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 $i = 1, \ldots, n$. 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.

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

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 $p = m/n$):

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

$\hat{b}_{ITT}$ is an unbiased estimator. $\mathrm{E}[\hat{b}_{ITT}] = b$, 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:

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

where M is defined as the design matrix.

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

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

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:

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

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

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

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

However, equation (1) is nothing like a regression equation, in actuality. The most important differences are in the δs. 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 δs 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 Xs in the Neyman framework are the stochastic element, and conditional on treatment assignment, the Ys 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 $cov(y_i, y_j) = 0 = cov(\epsilon_i, \epsilon_j) \qquad i \neq j$

7. *Normal Errors* (optional): $Y \sim \mathbb{N}(X\beta, \sigma^2)$

- 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

PS C236A/ Stat C239A
Section 2 Notes

1 Expectation of a random variable

A random variable X has *expectation* and *variance*, denoted $E(X)$ and $var(x)$ respectively:

$$var(x) = E\{[X - E(X)]^2\} = E(X^2) - [E(X)^2]$$

The *standard error* of X is $\sqrt{var(X)}$. The standard error is abbreviated as SE.

The expected value or expectation of a random variable is its average value, where "average value" means a value weighted according to the probability distribution. For a discrete random variable, its expected value can be interpreted as a weighted average of all possible outcomes. In this context, the weight assigned to each particular outcome is equal to the probability of that outcome occurring.

Some rules:

1. The Discrete Case: suppose $P\{X = x_i\} = p_i$ for $i = 1, 2, ...$ and $\sum_i p_i = 1$. Then $E(X) = \sum_i x_i p_i$, $E(X^2) = \sum_i x_i^2 p_i$, and so forth. More generally, $E\{g(X)\} = \sum_i g(x_i)p_i$
2. The absolutely continuous case: if X has a density f, i.e., $P\{X \leq x\} = \int_{-\infty}^{x} f(u)du$, if then $E(X) = \int_{-\infty}^{\infty} uf(u)du$, $\int_{-\infty}^{\infty} u^2 f(u)du$, and so forth. More generally $\int_{-\infty}^{\infty} g(u)f(u)du$.
3. If a is a real number, then $E(aX) = aE(X)$.
4. $E(X + Y) = E(X) + E(Y)$.
5. If a is a real number, then var$(aX) = a^2$var(x).
6. var$(X + Y) =$ var$(X) +$ var$(Y) + 2 \cdot$ cov(X, Y).

2 Conditional distributions and expectations

The formal definition of conditional probability is:

$$P\{A|B\} = \frac{P\{A \text{ and } B\}}{P\{B\}}$$

The interpretation is a bit opaque, but essentially, $P\{A|B\}$ is a new probability on the sample space. Probabilities outside B are reset to 0; inside B, probabilities are renormalized so the sum is 1.

The *conditional distribution* of Y given X is the distribution of Y, given the value of the value of X. In the discrete case, this is just $P\{Y = y|X = x\}$. The conditional expectation of Y given X is

$$E(Y|X = x) = \sum_y yP\{Y = y|X = x\}$$

In the absolutely continuous case, the pair (X, Y) has the density f, i.e.,

$$P\{X \leq x \text{ and } Y \leq y\} = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv$$

Then X has density g and Y has density h:

$$g(x) = \int_{-\infty}^{\infty} f(x, v) dv, h(y) = \int_{-\infty}^{\infty} f(u, y) du$$

Furthemore, Y has a conditional density given that $X = x$, then $h(y|x) = f(x, y)/g(x)$. Said another way, the conditional distribution of Y given $X = x$ has the density $h(y|x)$. For instance,

$$P\{Y \leq w|X = x\} = \int_{-\infty}^{w} h(y|x) dy, E(Y|X = x) = \int_{-\infty}^{\infty} yh(|x) dy$$

What is fixed? The conditional distribution of Y given $X = x$ is possibly a different probability distribution for every value of x. When we want to describe the entire famiy of distributions, we write the distribution of $E(Y|X)$. When we describe a particular conditional expectation, we use $E(Y|X = x)$.

