# 36-617: Applied Linear Regression

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

# Announcements

- Quiz next Monday (not today!)
  - Focus on Sheather 10.1
- HW06 is out… when should it be due?
  - Nonparametric regression & causal reasoning
- Reading
  - This week: Causal Reasoning
    - G&H Ch's 9 & 10 (see pdf in week08 area)
  - Next week: Intro to Multilevel Models
    - Sheather, 10.1
    - Intercepts: G&H Ch 12 (see pdf in week08 area)
    - Slopes: G&H Ch 13 (see pdf in week08 area)

# Outline

- 15 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!)
- 16 More sophisticated tools for causal inference [G&H Ch 10]
  - Observational Studies
  - Instrumental Variables
  - Matching and propensity scores
  - Regression discontinuity designs

# 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

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

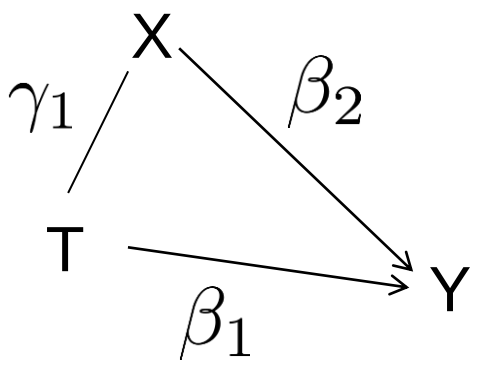
- However, if there is a "confounding" variable $x_i$ , the correct $\hat{\beta}_1$ should come from

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

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

# Causal inference - Confounders

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

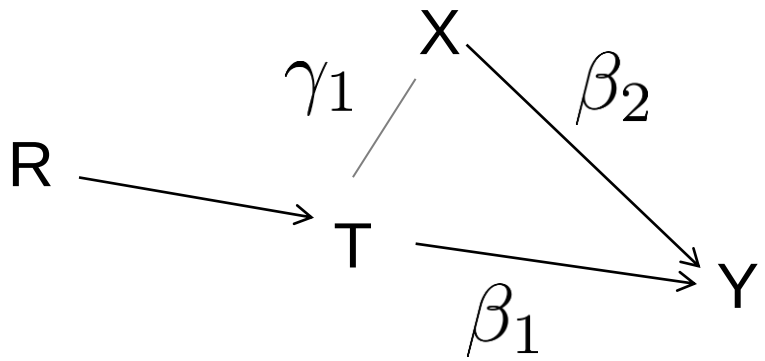



- $\beta_2 = 0$: $X$ not really a confounder!
- $\gamma_1 = 0$: No selection effect!

- If we omit X (or it is hidden!) then we only get the right answer from y = $\beta_0$ + $\beta_1$ T + $\epsilon$, if $\beta_2$ or $\gamma_1$ is zero.

# Causal inference – randomized trials

- If R is a random treatment assignment (coin flip!), then $\gamma_1$ 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($\beta_1$) !

# Causal inference – Estimating ACE

- We can get an unbiased estimate of ACE in any of the following ways
  - *If there are no confounders*, estimate $\beta_1$ 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 $\beta_1$ 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 $\beta_1$ from
$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

Even if there are no con-founders it is a good idea to include baseline covariates in this regression, to reduce $SE(\hat{\beta}_1)$

# 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 $\beta_1$ 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

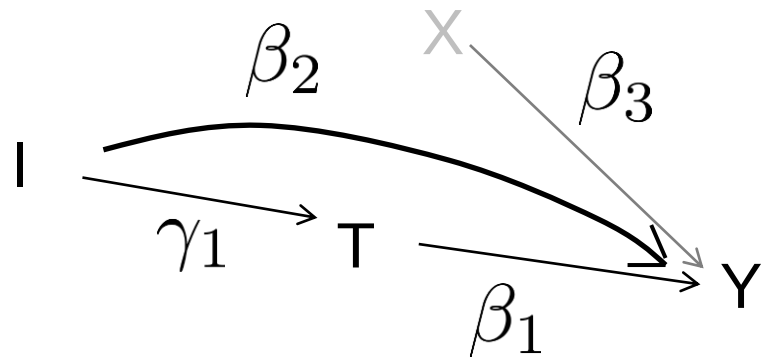
- Sometimes hard to say exactly what $T_i$ is
  - Try to make an analogy from the observational study to the "ideal" randomized trial to see what $T_i$ is (or even if there could be a $T_i$!). Examples:
    - If the ideal experiment involves randomly assigning classrooms to different math curricula, 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 ($\beta_2$=0)
    - I is independent of X

