36-463/663: Multilevel & Hierarchical Models

Causal Inference

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

Outline

- Causal Inference [G&H Ch 9]
  - The Fundamental Problem
  - Confounders, and how Controlled Randomized Trials control them
  - Adjusting an analysis for pre-treatment covariates (but not post-treatment ones!)
  - Observational Studies
- More sophisticated tools for causal inference [G&H Ch 10]
  - Instrumental Variables
  - Matching and propensity scores
  - Regression discontinuity designs

Causal Inference

- Want to test a new pain reliever for headaches
- Have 200 patients i=1,…,200.
  - T_i=1 (patient gets drug) for i=1..100,
  - T_i=0 (patient gets nothing) for i=101..200.
- Suppose drug is worthless, but
  - i=1..100 are healthy and
  - i=101..200 all have flu, colds, etc.
  - *How will the drug look?*
- Suppose drug is effective, but
  - i=1..100 have colds & flu, and
  - i=101..200 are healthy.
  - *How will the drug look now?*
- What is wrong with these examples?

Causal Inference—The Fundamental Problem

- We really would like to see the difference between pain level “with the drug” vs pain level “without”, *for each individual patient.*

$$\begin{aligned} y_i^0 &= \text{outcome without treatment} \\ y_i^1 &= \text{outcome with treatment} \\ y_i^1 - y_i^0 &= \text{treatment effect for unit } i \end{aligned}$$

- But we cannot try the drug, and then go back in time and try without the drug.
  - *For each patient i, can see either y_i^0 or y_i^1 but not both!*

Causal Inference—The Fundamental Problem

- If we average the individual treatment effect over all patients, get the average causal effect (ACE):

$$\begin{aligned}\text{ACE} &= \frac{1}{N}\sum_{i=1}^{N}(y_i^1 - y_i^0) = \frac{1}{N}\sum_{i=1}^{N} y_i^1 - \frac{1}{N}\sum_{i=1}^{N} y_i^0 \\ &= E[y^1] - E[y^0]\end{aligned}$$

- Most studies try to estimate ACE. A good way to do this would be:
  - ❑ Estimate $E[y^1] \approx \bar{y}^1$ from unbiased sample $y_1^1, \ldots y_{n_1}^1$
  - ❑ Estimate $E[y^0] \approx \bar{y}^0$ from unbiased sample $y_1^0, \ldots y_{n_0}^0$

Causal Inference—The Fundamental Problem

- The problem with the examples we started with was that ***<u>the samples were not unbiased</u>***.
- There are basically two ways to deal with bias
  - ❑ Design a study for which the samples are guaranteed to be unbiased
  - ❑ Do some statistical adjustment to account for the bias
- To understand how to design an "unbiased" study, we first consider how "bias" arises…

Causal inference - Confounders

- If some patients have $T_i = 1$ and others have $T_i = 0$, we know that $E[y^1] - E[y^0] \approx \hat{\beta}_1$ in the regression

$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

- However, if there is a "confounding" variable x_i , the correct $\hat{\beta}_1$ should come from

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i$$

- How bad can the bias be if we omit x_i?

Causal inference - Confounders

We suppose the correct model is

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i \qquad (1)$$

but we fit instead

$$y_i = \beta_0^* + \beta_1^* T_i + \epsilon_i^* \qquad (2)$$

Note that x_i also has some relationship with T_i that can be expressed as a linear regression:

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i \qquad (3)$$

If we substitute (3) into (1) and do a little rearranging, we get

$$y_i = (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) T_i + (\epsilon_i + \beta_2 \nu_i) \qquad (4)$$

Equating coefficients in (2) and (4), we see

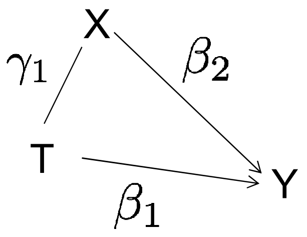
$$\beta_1^* = \beta_1 + \beta_2 \gamma_1 \qquad (5)$$

Thus, estimating $E[y^1] - E[y^0] \approx \hat{\beta}_1^*$ will be biased, *unless*

- $\gamma_1 = 0$, i.e. x_i is independent of treatment assignment T_i
- $\beta_2 = 0$, i.e. x_i has no influence on y_i after considering T_i (x_i not really a confounder!)

