36-463/663: Multilevel & Hierarchical Models

Design and Power

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Outline

- Effect size, sample size and power for a simple treatment effect
  - Digression: The value of a baseline covariate
- Estimating a mean from clustered data
- Power for more complex multi-level models:
  - od.exe (for balanced designs)
  - Fake-data simulations (for unbalanced designs and "unusual" assumptions)

Effect size, sample size and power for a simple treatment effect

- Let n units, $i=1, \ldots, n$, be randomly assigned to treatment ($T_i=1$) or control ($T_i=0$), with outcome y_i.
- The treatment effect is β_1 in the model

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i,\ \ \varepsilon_i \sim N(0, \sigma^2)$$

- How large does J have to be, to "detect" the treatment effect?

$$0 \overset{?}{\in} (\hat{\beta}_1 - 1.96 SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 SE(\hat{\beta}_1))$$

Effect size, sample size and power for a simple treatment effect (cont'd)

- More formally, we are testing
  - H0: $\beta_1 = 0$, vs
  - H1: $\beta_1 \neq 0$

  with the test statistic $S = |\hat{\beta}_1| / (SE(\hat{\beta}_1))$,

  and z_α = 1.96 is the (two-sided) α = 0.05 cutoff of the normal distribution.
- The *level* of the test is

$$P[|\hat{\beta}_1|/SE(\hat{\beta}_1) > z_\alpha \mid \beta_1 = 0] \approx \alpha = 0.05$$

- The *power* of the test at *effect size* $b \in$ H1 is

$$P[|\hat{\beta}_1|/SE(\hat{\beta}_1) > z_\alpha \mid \beta_1 = b]$$

Effect size, sample size and power for a simple treatment effect (cont'd)

- A power calculation typically involves finding the ***sample size*** that leads to a certain power, at level α and effect size *b*.

- To do this we need a formula or other method to relate $SE(\hat{\beta}_1)$ to sample size.

- In the simple linear regression case it is not too hard to derive a formula…

Effect size, sample size and power for a simple treatment effect (cont'd)

- Our regression

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2)$$

- can be rewritten $y = X\beta + \epsilon$, where *y* and ϵ are column vectors of length *n* and

$$X = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

(the first column has *n* 1's and the second column has n_T 1's)

Effect size, sample size and power for a simple treatment effect (cont'd)

- Consulting a linear regression reference,

$$SE(\hat{\beta}_1) = \sqrt{s^2 (X^T X)^{-1}_{22}}$$

- We calculate

$$X^T X = \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} n & n_T \\ n_T & n_T \end{bmatrix}$$

- And after some further calculation

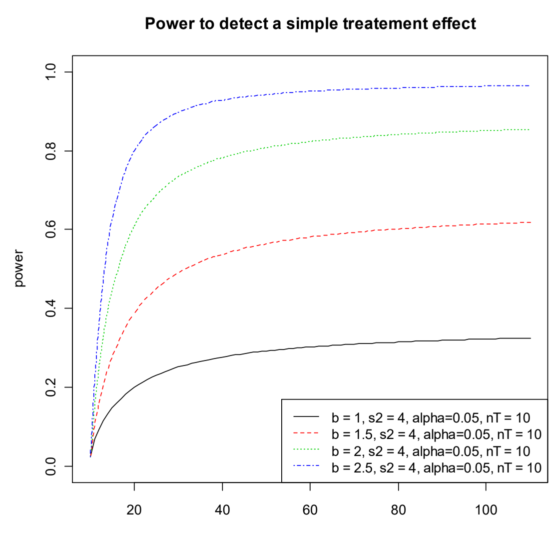
$$SE(\hat{\beta}) = s\sqrt{\frac{1}{n_T} + \frac{1}{n - n_T}}$$

Effect size, sample size and power for a simple treatment effect (cont'd)

- Power as a function of total sample size, for various effect sizes, is shown at right.

- Although we specified effect size here, only the ratio *b/SE(*$\hat{\beta}$*)* really matters.