# Instrumental Variables

- The regression equations are

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

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

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

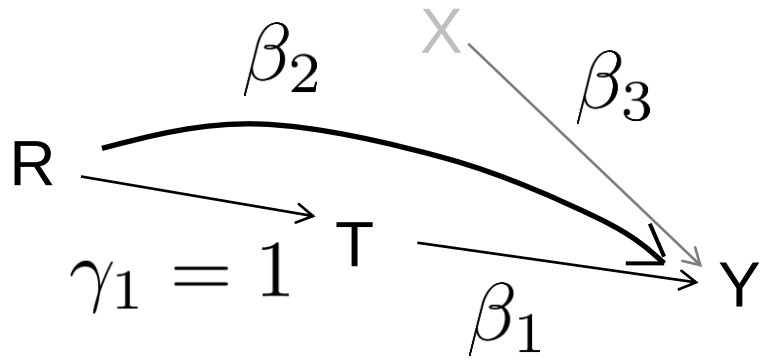
- And so if we fit the regressions

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

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 $\beta_2 = 0$ & $\gamma_1 = 1$.

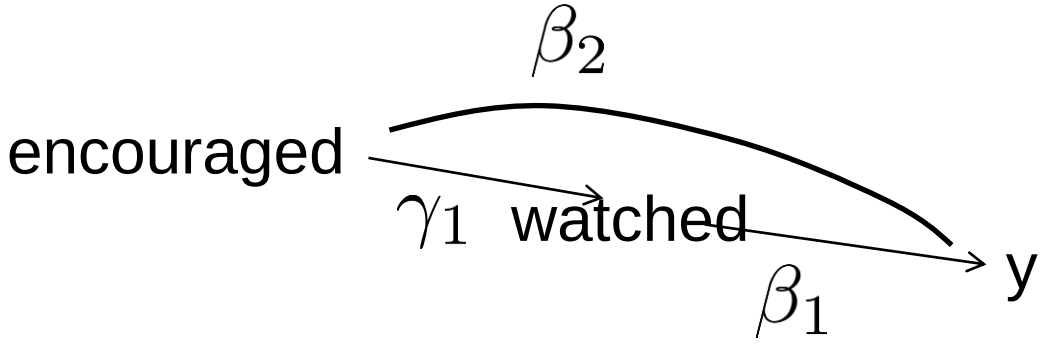
# Example – just to give the flavor of instrumental variables

- 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 in TSLS

```
> 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
```

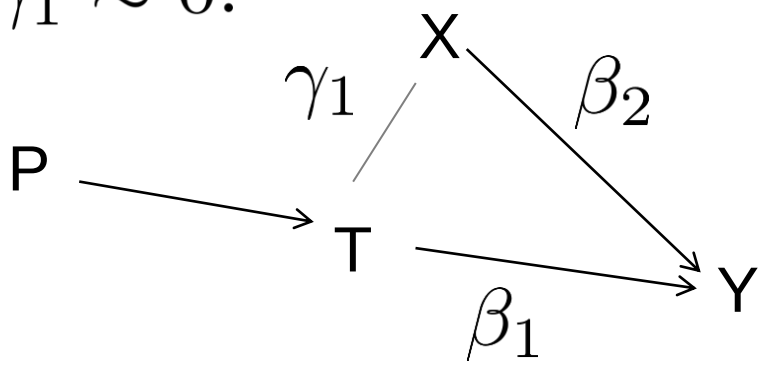
The covariates get put In both regressions

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 $\gamma_1 \approx 0$.

