Final Exam Prep Sessions 6, 7, 8

MSBA 6440: Causal Inference via Experimentation

Danny Moncada (monca016)

April 27, 2020

Session 6: Differences-in-Differences

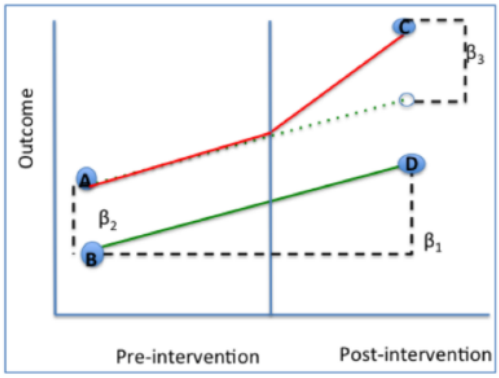
$$ln(y_{st}) = \boldsymbol{A}_s + \boldsymbol{B}_t + g \cdot \boldsymbol{Z}_{st} + p \cdot \boldsymbol{C}_{st} + e_{st} \quad (1)$$

where s indexes states and t indexes time, $t = 1999, \ldots, 2008$; y_{st} is the number of HIV cases for state s at time t; $\boldsymbol{A}_s$ is a vector of 33 state fixed effects; $\boldsymbol{B}_t$ is a vector of time fixed effects; $\boldsymbol{Z}_{st}$ is a vector of state demographics features, socio-economic indictors, and Internet availability, which includes age proportion, ethnicity proportion, population size, education attainment proportion, median income levels, and number of high speed Internet lines; $\boldsymbol{C}_{st}$ is the binary indicator for Craigslist entry, that is, $\boldsymbol{C}_{st} = 1$ if the state has Craigslist in a particular year, zero otherwise; and e_{st} is an error term. The coefficient p is the difference-in-difference estimate of the effect of Craigslist's entry on the incidence of HIV. If $p > 0$, then site entry has caused an increase in HIV prevalence.

Figure 1: Internet's Dirty Secret: Assessing the Impact of Online Intermediaries on HIV Transmiission

Intuition behind Difference-in-Differences

Coefficient	Calculation	Interpretation
β_0	B	Baseline average
β_1	D-B	Time trend in control group
β_2	A-B	Difference between two groups pre-intervention
β_3	(C-A)-(D-B)	Difference in changes over time



Difference-in-Differences Estimator

- **Accounts for Subject-Level Confounds and Inter-Temporal Confounds.**
  - First we take the after-before difference in each group.
  - Next, calculate the difference in those inter-temporal differences.

Average outcome in the "treatment group" After – Before — Average outcome in the "control group" After – Before

"Difference in Differences"

Difference-in-Differences Estimator

- After-before difference in each group accounts for any individual-level confounds that are time-invariant (same as a individual-level fixed effect)
- Calculate the difference in those differences. This accounts for inter-temporal (time-variant) confounds that are common to both individuals.

Average outcome in the "treatment group" After – Before **–** Average outcome in the "control group" After – Before

"Difference in Differences"

Regression Setup

- **We Estimate the Following Regression:**

$$Viewership_{it} = \alpha + \beta_1 Premium_i + \beta_2 After_t + \beta_3 Premium_i \cdot After_t + u_{it}$$

- E[Viewership | Premium=1, After=1] = $\alpha + \beta_1 + \beta_2 + \beta_3$
  - E[Viewership | Premium=1, After=0] = $\alpha + \beta_1$
  - E[Viewership | Premium=0, After=1] = $\alpha + \beta_2$
    - E[Viewership | Premium=0, After=0] = α

(E[Views | Prem = 1, After = 1] - E[Views | Prem = 0, After = 1])

–

(E[Views | Prem = 1, After = 0] - E[Views | Prem = 0, After = 0]) $= \beta_3$

```
setwd("~/MSBA 2020 All Files/Spring 2020/MSBA 6440 - Causal Inference via Ecnmtrcs Exprmnt/Week 6 - Dif
```

```
#### Load the data ####
MyData = read.csv("TSTV-Obs-Dataset.csv")


#how long is the period of observation?
max(MyData$week)-min(MyData$week)
```

```
## [1] 13
```

```r
#How many subjects got TSTV? (Treated)
length(unique(MyData$id[MyData$premium==TRUE]))
```

```
## [1] 8348
```

