

PS C236A/ Stat C239A

How do we get there from here?

Freedman: *On regression adjustment to experimental data*

Erin Hartman

September 22, 2009

1 Potential Outcomes

Since we are discussing the David Freedman article, first we will set up potential outcomes in his notation. Index subjects by $i = 1, \dots, n$. Let T_i be the response of the subject i if i is assigned to treatment, and let C_i be the response of the subject i if i is assigned to control. For now, these are fixed numbers. Remember, this is a missing data problem, so the investigator can only choose to observe either T_i or C_i , but not both. Let X_i be the assignment variable: $X_i = 1$ if subject i receives treatment, and $X_i = 0$ if subject i is assigned to control. Therefore, we get the observed response as follows:

$$Y_i = X_i T_i + (1 - X_i) C_i$$

Notice here how this relates to the statement in the class lecture notes:

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$$

In the class notes, T_i refers to treatment assignment, not potential outcome under treatment, as in the Freedman framework. The potential outcomes are referred to as Y_{i1} and Y_{i0} , instead of T_i and C_i .

2 \hat{b}_{ITT}

\hat{b}_{ITT} refers to the “intention-to-treat” effect. It is defined as the average response if all subjects are assigned to treatment minus the average response of all subjects assigned to control. We will assume that we have an experiment in which m out of n people are chosen at random for treatment, and the remaining $n - m$ are assigned to the control.

$$\hat{b}_{ITT} = \left(\frac{1}{m} \sum_i \{Y_i : X_i = 1\} \right) - \left(\frac{1}{n - m} \sum_i \{Y_i : X_i = 0\} \right)$$

note here that $m = \sum X_i$ is the size of the treatment group.

Let \hat{b}_{SR} refer to coefficient of X of a regression of Y on X and an intercept. So long as $X_i = 0$ or 1 , then we get the following (let $p = m/n$):

$$\begin{aligned}
\hat{b}_{ITT} &= \frac{\sum X_i Y_i}{\sum X_i} - \frac{\sum (1-X_i) Y_i}{\sum (1-X_i)} \\
&= \frac{\text{ave}(XY)}{\text{ave}(X)} - \frac{\text{ave}(Y) - \text{ave}(XY)}{1 - \text{ave}(X)} \\
&= \frac{\text{ave}(XY) - \text{ave}(X)\text{ave}(Y) - \text{ave}(X)\text{ave}(Y) + \text{ave}(X)\text{ave}(XY)}{p(1-p)} \\
&= \frac{\text{ave}(XY) - \text{ave}(X)\text{ave}(Y)}{p(1-p)} \\
&\quad \text{note: } \text{cov}(X, Y) = \text{ave}(XY) - \text{ave}(X)\text{ave}(Y) \\
&\quad \text{var}(X) = \text{ave}(X^2) - [\text{ave}(X)]^2 = p(1-p) \\
&= \frac{\text{cov}(X, Y)}{\text{var}(X)} \\
&= \hat{b}_{SR}
\end{aligned}$$

\hat{b}_{ITT} is an unbiased estimator. $E[\hat{b}_{ITT}] = b$, where b is defined as the average treatment effect. This is because with simple random samples, the sample average is an unbiased estimator for the population average.

3 Nominal Variance

From OLS, we know that the nominal variance of \hat{b}_{SR} is:

$$\text{var}(\hat{b}_{SR}) = \sigma^2 (M^T M)^{-1}$$

where M is defined as the design matrix.

However, the nominal variance of \hat{b}_{ITT} is:

$$\text{var}(\hat{b}_{ITT}) = \frac{\hat{v}_T}{m} + \frac{\hat{v}_C}{n-m}$$

where \hat{v}_T is defined as the sample variance of the treatment group and \hat{v}_C is the sample variance of the control group.

How will these compare? They can be very different. The OLS nominal variance assumes homoskedastic errors, where as the ITT estimator adjusts for heteroskedasticity between treatment and control groups.

Is it reasonable to assume homoskedastic errors? Nothing in the design of the experiment guaranteed it.

