Exploring Relationships between Sample Size, Power, and MDE through Power Simulations

Jerry Hong

October 16, 2024

Abstract

In statistical inference and hypothesis testing, understanding the dynamics of sample size, power and effect size are important when conducting such studies. Each of these attributes directly influence one another through both direct and inverse means. What follows is a series of power simulations showcasing the relationships among these factors through both a synthetic data generation and a household survey data set. These simulations can allow us to find one of the unknown factors when the other two are known. Mainly, finding the minimum detectable effect (MDE) and the sample are most common to find through these simulations.

1 Power Simulations from Synthetic Data

My methodology for these simulations is through a randomized data generating process (DGP). Through my DGP, I will generate a sample of outcome values from some constant, a treatment variable, and a value for random noise. The random covariates are from a normal distribution with mean 0 and standard deviation of 1.

1.1 Power Analysis With and Without Covariates

Throughout these simulations, I will hold the level of power at 0.8 as this is the traditional value of statistical power where, under certain assumptions, including but not limited to the particular sample size and true effect size, these tests have an 80% probability of rejecting the null hypothesis at a 5% significance level. I run 1000 simulations at different sample sizes from 100 to 1000 stepped by 50 and calculate the statistical power. Each simulation will generate the outcome variable y from the randomly generated values of my inputs. Each loop simulates a randomized treatment experiment and checks if the treatment effect is statistically significant where p-value ≤ 0.05. From there, the power is calculated as a proportion of simulations when the treatment effect is significant. I will graph this relationship and indicate the sample size necessary for a statistical power of 0.8 given the effect size set at 0.2. Each plot will have a quadratic fit over the simulation with an estimated sample size at power level 0.8. I will compare simulations from models with and without covariates and explore how might sample size be impacted at the same power level.