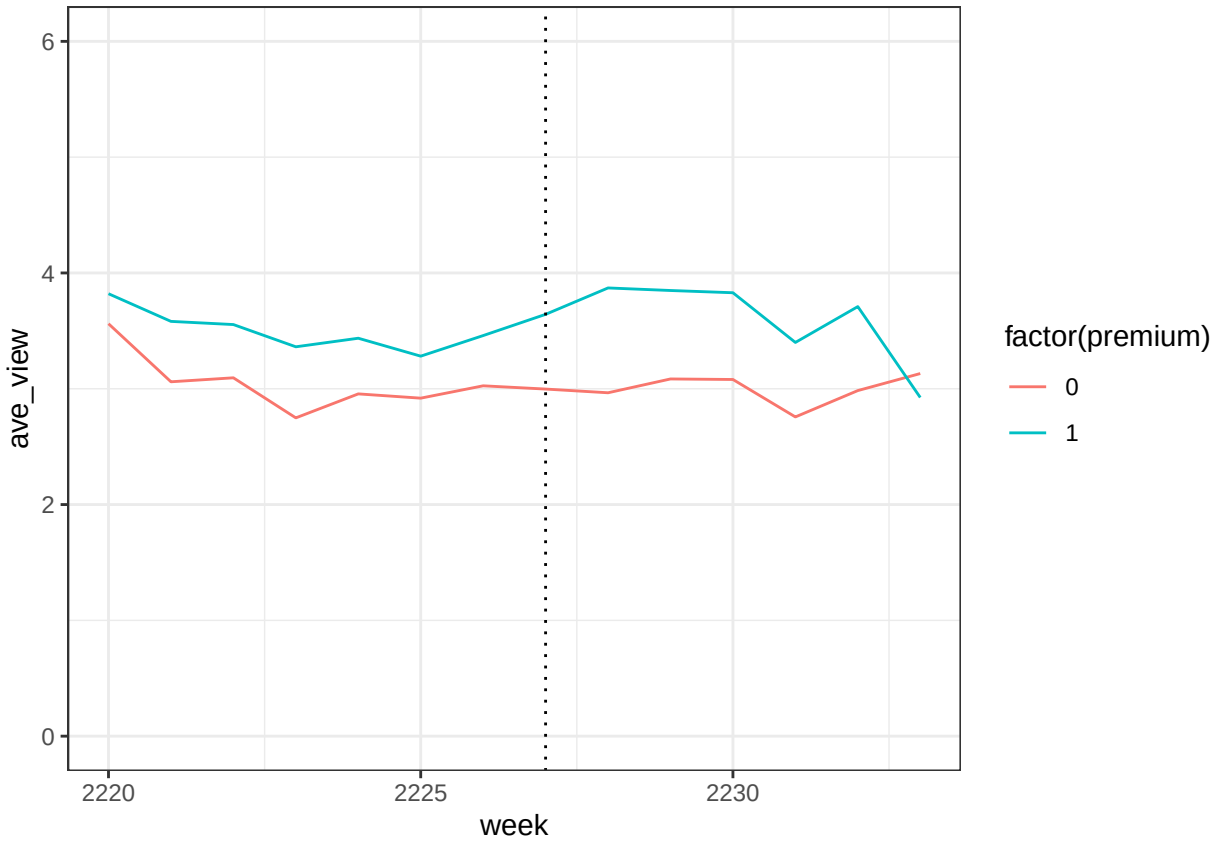
```r
#How many subjects did not get TSTV? (Control)
length(unique(MyData$id[MyData$premium==FALSE]))
```

```
## [1] 41686
```

```r
#In what 'week' does the "treatment" begin?
min(unique(MyData$week[MyData$after==TRUE]))
```

```
## [1] 2227
```

```r
# As descriptive visualization, let's look at average weekly viewership for both premium and regular vi
Week_Ave = MyData %>% group_by(week, premium) %>%
  summarise(ave_view = mean(view_time_total_hr)) %>% ungroup()
ggplot(Week_Ave, aes(x = week, y = ave_view, color = factor(premium))) +
  geom_line() +
  geom_vline(xintercept = 2227, linetype='dotted') +
  ylim(0, 6) + xlim(2220,2233) +
  theme_bw()
```



Wait, Isn't This Just a Fixed Effect Regression?