4 Rewriting potential outcomes in a familiar framework

Let's rewrite the potential outcome framework to look something like the regression framework:

$$Y_i = a + b(X_i - p) + \delta_i \tag{1}$$

Now we have an a that appears to be similar to an intercept, a b that is similar to a treatment effect coefficient, and a δ_i that looks something like an error term. We mean deviate X for ease of the asymptotic proof that we will do later, but it doesn't change the estimators. Define:

$$a = p\bar{T} + (1 - p)\bar{C}$$

$$b = \bar{T} - \bar{C}$$

$$\delta_i = \alpha_i + \beta_i(X_i - p)$$

$$\alpha_i = p(T_i - \bar{T}) + (1 - p)(C_i - \bar{C}) \quad \beta_i = (T_i - \bar{T}) - (C_i - \bar{C})$$

However, equation (1) is nothing like a regression equation, in actuality. The most important differences are in the δ s. It is important to note that the randomness in δ_i is entirely due to the randomness in X_i , which means that the error term is *strongly* dependent on the explanatory variable, in fact, it is partially determined by X_i . The δ s are not IID, and they do not have mean 0. However, they do sum to zero. In effect, we get weak forms of orthogonality without having independence.

We should note here that the assignment variables are a little dependent because their sum is fixed. However, they are exchangeable and behave similar to Bernoulli variables when n is large.

Observables and Unobservables It is important to note that our estimators are defined in terms of observable random variables like X_i , Y_i and, in the multiple regression framework, Z_i . The unobservable parameters, namely T_i and C_i do not enter into the formulas of things we estimate.

What is random in this framework? The only stochastic element of this framework is treatment assignment. This is very different than OLS, where we assume that the disturbance, ϵ_i is random. In this framework, conditional on X_i , the Y_i are fixed, as are the *error terms*, δ_i .

5 Where does the bias come from?

The multiple regression model:

$$Y_i = a + b(X_i - p) + \theta Z_i + \delta'_i$$

$$\delta'_i = \delta_i - \theta Z_i = (\alpha_i - \theta Z_i) + \beta_i(X_i - p)$$

$$\theta = \frac{1}{n} \sum_{i=1}^n \alpha_i Z_i$$

In multiple regression, where we add a term Z_i , a covariate that is measured pre-treatment, the bias comes from the fact that the regression model assumes that the effects are not only linear and additive, but constant across all subjects. We can see from the above notation that the effects are not guaranteed to be constant across subjects. Each subject i can have a different value for $T_i - C_i$ (the unit treatment effect). This violates the basic assumption needed to prove that regression estimates are unbiased.

6 Randomization and the OLS assumptions

Below are the OLS assumptions, and a discussion of how randomization applies to each.

1. *Linear in Parameters*: Y is related to the independent variables and the error term as $Y = X\beta + \epsilon$
 - Randomization does not guarantee linearity, nor does the ITT estimator require it.
2. The X 's are fixed at take on ≥ 2 values
 - In fact, the X s in the Neyman framework are the stochastic element, and conditional on treatment assignment, the Y s and the "error terms" are fixed.
3. *Full Rank* (in multiple regression): There is no perfect collinearity among any of the independent variables
4. *Zero Conditional Mean*: $E(\epsilon|X) = 0$
 - We saw that this is not guaranteed. It is the case that we can sum to zero, but it isn't necessarily the case that the expectation is zero. We get orthogonality but not necessarily independence. We see that the "error term" is in part dependent on X , so it is clear that it is not independent.
5. *Homoskedasticity*: $Var(\epsilon|X) = \sigma^2$
 - Randomization does not guarantee homoskedasticity. We saw that the ITT estimator adjusts for heteroskedasticity in the nominal variance equation, but that the OLS estimator does not. The Neyman model does not require a constant treatment effect for all i . Each subject can have a different value for $T_i - C_i$.
6. *Random Sampling*: Y_i is an *iid* random sample, although this can be relaxed to $cov(y_i, y_j) = 0 = cov(\epsilon_i, \epsilon_j) \quad i \neq j$
7. *Normal Errors* (optional): $Y \sim N(X\beta, \sigma^2)$
 - Randomization definitely doesn't guarantee this.

"Practitioners will doubtless be heard to object that they know all this perfectly well. Perhaps, but then why do they so often fit models without discussing assumptions?" - David Freedman