Using Covariates in Experiments: Design and Analysis

STA 320
Design and Analysis of Causal Studies
Dr. Kari Lock Morgan and Dr. Fan Li
Department of Statistical Science
Duke University

Covariates

- Pre-treatment variables
- X: n x k covariate matrix
- GOAL: balance between treatment groups

Covariates

- Randomization *should* balance all covariates (observed and unobserved) *on average*…
- … but covariates may be imbalanced by random chance, and sometimes better balance is desired

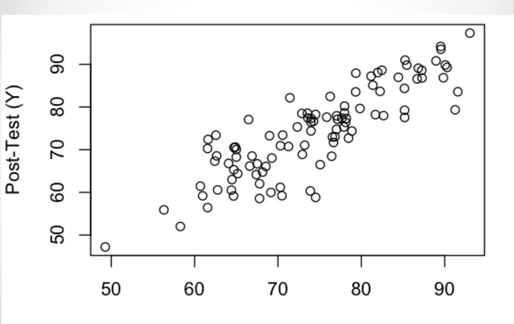
Covariate Balance

- Why is covariate balance important?
- Better covariate balance…

1) provides more meaningful estimates of the causal effect

2) increases precision (reduces variance) of estimator, if covariates correlated with outcome (outcome less variable for similar values of covariates)

Precision



Covariates

Two options:

- Option 1: force better balance on important covariates ***by design***
- Option 2: correct imbalance in covariates ***by analysis***

Covariate Balance

- By **design**:
  - stratified randomized experiments
  - paired randomized experiments
  - rerandomization (Wed)
- By **analysis**:
  - outcome: gain scores
  - separate analyses within subgroups
  - regression
  - model-based imputation

Stratified Experiments

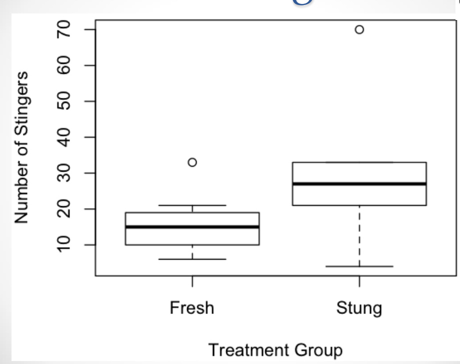
- Units are stratified (grouped, blocked) according to covariate(s)
- Subdivide sample into J homogeneous strata (blocks)
- Randomize units to treatment groups within each strata
- Often used with important categorical covariates (or discretized quantitative)
- (similar to stratified sampling)

Bee Stings

- If you are stung by a bee, does that make you more likely to get stung again? (Might bees leave behind a chemical message that tells other bees to attack you?)
- Scientists dangled 16 muslin-wrapped cotton balls over a beehive, where half of the balls had been previously stung and the other half were fresh.
- Outcome: total number of new stingers
- This was repeated for a total of nine trials.

Free, J.B. (1961) "The stinging response of honeybees," Animal Behavior, vol. 9, pp 193-196.

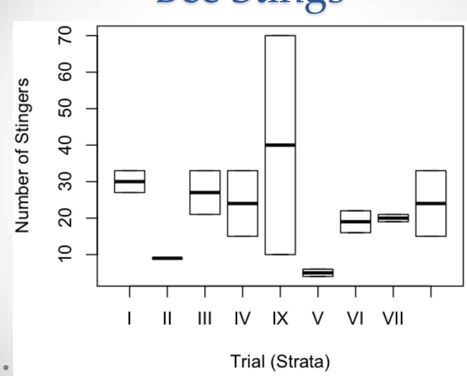
Bee Stings



Bee Stings

- Scientists expect the number of stings to vary by trial (different number of bees in the hive, different times of day, different weather, etc.)
- Each trial is a different strata
- J = 9 strata

Bee Stings



Stratified Experiment

- What to use for a test statistic (Fisher)?
- Lots of options. A common one:

