1. Randomized Controlled Trial – It begins...

You have a randomized controlled trial (RCT) with potential outcomes $Y_i(0)$, $Y_i(1)$ and treatment D_i. Let $P(D_i = 1) = 0.5$ and let $i = 1, ..., N$. You observe $Y_i = Y_i(0) \cdot (1 - D_i) + Y_i(1) \cdot D_i$.

(a) Consider a regression of Y_i on D_i. Prove that the regression of Y_i on D_i estimates ATE. Does it estimate TOT too? If so, prove that.

Taking a hint from Joel, we have linear CEF given D_i is binary meaning we have a saturated model.

$$\mathbb{E}\left[Y_i \mid D_i\right] = \alpha + \tau D_i$$

Then the average treatment effect estimate τ from this regression is an unbiased estimator for

$$\begin{aligned}
\tau &= (\alpha + \tau) - \alpha \\
&= \mathbb{E}\left[Y_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i \mid D_i = 0\right] \\
&= \mathbb{E}\left[Y_i(0)(1 - D_i) + Y_i(1)D_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0)(1 - D_i) + Y_i(1)D_i \mid D_i = 0\right] \\
&= \mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 0\right]
\end{aligned}$$

and because D is randomly assigned, $Y_i(1), Y_i(0) \perp D_i$, so $\mathbb{E}\left[Y_i(0) \mid D_i = 0\right] = \mathbb{E}\left[Y_i(0)\right]$ and $\mathbb{E}\left[Y_i(1) \mid D_i = 1\right] = \mathbb{E}\left[Y_i(1)\right]$

$$\begin{aligned}
&= \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] \\
&= \overline{\tau}_{ATE}
\end{aligned}$$

Additionally, because treatment assignment is random, the average treatment effect becomes

$$\begin{aligned}
\overline{\tau}_{ATE} &= \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] \\
&= \mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 1\right] \\
&= \overline{\tau}_{TOT}
\end{aligned}$$

and is also estimated by the regression.

(b) Does setting $P(D_i = 1) = 0.5$ maximize the power (i.e. minimize the standard error of the estimated treatment effect) in your RCT? Prove. (You may take the formula for $Var(\hat{\beta}_{OLS})$ as given; feel free to look it up if you do not recall it.)

Given a simple univariate OLS, and given D is distributed with binomial, whose variance can be asmptotically approximated by $np(1-p)$,

$$\begin{aligned} Var(\hat{\beta_{OLS}}) &= \frac{\sigma_u^2}{SST_D} \\ &= \frac{\sigma_u^2}{np(1-p)} \end{aligned}$$

Minimizing the above expression w.r.t. p is equivalent to maximizing $p(1-p)$ w.r.t. p. Hence, $p^* = 0.5$

(c)* Suppose you also observe covariates X_i. You are concerned about heteroskedaticity and estimate a regression of $\hat{\varepsilon}_i^2$ on X_i, where $\hat{\varepsilon}_i$ are OLS residuals from regressing Y_i on X_i. Using the fitted values from this regression you form weights $w_i = 1/\hat{\sigma}_i$ and reweight your data. Prove that a regression of Y_i^w on D_i^w, where a w superscript signifies a reweighted variable, does not necessarily estimate ATE. Does it estimate TOT?

(d) Assuming your model of heteroskedasticity is correct, name one advantage and one disadvantage of the weighted regression (versus OLS). You may rely on results you've learned from previous econometrics courses.

Advantage: In ARE 212, we learned that under heteroskedasticity, OLS is no longer the minimum variance estimator among the class of linear and unbiased estimators. Weighting solves above issue and reduces standard errors on the coefficients.

Disadvantage: Weighted least squares requires one to know exactly what the weights should be (given the interpretation you're looking for). Otherwise, picking the wrong weights (for the question) will risk skewing the results in unwanted ways. For instance, even if you have correctly modeled the heteroskedasticity, if you have an outlier observation that shares regressor values that are similar to other observations with low error, then WLS would give this outlier too much weight. In an attempt to approach the right weights, one can use iterative WLS. Using WLS may also lead to a misinterpretation of the treatment effect since outlier treated observations will be given less weight when we may care most about them. All depends on our intended interpretation!

