Covariate Adjustment and Statistical Power

Tara Slough

EGAP Learning Days X

Covariate Adjustment

- Covariate adjustment = "controlling" for variables in multiple regression.
- Regression model without covariate adjustment:

$$Y_i = \beta_0 + \beta_1 Z_i + \epsilon_i \tag{1}$$

- Regression model with covariate adjustment

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \epsilon_i \tag{2}$$

- Z_i is the treatment, X_i is a covariate

Justification for "controls" in observational research

- In observational research (not quasi-experimental):
  - Some $X_1 \rightarrow Z$ and $X_1 \rightarrow Y$
  - We care about estimating the causal effect of Z, so we need to adjust for X_1
  - But there may be some unobserved/unmeasured $u_1 \rightarrow Z$ and $u_1 \rightarrow Y$.
  - We can't control for u_1 if we can't observe/measure it. This induces **omitted variable bias**.
- In experimental research:
  - By random assignment, $Z \perp X$. It still is the case that $X \rightarrow Y$
  - By random assignment, $Z \perp u_1$. It still is the case that $u_1 \rightarrow Y$

Justification for covariate adjustment in experiments

- Recall that:
  - By random assignment, $Z \perp X_1$. It still is the case that $X_1 \rightarrow Y$
- So if we adjust for X_1 we can mop up (reduce) variance in Y.
- Improves precision in the detection of treatment effects of Z
- Covariate adjustment can also increase precision in observational research
  - But can also be quite costly...

The cost of covariate adjustment

- "Bad" control: Suppose that
  - $Z \rightarrow Y$
  - $Z \rightarrow X_2$
  - $X_2 \rightarrow Y$
- If we control for X_2 (a function of Z), we can induce bias in our estimate of the causal effect of Z
  - In *experimental* or *observational* research
  - One form of post-treatment bias
- How do we avoid "bad" controls:
  - Do not control/adjust for anything temporally *after* treatment (no post-treatment controls)

Implications

- Not unambiguously good to dump in more and more controls
- Robustness tests in published literature often don't make sense
- Does it make sense to ask someone if they have "controlled" for some X in an experiment?

False Negatives and Power

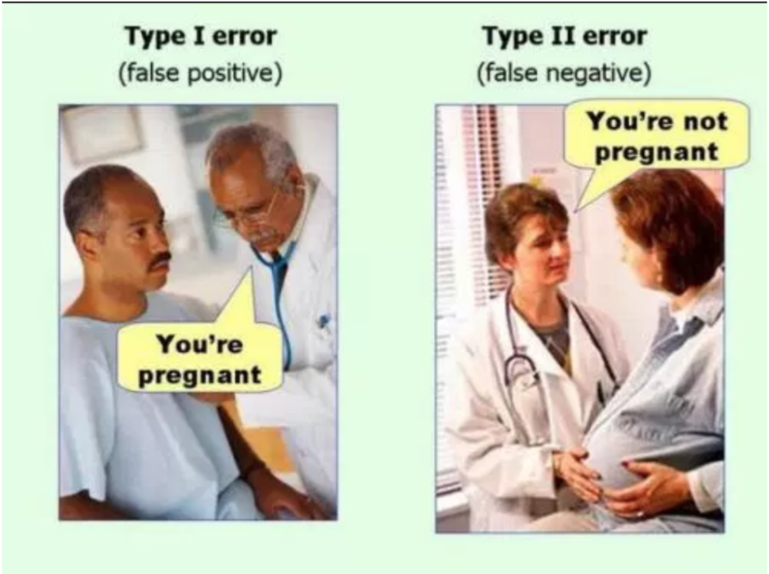


Figure 1: Illustration of error types.

What is statistical power and why should we care?

What is power?

- Probability of rejecting null hypothesis, given true effect $\neq 0$.
- Informally: our ability to detect a non-zero effect given that it exists.
- Formally: 1 - Type II error rate

Why do we care?

- [Null findings should be published.]
- But: hard to learn from an under-powered null finding.
- Avoid “wasting” money/effort.

General Approach to Power Calculations

- Ex-ante:
  - Analytical power calculations: plug and chug
    - Only derived for some estimands (ATE/ITT)
    - Makes strong assumptions about DGP/potential outcomes functions
  - By simulation
    - Create dataset and simulate research design
    - You make your own assumptions, but assumptions are made(!)
    - `DeclareDesign` approach
- Ex-post:
  - We don't really do this but probably should.
  - Still requires assumptions.