$\bar{Y}_T^{obs}(j)$ = average observed Y for treated units in the j^{th} strata

$\bar{Y}_C^{obs}(j)$ = average observed Y for control units in the j^{th} strata

For each strata: $\bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j)$

- How to combine?

$$T = \sum_{j=1}^{J} \lambda_j \left(\bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j) \right)$$

Stratified Experiment

$$T = \sum_{j=1}^{J} \lambda_j \left(\bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j) \right)$$

- What to use for the weights?
- Weights must sum to 1
- Multiple options, but one common possibility is to weight by the sample size of each strata, N(j):

$$T = \sum_{j=1}^{J} \frac{N(j)}{N} \left(\bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j) \right)$$

Bee Stings

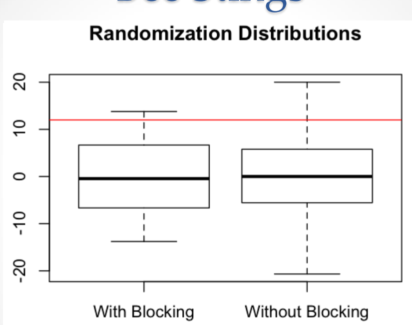
- In this example, the trials are all the same sample size, so just a simple average of the treatment effects for each trial:

$$T^{obs} = 12$$

Inference

- Fisher randomization test
- Easier to simulate randomizations, rather than enumerate all possible allocations
- For each simulated randomization, just randomize within strata
- (Always for randomization test, just simulate randomization scheme actually used in experiment)

Bee Stings



p-value = 0.04

Stratified Experiment

- What to use for an estimator (Neyman)?
- In general…

$$\hat{\tau} = \sum_{j=1}^{J} \lambda_j \hat{\tau}_j$$

$\hat{\tau}_j$ is the estimate within strata j

λ_j is the weight given to strata j

- One common option:

$$\hat{\tau} = \sum_{j=1}^{J} \frac{N(j)}{N} \left(\bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j) \right)$$

Stratified Experiments

$$\text{var}\left(\hat{\tau}\right) = \sum_{j=1}^{J} \lambda_j^2 \, \text{var}\left(\hat{\tau}_j\right)$$

$$\text{var}\left(\sum_{j=1}^{J} \frac{N(j)}{N}\left(\bar{Y}_T^{obs}(j) - \bar{Y}_C^{obs}(j)\right)\right)$$

$$= \sum_{j=1}^{J} \frac{N(j)^2}{N^2}\left(\frac{s_{T,j}^2}{N_T(j)} + \frac{s_{C,j}^2}{N_C(j)}\right)$$

Stratified Experiments

- No harm – can only help
- Can stratify on more than one covariate
- Strata can be any size
- "block what you can; randomize what you cannot"

Paired Experiments

- Units are matched into pairs (based on covariate(s))
- Special case of stratified randomized experiment with $N_j = 2$ for each strata
- Useful when expect difference in potential outcomes within a pair to be much smaller than differences across pairs

Wetsuit Advantage?

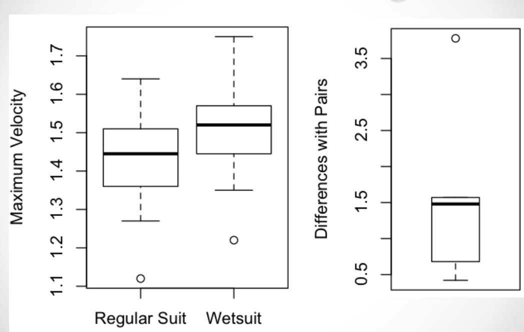
- The 2008 Olympics were full of controversy about whether the new wetsuits provide an unfair advantage
- Can a wetsuit really make someone swim faster? How much faster?