(e) Suppose that it costs \$1 to generate each control observation and \$4 to generate each treated observation (i.e. the treatment itself is expensive). You have a total budget of \$150. To maximize power in your RCT, how many observations will you assign to treatment, and how many will you assign to control? (To be clear, your estimator here will be an OLS regression of Y_i on D_i, not the re-weighted regression.)

Utilizing equation in part(b), we want to maximize $np(1-p)$, where p is probability of treatment.

Define budget constraint:

$$\begin{aligned} 150 &= 4np + n(1-p) \\ &= n(3p+1) \end{aligned}$$

Given above budget constraint, we can write the following maximization problem:

$$\begin{aligned} &\max_{p,n} np(1-p) \ \text{ s.t. } \ n(3p+1) = 150 \\ \iff &\max_{p} \frac{150}{3p+1} p(1-p) \\ \iff &\max_{p} \frac{150}{3p+1} (p - p^2) \end{aligned}$$

Solving the FOC w.r.t. p, we achieve the following:

$$p = +\frac{1}{3} \ \text{ or } \ -1$$

p clearly cannot be negative. Hence, $p^* = \frac{1}{3}$, $N^* = 75$, $N^*(Trt) = 25$, and $N^*(Cont) = 50$

2. Randomized Controlled Trial – Here we go again!

We continue to use the RCT discussed above, but you no longer have any covariates X_i (and you can forget about the weights and costs). You may assume $P(D_i = 1) = 0.5$. Suppose Y_i is missing for some observations (but D_i is complete for all observations). might happen because, for example, some participants in the RCT did not respond to the survey collecting the outcome data.

(a) First assume Y_i are missing at random. Propose a simple test for this hypothesis. Does a regression of Y_i on D_i (using complete observations) estimate ATE if your assumption is correct?

Let M_i be an indicator for of Y_i, i.e., $M_i = 1$ if Y_i is missing and $M_i = 0$ if Y_i is not missing. In this setting, missing at random would mean that the missing data process is independent of Y_i and D_i. Because we have no other data to compare the distribution of Y_i to, we cannot test if the missing data process is independent of Y_i. However, by running a regression of M_i on D_i, we can test for linear correlation between missingness and treatment assignment. So a simple test could include runing the following regression

$$M_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

and checking if β_1 is significantly different from zero. If β_1 is significantly different from zero, we have evidence that the Missing At Random assumption is violated.

Under the Missing At Random assumption, the regression of Y_i on D_i still estimates ATE.

(b) If the Y_i are missing at random, does your test have the correct size? That is, does it reject at a rate $\alpha = 0.05$? Briefly explain.

If the data are missing at random, the test of Missing at Random has the correct size because the regression would only include the complete observations.

(c) If the null hypothesis is false (i.e. the Y_i are not missing at random), will your test necessarily reject as $N \to \infty$?

One of the ways that the missingness could not be random is if the missing data process is related to the value of Y_i. Because we only observe the Y_i we collect, and the simple test above only tests for a relationship between missingness and the treatment, any missingness related to Y_i would not be detected and the test will not necessarily reject at $N \to \infty$.

3. The Roy Model (Selection)

We now consider a gentle introduction to the selection model known as the Roy Model. discussed this model briefly in class in the context of choosing fishing versus hunting. The key assumption in the Roy Model is that individuals select into treatment based on their gains (or expected gains) from treatment. This assumption comes naturally from a framework in which individuals maximize utility, but, to be clear, it is an assumption (i.e. there is no guarantee it holds in your real-world data). To fix ideas, assume a binary treatment D_i and potential outcomes $Y_i(0)$ and $Y_i(1)$. Let the treatment gain for individual i be equal to i's treatment effect, $\tau_i = Y_i(1) - Y_i(0)$. Assume i selects into treatment if and only if $\tau_i > 0$ (i.e. $D_I = \mathbf{1}(\tau_i > 0)$).

For part (d) of this question you may take as known a result we refer to as the inverse Mills ratio — if $A \sim N(0,1)$, then $E[A \mid A > c] = \frac{\phi(c)}{1-\Phi(c)}$ (we will show this in lecture).