Causal inference - Confounders

- If X is a confounder, the total effect of T on Y is $\beta_1 + \beta_2\gamma_1$:



- $\beta_2 = 0$: X not really a confounder!
- $\gamma_1 = 0$: No selection effect!

- If we omit X (or it is hidden!) then we only get the right answer from y = β_0 + β_1 T + ϵ, if β_2 or γ_1 is zero.

Causal inference – Estimating ACE

- We can get an unbiased estimate of ACE in any of the following ways
  - *If there are no confounders*, estimate β_1 in
$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$
  - *If there are confounders*, ***find them all***, include them as x's, and then estimate β_1 in
$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_K x_{Ki} + \epsilon_i$$
  - *Design the experiment* so that all ***confounders*** x_i are ***independent of treatment*** assignment T_i and then estimate β_1 from
$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

Causal inference – randomized trials

- In a ***randomized experiment***, each unit i is assigned T_i = 1 (treatment) or T_i = 0 (no tx) randomly (e.g. by random coin toss!).
  - This forces every potential confounder x_i to be independent of T_i, whether we "discover" x_i or not! $(\gamma_1 = 0)$
  - From a randomized experiment we can always estimate ACE by estimating β_1 in

$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

Causal inference – randomized trials

- In many settings you can't completely randomize
  - A study of effectiveness of a new math curriculum might involve several schools.
    - Can't put all math classes in all schools together in one "pot" and randomly assign some to new math curriculum
    - Instead assign ½ the classes to the new math program and ½ to the old math program within each school
    - Since schools contain other factors that affect math performance, school becomes an x_i and we can estimate the ACE for the new math program from

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i$$

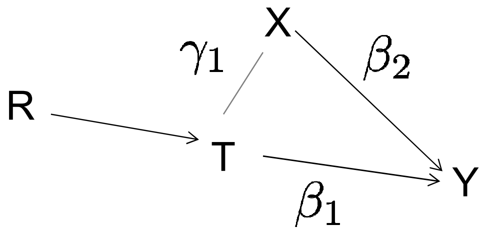
- A lot of experimental design is like this…

Causal inference – pre-treatment covariates in randomized trials

- Even in a randomized experiment, if we can identify a confounder x_i, it is good to include it in the model.
- Estimating ACE = $\hat{\beta}_1$ from
$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$
is unbiased, but not efficient (more uncertainty)
- Estimating ACE = $\hat{\beta}_1$ from
$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i$$
will be more efficient (less uncertainty).

Causal inference – randomized trials

- If R is a random treatment assignment (coin flip!), then γ_1 must equal zero!



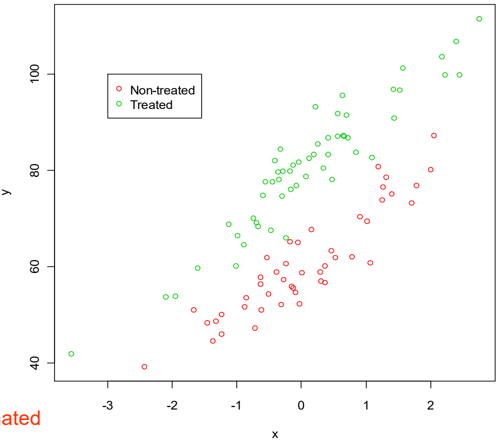
- $\gamma_1 = 0$: No selection effect!

- We can now get the right treatment effect from
$$y = \beta_0 + \beta_1 T + \epsilon.$$
- It is still worth including X in the model if possible,
$$y = \beta_0 + \beta_1 T + \beta_2 X + \epsilon$$
because including X will reduce SE(β_1) !

Randomized trials – pre-treatment covariates – uniform tx effect

```
> x <- rnorm(n)
> y <- 60 + 10*x + 5*rnorm(n)
  # x is a confounder
> T <- rbinom(100,1,.5)
  # treatment by random experiment
> y <- ifelse(T==1,y+20,y)
  # add treatment effect for treated
> plot(x,y,col=T+2)
> legend(-3,100,pch=c(1,1),col=2:3,
      legend=c("Non-treated","Treated"))
> (ACE <- mean(y[T==1]) - mean(y[T==0]))
[1] 20.26647
>
> summary(lm(y ~ T))$coef[,1:2]
            Estimate Std. Error
(Intercept) 60.63675   1.854682
T           20.26647   2.523902
>
> summary(lm(y ~ T + x))$coef[,1:2]
            Estimate Std. Error
(Intercept) 60.13741  0.6815005
T           19.49961  0.9275130
x           10.49448  0.4182943
```