Premium is a Group FE, After is a Time FE:

$$Viewership_{it} = \alpha + \beta_1 Premium_i + \beta_2 After_t + \beta_3 Premium_i \cdot After_t + u_{it}$$

- The premium dummy absorbs everything that's systematically different about the treatment group (relative to the non-premium control).
- The after dummy absorbs anything that's systematically different about the post period (relative to the pre period), e.g., think common unobserved factors affecting the whole market.
- In fact: nothing stops us from replacing Premium with a set of customer dummies, and After with a set of week dummies! FE regression lets us have staggered treatments.

```
#### Difference in Differences Regression ####
# Interpret the treatment effect
did_basic = lm(log(view_time_total_hr+1) ~ premium*after, data=MyData)
summary(did_basic)
```

```
##
## Call:
## lm(formula = log(view_time_total_hr + 1) ~ premium * after, data = MyData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28421 -0.69919  0.07235  0.63026  2.05054
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.122544   0.001491  752.67   <2e-16 ***
## premium        0.116126   0.003613   32.14   <2e-16 ***
## after         -0.029016   0.002094  -13.86   <2e-16 ***
## premium:after  0.074558   0.005042   14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7695 on 652795 degrees of freedom
## Multiple R-squared:  0.006149,   Adjusted R-squared:  0.006145
## F-statistic:  1346 on 3 and 652795 DF,  p-value: < 2.2e-16
```

```
# a 1% increase in viewing time is 7% increase going from control to premium
```

```r
# Let's try replacing the treatment dummy with subject fixed effects.
# What happened to the estimate of premium?
did_fe = plm(log(view_time_total_hr+1) ~ premium*after, data = MyData,
             index=c("id"), effect="individual", model="within")
# Consumers are identified by id; here we want to see each individual subject and their fixed
# effect.

summary(did_fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(view_time_total_hr + 1) ~ premium * after,
##     data = MyData, effect = "individual", model = "within", index = c("id"))
##
## Unbalanced Panel: n = 50034, T = 1-14, N = 652799
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -2.583793 -0.252482  0.016201  0.296623  2.358606
##
## Coefficients:
##                 Estimate Std. Error t-value  Pr(>|t|)
## after         -0.0096263  0.0013662 -7.0462  1.84e-12 ***
## premium:after  0.0668180  0.0032670 20.4521 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    147680
## Residual Sum of Squares: 147580
## R-Squared:      0.00069803
## Adj. R-Squared: -0.082253
## F-statistic: 210.52 on 2 and 602763 DF, p-value: < 2.22e-16
```

```r
# Similar to output we had before; there is no coefficent for premium.  Because we are at the
# individual user fixed effect, it is perfectly correlated with premium/not premium; this washes
# this cofficient and removes it from the model.

# Further add week fixed effects
did_sfe_tfe = plm(log(view_time_total_hr+1) ~ premium*after,
                  data = MyData, index=c("id", "week"),
                  effect="twoway", model="within")
summary(did_sfe_tfe)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = log(view_time_total_hr + 1) ~ premium * after,
##     data = MyData, effect = "twoway", model = "within", index = c("id",
##         "week"))
##
## Unbalanced Panel: n = 50034, T = 1-14, N = 652799
```

```
##
## Residuals:
##      Min   1st Qu.    Median   3rd Qu.      Max.
## -2.594527 -0.252892  0.017542  0.295771  2.273132
##
## Coefficients:
##                Estimate Std. Error t-value  Pr(>|t|)
## premium:after 0.0682979  0.0032553   20.98 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    146610
## Residual Sum of Squares: 146510
## R-Squared:      0.00072974
## Adj. R-Squared: -0.082241
## F-statistic: 440.172 on 1 and 602751 DF, p-value: < 2.22e-16
```

```r
# We lose the "after" coefficient because the weeks are perfectly correlated with this.


# Let's try dynamic DiD instead.
did_dyn_sfe_tfe <- lm(log(view_time_total_hr+1) ~ premium +
                        factor(week) + premium*factor(week), data = MyData)
# for every week we have a sigma, and for every week we have a beta term/coefficient.
summary(did_dyn_sfe_tfe)
```

