I. What is Impact Evaluation

- Econometricians are mainly concerned with revealing causal relationships, which is actually very difficult and often requires more than a little creativity. Omitted variable bias, in particular, makes this task a hard one and has prevented us from making many credible causal claims so far this semester.

- Impact evaluation seeks to identify the *causal* effect (impact) of an intervention on some selected outcomes, and to quantify these changes in outcomes

- How can we achieve this? The ideal would be to give the policy to certain people, see how they turn out, then go back in time and see what would have happened to those very same people had we not given them the policy. Obviously time travel is not possible. Why can't we just compare these people's outcomes to another group that didn't receive the policy? Well these two groups could be completely different, and I may have difficulties isolating the effect of the program from the other differences.

    - Ex: Let's say we want to know the effect of a new health insurance program on health outcomes. Our first thought is to compare the health of those who decide to purchase insurance coverage to the health of those without insurance coverage. What's the problem here? Well these two populations are likely completely different (think about who seeks out health insurance). How can we be sure that the differences we find among these two groups are solely due to health insurance?

- The key issue in measuring impact is to establish a **counterfactual** against which the changes in outcomes for one group induced by the intervention can be measured. The counterfactual should allow the researcher to *convincingly* measure what would have happened to the beneficiaries in the absence of the intervention.

- The techniques we investigate below aim to do just that: they establish a good counterfactual group for those who receive the intervention/program. We can evaluate the impact of this program on outcomes of interest.

Common Challenges to Causality

As suggested above and as we've discussed in class, there are a number of common challenges with *isolating* the impact of X on Y:

1. **Reverse Causality** ($Y \Rightarrow X$) means that the outcome variable (Y) actually affects the realization of X. Suppose we want to look at the effect of education spending on GDP at the country level. In this example, we might have reverse causality if countries with larger increases in income decide to increase their spending on education, because they can afford it. So it's not that education causes growth, but that growth leads to more expenditure in education.

2. **Simultaneity** ($Y \Rightarrow X$ and $X \Rightarrow Y$) means that the outcome variable causes selection into the treatment (X) and control, but treatment status also has effects on the outcome. For our purposes, this is a special case of reverse causality. In the example above, country income might lead to more investment in education, which might simultaneously increase country income.

3. **Omitted Variables Bias** ($Z \Rightarrow X$ and $Z \Rightarrow Y$) means there is some unobserved variable Z that is correlated with both selection into treatment and the outcome. In our example, improving health may cause countries to both increase their income and spend more on education. This would mean that we would observe both large increases in education spending and large increases in income when there was improving health, even if there was no causal relationship between education spending and income.

II. Potential Outcomes Framework

Suppose we have a program or intervention—which we commonly refer to as the "treatment"—such that you either receive the treatment or you do not. We represent the treatment with a variable T that is 1 for those that receive the treatment and 0 for those that do not. We refer to those that do not receive the treatment as the "control".

If we are interested in the effect of a treatment on some outcome variable Y, the ideal situation would be to observe how the outcome changes when the same units receive both the treatment and the control. We could then estimate the effect of the treatment as $E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$, the mean difference in outcomes between the units when they are treated and when they are not.

Unfortunately, in the real world we never observe the same units both receiving treatment and not at the same time: units either receive the treatment or they do not. We therefore only observe the outcome under treatment for units that are treated, and the outcome under control for units that are not treated. In the potential outcomes framework, we think of four possible conditions we could hope to observe:

$$E[Y_i^T|T_i = 1] \text{ Outcome under treatment for treated} \tag{1}$$

$$E[Y_i^T|T_i = 0] \text{ Outcome under control for treated} \tag{2}$$

$$E[Y_i^C|T_i = 0] \text{ Outcome under control for control} \tag{3}$$

$$E[Y_i^C|T_i = 1] \text{ Outcome under treatment for control} \tag{4}$$

We can observe (1) and (3) but never (2) or (4), the counterfactuals we would have expected in the treatment and control groups if they had not or had received the treatment, respectively.

When we estimate the impact of a program we are therefore constrained to estimating $E[Y_i^T|T_i = 1] - E[Y_i^C|T_i = 0]$. When does this recover the counterfactual we would have hoped to observe? Only if there are no differences (on average) between the units that did and did not receive the treatment.