Power: The quantity

- Is a probability
  - Probability of rejecting null hypothesis (given true effect $\neq 0$)
  - Thus power $\in (0, 1)$
  - Standard thresholds: 0.8 or 0.9
    - What is the interpretation of power of 0.8?

Analytical Power Calculation: The ATE

- Two-tailed hypothesis test:

$$\text{Power} = \Phi\left(\underbrace{\frac{|\tau|\sqrt{N}}{2\sigma}}_{\text{Variable}} - \underbrace{\Phi^{-1}(1-\frac{\alpha}{2})}_{\text{Constant}}\right) \tag{3}$$

Components:

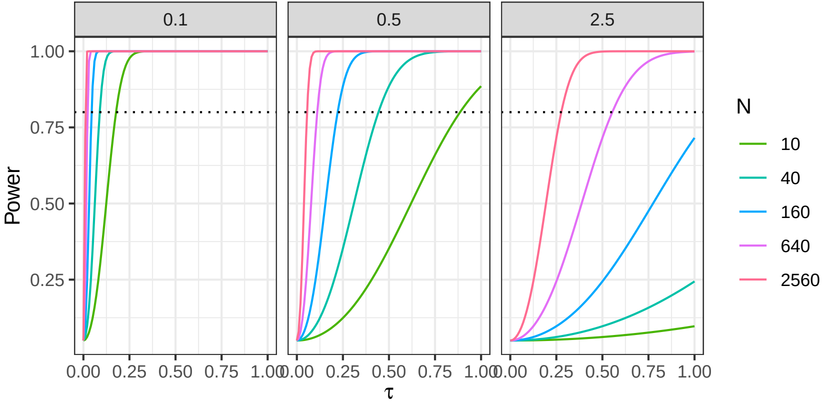
- Φ: Standard normal CDF is monotonically increasing
- τ: the effect size
- N: the sample size
- σ: the standard deviation of the outcome
- α: the significance level (typically 0.05)

Power: Comparative Statics

Power is:

- Increasing in $|\tau|$
- Increasing in N
- Decreasing in σ

Panels are increasing values of σ



Limitations to the Power Formula

- Limited to ATE/ITT
- Makes specific assumptions about the data generating process
- Incompatible with more complex designs

Alternative: Simulation

- Define the sample, assignment procedure
- Define the potential outcomes function
- Create data, estimate
- Do this many times; evaluate how many times

Power Simulation: Intuition

```
power_sim <- function(N, tau){
  Y0 <- rnorm(n = N)
  Z <- complete_ra(N = N)
  Y1 <- Y0 + Z * tau
  Yobs <- Z * Y1 + (1 - Z) * Y0
  estimator <- lm_robust(Yobs ~ Z)
  pval <-  estimator$p.value[2]
  return(pval)
}

sims <- replicate(n = 500,
                  expr = power_sim(N = 80, tau = .25))
sum(sims < 0.05)/length(sims)
```

```
## [1] 0.188
```

Power and Clustered Designs

- Given a fixed N, a clustered design is weakly less powered than a non-clustered design
  - The difference is often substantial
- To increase power
  - Better to increase number of clusters than number of units per cluster
  - How big of a hit to power depends critically on the intra-cluster correlation: ratio of variance within clusters to total variance
- Note: We have to estimate variance correctly:
  - Clustering standard errors (the usual)
  - Randomization inference