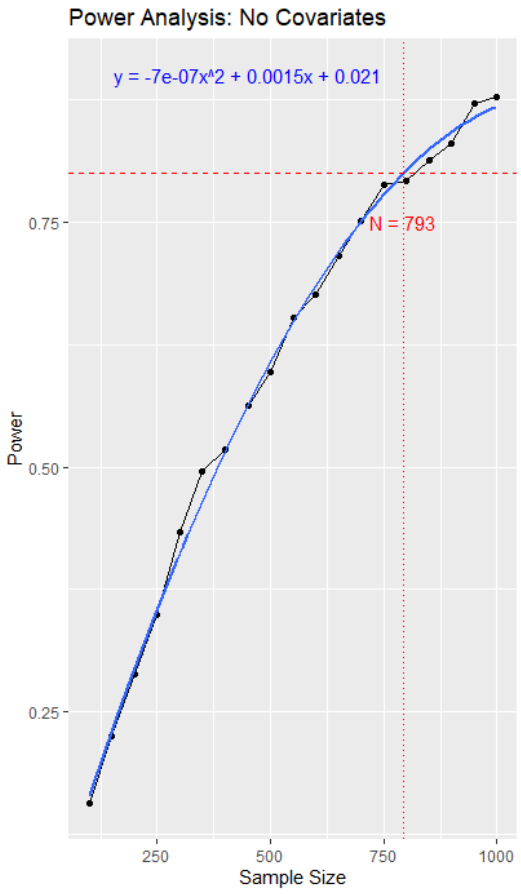

Figure 1: Power Simulation with No Covariates

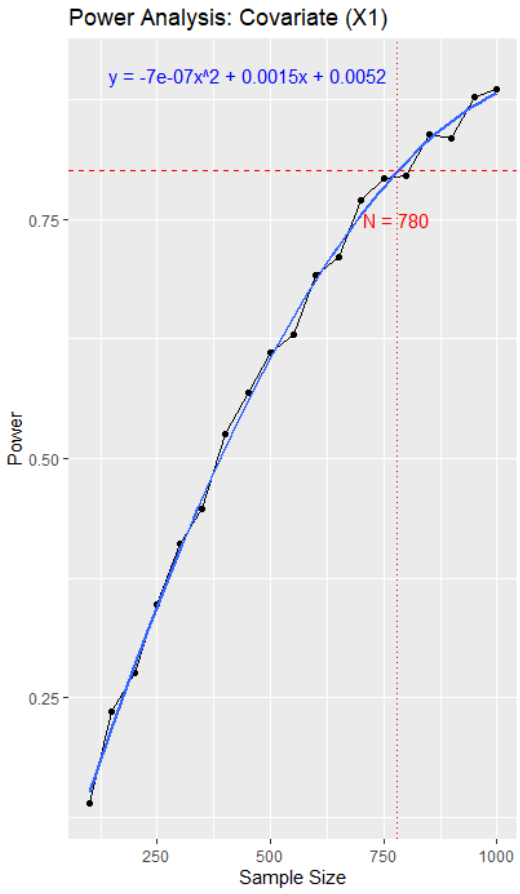

Figure 2: Power Simulation with One Covariate (X1)

Based on Figure 1, it seems that the necessary sample size at the power level of 0.8 is 793 or around 800, which makes sense given there are no covariates involved. Granted, this is from one run, and thus, I would need to run this enough times to ensure consistency.

Figure 2 shows the sample size if I add one random covariate to the model. The covariate $X1$ has a coefficient of 1. Usually, by adding a covariate to the model, we expect the sample size to somewhat decrease as it can help improve the model's explanatory power. Now the estimated sample is around 780, which is not as significant of decrease, but it requires less samples than the base model nonetheless.

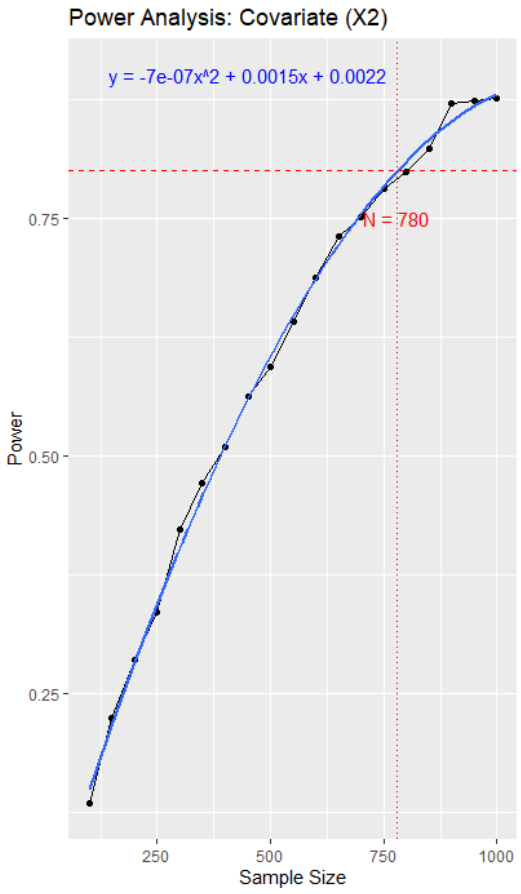

Figure 3: Power Simulation with One Covariate (X1)