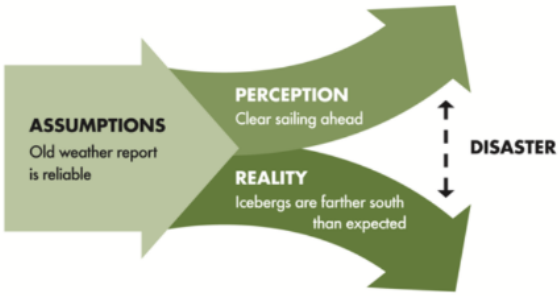
```
##
## Call:
## lm(formula = log(view_time_total_hr + 1) ~ premium + factor(week) +
##     premium * factor(week), data = MyData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35401 -0.70039  0.06861  0.62780  2.03182
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.284204   0.004361 294.479  < 2e-16 ***
## premium                     0.065613   0.010272   6.388 1.68e-10 ***
## factor(week)2221           -0.116254   0.005927 -19.614  < 2e-16 ***
## factor(week)2222           -0.131373   0.005858 -22.428  < 2e-16 ***
## factor(week)2223           -0.241444   0.005830 -41.416  < 2e-16 ***
## factor(week)2224           -0.199708   0.005810 -34.372  < 2e-16 ***
## factor(week)2225           -0.216551   0.005794 -37.375  < 2e-16 ***
## factor(week)2226           -0.185174   0.005794 -31.958  < 2e-16 ***
## factor(week)2227           -0.176850   0.005803 -30.474  < 2e-16 ***
## factor(week)2228           -0.193413   0.005814 -33.266  < 2e-16 ***
## factor(week)2229           -0.159218   0.005825 -27.336  < 2e-16 ***
## factor(week)2230           -0.162324   0.005835 -27.818  < 2e-16 ***
## factor(week)2231           -0.260881   0.005847 -44.621  < 2e-16 ***
## factor(week)2232           -0.192753   0.005860 -32.896  < 2e-16 ***
## factor(week)2233           -0.190512   0.005875 -32.426  < 2e-16 ***
## premium:factor(week)2221    0.061274   0.014178   4.322 1.55e-05 ***
## premium:factor(week)2222    0.053423   0.014022   3.810 0.000139 ***
```

```
## premium:factor(week)2223  0.078268   0.013944   5.613 1.99e-08 ***
## premium:factor(week)2224  0.060519   0.013882   4.360 1.30e-05 ***
## premium:factor(week)2225  0.033752   0.013828   2.441 0.014651 *
## premium:factor(week)2226  0.050495   0.013813   3.656 0.000257 ***
## premium:factor(week)2227  0.106577   0.013818   7.713 1.23e-14 ***
## premium:factor(week)2228  0.181185   0.013827  13.104  < 2e-16 ***
## premium:factor(week)2229  0.163413   0.013834  11.813  < 2e-16 ***
## premium:factor(week)2230  0.159237   0.013841  11.505  < 2e-16 ***
## premium:factor(week)2231  0.142558   0.013850  10.293  < 2e-16 ***
## premium:factor(week)2232  0.158616   0.013857  11.446  < 2e-16 ***
## premium:factor(week)2233 -0.035631   0.013867  -2.569 0.010186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7673 on 652771 degrees of freedom
## Multiple R-squared:  0.01174,    Adjusted R-squared:  0.0117
## F-statistic: 287.2 on 27 and 652771 DF,  p-value: < 2.2e-16
```

```
# Prior to week 2226, we want the cofficients to be zero (before treatment began).
# The cofficients DO increase after the treatment period, so that's a positive sign.
# Weeks 16:28, this is the interaction effect between premium and week fixed effect.
```

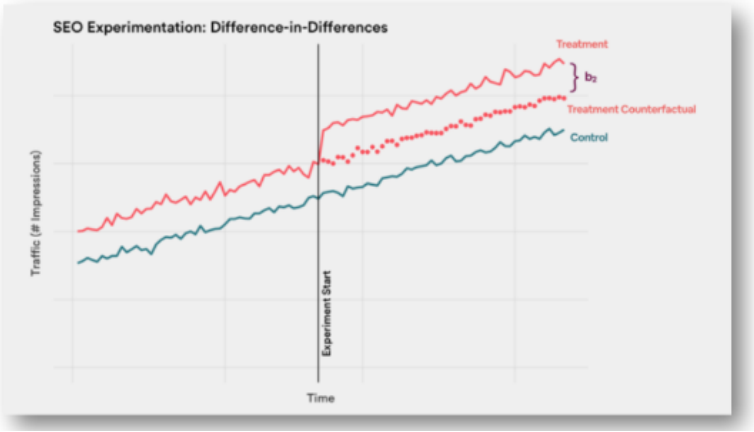
Difference-in-Differences Assumptions

- We have to make a few assumptions for this to work.
  - We assume parallel trends between the two groups on the outcome variable.
  - We also assume that treated subjects are not influencing control subjects.