ACE is estimated better when covariate in the model



- x is a pretest score
- y is a post-test score, of course affected by x
- T is treatment (new curriculum)

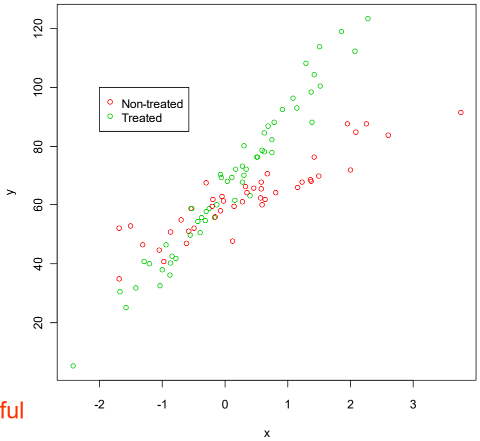
Randomized trials – pre-treatment covariates – nonuniform tx effect

```
> n <- 100
> x <- rnorm(n)
> y <- 60 + 10*x + 5*rnorm(n)
> T <- rbinom(100,1,.5)
> y <- ifelse(T==1,y+5+15*x,y)
> plot(x,y,col=T+2)
> legend(-2,100,pch=c(1,1),col=2:3,
     legend=c("Non-treated","Treated"))
> (ACE <- mean(y[T==1]) - mean(y[T==0]))
[1] 5.684276
> summary(lm(y ~ T))$coef[,1:2]
             Estimate Std. Error
(Intercept) 62.599809   3.164975
T            5.684276   4.229376
> (coef <- summary(lm(y ~ T + x +
   T:x))$coef[,1:2])
             Estimate Std. Error
(Intercept) 59.205524  0.8095489
T            6.149310  1.0646086
x            9.499872  0.6574682
T:x         15.653435  0.9527179
> mean(coef[2,1] + coef[4,1]*x)
[1] 9.631048
```

Tx affects not only the intercept but also the slope!

ACE not all that meaningful

Here's a kind of ACE that may be useful…



- x is a pretest score
- y is a post-test score, of course affected by x
- T is treatment (new curriculum)

Randomized trials – ***do not include*** post-treatment covariates!

```
> n <- 100
> x <- rnorm(n)
> y <- 60 + 10*x + 5*rnorm(n)
> T <- rbinom(100,1,.5)
> y <- ifelse(T==1,y+20,y)
> z <- ifelse(T==1,rnorm(100,3),
   rnorm(100,-3))
> plot(x,y,col=T+2)
> legend(-2,100,pch=c(1,1),col=2:3,
    legend=c("Non-treated","Treated"))
> (ACE <- mean(y[T==1]) -
   mean(y[T==0]))
[1] 22.43931
> summary(lm(y ~ T))$coef[,1:2]
              Estimate Std. Error
(Intercept) 58.11903   1.660045
T           22.43931   2.347659
```

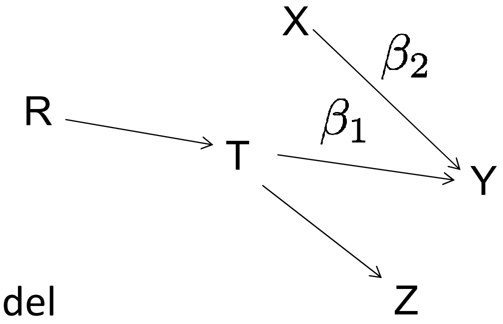
```
> summary(lm(y ~ T +
   x))$coef[,1:2]
              Estimate Std. Error
(Intercept) 59.85651  0.7068169
T           20.78911  0.9959064
x           10.58185  0.4983279
> summary(lm(y ~ T + x +
   z))$coef[,1:2]
               Estimate Std. Error
(Intercept) 64.884033  1.9499540
T           10.505663  3.8573971
x           10.416234  0.4859765
z            1.608895  0.5843686
```

Including z in the model completely dilutes the effect of T that we are trying to estimate!