Wetsuit Advantage

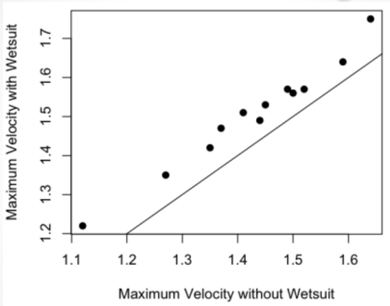
- Twelve competitive swimmers and triathletes swam 1500m at maximum speed twice each, once wearing a wetsuit and once wearing a regular suit
- Maximum velocity (m/sec) recorded (one of several possible outcomes)
- The order of the trials was randomized
- Each person is one "pair"

e Lucas, R., Balildan, P., Neiva, C., Greco, C., and Denadai, B., "The effects of wetsuits on physiological and biomechanical indices during swimming," Journal of Science and Medicine in Sport, 2000; 3(1): 1-8

Wetsuit Advantage?



Wetsuit Advantage?



Paired Experiment

- Test statistic / estimate: average of differences across all pairs
- (note: this is the same as the difference of the averages within treatment groups)
- Randomization test: randomize sign of each difference
- Neyman inference: analyze differences as a single variable

Wetsuit Advantage?

```
> t.test(nw, w, paired=TRUE)

	Paired t-test

data:  nw and w
t = -12.3182, df = 11, p-value = 8.885e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09134756 -0.06365244
sample estimates:
mean of the differences
                -0.0775

> t.test(nw, w)

	Welch Two Sample t-test

data:  nw and w
t = -1.3688, df = 21.974, p-value = 0.1849
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.19492937  0.03992937
sample estimates:
mean of x mean of y
 1.429167  1.506667
```

Paired Experiments

- Analysis is easy!
- When variability within pairs is much smaller than variability across pairs, can get huge gains in precision
- Better precision translates into higher power for tests (lower p-values) and narrower confidence intervals

Covariate Balance

- By **design**:
  - stratified randomized experiments
  - paired randomized experiments
  - rerandomization (Wed)
- By **analysis**:
  - outcome: gain scores
  - separate analyses within subgroups
  - regression
  - model-based imputation

Regression

$$Y_i^{obs} = \alpha + \tau W_i + \beta' X_i + \varepsilon_i$$

- Regress the observed outcomes on relevant covariates AND the treatment assignment indicator, W
- Different perspective: potential outcomes are considered random
- Use OLS (ordinary least squares)

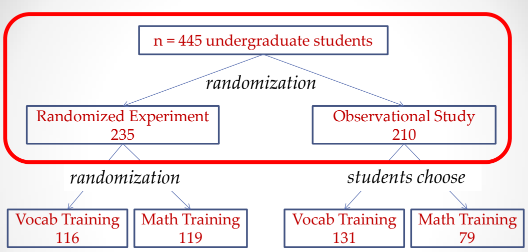
Regression

- Pros:
  - Easy to implement
  - Easy way to incorporate many covariates
  - Can be done after the fact if imbalance observed
- Cons:
  - Not completely unbiased

Regression

- Estimate biased for finite samples, but unbiased asymptotically (consistent)
- Because randomized experiment, bias is usually small and negligible
- Asymptotic unbiasedness holds even if the regression model is inaccurate
- (only because of randomization – not true in observational studies)

Shadish Data

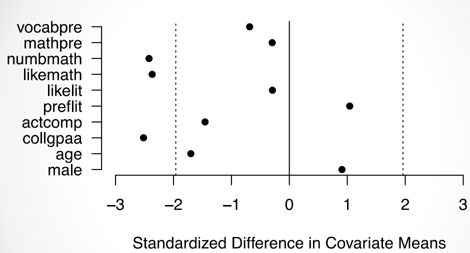


Shadish, M. R., Clark, M. H., Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *JASA*. **103**(484): 1334-1344.