We can see this by adding and subtracting $E[Y_i^T|T = 0]$ to the above. Rearranging gives us:

$$\underbrace{(E[Y_i^T|T = 1] - E[Y_i^T|T = 0])}_{\text{True effect}} + \underbrace{(E[Y_i^T|T = 0] - E[Y_i^C|T = 0])}_{\text{Selection bias}}$$

or in other words, the *true* impact of the program plus any latent differences between the two groups, which we call *selection bias*: factors that leads to the different groups being *selected* in some way into the groups, as opposed to randomly assigned. So how can we recover the true effect? One way we can accomplish this is by **randomizing** the treatment. With a large enough randomly assigned sample, we will have $E[Y_i^T|T_i = 0] = E[Y_i^C|T_i = 0]$, recovering the missing counterfactual.

What randomization accomplishes is to help make sure that treatment status is not correlated with other variables which might affect the outcome of interest, so we can be sure we are isolating the effect of the treatment rather than differences between the treatment and control groups. Of course, even under randomization we may still have some variables that are significantly different across treatment and control groups by random chance in a given sample[1]. The groups will only be identical in expectation, across repeated samples.

III. Introducing the Randomized Controlled Trial (RCT)

i. What is a Randomized Controlled Trial (RCT)

- When feasible, randomization is the most rigorous approach to construct a treatment and a control group from among an eligible population (see Duflo, 2006; Banerjee and Duflo, 2009).
- What do we mean by randomization? Well if we want to know the effect of health insurance for example, we could go out and select a sample of people who are uninsured. We would then provide health insurance to a randomly chosen subset of this sample. We can then interview everyone in our sample 6 months later, and compare the health outcomes of those who received insurance (the treatment group), to those who did not (the control group).
- Why does this work? Because randomization over a sufficiently large number of units creates statistically identical treatment and control groups. If randomization is properly done, and the samples are large enough, the two groups should not have any statistically different features that could be correlated with our treatment and our outcome variable. Impact can then be measured by simple difference in the outcome variable between treatment and control groups.

ii. Causal Effect

With randomization, and after verification that there are on average no statistically significant differences between units[2] in the treatment and control groups based on observable characteristics (see below), the measure of impact can be obtained by simple difference in the average outcome between treatment and control groups:

$$Impact = \bar{Y}_T - \bar{Y}_C$$

where $\bar{Y}_T$ is the average outcome for individuals in group T, and $\bar{Y}_C$ the average outcome for individuals in group C.

In a regression framework, we can retrieve the causal effect by simply regressing our outcome of interest on our treatment variable:

$$Y_i = a + \beta_1 T_i + u_i$$

[1] When testing for significant differences in variable means between two groups, if we use a 5% threshold for a significant difference this implies that we should expect to see a significant difference in 5% of our variables on average.

[2] These could be individuals, households, schools, businesses, cities, etc.

This regression formula works because:

$$\begin{aligned}\bar{Y}_C &= E[Y_i|i \text{ in Control group}]\\ &= E[a+\beta_1 T_i + u_i|i \text{ in Control group}]\\ &= E[a|i \text{ in Control group}] + E[\beta_1 T_i|i \text{ in Control group}] + E[u_i|i \text{ in Control group}]\\ &= a+0+0\\ &= a\end{aligned}$$

and

$$\begin{aligned}\bar{Y}_T &= E[Y_i|i \text{ in Treatment group}]\\ &= E[a+\beta_1 T_i + u_i|i \text{ in Treatment group}]\\ &= E[a|i \text{ in Treatment group}] + E[\beta_1 T_i|i \text{ in Treatment group}] + E[u_i|i \text{ in Treatment group}]\\ &= a+\beta_1+0\\ &= a+\beta_1\end{aligned}$$

so

$$\begin{aligned}\bar{Y}_T - \bar{Y}_C &= a+\beta_1-a\\ &= \beta_1\end{aligned}$$

The regression will provide us with an estimate of β_1, as well as standard errors. We can use this to make/test hypotheses about the significance of the treatment variable in explaining our outcome variable of interest.

