}^2 \sum S_{t-1,i} (n \sum T_i S_{t+2,i} - \sum S_{t+2,i} \sum T_i) (\sum T_i \sum T_i S_{t-1,i} - \sum S_{t-1,i} \sum T_i^2) \\
= \; & \sum S_{t-1,i}^2 (n \sum T_i S_{t+2,i} - \sum S_{t+2,i} \sum T_i) \sum S_{t-1,i} (\sum T_i \sum T_i S_{t-1,i} - \sum S_{t-1,i} \sum T_i^2) \\
= \; & n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}}) \sum S_{t-1,i} (\sum T_i \sum T_i S_{t-1,i} - \sum S_{t-1,i} \sum T_i^2)
\end{aligned}$$

Thus,

$$\frac{\sum S_{t-1,i} (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i})}{n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}})} = \frac{\sum S_{t-1,i} (\sum T_i \sum T_i S_{t-1,i} - \sum S_{t-1,i} \sum T_i^2)}{n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})^2}$$

And so,

$$\begin{aligned}
\hat{\beta} \; = \; & \frac{-n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}}) - \sum S_{t-1,i} (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i})}{-n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})^2 - \sum S_{t-1,i} (\sum T_i \sum T_i S_{t-1,i} - \sum S_{t-1,i} \sum T_i^2)} \\
= \; & \frac{(-n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}}))(1 + \frac{\sum S_{t-1,i} (\sum S_{t+2,i} \sum S_{t-1,i} T_i - \sum T_i S_{t+2,i} \sum S_{t-1,i})}{n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}})})}{(-n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})^2)(1 + \frac{\sum S_{t-1,i} (\sum T_i \sum T_i S_{t-1,i} - \sum S_{t-1,i} \sum T_i^2)}{n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})^2})} \\
= \; & \frac{-n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}})}{-n \sum S_{t-1,i}^2 \sum (T_i - \overline{T})^2} = \frac{\sum (T_i - \overline{T})(S_{t+2,i} - \overline{S_{t+2}})}{\sum (T_i - \overline{T})^2}
\end{aligned}$$

That is, the $\hat{\beta}$ in this model is the same as the $\hat{\beta}$ in the model without S_{t-1}. Thus, $\hat{\beta}$ is an unbiased estimate of β.

b) The process for part a) can be followed for part b). However, the correlation between T_i and $S_{t+1,i}$ is not zero, and in fact, is probably positive. Similar to 4d), the result will then be biased.