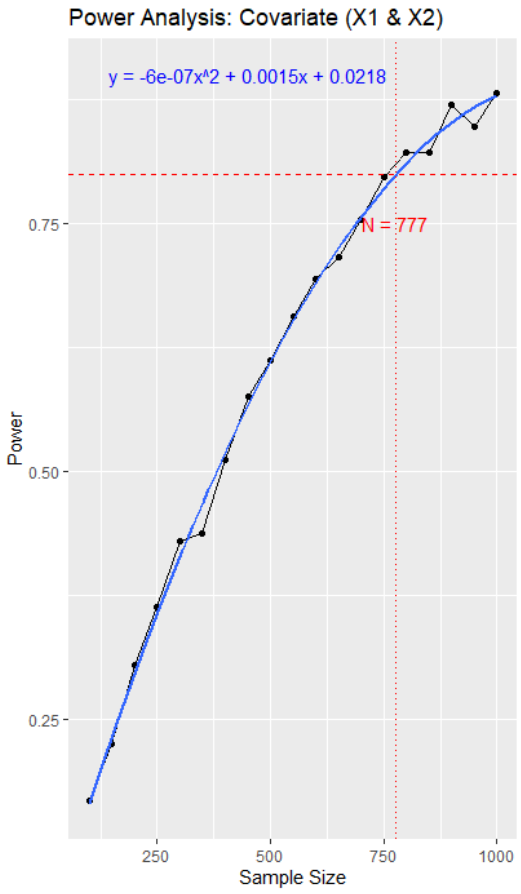

Figure 4: Power Simulation with Both Covariates

Figure 3 includes the covariate $X2$, but with a coefficient of 5. I expect that given the higher coefficient, the estimated sample size should decrease. However, we see that the sample size is identical with Figure 2 at 780. This may be due to margin of error given this was only one run. Figure 4 combines both covariates, and we see this led to the lowest estimated sample size among the four models. This supports the idea that adding more covariates to the model will decrease the estimated sample size. Of course, these values are within margins of each other, so more simulated runs is necessary to further support this idea.

1.2 Finding the MDE from a Fixed Sample Size

Using a fixed sample size of 500, I will explore the relationship between the MDE and Power and find the estimated MDE at power level 0.8. I will estimate this value in both the unconditional and conditional covariate specifications.

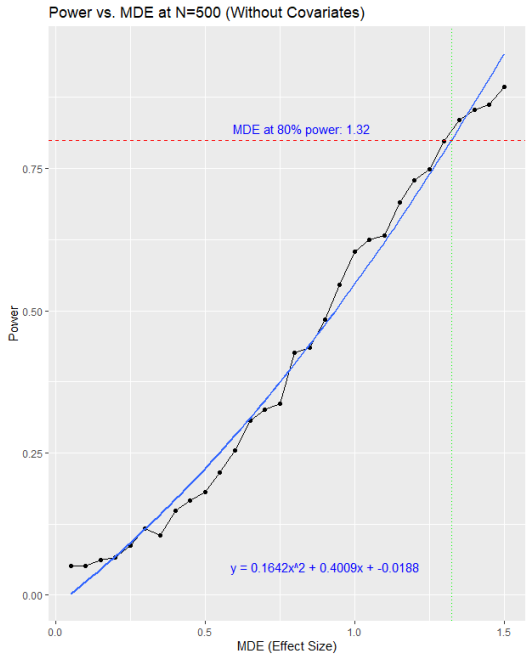

Figure 5: Power Simulation with No Covariates (MDE)

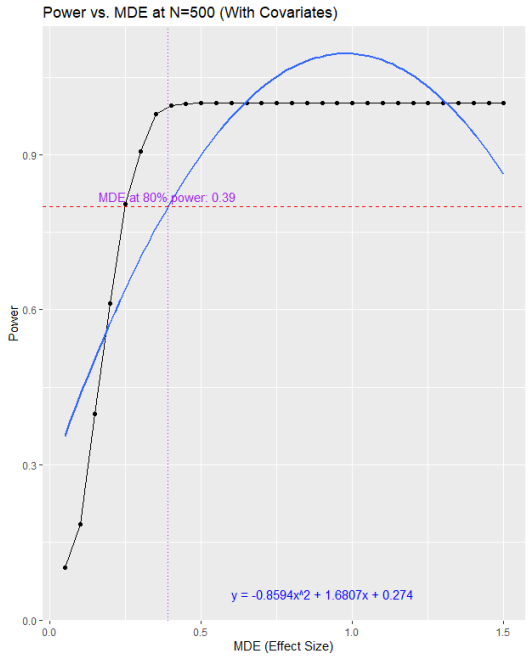

Figure 6: Power Simulation with Both Covariates (MDE)

2 Power Simulations using Household Data

I expand the ideas from the previous simulations and apply them through real data rather synthetic data. This data set is from a causal experiment evaluating the effect of a job skills training program in a lower-income country. The outcome variable of choice is "calculated.EPC" which is the earnings per capita per month. In the figures below, I will first estimate the sample size to detect an MDE of 0.2 standard deviations at 80% power with this outcome variable. This is in both the unconditional and conditional covariate specifications. Next, I will use the sample size of the conditional specification and estimate the MDE at 80% power.