The distinction between X and x is crucial. When we write X, we mean X is a random variable. When we write x, we mean that x is a realization of the random variable and therefore is fixed. So for example, $E(Y|x)$ is a real number obtained by calculating the appropriate sum or integral and there is nothing random about it. $E(Y|X)]$, however, is a random variable whose value depends on the value of X.

3 What is independence?

Suppose we make two draw at random from the box $\boxed{1}\boxed{2}\boxed{2}\boxed{5}$. Let X be the first draw, and Y the second.

1. Suppose the draws are made with replacement:

   If $X = 1$, the chance that $Y = 5$ is 1/4.

   If $X = 2$, the chance that $Y = 5$ is 1/4.

   If $X = 2$, the chance that $Y = 5$ is 1/4.

   This is *independence*: $P\{Y = y|X = x\}$ is the same for all x. This definition only applies to the discrete case and the equality has to hold for each y. Note that this also implies that $E(Y|X) = E(X)$.

   Factorization: Discrete random variables X and $Y$are independent provided $P\{X = x \text{ and } Y = y\} = P\{X = x\}P\{Y = y\}$ for all x and y.

2. In the box example, if the draws are made without replacement, the two random variables are dependent: $P\{Y = y|X = x\}$ may be different for different x's.

   If $X = 1$, the chance that $Y = 5$ is 1/3.

   If $X = 2$, the chance that $Y = 5$ is 1/3.

   If $X = 5$, the chance that $Y = 5$ is 0.

In the absolute continuous case, the above definition doesn't work, since the $P\{X = x\} = 0$ for all values of x and 0/0 is not defined. Suppose the pair (X, Y) has a joint density f. The independence condition here is that $h(y|x)$ is the same for all x, where h is the conditional density of Y given $X = x$.

Factorization: Absolutely continuous variables X, Y are independent provided the joint density f factors: $f(x, y) = g(x)h(y)$ for all x, y.

Notation: $X \perp\!\!\!\perp Y$ means that X and Y are independent.

4 Sums of independent variables

If $X \perp\!\!\!\perp Y$ then

1. $E(XY) = E(X)E(Y)$
2. cov$(X, Y) = 0$
3. var$(X + Y) =$ var$(X) +$ var(Y)

Suppose $X_1, X_2, ...$ are independent and identically distributed (IID). Let $E(X_i) = \mu$ and var$(X_i) = \sigma^2$. Let $S_n = X_1 + ... + X_n$. Then

1. $E(S_n) = n\mu$
2. var$(S_n) = n\sigma^2$.

In other words, (1) the sum of IID random variables has the expected value equal to n times the common expected value of the summands. (2) The standard error of the sum is $\sqrt{n}$ times the common standard error of the summands.

Let $\bar{X} = S_n/n$ be the average of $X_1, X_2, ..., X_n$. Then $E(\bar{X}) = n\mu/n = \mu$and var$(\bar{X}) = n\sigma^2/n^2 = \sigma^2/\sqrt{n}$. Thus, $\bar{X}$ has expectation μ and a standard error $\sigma/\sqrt{n}$. What happens when the sample without replacement? Then the variance of the sample mean is var$(\bar{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$.

5 Potential outcomes

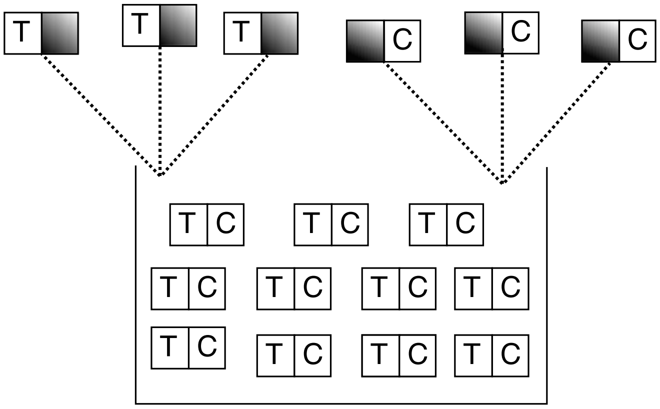

Figure 1: A randomized controlled experiment comparing treatments T and C. There is a ticket fo reach subject. The ticket has two numbers: one shows the subject's response to treatment T; the other, to treatment C. Only one of the two numbers can be observed.