c) The estimate will be biased asymptotically. The non-zero correlation between T_i and $S_{t+1,i}$ will exist asymptotically, and so, the estimate in expectation will be biased.

```
#####################
#  PS C236A / STAT C239A
#
#  John Henderson
#   HW 1 - Solutions: 6 - 9
#
# Sept. 29, 2012
#####################

rm(list=ls())
# 6. Potential Outcomes from an Experiment

yt=c(2,6,33,17,2,54)
yc=c(1,2,13,14,10,3)

# (a)  unit-level and average effects
tau_i=yt-yc
print(tau_i)
#1  4 20  3 -8 51

tau_bar=mean(tau_i)
#11.833

# The ATE is reasonable, although a median treatment effect is helpful
#  since there are two large outliers, and the response surface is
#  thus skewed.

# (b) variance
var_yt=sum((yt-mean(yt))^2)/6
var_yc=sum((yc-mean(yc))^2)/6
cov_y=sum((yt-mean(yt))*(yc-mean(yc)))/6

var.usual=6/(6-1)*(var_yt/3 + var_yc/3)
var.true=6/(6-1)*(var_yt/3 + var_yc/3) + (1/(6-1))*(2*cov_y - var_yt - var_yc)

#Note that the true variance is smaller than the "usual" variance


# (c)

exp.fun=function(y_t=yt, y_c=yc, n.treat){
  #create a treatment vector
   treat.assign = matrix(0,ncol=1,nrow=length(y_t))
   #Randomly sample and assign treatment
   treat.assign[sample(1:length(treat.assign),n.treat)] = 1
   #produce observed values
   y_t.obs = y_t[treat.assign==1]
   y_c.obs = y_c[treat.assign==0]
   #calculate estimate average treatement effect
   ate = mean(y_t.obs) - mean(y_c.obs)
   #calculate the standard error
```

```
   ate.se = sqrt(var(y_t.obs)/length(y_t.obs) + var(y_c.obs)/length(y_c.obs))
   #create a list with our desired outputs
   output = list(y_t.obs = y_t.obs, y_c.obs=y_c.obs, ate = ate, ate.se = ate.se)
   #return the output
   return(output)
}

#Lets make sure the function works
set.seed(1005)
exp.fun(yt,yc,3)


# (d)
##calculate every possible permutation
perms = combn(6,3)

#create a matrix to hold all the average treatment effectx
ates = matrix()

##loop over every combination and calculate the treatment effect
for (i in 1:ncol(perms)){
ates[i] =  mean(yt[perms[,i]]) - mean(yc[-1* perms[,i]])
}

#plot a histogram showing the distribution of the treatment effects
hist(ates,breaks=8, col="darkgreen")

#Calculate the variance
var(ates)
# 85.433

# where true variance is 86.9


# 7 Olken Data and ATT with OLS

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

#load data
load(file=url("http://sekhon.berkeley.edu/causalinf/data/hw1data.RData"))

# (a) Average differences

y = data$pct.missing
tr = data$treat.invite

tau_bar=mean(y[tr==1],na.rm=T)-mean(y[tr==0],na.rm=T)

#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(tr)
n.control = sum(1 - tr)

var.treated = var(y[tr == 1], na.rm = TRUE) / n.treat
var.control = var(y[tr == 0], na.rm = TRUE) / n.control
itt.se = sqrt(var.treated + var.control)


# lets write a function to collect this information together
# mean.dif
mean.dif=function(y,tr){
tau_bar=mean(y[tr==1],na.rm=T)-mean(y[tr==0],na.rm=T)

n.treat = sum(tr)
n.control = sum(1 - tr)

var.treated = var(y[tr == 1], na.rm = TRUE) / n.treat
var.control = var(y[tr == 0], na.rm = TRUE) / n.control
itt.se = sqrt(var.treated + var.control)

return(list("mean.dif"=tau_bar,"se"=itt.se))
}


# (b)  write an OLS function using matrix computations

ols.matrix=function(Y,X){

if(any(is.na(Y))){
stop("Missing values on the outcome variable, OLS estimates not identified.")
}

if(any(is.na(X))){
stop("Missing values on a covariate, OLS estimates not identified.")
}

n=length(Y)
p=dim(X)[2]

# OLS estiamtes
beta.ols=solve(t(X)%*%(X))%*%(t(X)%*%(Y))

# Projection matrix for SE estimates
h=X%*%solve(t(X)%*%X)%*%t(X)
i=(h-h)
diag(i)=1
hhat=i-h
e=hhat%*%Y

sigma2=sum(e^2)/(n-p)
```

```
varsB=sigma2*solve(t(X)%*%X)
se.ols=t(t(sqrt(diag(varsB))))

st=dim(X)[2]

var.names=c()
for(i in 2:st){
var.names=c(var.names,paste('v',i-1,sep=''))
} #END var.name loop

var.names=c('alpha',var.names)
outs=matrix(NA,length(beta.ols),2)
outs[,1]=beta.ols
outs[,2]=se.ols
rownames(beta.ols)=rownames(se.ols)=rownames(outs)=var.names
colnames(outs)=c('beta','se')

return(outs)

} # END ols.matrix function


X=cbind(1,tr)
Y=y
X=X[!is.na(Y),]
Y=Y[!is.na(Y)]

ols.matrix(Y,X)


# (c)

misses=array(TRUE,length(y))
X=cbind(1,tr,data$share.total.unskilled,data$head.edu,data$mosques,
  data$pct.poor,data$total.budget)
Y=y
mMat=cbind(Y,X)
for(j in 1:ncol(mMat)){
inds=which(is.na(mMat[,j]))
if(length(inds>0)){
misses[inds]=FALSE
}
}
Y=Y[misses]
X=X[misses,]

ols.matrix(Y,X)

# Note the 'data$unskilled.transformed' should be dropped due to
  collinearity with 'data$share.total.unskilled'
```

```
# 8. ATT via OLS regression

# use ols.matrix from above

misses=array(TRUE,length(y))
X=cbind(1,tr,data$mosques,data$unskilled.transformed)
Y=y
mMat=cbind(Y,X)
for(j in 1:ncol(mMat)){
inds=which(is.na(mMat[,j]))
if(length(inds>0)){
misses[inds]=FALSE
}
}
Y=Y[misses]
X=X[misses,]


# consider OLS fixed, and estimate the full model

# (a)

Tr=tr[misses]

betas=ols.matrix(Y,X)[,1]

Xtr=X[Tr==1,]
Xct=X[Tr==1,]
Xct[,2]=0


X_rep=rbind(Xtr,Xct)
Y_pred=X_rep%*%betas

mean(Y_pred[X_rep[,2]==1]-Y_pred[X_rep[,2]==0])
ols.matrix(Y_pred,X_rep)[2,1]


# (b) interactive model

misses=array(TRUE,length(y))
X=cbind(1,tr,data$unskilled.transformed,data$mosques,
  tr*data$unskilled.transformed)
Y=y
mMat=cbind(Y,X)
for(j in 1:ncol(mMat)){
inds=which(is.na(mMat[,j]))
if(length(inds>0)){
misses[inds]=FALSE
```

```
}
}
Y=Y[misses]
X=X[misses,]

betas=ols.matrix(Y,X)[,1]

Xtr=X[Tr==1,]
Xct=X[Tr==1,]
Xct[,5]=Xct[,2]=0

X_rep=rbind(Xtr,Xct)
Y_pred=X_rep%*%betas

#Y_pred_1=Xtr%*%betas
#Y_pred_0=Xct%*%betas

#mean(Y_pred_1-Y_pred_0)
mean(Y_pred[X_rep[,2]==1]-Y_pred[X_rep[,2]==0])



# END HW1_Answers.R
```