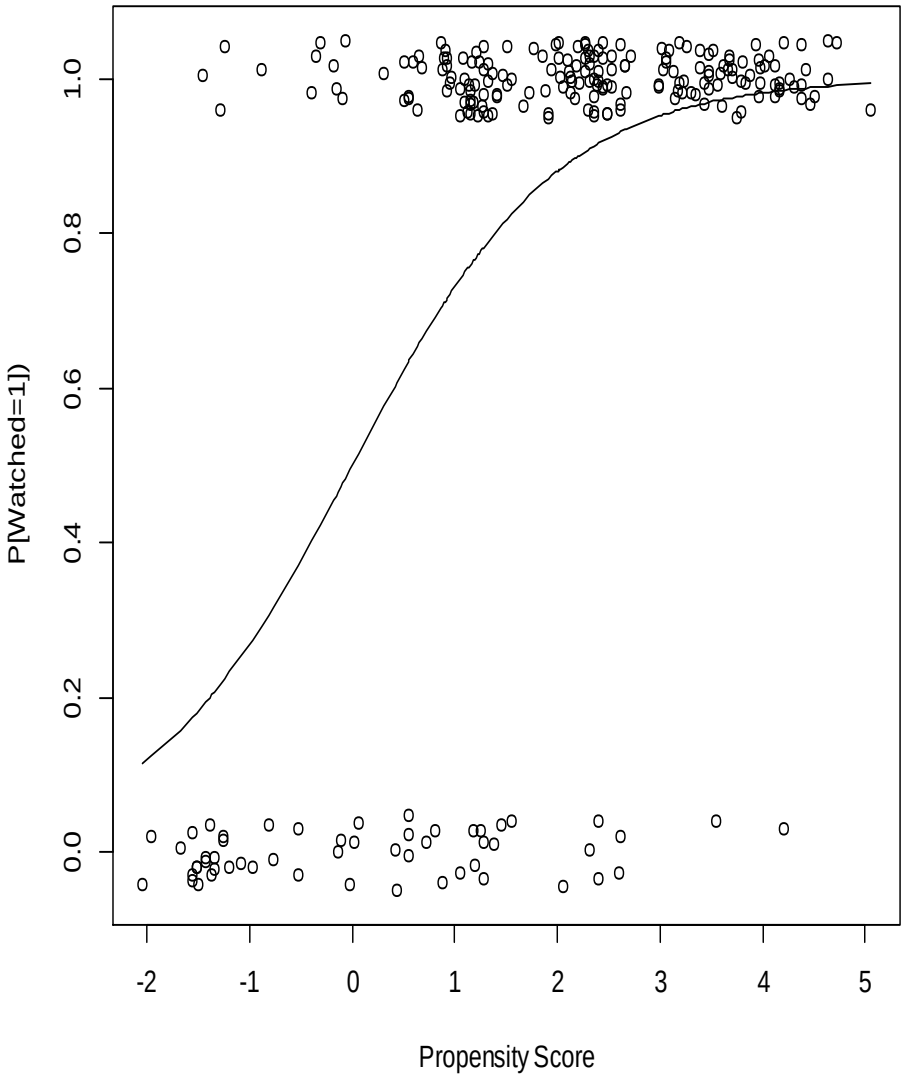



- $\gamma_1 = 0$: No selection effect!

- 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 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)))
```

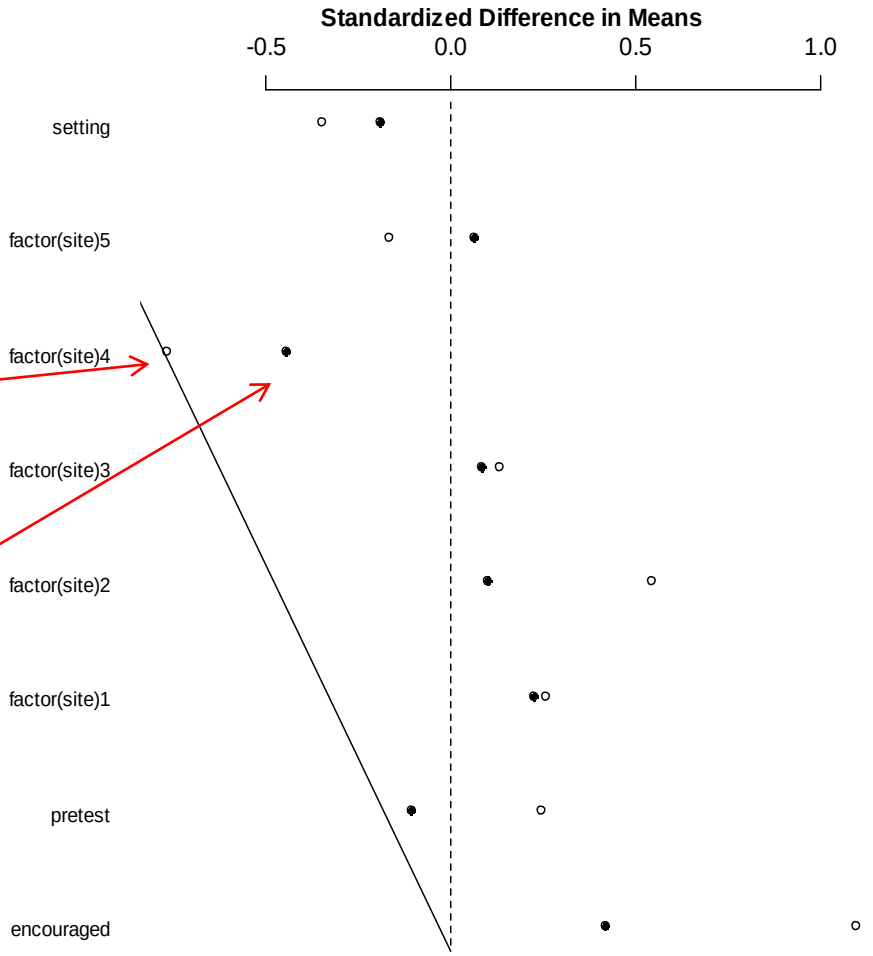


# Making the matched data set

```
> 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 $\gamma_1$ to zero.