- x is a pretest score
- y is a post-test score, of course affected by x
- T is treatment (new curriculum)
- z is a secondary effect of T

Causal inference – Post-tx covariates

- If R is a random treatment assignment (coin flip!), then γ_1 must equal zero!



- In the model

$$y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 Z + \epsilon$$

the estimate of β_1 will only include the influence of the part of T not explained by Z… That might not be much!

Observational Studies

- Often have the form of randomized trials
  - Treatment T_i
  - Covariate(s) x_i – possible confounders
- Want to know causal effect of T_i…
  - Can run same regressions as before to estimate β_1 Generally should include all known confounders

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_K x_{Ki} + \epsilon_i$$

  - But since we do not have control over T_i there could be hidden confounders (lurking variables)
  - Often associated with selection effects (why does someone volunteer for the treatment?)
  - Usually cannot make causal statements

Observational Studies

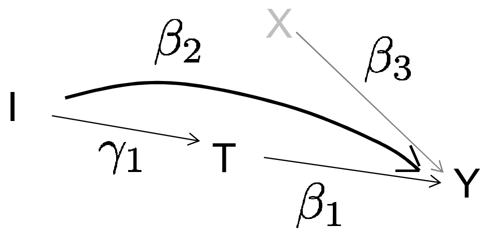
- Sometimes hard to say exactly what T_i is
  - Try to make an analogy from the observational study to the "ideal" randomized trial to see what T_i is (or even if there could be a T_i!)
    - If the ideal experiment involves randomly assigning classrooms to different math curricula, then T_i could be a cause
    - If the ideal experiment involves randomly assigning race or gender to people, then T_i probably is not a cause
  - The regression analyses can suggest whether a further randomized experiment is worth doing, but generally we cannot make causal inferences (lurking variables!)

Observational Studies

- Sometimes ***causal inferences*** can be made from observational studies. Here are four methods:
  - *Instrumental variables* – substitute for the coin flip in randomized trials to eliminate selection effects
  - *Propensity score matching* – rearrange the data to eliminate selection effects
  - *Regression discontinuity designs* – exploit random errors in selection effects
  - *Bounding the influence of confounders* – sometimes the effect (ACE) of T_i is so big, that we can calculate that no reasonable set of confounders could be responsible for it. *(This is basically how the link between smoking and lung cancer was made.)*

Instrumental Variables

- An instrumental variable I is another variable that “works like” randomization:



- Need
  - *Monotonicity:* $\gamma_1 \neq 0$
  - *Ignorable assignment:*
    - I affects Y only through T (β_2=0)
    - I is independent of X

Instrumental Variables

- The regression equations are

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \beta_2 \mathrm{I} + \epsilon && (1) \\ T &= \gamma_0 + \gamma_1 \mathrm{I} + \nu && (2) \end{aligned}$$

- Substituting (2) into (1), we get

$$y = (\beta_0 + \beta_1\gamma_0) + (\beta_1\gamma_1 + \beta_2)\mathrm{I} + (\text{error terms})$$

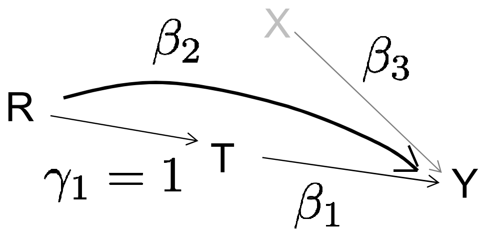
- And so if we fit the regressions

$$\begin{aligned} y &= \delta_0 + \delta_1 \mathrm{I} + \epsilon \\ T &= \gamma_0 + \gamma_1 \mathrm{I} + \nu \end{aligned}$$

we find $\beta_1 = (\delta_1 - \beta_2)/\gamma_1 = \delta_1/\gamma_1$, since $\beta_2 = 0$.

Coin-Flip is the perfect instrument!

- An instrumental variable I is another variable that "works like" randomization:



- Fit

$$\begin{aligned} y &= \delta_0 + \delta_1 \mathrm{I} + \epsilon \\ T &= \gamma_0 + \gamma_1 \mathrm{I} + \nu \end{aligned}$$

- $\beta_1 = (\delta_1 - \beta_2)/\gamma_1 = \delta_1$ since $\beta_2 = 0$ & $\gamma_1 = 1$.