2.1 Unconstrained Implementation Budget

In the unconditioned specification, there will be no covariates involved and it is a essentially a simulation like Figure 1. I will estimate the sample size for both models.

In the conditioned specification, I am using the covariates: `score`, `hhsize`, and `expenditure_rent_pc`. They represent from my understanding, the job skill score, household size, and expenditure on rent per capita respectively. Re-running the simulation led to an interesting result.

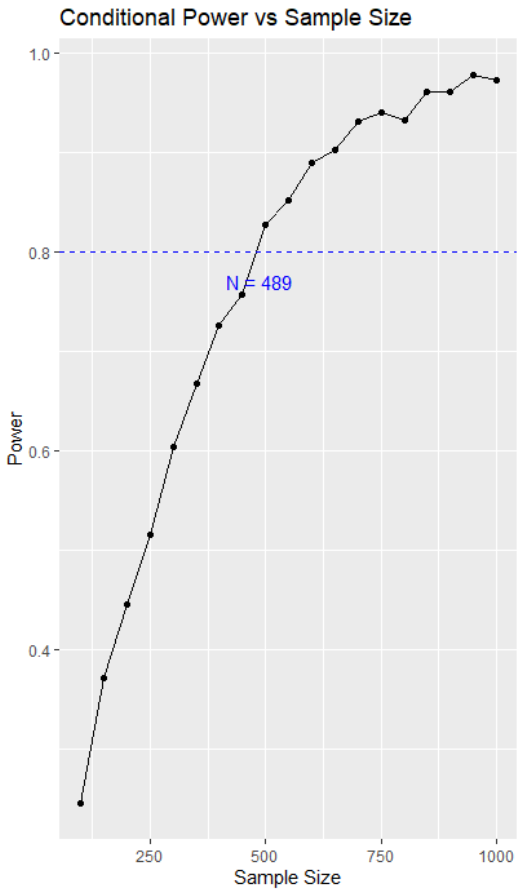

Figure 7: Power Simulation with No Covariates (MDE)

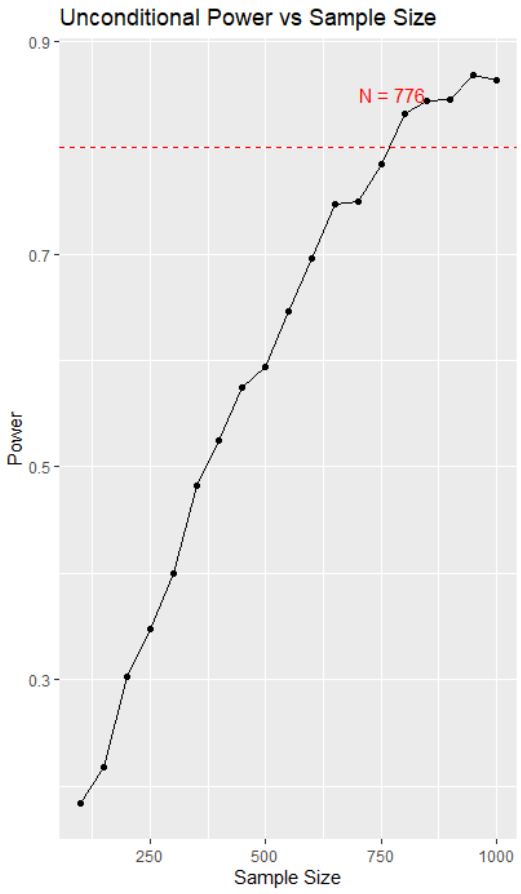

Figure 8: Power Simulation with Both Covariates (MDE)

In Figure 8, the estimated sample size at 80% power is around 776, with the covariate one in Figure 7 having a sample of 489. This is a rather striking comparison that reinforces the idea of incorporating strong covariates that can decrease the sample size considerably. This may imply that the covariates of choice have some significance in the MDE for this causal effect.