# Now we estimate of effect of watching Sesame Street just using matched dataset

```
> 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
> 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
```

Unmatched Tx Effect Est.

Matched Tx Effect Est.

# Propensity Scores: How did we do?

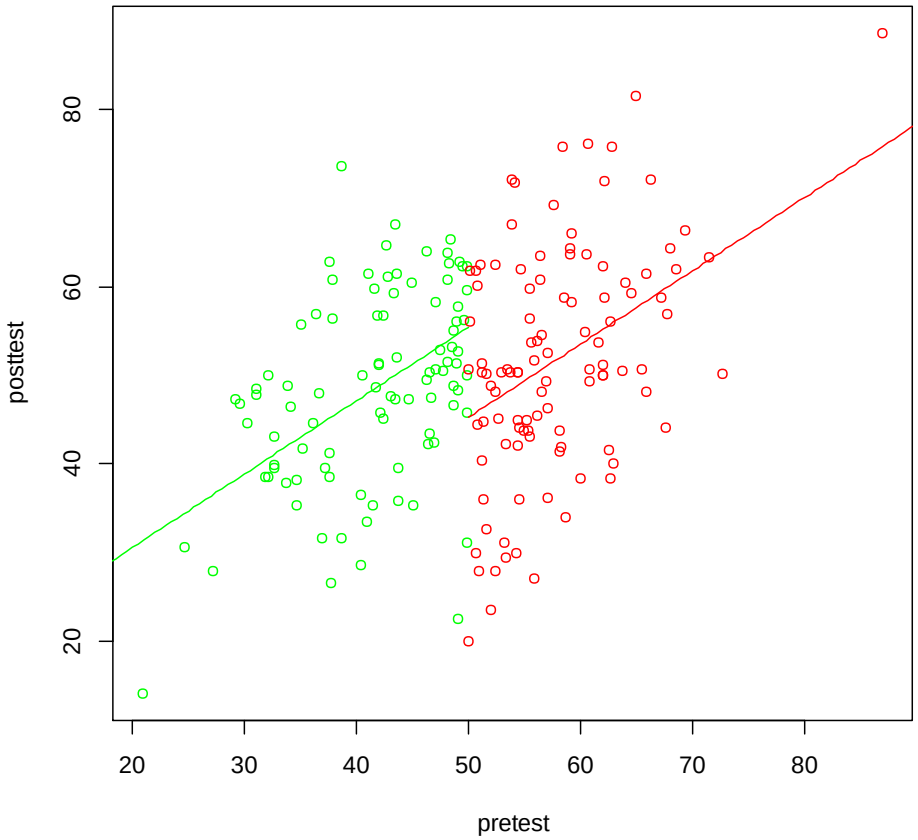
- The estimate of the effect of watching Sesame Street is a bit bigger for the matched data than for the non-matched data.
- It is not as big as the IV estimate, in part because the matching isn't very good yet. More effort needed to build a good logistic regression for the propensity scores!
- SE's 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)?*

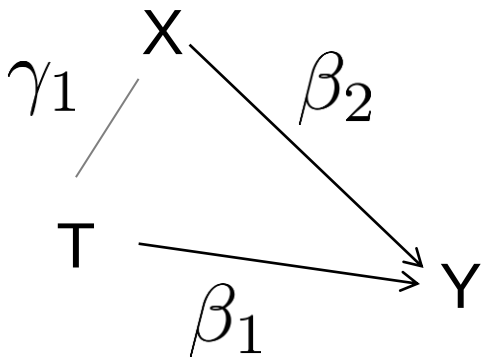
# Regression Discontinuity Designs

- *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 $\beta_2$. This also improves SE's for $\beta_1$.
- For subjects near the jump, $\gamma_1 \approx 0$, so $\beta_1$ represents a causal effect for them.
- *How far can we generalize $\beta_1$ 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
```

lowkids <- (pretest<50), same as T= Tx indicator

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

# Summary

- 15 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!)
- 16 More sophisticated tools for causal inference [G&H Ch 10]
  - Observational Studies
  - Instrumental Variables
  - Matching and propensity scores
  - Regression discontinuity designs