(a) First assume $Y_i(0) = c$. Show that a regression of Y_i on D_i estimates TOT.

The regression of Y_i on D_i is

$$\mathbb{E}\left[Y_i | D_i\right] = \alpha + \tau D_i$$

So the treatment effect estimated by the regression is

$$\begin{aligned}
\tau &= (\alpha + \tau) - \alpha \\
&= \mathbb{E}\left[Y_i \mid D_i = 1\right] - \mathbb{E}\left[Y_i \mid D_i = 0\right] \\
&= \mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 0\right] \\
&= \mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[c \mid D_i = 0\right] \\
&= \mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[c \mid D_i = 1\right] \\
&= \mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 1\right] \\
&= \overline{\tau}_{TOT}
\end{aligned}$$

(b)* Now let $Y_i(0)$ and $Y_i(1)$ be independent with marginal Uniform(0,1) distributions. Analytically derive $E[Y(1)_i \mid D_i = 1]$ and $E[Y(0)_i \mid D_i = 0]$. Does a regression of Y_i on D_i estimate TOT?

Because $Y_i(0)$ and $Y_i(1)$ are independent and symmetrically distributed, we know $\mathbb{P}[Y_i(0) < Y_i(1)] = 0.5$. Suppressing the i for a moment, let $Y_k \equiv Y_i(k)$, then the conditional CDF of Y_1) is

$$\begin{aligned}
F_{Y_1|Y_0>Y_1}(y) &= \mathbb{P}[Y_1 < y | Y_0 < Y_1] \\
&= \frac{\mathbb{P}[Y_0 < Y_1 < y]}{\mathbb{P}[Y_0 < Y_1]} \\
&= \frac{1}{1/2} \int_0^y \left[\int_0^{y_1} f(y_0, y_1)\, dy_0 \right] dy_1 \\
&= 2 \int_0^y \left[\int_0^{y_1} 1\, dy_0 \right] dy_1 \\
&= 2 \int_0^y y_1\, dy_1 \\
&= 2 \left[\frac{1}{2} y_1^2 \right]_{y_1=0}^{y} \\
&= y^2 \qquad y \in [0,1]
\end{aligned}$$

$$\begin{aligned}
f_{Y_1|Y_0>Y_1}(y) &= \frac{d}{dy} F_{Y_1|Y_0>Y_1}(y) \\
&= 2y \qquad y \in [0,1],\ (0 \text{ otherwise})
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[Y_1 \mid D_i = 1\right] &= \mathbb{E}\left[Y_1 \mid Y_1 > Y_0\right] \\
&= \int_{-\infty}^{\infty} y\ f_{Y_1|Y_0>Y_1}(y)\ dy \\
&= \int_0^1 y \cdot 2y\ dy \\
&= \int_0^1 2y^2\ dy \\
&= \frac{2}{3} y^3 \Big|_{y=0}^{1} \\
&= \frac{2}{3}
\end{aligned}$$

$$\mathbb{E}\left[Y_0 \mid D_i = 0\right] = \mathbb{E}\left[Y_0 \mid Y_0 > Y_1\right]$$

By symmetry of the independent and identical distributions, we know this is equal to the previous expectation

$$\mathbb{E}\left[Y_0 \mid D_i = 0\right] = \frac{2}{3}$$

So the ATE is $\mathbb{E}\left[Y_1 \mid D_i = 1\right] - \mathbb{E}\left[Y_0 \mid D_i = 0\right] = 0$.

$$\begin{aligned}
\overline{\tau}_{TOT} &= \mathbb{E}\left[Y_i(1) - Y_i(0) | D_i = 1\right] \\
&= \mathbb{E}\left[Y_i(1) - Y_i(0) | \tau_i > 0\right] \\
&= \mathbb{E}\left[Y_i(1) - Y_i(0) | Y_i(1) - Y_i(0) > 0\right]
\end{aligned}$$

Since we are conditioning on only values where $Y_i(1) - Y_i(0) > 0$, the expectation of $Y_i(1) - Y_i(0)$ must be positive.

$$\overline{\tau}_{TOT} > 0$$

So the regression of Y on D does not give us TOT.