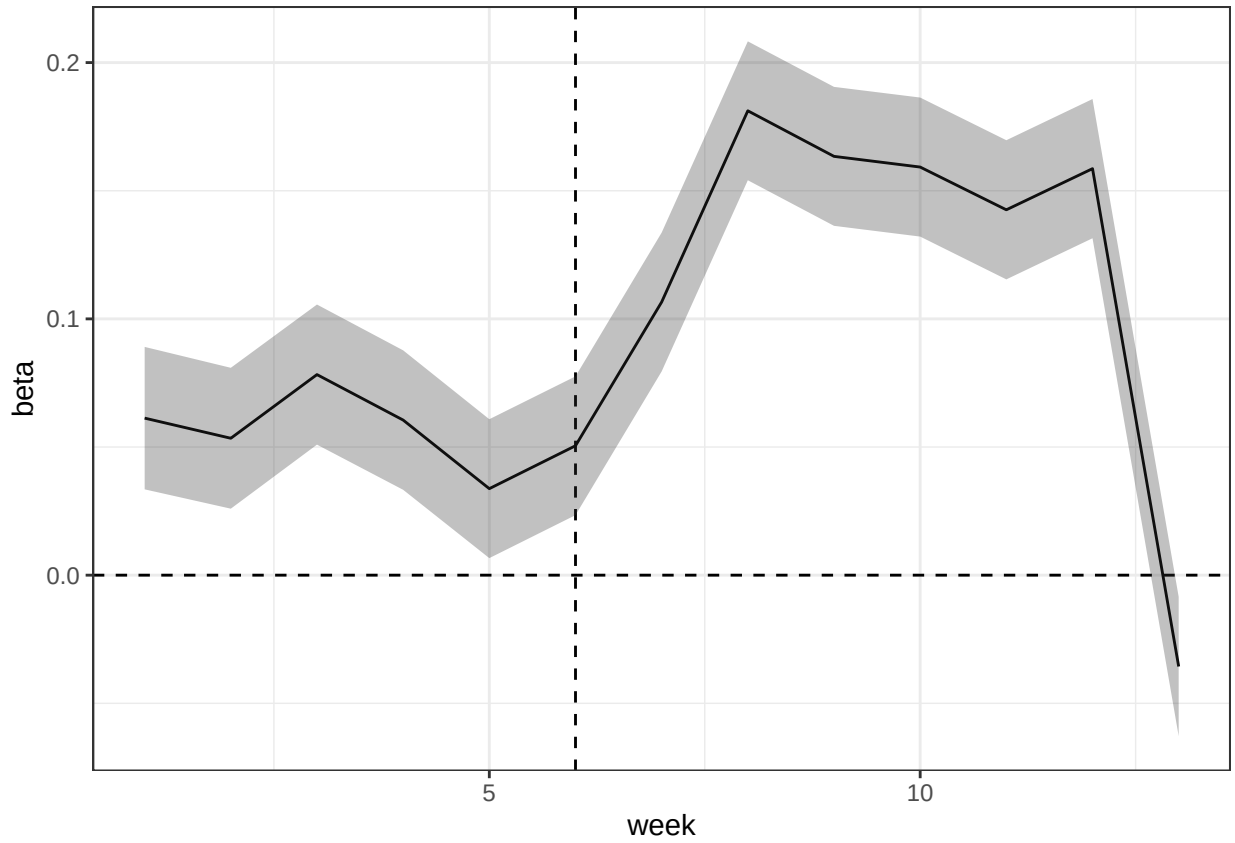


Difference-in-Differences Assumptions

- Why Parallel Trends?
  - Because, the control group needs to be a good counterfactual for the treated group!




```
# Let's retrieve the coefficients and standard errors, and create confidence intervals
model = summary(did_dyn_sfe_tfe)
coefs_ses = as.data.frame(model$coefficients[16:28,c("Estimate", "Std. Error")])
colnames(coefs_ses) = c("beta", "se")
coefs_ses = coefs_ses %>%
  mutate(ub90 = beta + 1.96*se, # Upper bound, 2 standard deviations
         lb90 = beta - 1.96*se, # lower bound, 2 SDs
         week = 1:nrow(coefs_ses))

# Let's connect the estimates with a line and include a ribbon for the CIs.
ggplot(coefs_ses, aes(x = week, y = beta)) +
  geom_line() +
  geom_hline(yintercept=0,linetype="dashed") +
  geom_vline(xintercept=6,linetype="dashed") +
  geom_ribbon(aes(ymin = lb90, ymax = ub90), alpha = 0.3) +
  theme_bw()
```




```
# If the parallel assumption was held, then the line prior to the treatment effect would
# be at zero.  But it is not, so this assumption is violated (but the effect is small).

# Time for our placebo test...
# Let's limit to pre-period data, and shift the treatment date back in time, artificially,
# and see if we see sig differences pre treatment.
# Again, recall first week when treatment starts

MyDataPre <- MyData[MyData$after==0,]
max(MyDataPre$week)
```

```
## [1] 2226
```

```
MyDataPre$after <- MyDataPre$week > 2224
did_log_basic_placebo <- lm(data=MyDataPre,log(view_time_total_hr+1)~premium+after+premium*after)
summary(did_log_basic_placebo)
```

```
##
## Call:
## lm(formula = log(view_time_total_hr + 1) ~ premium + after +
##     premium * after, data = MyDataPre)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25939 -0.66929  0.06303  0.61911  2.03342
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.139663   0.001763 646.302   <2e-16 ***