Covariates

- Vocab and math pre-test scores
- Number of math classes taken
- How much do you like math?
- How much do you like literature?
- Do you prefer math or literature?
- ACT score
- College GPA
- Age
- Gender
- …

Shadish Data

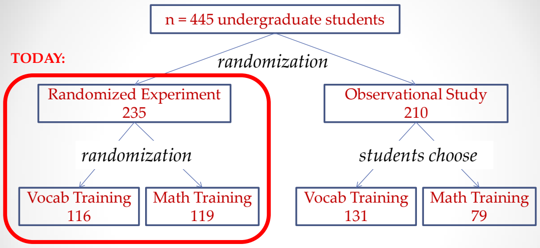


k = 10 covariates

Shadish Data

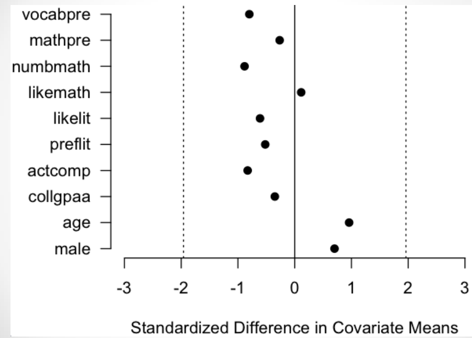
- How might we have prevented this by design???

Shadish Data



- Outcomes: scores vocab and math tests
- Causal questions: How does vocab (math) training effect scores on the vocab (math) test?

Shadish Data



Shadish Data: Math

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.67500    1.86762   1.968  0.05034 .
vocabpre     0.04698    0.05244   0.896  0.37125
mathpre      0.02795    0.07631   0.366  0.71451
numbmath     0.58874    0.22335   2.636  0.00898 **
likemath     0.30453    0.11861   2.568  0.01090 *
likelit     -0.31980    0.10944  -2.922  0.00384 **
preflit     -0.02738    0.37052  -0.074  0.94115
actcomp      0.28241    0.06441   4.385 1.79e-05 ***
collgpaa     0.50916    0.27036   1.883  0.06097 .
age         -0.05413    0.04126  -1.312  0.19087
male        -0.07339    0.41918  -0.175  0.86117
vm          -3.99171    0.35768 -11.160  < 2e-16 ***
```

Model-Based Imputation

- Impute missing potential outcomes
- Use information from control units to impute Y(0) for treated units
- Use information from treated units to impute Y(1) for control units
- Ideally, this should reflect the uncertainty in the imputation
- (similar to multiple imputation)

Model-Based Imputation

Y(1)	Y(0)	X
observed	?	observed
observed	?	observed
observed	?	observed
observed	?	observed
observed	?	observed
?	observed	observed
?	observed	observed
?	observed	observed
?	observed	observed
?	observed	observed

Model-Based Imputation

- Options without covariates:
  - Impute observed mean (doesn't reflect uncertainty)
  - Sample with replacement (bootstrap)
- Options with covariates:
  - Sample with replacement from "donor pools" with similar covariate values
  - Build models to predict potential outcomes based on covariates

Model-Based Imputation

Y(1)	Y(0)	X
observed	?	observed
observed	?	observed
observed	?	observed
observed	?	observed
observed	?	observed
?	observed	observed
?	observed	observed
?	observed	observed
?	observed	observed
?	observed	observed

Use treated units to model Y(1) | X

Use control units to model Y(0) | X

Model-Based Imputation

- Build models to predict potential outcomes based on covariates:
  - o Regression can work, but doesn't fully reflect uncertainty
  - o BEST: be fully Bayesian and draw from posterior distribution for Y(1) or Y(0) given the covariates

Summary: Using Covariates

- By **design**:
  - o stratified randomized experiments
  - o paired randomized experiments
  - o rerandomization (Wed)
- By **analysis**:
  - o outcome: gain scores
  - o separate analyses within subgroups
  - o regression
  - o model-based imputation

To Do

- Read Ch 7, 8, 9, 10