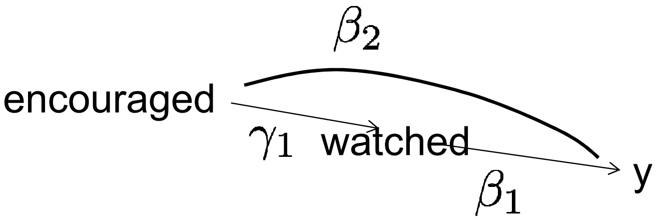
Example – just to give the flavor of instrumental variables

- What is the effect of watching Sesame Street on childrens' letter-recognition skills?

```
pretest     - letter skills test before experiment
y           - letter skills test after experiment
encouraged  - 1 = encouraged to watch; 0 = not
watched     - 1 = did watch Sesame Street; 0 = not
site        - 1,2,3,4,5: combos of age, SES,
                   language, urbanicity
setting     - 1 = at home; 0 = at school
```

Example – Simple IV Estimate

- What we can actually manipulate is "encouraging" kids to watch



- We might be interested in two things:
  - The effect of "encouraged" on post-test score y
    - (the "intention to treat", ITT, analysis)
  - The effect of actually watching, on post-test score y
    - (the "instrumental variables", IV, analysis)

Simple IV analysis– Intention to Treat (ITT), and IV estimates

- ITT effect of “encouraged” on post-test y

```
> fit.1b <- lm(y ~ encouraged)
> coef(fit.1b) # the ITT effect
(Intercept)  encouraged
  24.920455    2.875598
```

This is the effect of encouragement on the post-test score

- IV effect of “watched” on post-test y

```
> fit.1a <- lm(watched ~ encouraged)
> coef(fit.1a)
(Intercept)  encouraged
  0.5454545   0.3624402
> coef(fit.1b)[2]/coef(fit.1a)[2]
encouraged
  7.933993
```

$\hat{\delta}_1/\hat{\gamma}_1$

This is the effect of watching S.Street on the post-test score

IV’s – Two-stage least-squares

- The “Ratio” estimate $\hat{\delta}_1/\hat{\gamma}_1$ is the “Wald Estimate”.
- A more popular method is called “Two-stage least-squares” (TSLS):

```
> coef(fit.2a <- lm (watched ~ encouraged))
(Intercept)  encouraged
  0.5454545   0.3624402
> watched.hat <- fit.2a$fitted
> coef(fit.2b <- lm (y ~ watched.hat))
(Intercept) watched.hat
  20.592822    7.933993
```

In TSLS, second regression Uses fitted values from first regression..

This TSLS estimate is identical to the Wald estimate on the previous slide.

- There is a function tsls() in library(“sem”) that does tsls estimates automatically.

IV's – Including covariates

```
> fit.3a <- lm (watched ~ encouraged +
+   pretest + factor(site) + setting)
> watched.hat <- fit.3a$fitted
> fit.3b <- lm (y ~ watched.hat +
+   pretest + factor(site) + setting)
> coef(fit.3b)
  (Intercept)   watched.hat       pretest
         1.22         14.03          0.70
factor(site)2 factor(site)3 factor(site)4
         8.40         -3.94          0.94
factor(site)5       setting
         2.76          1.60
```

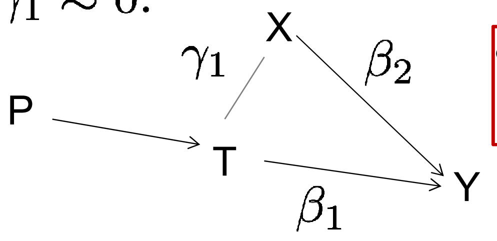
The covariates get put
In both regressions

The IV estimate of the effect of watching Sesame Streetm after controlling for covariates.

- SE's are more work; see G&H or use tsls() function…

Causal inference – Propensity Scores

- The propensity score P is used to rearrange the data so that $\gamma_1 \approx 0$.