(c) Write some Monte Carlo simulation code in Stata, R, or another package of your choice to confirm your answer to part (*b*) (or, if you skipped part (*b*), to determine the answer numerically). What is your intuition as to why the regression does or does not estimate TOT now?

Preamble which includes clearing the environment and setting seed is omitted.

Hide

```
# DGP & Potential Outcomes and Observed Outcome
N <- 1000
Y_0c <- runif(N,0,1)
Y_1c <- runif(N,0,1)
df_c <- data.frame(Y_0c, Y_1c, D_c = Y_1c > Y_0c) %>%
          mutate(Y_obs_c = D_c*Y_1c + (1-D_c)*Y_0c)

# Calculation of E[Y_i(1) | D_i = 1]
print(paste0("E[Y_i(1) | D_i = 1] = ", mean(df_c$Y_obs_c[df_c$D_c == 1])))
```

```
[1] "E[Y_i(1) | D_i = 1] = 0.674973633830228"
```

Hide

```
# Calculation of E[Y_i(0) | D_i = 0]
print(paste0("E[Y_i(0) | D_i = 0] = ", mean(df_c$Y_obs_c[df_c$D_c == 0])))
```

```
[1] "E[Y_i(0) | D_i = 0] = 0.678957071035294"
```

Hide

```
# Naive Calculation of Treatment Effect
Reg_c <- mean(df_c$Y_obs_c[df_c$D_c == 1]) - mean(df_c$Y_obs_c[df_c$D_c == 0])
print(paste0("Regression of observed Y on D would predict Treatment Effect of:", Reg_c))
```

```
[1] "Regression of observed Y on D would predict Treatment Effect of:-0.00398343720506633"
```

Hide

```
# Treatment Effect on Treated (Hypothetical)
TOT_c <- mean(df_c$Y_obs_c[df_c$D_c == 1]) - mean(df_c$Y_0c[df_c$D_c == 1])
print(paste0("Hypothetical Treatment Effect on Treated:", TOT_c))
```

```
[1] "Hypothetical Treatment Effect on Treated:0.31622901905545"
```

Due to selection (lack of randomization of treatment), the treated group were of individuals who, if they could observe ex-ante their hypothetical outcomes in both states (treated and not treated) and compare, would clearly benefit from treatment. Hence, TOT will be positive.

If you now think about the group that selected into control / no treatment, it's the exact opposite. Their outcome is better off without treatment. If they were forced into treatment (instead of the others), their TOT would be negative.

Hence, expectation of observed outcomes at each level of D is higher than expectation of potential outcomes had there been randomization or equivalently had there been a way to observe everyone's potential outcomes in both treatment and control states.

This regression of Y on D would estimate no effect (if not negative), even though the hypothetical TOT is clearly positive.

(d)* Now let $Y_i(0)$ and $Y_i(1)$ be independent with marginal normal distributions with $\mu = 0$ and $\sigma^2 = 0.5$. Analytically derive $\mathbb{E}\left[Y_i(1) \,|\, D_i = 1\right]$ and $\mathbb{E}\left[Y(0)_i \,|\, D_i = 0\right]$. Does a regression of Y_i on D_i estimate TOT?

Because $Y_i(0)$ and $Y_i(1)$ are independent and symmetrically distributed, we know $\mathbb{P}[Y_i(0) < Y_i(1)] = 0.5$. Suppressing the i for a moment, let $Y_k \equiv Y_i(k)$, then the conditional CDF of Y_1 is

$$\begin{aligned}
F_{Y_1|Y_0>Y_1}(y) &= \mathbb{P}[Y_1 < y | Y_0 < Y_1] \\
&= \frac{\mathbb{P}[Y_0 < Y_1 < y]}{\mathbb{P}[Y_0 < Y_1]} \\
&= \frac{1}{1/2} \int_{-\infty}^{y} \left[\int_{-\infty}^{y_1} f(y_0, y_1)\, dy_0 \right] dy_1 \\
&= 2 \int_{-\infty}^{y} \left[\int_{-\infty}^{y_1} \phi(y_0)\phi(y_1)\, dy_0 \right] dy_1 \\
&= 2 \int_{-\infty}^{y} \left[\int_{-\infty}^{y_1} \phi(y_0)\, dy_0 \right] \phi(y_1)\, dy_1 \\
&= 2 \int_{-\infty}^{y} \Phi(y_1)\phi(y_1)\, dy_1
\end{aligned}$$