Say we are interested in inferring the effect of some cause T on a parameter $\overline{Y}$ of the distribution of outcome Y in population A relative to treatment C (control). Population A is composed of a finite number of units and $\overline{Y}_{A,T}$ is simply a summary of the distribution of that population when exposed to T, such as the mean. If treatment C (control) were to be applied to population A, then we would observe $\overline{Y}_{A,C}$. In other words, we observe $\overline{Y}_{A,T}$ and in the counterfactual world, we would observe $\overline{Y}_{A,C}$. The causal effect of T relative to C for population A is a measure of the difference between $\overline{Y}_{A,T}$ and $\overline{Y}_{A,C}$, such as $\overline{Y}_{A,T} - \overline{Y}_{A,C}$. Of course, we can only observe the parameter that summarizes the actual world and not the counterfactual world.

The key insight of statistical models of causation is that under special circumstances we can use another population, B, that was exposed to control, to act as the counterfactual of A. If we believe that $\overline{Y}_{A,C} = \overline{Y}_{B,C}$, then we no longer need to rely on a unobserved counterfactual world to make causal inferences, we simply can simply look at the difference between the observed $\overline{Y}_{A,T}$ and $\overline{Y}_{B,C}$. In most cases $\overline{Y}_{A,C} \neq \overline{Y}_{B,C}$, however, so any inferences made by comparing the two populations will be *confounded*. What are the special circumstances that allow us to construct a suitable counterfactual population and make unconfounded inferences? As discussed below, the most reliable method is through randomization of treatment assignment, but counterfactual inferences with observational data is possible—albeit more hazardous—as well. Randomization of treatment motivates the most popular counterfactual model for causation: the Neyman-Rubin model.

Let Y_{iT} denote the potential outcome for unit i if the unit receives treatment, and let Y_{iC} denote the potential outcome for unit i in the control regime. The treatment effect for observation i is defined by $\tau_i = Y_{iT} - Y_{iC}$ Causal inference is a missing data problem because Y_{iT} and Y_{iC} are never both observed. This remains true regardless of the methodology used to make inferential progress—regardless of whether we use quantitative or qualitative methods of inference. The fact that we cannot observe both potential outcomes at the same time is commonly referred to as the "fundamental problem of causal inference".

Let T_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise. The observed outcome for observation i is then:

$$Y_i = T_i Y_{iT} + (1 - T_i) Y_{iC}$$

The average causal effect τ is the difference between the expected values $E(Y_T)$ and $E(Y_C)$. We only observe the conditional expectations $E(Y_T|T = 1)$ and $E(Y_C|T = 0)$, not the unconditional expectations required for obtaining τ. Until assume that $E(Y_T|T = 1) = E(Y_T)$ and $E(Y_C|T = 0) = E(Y_C)$, we cannot calculate the average treatment effect.

To estimate the average treatment effect, we require the assumption of *independence*. The singular virtue of experiments is that physical randomization of an intervention ensures independence between treatment status and potential outcomes.. With the independence assumption, the average treatment effect can be estimated from observables using the following expression:

$$\tau = E(Y_{iT}|T = 1) - E(Y_{iC}|T = 0) = E(Y_{iT}) - E(Y_{iC})$$

Under randomization, the assumption that T_i is independent of Y_{iT} and Y_{iC} is plausible, making the treatment and control groups exchangeable in expectation.