- *b/SE(*$\hat{\beta}$*) = "standardized effect size"*



$$P[|\hat{\beta}_1|/SE(\hat{\beta}_1) > z_\alpha \mid \beta_1 = b]$$

Estimating a mean in a clustered sample

- Now suppose we wish to estimate a population mean β_o using $\overline{y}$ from clustered data, with *J* clusters of size *m*, for a total sample size of *n=Jm*.
- Under the model

$$\begin{aligned} y_i &= \alpha_{j[i]} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0,\sigma^2),\ i=1,\ldots,n \\ \alpha_j &= \beta_0 + \eta_j,\ \eta_j \overset{iid}{\sim} N(0,\tau^2),\ j=1,\ldots,J \end{aligned}$$

we can easily calculate that

$$\mathrm{SE}(\overline{y}) = \mathrm{SE}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \sqrt{\sigma^2/n + \tau^2/J}$$

Estimating a mean in a clustered sample (cont'd)

- We can rewrite this as

$$\mathrm{SE}(\overline{y}) = \sqrt{\sigma^2/n + \tau^2/J} = \sqrt{\frac{\sigma^2_{tot}}{n}\underbrace{[1+(m-1)ICC)]}}$$

where $\sigma^2_{tot} = \sigma^2 + \tau^2$, and

$$ICC = \frac{\tau^2}{\sigma^2+\tau^2}$$

This is called the "design effect", or DEFF

- This tells us:
  - SE for estimating β_0 from $\overline{y}$ depends on both number of clusters *J* and number of observations *m* per cluster
  - Bigger τ^2 → higher ICC → smaller effective sample size for estimating β_0 from $\overline{y}$.

Power for balanced multi-level models

- Consider a multi-level model for detecting a treatment effect, such as

$$y_i = \alpha_j + \epsilon_i, \ \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$
$$\alpha_j = \beta_0 + \beta_1 T_j + \eta_j, \ \eta_j \stackrel{iid}{\sim} N(0, \tau^2)$$

- If the data are balanced
  - Same number of observations in each cluster
  - Same number of treatement as control cases, etc.

  then there are tractable power formulae.

Power for balanced multi-level models (cont'd)

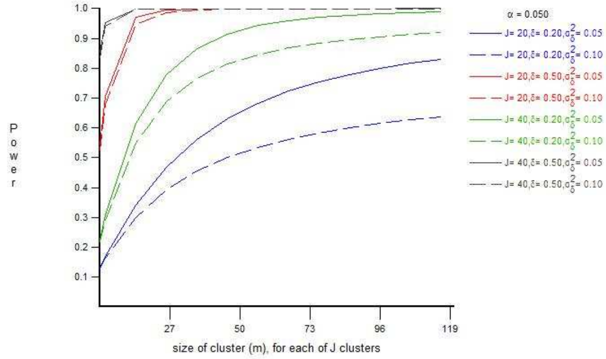
- For balanced multilevel models, these papers work out the ugly formulae [latest in a long line of such efforts]
  - Raudenbush, S. & Liu, X (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 2,* 199-213
  - Snijders, T. & Bosker, R. (1993). Standard errors and sample sizes in two-level research. *Journal of Educational Statistics, 18,* 237-259.
- Fortunately there is a small computer program that does the calculations…
  - http://sitemaker.umich.edu/group-based/optimal_design_software

Power for balanced multi-level models (cont'd)

- Power for detecting β_1 in

$$\begin{aligned} y_i &= \alpha_j + \epsilon_i, \\ \alpha_j &= \beta_0 + \beta_1 T_j + \eta_j \end{aligned}$$

  as calculated by od.exe



- J = # clusters
- m = persons/cluster
- $\delta = \text{standardized effect size} \approx \hat{\beta}_1 / SE(\hat{\beta}_1)$
- Half of sample to Tx, half to Ctrl.

Note that number of clusters (J) has a bigger effect on power than number of observations per cluster (m). This is very typical...