2.2 Constrained Budget

Taking the sample size from the conditioned specification, I will impose a constraint to find the minimum detectable effect at power level 0.8. Figure 9 shows that the simulated values as it we increase the MDE, the power plateaus slightly above 0.9 power.

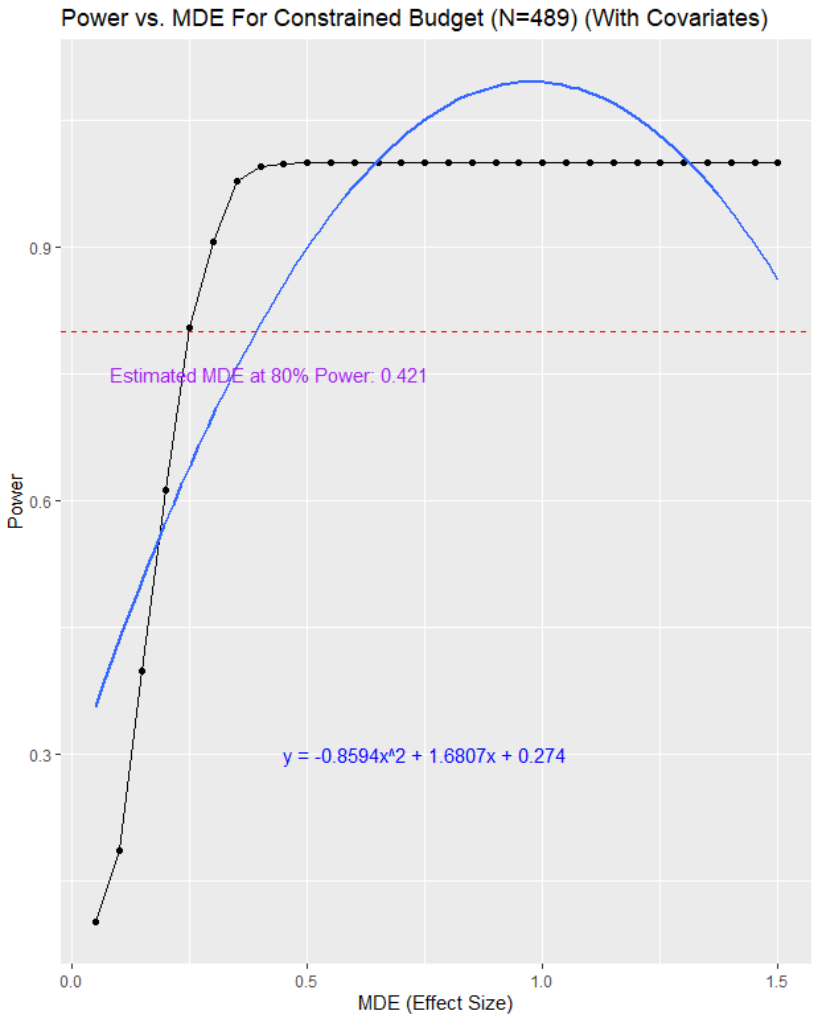

Figure 9: Constrained Power Simulation with Covariates at Fixed Sample Size (N=489)

However, from the quadratic fit, the estimated MDE at the power level 0.8 is about 0.42. With the budget constrained at a sample size of 489, the intervention of the job skills program would need to generate an effect of at least 0.42 in order to be detected at 80% statistical power. In addition, this might lead to the intervention appearing less effective if smaller effects are present but undetected due to the constrained sample size.

Disclaimer

The figures and models presented in this document have been generated with the assistance of artificial intelligence tools. The findings are reviewed and verified by the author(s) to ensure accuracy and clarity.