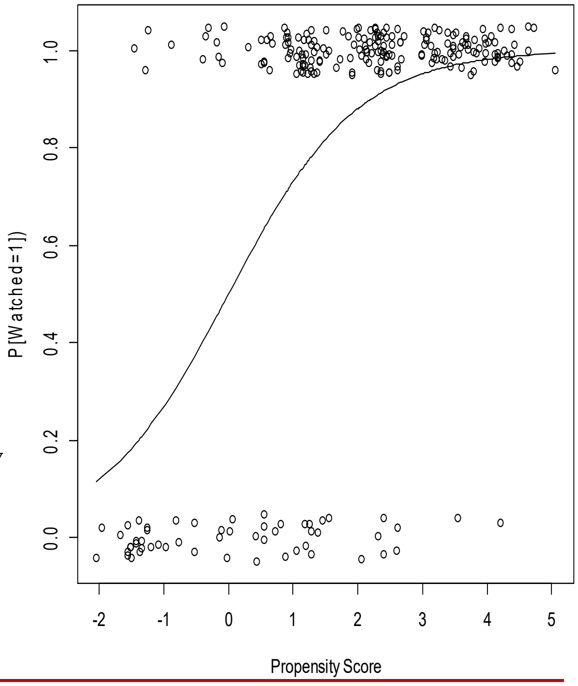


- $\gamma_1 = 0$: No selection effect!

- Use logistic regression to predict T as well as possible from all the X's. P(T=1) from this logistic regression is the *propensity score*.
- For each unit in with T=1, match it to a unit with T=0 with the same (or similar) propensity score.
    - Discard non-matching units at the end of the process

Making the propensity scores

```
> big.sesame <- cbind(y, sesame,
+ watched, encouraged, pretest)
> p.fit <- glm(watched ~
+ encouraged + pretest +
+ factor(site) + setting,
+ family = binomial,
+ data=big.sesame)
> p.scores <- predict(p.fit,
+ type="link")
> plot(p.scores, jitter(watched,
+ amount=0.05), xlab="Propensity
Score",ylab="P[Watched=1])")
> o.scores <- sort(p.scores)
> lines(o.scores, exp(o.scores)
+ / (1 + exp(o.scores)))
```



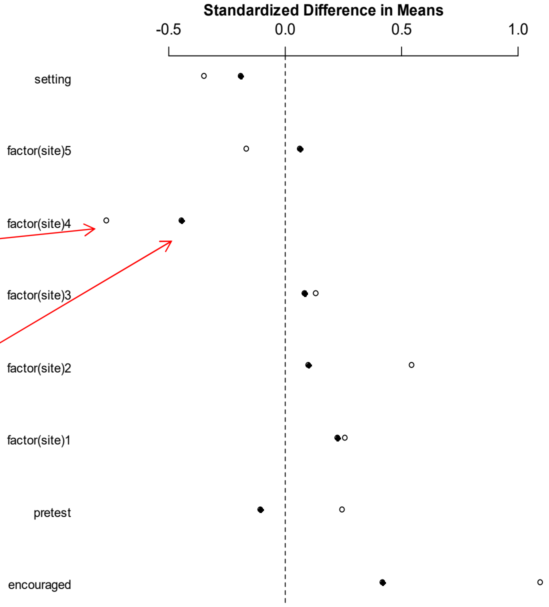
Making the matched data set

```
> matches <- matching(z =
+ watched, score = p.scores)

> matched <- big.sesame[
+ matches$matched,]

> dim(big.sesame)
[1] 240  32
> dim(matched)
[1] 108  32

> b.stats <-
+ balance(big.sesame,
+ matched, p.fit)
> plot(b.stats)
```



Diff between Tx vs Ctrl In *unmatched* data.

Diff between Tx vs Ctrl In *matched* data.

(The `matching()` *and* `balance()` *functions are from* `library(arm)`*.)*

Is $\gamma_1 \approx 0$ in the Matched Data Set?

```
> display(glm(formula = watched ~ encouraged + pretest +
+ factor(site) + setting, family = binomial, data =
+ matched))
              coef.est  coef.se
(Intercept)     0.63      0.96
encouraged      1.14      0.48
pretest        -0.02      0.04
factor(site)2  -0.03      0.78
factor(site)3  -0.66      0.62
factor(site)4  -1.32      0.58
factor(site)5  -0.93      0.81
setting         0.00      0.47
---
  n = 108, k = 8
  residual deviance = 138.5, null deviance = 149.7
(difference = 11.2)
```

We did pretty well except for these Two predictors.

More effort chosing variables and interactions from among the 32 available in the data set would probably generate propensity scores that drive γ_1 to zero.