In R:

```
lm(Y~treatment, data=df)
```

Because T is a dummy variable indicating treatment, we can interpret the coefficient β_1 as we usually do for a dummy variable: it is an intercept shifter for being in the dummy category compared to the left-out category. In this case, this is exactly the difference in means for Y between the treatment and control groups.

iii. Key Assumption

The key assumption to recover a causal impact of the treatment is that if it were not for the treatment, the control and the treatment population would be statistically identical, i.e., have identical expected values for the outcome of interest, regardless of whether they are assigned to treatment/control.

$$E[Y_i|i \text{ in Treatment group}, T] = E[Y_j|j \text{ in Control group}, T]$$

This says that if we hadn't yet administered treatment we would expect our outcomes (the y variables) to be the same across both groups.

In a regression framework, the key assumption is SLR4:

$$E[u_i|T_i=0] = E[u_i|T_i=1] = 0$$

In other words, there are no other variables correlated with the outcome Y that are correlated with treatment status. This ensures that $\hat{\beta}_1$ is an unbiased estimator of β_1.

iv. Tests for Validity of Assumption

We cannot test

$$E[u_i|T_i = 0] = E[u_i|T_i = 1] = 0$$

But we can check to see if the observable characteristics among treatment and control groups are the same on average.

Ex: if we want to know the effects of health insurance on health outcomes, and we want to provide some assurance that absent the treatment, our treatment and control groups would have the same outcome, we can show that the average level of education, age, and income (among other variables) are the same across treatment and control groups before the treatment.

What we are doing is testing for "the equality of means of the **observed** characteristics," with the idea that if there is no statistical difference in the observable characteristics (eg. age, education, income), then this provides plausible evidence that there is no difference in the **unobserved** characteristics as well (eg. ability, motivation, level of hypochondria)

Formally, we want to test that for any observed variable (e.g., age, education, income):

$$E[x_i|i \text{ in Treatment group }] = E[x_i|i \text{ in Control group}]$$

cannot be rejected prior to the treatment being implemented.

In R we can test this using the t.test() command:

```
t.test(df[df$treatment==0,]$age,df[df$treatment==1,]$age)
t.test(df[df$treatment==0,]$education,df[df$treatment==1,]$education)
```

v. Adding Covariates

When we do an RCT, we always want to run a simple regression of the outcome of interest on our treatment variable, as we did above. But we can run additional regressions where we also add observable covariates (additional X / left-hand side variables) to:

1. Add precision to the estimation
2. Verify, as a robustness check, that $\hat{\beta}$ is invariant to the introduction of covariates in the regression

Why does adding covariates add precision? Think about the formula for the variance/standard error of our estimator:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x(1-R_j^2)}$$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x(1-R_j^2)}}$$

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_i^n \hat{u}_i^2$$

If we include more $X's$ in our regression, we can reduce $\hat{u}_i^2$, i.e. the unexplained variation in Y goes down, which means $se(\hat{\beta}_1)$ decreases, which means our $\hat{\beta}$ can be estimated more precisely (think about the confidence interval for our estimate getting smaller).

Why don't we expect $\hat{\beta}$ to change? We don't expect our estimate to change precisely because we have randomly selected people to be in the control and treatment group. In other words the observable characteristics of the treatment group shouldn't be correlated with any features of the policy/reform/intervention we gave to the treatment group. But in a given sample, we might by chance observe some statistically significant differences by treatment status for some variables. We can then control for those differences in the regression to again isolate the effect of the treatment.

In a regression framework, we can simply add covariates :

$$Y_i = a + \beta_1 T_i + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i$$

In R:

```
lm(Y~treatment +x_1 +x_2 +x_3 +...+x_k, data=df)
```

vi. Heterogeneity

Finally, we can also measure heterogeneity of the program effect for individuals with specific characteristics (such as gender, age, socio-economic status, etc.) by interacting these characteristics with the treatment variable. The implication is that the program has a differential effect on certain subgroups of the population.

In a regression framework, we can simply add an interaction :

$$Y_i = a + \beta_1 T_i + \beta_2 x_{2i} + \beta_3 T_i \times x_{2i} + u_i$$

If the variable x_2 represents a dummy for being female for example, then β_3 gives us the differential effect of the treatment for females relatives to males.

vii. Example

Let's practice evaluting impact with an RCT, using a real-life example.