$$\begin{aligned}
f_{Y_1|Y_0>Y_1}(y) &= \frac{d}{dy} F_{Y_1|Y_0>Y_1}(y) \\
&= 2 \frac{d}{dy} \int_{-\infty}^{y} \Phi(y_1)\phi(y_1)\, dy_1 \\
&= 2\Phi(y)\phi(y)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[Y_1 \,|\, D_i = 1\right] &= \mathbb{E}\left[Y_1 \,|\, Y_1 > Y_0\right] \\
&= \int_{-\infty}^{\infty} y\ f_{Y_1|Y_0>Y_1}(y)\ dy \\
&= \int_{-\infty}^{\infty} y\ 2\Phi(y)\phi(y)\ dy \\
&= \frac{\sigma}{\sqrt{\pi}} \quad \text{by mathematica}
\end{aligned}$$

$$\mathbb{E}\left[Y_0 \,|\, D_i = 0\right] = \mathbb{E}\left[Y_0 \,|\, Y_0 > Y_1\right]$$

By symmetry of the independent and identical distributions, we know this is equal to the previous expectation

$$= \frac{\sigma}{\sqrt{\pi}}$$

So the ATE is 0.

By the same reasoning as in part (b),

$$\begin{aligned}\overline{\tau}_{TOT} &= \mathbb{E}\left[Y_i(1) - Y_i(0) | D_i = 1\right] \\ &= \mathbb{E}\left[Y_i(1) - Y_i(0) | \tau_i > 0\right] \\ &= \mathbb{E}\left[Y_i(1) - Y_i(0) | Y_i(1) - Y_i(0) > 0\right] \\ &> 0\end{aligned}$$

So the regression of Y on D does not give us TOT.

(e) Write Monte Carlo simulation code to confirm your0answer to part (d) (or, if you skipped part (d), to determine the answer numerically). What happens if you generate an error ε_i that's normally distributed with $\mu = 0$ and $\sigma^2 = 1$ and add it to both $Y_i(0)$ and $Y_i(1)$ to create a positive correlation between the two. Does the regression estimate TOT, and if not, is it upwardly biased or downwardly biased?

Preamble which includes clearing the environment and setting seed is omitted.

Hide

```
# DGP of Potential and Observed Outcomes with Multivariate Normal Distribution
N <- 1000
Sigma <- matrix(c(0.5,0,0,0.5),2,2)
df_e <- data.frame(mvrnorm(N, rep(0, 2), Sigma)) %>%
          rename(Y_0e = X1, Y_1e = X2) %>%
          mutate(D_e = Y_1e > Y_0e,
                 Y_obs_e = D_e*Y_1e + (1-D_e)*Y_0e)

# Calculation of E[Y_i(1) | D_i = 1]
print(paste0("E[Y_i(1) | D_i = 1] = ", mean(df_e$Y_obs_e[df_e$D_e == 1])))
```

```
[1] "E[Y_i(1) | D_i = 1] = 0.413801008593895"
```

Hide

```
# Calculation of E[Y_i(0) | D_i = 0]
print(paste0("E[Y_i(0) | D_i = 0] = ", mean(df_e$Y_obs_e[df_e$D_e == 0])))
```

```
[1] "E[Y_i(0) | D_i = 0] = 0.376201414328847"
```

Hide

```
# Naive Calculation of Treatment Effect
Reg_e <- mean(df_e$Y_obs_e[df_e$D_e == 1]) - mean(df_e$Y_obs_e[df_e$D_e == 0])
print(paste0("Regression of observed Y on D would predict Treatment Effect of:", Reg_e))
```

```
[1] "Regression of observed Y on D would predict Treatment Effect of:0.0375995942650486"
```