## premium             0.119725   0.004271  28.033   <2e-16 ***
## afterTRUE          -0.056323   0.003198 -17.610   <2e-16 ***
## premium:afterTRUE -0.011916   0.007750  -1.538    0.124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.759 on 320873 degrees of freedom
## Multiple R-squared:  0.004546,   Adjusted R-squared:  0.004536
## F-statistic: 488.4 on 3 and 320873 DF,  p-value: < 2.2e-16
```

Session 7: Instrumental Variables (ivreg)

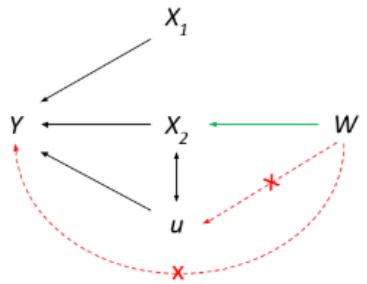
What's the Basic Idea?

- Think of your treatment variable, X, as a piece of fruit.
  - Instruments are like a knife that lets you separate "good" variation from bad, so you can then use it in your regression.

Mathematically

- You want to estimate the effect of X_2 on Y (equation).
  - But, you believe X_2 is also correlated with the error term i.e., it has unobserved confounders, is mis-measured, or y also influences X_2.
  - *If we have a W that is correlated with X_2 but NOT u or Y, then we can simply regress X_2 onto W and use the predicted values of X_2 in our final regression.*

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + u$$



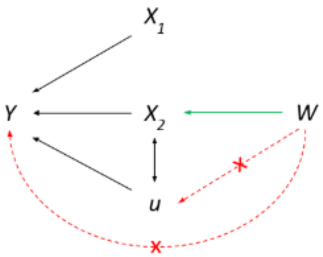
$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 + X_2 + u$$

2SLS: Two-Stage Least Squares

1. Regress X_2 onto its instrument(s), W, and recover predicted values of X_2 from the first stage, i.e., decompose X_2 into part explained by W, and "the rest," which includes the bad part.
2. Regress Y onto predicted values of X_2.

$$X_2 = \pi_0 + \pi_1 \cdot W + \epsilon$$

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot \hat{X_2} + u$$



- Note: this gives you a consistent estimate of Beta$_2$, but SEs will be wrong, because the software does not recognize you are using "predicted" variables.

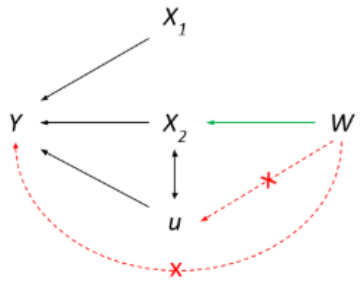
$$X_2 = \pi_0 + \pi_1 \cdot W + \epsilon$$