In order to increase school enrollment among the youth in poor rural families, the Mexican government introduced the PROGRESA program in 1998. Household enrollment into the program was randomized. For treatment hosueholds, the program gives cash transfers to mothers in poor households if their children attend school regularly and receive periodical medical checkups. We focus on a subset of data for children 12-13 years old that have just completed primary school.

1. *Estimate the impact of the program on school enrollment.*

   Section III.ii above

   This question is asking for the causal effect of the program on school enrollment. Simply run a regression of the outcome of interest on the treatment variable.

```
Call:
lm(formula = enroll98 ~ program, data = progresa)

Residuals:
<Labelled double>: Enrolled in school for academic year 1998
     Min       1Q   Median       3Q      Max
-0.85111  0.14889  0.14889  0.22170  0.22170

Labels:
 value label
     0    no
     1   yes

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.77830    0.01869  41.638  < 2e-16 ***
program      0.07280    0.02545   2.861  0.00432 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3849 on 919 degrees of freedom
Multiple R-squared:  0.00883,Adjusted R-squared:  0.007751
F-statistic: 8.187 on 1 and 919 DF,  p-value: 0.004316
```

Always remember to comment:

- Sign: There is a positive predicted effect of the Progresa program on school enrollment.
- Size: Being in the Progresa treatment group is associated with a 0.072808 increase in predicted school enrollment for 12-13 year olds in the year 1998. (Since enroll98 is a binary variable, we can think of this as a 7.28 percentage point increase in the likelihood of attending the school due to being randomly enrolled in the program.)
- Significance: The t-statistic is $|2.86| > |1.96|$, and the p-value is $= 0.004$. We reject the null hypothesis in favor of the alternative at the 5% significance level. As a result, we conclude that there is statistical evidence against the null hypothesis that Progresa has no effect on expected school enrollment.

2. *What are the conditions for the validity of the method to measure the causal impact of the program*?

   Section II.iii above

   This question is asking about the key assumption that allow us to say we have estimated a causal impact of the program. You can either write:

$$E[Y_i | i \text{ in Treatment group}, T] = E[Y_j | j \text{ in Control group}, T]$$

   or

$$E[u_i | T_i = 0] = E[u_i | T_i = 1] = 0$$

And explain that the treatment is randomly assigned and hence: if it were not for the treatment, the control and the treatment population would be statistically identical, i.e., the treatment is not correlated with any unobservables in the error term.

3. *Proceed with the appropriate tests that support the validity of the randomization*

   Section III.iv above

   We can test for the equality of means of the observed characteristics between the treatment and control group (null is that the means are equal). The below output reports results of tests for equality of means for sex and age of the child, education of the head of household, household size, household expenditures, and distance to secondary school.

```
> t.test(progresa[progresa$program==0,]$male, progresa[progresa$program==1,]$male)

Welch Two Sample t-test

data:  progresa[progresa$program == 0, ]$male and progresa[progresa$program == 1, ]$male
t = 0.38682, df = 896.31, p-value = 0.699
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05213759  0.07773457
sample estimates:
mean of x mean of y
0.5117925 0.4989940

> t.test(progresa[progresa$program==0,]$age97, progresa[progresa$program==1,]$age97)

Welch Two Sample t-test

data:  progresa[progresa$program == 0, ]$age97 and progresa[progresa$program == 1, ]$age97
t = -0.51895, df = 896.34, p-value = 0.6039
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.08210035  0.04776225
sample estimates:
mean of x mean of y
 11.48585  11.50302

> t.test(progresa[progresa$program==0,]$h_edu, progresa[progresa$program==1,]$h_edu)

Welch Two Sample t-test

data:  progresa[progresa$program == 0, ]$h_edu and progresa[progresa$program == 1, ]$h_edu
t = -1.3205, df = 917.72, p-value = 0.187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.49800435  0.09739314
sample estimates:
mean of x mean of y
 2.334906  2.535211

> t.test(progresa[progresa$program==0,]$hhsize, progresa[progresa$program==1,]$hhsize)

Welch Two Sample t-test

data:  progresa[progresa$program == 0, ]$hhsize and progresa[progresa$program == 1, ]$hhsize
t = -0.87808, df = 879.3, p-value = 0.3801
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4133482  0.1578150
sample estimates:
mean of x mean of y
 7.500000  7.627767

> t.test(progresa[progresa$program==0,]$exp98, progresa[progresa$program==1,]$exp98)

Welch Two Sample t-test

data:  progresa[progresa$program == 0, ]$exp98 and progresa[progresa$program == 1, ]$exp98
t = -1.1645, df = 890.96, p-value = 0.2445
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.02483   3.57963
sample estimates:
mean of x mean of y
 104.7439  109.9665

> t.test(progresa[progresa$program==0,]$distsec, progresa[progresa$program==1,]$distsec)

Welch Two Sample t-test

data:  progresa[progresa$program == 0, ]$distsec and progresa[progresa$program == 1, ]$dists
t = 0.65777, df = 852.68, p-value = 0.5109
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2071203  0.4159167
sample estimates:
mean of x mean of y
 2.679601  2.575203
```