Clustering and Power: Variables

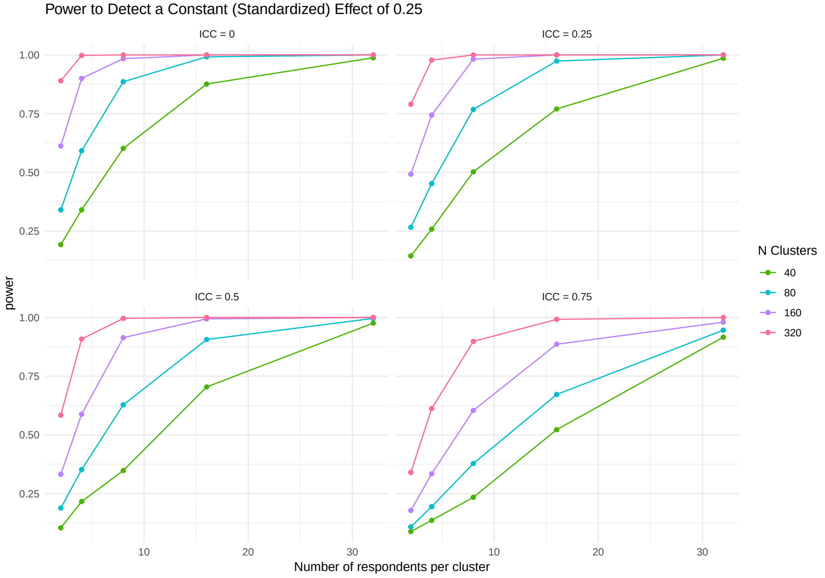
Variables

- Number of clusters $\in \{40, 80, 160, 320\}$
  - Clustered standard errors not consistent for fewer clusters
- Number of units per clusters $\in \{2, 4, 8, 16, 32\}$
- Intra-cluster correlation $\in \{0, .25, .5, .75\}$

Constants:

- $\tau = 0.25$ (standardized effect)

Demonstration of Clustering and Power

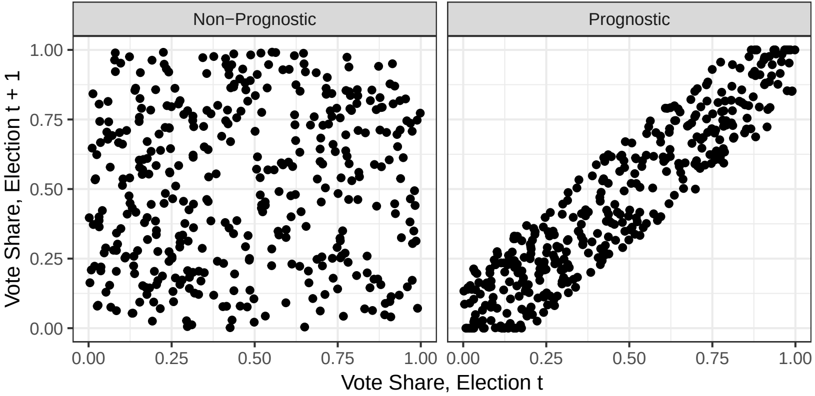


A Note on Clustering in Observational Research

- Often overlooked, leading to (possibly) wildly understated uncertainty
  - Frequentist inference based on ratio $\frac{\hat{\beta}}{\hat{se}}$
  - If we underestimate $\hat{se}$, we are much more likely to reject H_0. (Type-I error rate is too high.)
- Consider research on macro-economic conditions $\Rightarrow$ Voteshare for incumbent party with survey data
  - If treatment is macro-economic conditions, we should cluster at the *election* level
  - How many elections have there been in a given country?
  - Clustered SEs consistent for n $>$ 40 or 50 clusters
- Many observational designs much less powered than we think they are!

Why does covariate adjustment improve power?

- Mops up variation in the dependent variable
  - If prognostic, covariate adjustment can reduce variance dramatically: $\downarrow$ Variance $\Rightarrow$ $\uparrow$ Power
  - If non-prognostic, minimal power gains



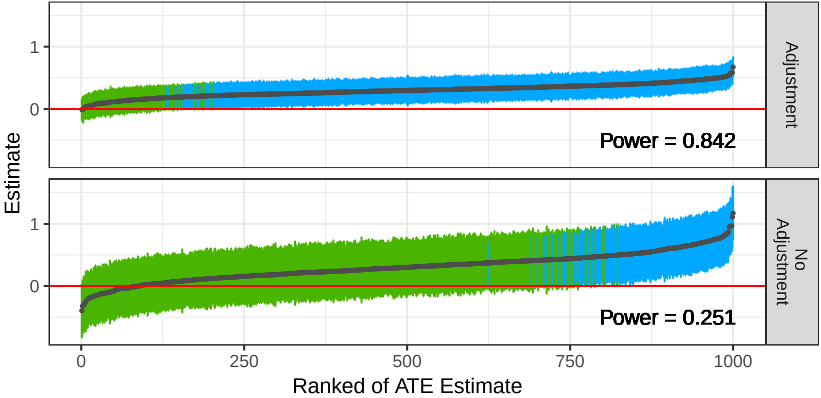
Covariate adjustment: Best Practices

- All covariates must be pretreatment
  - Never adjust for post-treatment variables
  - In an experiment looking at effects of leaflets on incumbent vote share, we should not "control" for turnout
- In practice, if all controls are pretreatment, you can add whatever controls you want
  - Until number of observations - number of controls < 20
- Missingness in pre-treatment covariates
  - Do not drop observations on account of pre-treatment missingness
  - Impute mean/median for pretreatment variable
  - Include missingness indicator and impute some value in the missing variable