Hide

```
# Treatment Effect on Treated (Hypothetical)
TOT_e <- mean(df_e$Y_obs_e[df_e$D_e == 1]) - mean(df_e$Y_0e[df_e$D_e == 1])
print(paste0("Hypothetical Treatment Effect on Treated:", TOT_e))
```

```
[1] "Hypothetical Treatment Effect on Treated:0.786879107752204"
```

Hide

```
e_i <- rnorm(N, 0, 1)
df_e_corr <- df_e %>%
                dplyr::select(Y_0e, Y_1e) %>%
                mutate(Y_0e_corr = Y_0e + e_i,
                       Y_1e_corr = Y_1e + e_i,
                       D_e_corr = Y_1e_corr > Y_0e_corr,
                       Y_obs_e_corr = D_e_corr*Y_1e_corr + (1-D_e_corr)*Y_0e_corr)

# Confirm Increased Correlation
print(paste0("Correlation before individual error terms is = ", cor(df_e$Y_0e, df_e$Y_1e)))
```

```
[1] "Correlation before individual error terms is = 0.00631320035556441"
```

Hide

```
print(paste0("Correlation with individual error terms is = ", cor(df_e_corr$Y_0e_corr, df_e_corr$Y_1e_corr)))
```

```
[1] "Correlation with individual error terms is = 0.659964281693646"
```

Hide

```
print("We see that correlation between Y_0 and Y_1 has increased significantly with inclusion of individual error
 terms.")
```

```
[1] "We see that correlation between Y_0 and Y_1 has increased significantly with inclusion of individual error t
erms."
```

Hide

```


# Calculation of E[Y_i(1) | D_i = 1]
print(paste0("E[Y_i(1) | D_i = 1] = ", mean(df_e_corr$Y_obs_e_corr[df_e_corr$D_e_corr == 1])))
```

```
[1] "E[Y_i(1) | D_i = 1] = 0.398831017828252"
```

Hide

```
# Calculation of E[Y_i(0) | D_i = 0]
print(paste0("E[Y_i(0) | D_i = 0] = ", mean(df_e_corr$Y_obs_e_corr[df_e_corr$D_e_corr == 0])))
```

```
[1] "E[Y_i(0) | D_i = 0] = 0.421032103303544"
```

Hide

```
# (Corr) Naive Calculation of Treatment Effect
Reg_e_corr <- mean(df_e_corr$Y_obs_e_corr[df_e_corr$D_e_corr == 1]) -
                mean(df_e_corr$Y_obs_e_corr[df_e_corr$D_e_corr == 0])
print(paste0("(Correlation Case)Regression of observed Y on D would predict Treatment Effect of:", Reg_e_corr))
```

```
[1] "(Correlation Case)Regression of observed Y on D would predict Treatment Effect of:-0.0222010854752918"
```

Hide

```
# (Corr) Treatment Effect on Treated (Hypothetical)
TOT_e_corr <- mean(df_e_corr$Y_obs_e_corr[df_e_corr$D_e_corr == 1]) -
                mean(df_e_corr$Y_0e_corr[df_e_corr$D_e_corr == 1])
print(paste0("(Correlation Case)Hypothetical Treatment Effect on Treated:", TOT_e_corr))
```

```
[1] "(Correlation Case)Hypothetical Treatment Effect on Treated:0.786879107752204"
```

Adding the error and making the two outcomes correlated only shifts both in the same direction. Again, this regression is downward-biased compared to actual TOT since both groups self-selected into the treatment and control.

(f) If there is selection bias in a regression of Y_i on D_i, is it positive or negative? How does the sign accord with your intuition about the general form of selection bias into, for example, getting a college degree (i.e. a case in which Y_i is later-life earnings and D_i is whether the individual got a college degree). What does this say about the most basic form of the Roy Model?

Because of selection bias, the treated are precisely those who will experience a larger positive effect from the treatment than those who choose not to be treated. Not only that, but those in the control group selected to be in the control precisely because that is where *they* would be better off. Thus, The selection bias in this regression of Y on D will be negative.