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot \hat{X_2} + u$$

Two Requirements

1. Relevance (Strength): W has to be highly correlated with X_2.

2. Exclusion (Exogeneity): W cannot be correlated with the error term, u, or Y (except through X_2). This is otherwise known as the "validity" requirement.




```
# Author: Gordon Burtch and Gautam Ray
# Course: MSBA 6440
# Session: Instrumental Variables
# Lecture 7

setwd("~/MSBA 2020 All Files/Spring 2020/MSBA 6440 - Causal Inference via Ecnmtrcs Exprmnt/Week 7 - Ins

MyData1<-read.csv("MROZ.csv")
MyData <- MyData1[MyData1$lfp==1,] #restricts sample to lfp=1

# OLS Model of Wage on Education

ols <- lm(log(wage)~educ+exper+expersq, data=MyData)

# 2SLS Model 'by hand'

educ.ols <- lm(educ~exper+expersq+motheduc, data=MyData)
educHat <- fitted(educ.ols)
wage.2sls <- lm(log(wage)~educHat+exper+expersq, data=MyData)

# IVREG

wage.ivreg <- ivreg(log(wage)~educ+exper+expersq|exper+expersq+motheduc, data=MyData)

stargazer(ols,wage.2sls,wage.ivreg,
          type="text",title="OLS vs 2SLS vs IVREG",
          column.labels = c("OLS","2SLS","IVREG"))
```

```
##
## OLS vs 2SLS vs IVREG
## ============================================================
##                                      Dependent variable:
##                                  -----------------------------
```

```
##                                         log(wage)
##                                    OLS         instrumental
##                                                  variable
##                                OLS      2SLS      IVREG
##                                (1)      (2)        (3)
## ---------------------------------------------------------------
## educ                         0.107***              0.049
##                               (0.014)              (0.037)
##
## educHat                                 0.049
##                                        (0.039)
##
## exper                        0.042***  0.045***   0.045***
##                               (0.013)  (0.014)    (0.014)
##
## expersq                      -0.001**  -0.001**   -0.001**
##                              (0.0004)  (0.0004)   (0.0004)
##
## Constant                     -0.522***  0.198      0.198
##                               (0.199)  (0.493)    (0.473)
##
## ---------------------------------------------------------------
## Observations                   428      428        428
## R2                            0.157    0.046      0.123
## Adjusted R2                   0.151    0.039      0.117
## Residual Std. Error (df = 424)  0.666    0.709      0.680
## F Statistic (df = 3; 424)    26.286*** 6.751***
## ===============================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
```

```r
# Setting "diagnostics = TRUE" let's us assess a hausman test, weak IV stats
# and overidentifying tests of instrument exclusion.
summary(wage.ivreg,diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = log(wage) ~ educ + exper + expersq | exper +
##     expersq + motheduc, data = MyData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.10804 -0.32633  0.06024  0.36772  2.34351
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1981861  0.4728772   0.419  0.67535
## educ         0.0492630  0.0374360   1.316  0.18891
## exper        0.0448558  0.0135768   3.304  0.00103 **
## expersq     -0.0009221  0.0004064  -2.269  0.02377 *
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1 424    73.946  <2e-16 ***
## Wu-Hausman         1 423     2.968  0.0856 .
```

```
## Sargan             0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6796 on 424 degrees of freedom
## Multiple R-Squared: 0.1231,  Adjusted R-squared: 0.1169
## Wald test: 7.348 on 3 and 424 DF,  p-value: 8.228e-05
```

```
wage.ivreg2 <- ivreg(log(wage)~educ+exper+expersq|exper+expersq+motheduc+fatheduc, data=MyData)
summary(wage.ivreg2,diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = log(wage) ~ educ + exper + expersq | exper +
##     expersq + motheduc + fatheduc, data = MyData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0986 -0.3196  0.0551  0.3689  2.3493
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0481003  0.4003281   0.120  0.90442
## educ         0.0613966  0.0314367   1.953  0.05147 .
## exper        0.0441704  0.0134325   3.288  0.00109 **
## expersq     -0.0008990  0.0004017  -2.238  0.02574 *
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   2 423    55.400  <2e-16 ***
## Wu-Hausman         1 423     2.793  0.0954 .
## Sargan             1  NA     0.378  0.5386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6747 on 424 degrees of freedom
## Multiple R-Squared: 0.1357,  Adjusted R-squared: 0.1296
## Wald test: 8.141 on 3 and 424 DF,  p-value: 2.787e-05
```

Example 1: Vietnam Veterans and Civilian Earnings (Angrist 1990)

- Vietnam era draft lottery assigned a Random Sequence Number (RSN) from 1-365 based on birthdate. Men with RSN below a ceiling were called for induction i.e., they were "draft-eligible."

- Draft eligibility ceilings: 195 in 1970, 125 in 1971, and 95 in 1972 were announced once Department of Defense needs were known.

- Instrument: Dummy variables for sequence of RSN, D1: RSN 1-5, D2, D3,…, D73: RSN 360-365. X: Vietnam Veteran Status.