Example of the Benefits of Covariate Adjustment

Consider the following:

$$X_i \sim \mathcal{N}(0, 1)$$
$$Y_i = X_i + 0.5 \times \mathcal{N}(0, 1) + \tau Z_i$$



Blocking

- Blocking: randomly assign treatment within blocks
- "Ex-ante" covariate adjustment
- Two benefits of blocking
  - Higher precision/efficiency $\rightarrow$ more power
  - Reduce "conditional bias": Association between treatment assignment and POs
- Benefits of blocking over covariate adjustment clearest in small experiments

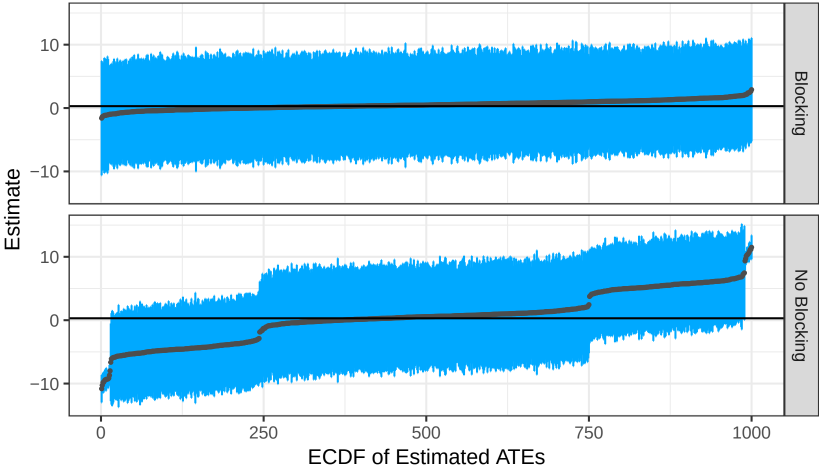
Example

- (Very small) experiment where blocks “explain” most variation in DV

```
block_sim <- function(){
  blocks <- rep(1:2, each = 4)
  Y0 <- rep(c(0, 10), each = 4) + rnorm(8)
  Zcomplete <- complete_ra(N = 8)
  Zblocked <- block_ra(blocks = blocks)
  Yobs1 <- Y0 + Zcomplete * .5
  Yobs2 <- Y0 + Zblocked * .5
  m1 <- lm(Yobs1 ~ Zcomplete)
  m2 <- lm(Yobs2 ~ Zblocked)
  return(c(coeftest(m1,
     vcov. = vcovHC(x = m1, type = "HC2"))[2,1:2],
           coeftest(m2,
     vcov. = vcovHC(x = m2, type = "HC2"))[2,1:2]))
}
```

Blocking Simulation Results

- Two benefits:
  - (Slight) efficiency gains – still have *huge* CIs
  - Reduction in conditional bias – kinks in line



Costs to Covariate Adjustment, Blocking

- Blocking
  - Sometimes harder to analyze correctly
  - If you block randomize and forget what the blocks are and blocks are anything but exactly vanilla, not great...
- Covariate Adjustment
  - Adjusting on a post-treatment variable is a big problem
  - Freedman's bias as n of observations decreases and K covariates increases

Comment on Power

- Know the dependent variable
  - What is the plausible range of variation?
  - Example 1: Effect of an intervention on corruption, measured in terms of public works projects
    - DV: Timing of contract completion (idea: corrupt projects take longer)
    - But do contracts ever complete early?
  - Example 2: Effect of a bias-reducing intervention
    - DV: Some behavioral measure of bias, only exhibited by 4% of participants in control
- An otherwise well-powered design with limited possible movement in the DV may not be powered to detect effects

A Note in Power in Factorial Designs:

- The usual regression-based estimator for factorial designs with T_1 and T_2 is:

$$Y_i = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_1 T_2$$

- Or consider the estimator that doesn't include the interaction:

$$Y_i = \gamma_0 + \gamma_1 T_1 + \gamma_2 T_2$$

Notes:

- The second estimator is generally well powered (estimand is subtly different)
- In the first estimator, β_3 is not very well powered, generally

Conclusion: How to improve your power:

1. Increase the N
   - If clustered, increase n clusters if at all possible
2. Strengthen the treatment (increase $|\tau|$)
3. Improve precision:
   - Covariate adjusment
   - Blocking
   - (Indexing)
4. Examine your DV for possible threats to power