How do we do estimating effect of watching Sesame Street?

```
> coef(lm(y ~ watched + encouraged + pretest + factor(site) +
+             setting,data=big.sesame))
  (Intercept)       watched    encouraged       pretest
factor(site)2
         4.52          9.04          1.71          0.73
8.55
factor(site)3 factor(site)4 factor(site)5       setting
        -4.52         -0.78          1.29          1.33
```

Unmatched Tx Effect Est.

```
> coef(lm(y ~ watched + encouraged + pretest + factor(site) +
+             setting,data=matched))
  (Intercept)       watched    encouraged       pretest
factor(site)2
         3.06         10.47          0.25          1.04
9.02
factor(site)3 factor(site)4 factor(site)5       setting
        -5.43         -3.71         -1.20          0.68
```

Matched Tx Effect Est.

Propensity Scores: How did we do?

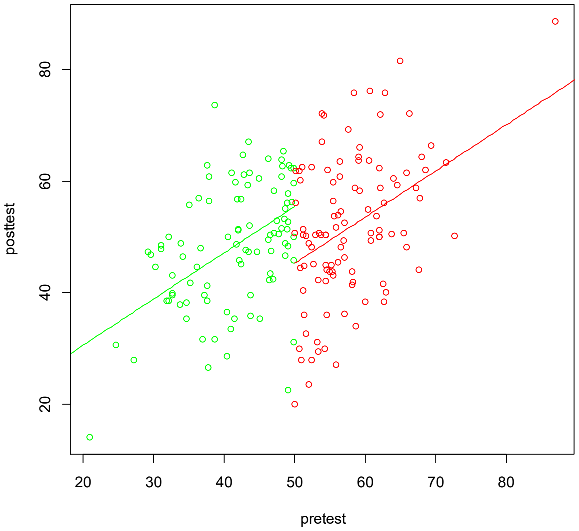
- The estimate of the effect of watching Sesame Street is a bit bigger for the matched data than for the non-matched data.
- It is not as big as the IV estimate, in part because the matching isn't very good yet. More effort needed to build a good logistic regression for the propensity scores!
- SE's are again problematic (we are using the data twice). See Gelman & Hill for details & solutions.

Regression Discontinuity Designs

- In the case of IV and Propensity Scores, we were looking for ways to break the relationship between X (covariates) and T (treatment)
- *What if X is intimately tied up with T?*
  - ❑ *Example*: Kids with low test scores (X low) get remedial math (T=1); Kids with high test scores (X high) get regular math (T=0).
  - ❑ *Can we still assess whether T causes a change in the end of year test scores (Y)?*

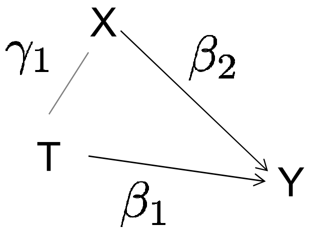
Regression Discontinuity Designs

- *Is the treatment effect the size of the jump?*
- For most of the data we can't make causal claim, because X is a confounder of T and Y.
- ***IF*** we can argue that people just either side of the cutoff are similar to each other, ***THEN*** the jump can represent a causal effect.



Regression Discontinuity Designs

- What does the RD design look like in terms of our regression diagram?



X = pretest
T = remedial math
Y = posttest

- All of the data can be used to get a really good estimate of β_2. This also improves SE's for β_1.
- For subjects near the jump, $\gamma_1 \approx 0$, so β_1 represents a causal effect for them.
- *How far can we generalize β_1 away from the jump?*

Regression Discontinuity Designs

- Estimation is very straightforward:

```
> display(fit <- lm(posttest ~ pretest + lowkids))
lm(formula = posttest ~ pretest + lowkids)
            coef.est coef.se
(Intercept)  3.84     7.06
pretest      0.83     0.12
lowkidsTRUE 10.17     2.52
---
n = 200, k = 3
residual sd = 10.97, R-Squared = 0.21
```

Our estimate, $\hat{\beta}_1$

.

Summary

- Causal Inference [G&H Ch 9]
  - The Fundamental Problem
  - Confounders, and how Controlled Randomized Trials control them
  - Adjusting an analysis for pre-treatment covariates (but not post-treatment ones!)
  - Observational Studies
- More sophisticated tools for causal inference [G&H Ch 10]
  - Instrumental Variables
  - Matching and propensity scores
  - Regression discontinuity designs