We fail to reject the null hypothesis that the differences in the observed characteristics between the treatment and control group are 0 at the 5% significance level (true for all these characteristics). This demonstrates that the treatment group and control group are statistically the same in terms of observed characteristics, which provides evidence that the average unobserved characterstics are the same across treatment and control groups as well. This supports the validity of the randomization.

4. *Add variables that may influence the decision to enroll in school in the regression. Do they actually explain enrollment? Do they affect the estimated coefficient of the program? Is that what you expected, and why?*

   Section III.v above

   We added all the additional explanatory variables we have:

```
Call:
lm(formula = enroll98 ~ program + distsec + exp98 + hhsize +
    h_edu + age97 + male, data = progresa)

Residuals:
<Labelled double>: Enrolled in school for academic year 1998
     Min       1Q   Median       3Q      Max
-0.97452  0.00730  0.12262  0.22414  0.69605

Labels:
 value label
     0    no
     1   yes

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.5773471  0.2849907   9.044  < 2e-16 ***
program      0.0740452  0.0243661   3.039  0.00244 **
distsec     -0.0288689  0.0051495  -5.606 2.74e-08 ***
exp98       -0.0004827  0.0001896  -2.546  0.01106 *
hhsize      -0.0049897  0.0058742  -0.849  0.39587
h_edu        0.0123590  0.0052988   2.332  0.01989 *
age97       -0.1483166  0.0242845  -6.107 1.50e-09 ***
male         0.0800760  0.0242560   3.301  0.00100 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   These variables are statistically significant at the 5% level, with the exception of household size. Indeed the t-statistics are well above 1.96 in absolute value (and/or the p-values are less than 0.05). This means that these variables help explain enrollment.

   These variables do not however significantly affect the estimated coefficient of the program (which moves from 0.072 to 0.074). This is what we expected because we have a randomized control trial and so the treatment shouldn't be significantly correlated with any of the observable characteristics. [3]

5. *Now, write a simple equation (without unnecessary variables) that allows you to test whether the impact of program is equal for boys and girls. Report the result and comment.*

   Section III.vi above

$$Y_i = a + \beta_1 program + \beta_2 male + \beta_3 program \times male + u_i$$

```
Call:
lm(formula = enroll98 ~ program + male + program:male, data = progresa)

Residuals:
<Labelled double>: Enrolled in school for academic year 1998
     Min       1Q   Median       3Q      Max
-0.88710  0.11290  0.18433  0.18474  0.26087

Labels:
 value label
     0    no
     1   yes

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.739130   0.026657  27.728   <2e-16 ***
program       0.076131   0.036073   2.110   0.0351 *
male          0.076538   0.037261   2.054   0.0403 *
program:male -0.004702   0.050717  -0.093   0.9262
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3835 on 917 degrees of freedom
Multiple R-squared:  0.01802,Adjusted R-squared:  0.0148
F-statistic: 5.608 on 3 and 917 DF,  p-value: 0.0008223
```

There is a -0.004702 differential effect in predicted enrollment for boys relative to girls in the program. The t-statistic of the interaction variable, is $|t| = .09 < 1.96$, so we fail to reject the null in favor of the alternative at the 5% significance. This suggests that the impact of the program is the same for boys and girls.

[3] Note: We concede that it does move ever so slightly - and that's because there might be some incidental/small/insignificant correlation between the treatment variable with one of the observable characteristics