Power for other multilevel designs

- Power calculation software tends to fail when
  - The design is severely unbalanced
  - The software can't handle your particular model
    - multi-level glm's for example!
    - nonstandard distributions (say, t- or gamma distributions for random effects, rather than normals, etc.)
  - You want to explore
    - *Robustness:* E.g., will I still be able to detect an effect if I am using slightly the wrong model?
    - *Utility:* What if I trade off the cost of making a wrong decision against the cost of collecting more data?
    - Etc.
- In all these cases, we may resort to fake-data simulation

Example: Our Cluster-Level Treatment Model

- We simulate this model

$$y_i = \alpha_j + \epsilon_i, \ \epsilon_i \stackrel{iid}{\sim} N(0,\sigma^2)$$

$$\alpha_j = \beta_0 + \beta_1 T_j + \eta_j, \ \eta_j \stackrel{iid}{\sim} N(0,\tau^2)$$

  1000 times, and fit it with

```
lmer(y ~ Tx + (1|cluster),data=fake$data)
```

- We record a "hit" each time

$$|\hat{\beta}_1|/SE(\hat{\beta}_1) > 1.96$$

- Estimated power is (# hits)/1000

Example: Our Cluster-Level Treatment Model (cont'd)

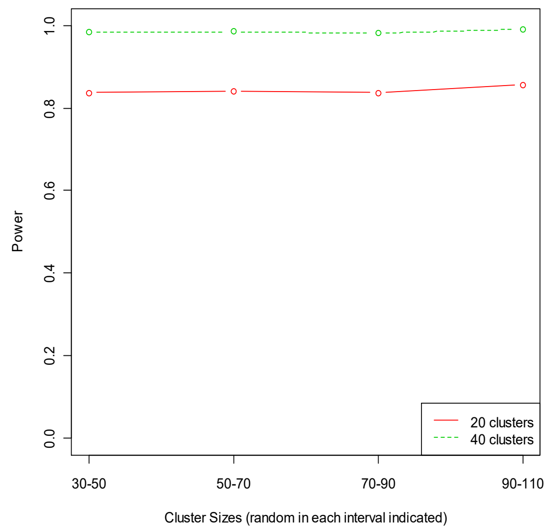
- We have taken
  - 4 unequal cluster sizes,
  - 50% assignment of clusters to treatment
  - fairly large variance components compared to the treatment effect (β_1=2).
- Comparable to the low end of the od.exe results
- Increasing the number of clusters should help…

```
n.sims <- 1000

hits <- 0

for (reps in 1:n.sims) {
  fake <- sim.one.data.set(
                       cl.sizes=c(3,5,7,9),
                       frac.T = .5,
                       b0     = 1,
                       b1     = 2,
                       sig    = 1,
                       tau    = 1.5)
  fake.lmer <-
    lmer(y ~ Tx + (1|cl),data=fake$data)
  t.stat <- coef(summary(fake.lmer))["Tx",
    "t value"]
  hits <- hits +
    ifelse(abs(t.stat) > 1.96, 1, 0)
}

(power <- hits/n.sims)

# [1] 0.371
```

This is not fast. It took about 100 seconds to run.

Example: Our Cluster-Level Treatment Model (cont'd)

- Results seem similar to or slightly better than "od.exe" calculation
- The power should be monotone, so any non-monotonicity here is Monte Carlo error.
- (Took about 30 min of simulation!)



Power – Some Final Thoughts

- Snijders & Bosker (Ch 10) has a more elaborate discussion, and Gelman & Hill (Ch 20) have more elaborate examples, but the messages are largely the same:
  - Power is relatively tractable if you have a balanced design and a lot of patience
  - For unbalanced designs, "unusual" assumptions, cost tradeoff considerations, etc., simulation-based power calculations are fine, but you still need patience!
- Baseline covariates that
  - are independent of Tx, but
  - Explain a lot of the variation in y

really improve power, in both linear models and mlm's!