ECON 485: Randomized Control Trials (RCT)

Isaac Shon

June 4, 2024

1 The "Gold Standard" of Empirical Research

The Randomized Control Trial (RCT) is an ideal research design for applied researchers wishing to estimate causal effects of treatments/policy interventions with strong internal validity, and is otherwise known as the "gold standard" approach for empirical research. In this design, the units in the sample are randomly assigned the treatment or intervention of interest. This approach necessarily requires that the researcher has *ex-ante* complete control over the treatment assignment procedure.

In a randomized control trial, the researcher randomly assigns units of the sample the treatment of interest. For each unit $i = 1, 2, ..., n$, a unit's treatment status is randomly assigned $T = 1$ or $T = 0$. After the randomized treatment assignment, we then compare the outcome of interest between the "treated" group and "untreated" group after the treatment is implemented. In the simplest regression framework, we can estimate the average treatment effect (ATE) of the intervention T through:

$$Y_i = \beta_0 + \beta_1 T_i + u_i \tag{1}$$

In a regression model with a binary treatment variable, our estimate of β_1 through OLS represents the difference in group means for our treated units ($Ti = 1$), in comparison to our untreated units ($T_i = 0$). However, the interpretation of the causal effect of T on Y can be extended to multi-valued treatments.

Because we randomly assigned the treatment to our units, OLS is actually an unbiased and consistent estimator of β_1, and we actually address selection and omitted variable bias. In other words, since our treatment or intervention is randomly assigned, our treatment dummy variable is independent of potential outcomes (i.e., $T_i \perp\!\!\!\perp (Y_{1i}, Y_{0i})$). We also ensure that $\mathbb{E}[u|T] = 0$ and that there is no systematic selection into the treatment. As such, including a set of covariates X_i for our units should not change our OLS estimate for β.

Adding covariates to our regression specification offers several practical advantages. First, by including other covariates to our model, we are able to reduce the standard error on our beta coefficient. This gives us more precision in our estimation of treatment effects. Second, by controlling for other variables, we can also explore potential heterogeneity of treatment effects. Without covariates, we do not know how our treatment effects change across different subgroups and we assume constant treatment effects. Additionally, including observed pre-treatment covariates in our model allows us to conduct balance tests to ensure that our randomization process was successfully administered.

2 Simulation Exercise

In this section, I provide an example of a randomized control trial using simulated data. The corresponding STATA do-file can be found in this assignment's folder. Here, I create a sample of 1,000 observations. I define 3 arbitrary covariates X_1, X_2, X_3, an error term ε that is drawn from a random distribution centered at 0, and a treatment variable T that randomly assigns observations values 0 or 1. Below is a summary table of the simulated data:

Table 1: Summary Table of Select Variables

	mean	sd	min	max
Y_i	28.51141	3.081336	18.78646	39.23702
T_i	.4956908	.2888258	.0007337	.999698
X_1	6.992366	1.062562	3.341623	10.12844
X_2	2.010333	.5081502	.0803124	3.465119
X_3	2.992982	.4894795	1.480373	4.622731
epsilon_i	.0262131	.9869443	-3.100515	2.85296
N	1000			

In this exercise, I define the following population model:

$$y = \beta_1 T + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + u, \tag{2}$$

where I set the population parameters $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = 2.5$, and $\beta_4 = 3$. OLS regression results are reported in the following table:

As can be seen from Table 2, the addition of covariates X_1, X_2, X_3 (drawn from separately-defined distributions) hardly changes our coefficient estimate, $\hat{\beta}_1$. At the same time, from models (1)-(4), we see an increasing reduction in the estimated standard errors for our coefficient estimate. This is because as we move from models (1)-(4), we allow some of the variation in our outcome variable y to be explained by the other covariates.

Table 2: Regression Results Using Simulated Data

	(1)	(2)	(3)	(4)
	OLS	OLS & 1 Covariate	OLS & 2 Covariates	OLS & 3 Covariates
T_i	1.009***	0.874***	0.832***	1.038***
	(0.334)	(0.245)	(0.203)	(0.108)
X_1		2.001***	1.956***	2.001***
		(0.066)	(0.054)	(0.028)
X_2			2.561***	2.470***
			(0.110)	(0.061)
X_3				3.052***
				(0.063)
_cons	28.011***	14.084***	9.272***	-0.099
	(0.196)	(0.487)	(0.439)	(0.297)
N	1000	1000	1000	1000

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

3 Conclusion

While RCTs are a powerful tool that may potentially deliver internally valid causal effects, there is still a need for researchers to remain cautious and exercise diligence with their particular research designs. More generally, in RCT studies, several major threats to identification of causal effects include:

- **Unsuccessful randomization of the treatment.** If the randomization is unsuccessful,we may introduce selection bias to our analysis. Balance tests of covariates can test for randomization.

- **Violation of SUTVA.** The potential outcome of a given unit should only depend on its own treatment status, not other units. Estimates of causal effects may be biased when treatment "spills over" to the control group, so the researcher must be careful in treatment implementation.

- **Sample attrition/drop-outs.** Drop-out of units reduce the statistical power of the study. However, more importantly this threat can also result in imbalances between the treatment and control groups. This can be addressed by observing more units or reduce the time passed post-intervention so more observations are retained.