Some Terminology

1. The effect of your treatment is "just-identified" if you have one instrument for it. Generally, if the number of instruments = number of endogenous regressors, the equation is just identified. A just-identified 2SLS is also called the IV estimator.

2. A 2SLS regression is "over-identified" if you have more instruments than endogenous regressors.

3. The regression is called "under-identified" if you do not have sufficient instruments for the number of endogenous regressors.

- **General Truths:** more instruments is better (up to a point), stronger instruments are better.

Evaluating Instrument Relevance

Relevance, i.e., corr $(W_i, X_i) \neq 0$

- Relevant instruments are highly correlated with the endogenous regressors even after controlling for the exogenous regressors. This requirement can be empirically tested using first stage regression model fit statistics.

- Stock and Watson's rule of thumb: the first-stage F-statistic testing the hypothesis that the coefficients on the instruments are jointly zero should be at least 10 (for a single endogenous regressor).

Evaluating Exclusion Restriction

Exclusion, i.e., corr $(W_i, e_i) = 0$

Excluded instruments are uncorrelated with the error term. This requirement needs a strong theoretical argument and can, in general, not be tested empirically. The theoretical argument has to be convincing.

1. First, describe how the instrument conceptually influences the endogenous regressor.
2. Second, rule out any direct effect of the instrument on the dependent variable or any effect on confounders.
3. Rule out any reverse effect of the dependent variable on the instrument, i.e., Y cannot influence W.

Figure 2: Cannot be tested empirically

Test Statistics

Weak IV Test: This is typically based on first-stage F-stat. Significance implies we reject the null of weak instruments.

Wu-Hausman Test: The null hypothesis is that IV yields equivalent estimates to OLS (and thus we should not use IV because we are losing power by doing so, and getting wider standard errors than necessary). Significance implies we reject the null of equivalence (and thus we should use IV).

Sargan Test: If we have more instruments than endogenous variables, the Sargan test evaluates equivalence of estimates between using all vs. a subset. Significance implies we reject the null of equivalence (and thus ***at least one instrument is invalid***). The test does not tell us which instrument is the problem, however.

```
# We will first simulate our treatment variable x, endogenous portion of x and its confounder, c
# Here, we refer to the endogenous variation in x as x*
# An easy way to make them confounded is to use the multivariate normal draw function, mvnorm.
xStarAndC <- mvrnorm(1000, c(20, 15), matrix(c(1, 0.5, 0.5, 1), 2, 2))
xStar <- xStarAndC[, 1]
c <- xStarAndC[, 2]

# If you are curious about syntax for mvnorm... ??MASS::mvnorm
# We pass it the number of obs to draw, the means of the two variables, and a covariance matrix.
# In this case, we simulated 1000 draws for two variables, mean 20 and 15, which are 50% correlated.
cov(xStar,c)
```

```
## [1] 0.4971877
```

```
# Now let's simulate our instrument, and make observed X a function of good variation (random stuff)
# and the bad variation, x*.
# By construction, z is a valid instrument for X now, because it is only correlated with the
# good variation, and it has no direct relationship on our eventual y (except through x).
z <- rnorm(1000)
x <- xStar + z

# Now let's simulate the data-generating process to recover y,
# a function of observed x, its confounder and an error term.
# Here, the true marginal effect of x on y is 2.
y <- 1 + 2*x + 10*c + rnorm(1000, 0, 0.5)

# Now lets check to make sure we have a problem of confounding.
cor(x, c)
```

```
## [1] 0.3399602
```

```r
cor(y, c)
```

```
## [1] 0.9700384
```

```r
# And let's check that the instrument is valid...
cor(x,z)
```

```
## [1] 0.6964734
```

```r
cor(c,z)
```

```
## [1] -0.02552544
```

```r
# Okay, let's run the 'true' regression first, controlling for the confounder.
ols_true <- lm(y ~ x + c)
# ... and let's run the endogenous regression, ignoring the confounder.
ols_endog <- lm(y ~ x)
stargazer(ols_true,ols_endog,type="text",
          title="True vs. Endogenous Regression",
          column.labels = c("True","Endogenous"))
```

```
##
## True vs. Endogenous Regression
## ================================================================================
##                                        Dependent variable:
##                                ---------------------------------------------
##                                                    y
##                                      True                      Endogenous
##                                      (1)                          (2)
## --------------------------------------------------------------------------------
## x                                  2.006***                     4.378***
##                                    (0.012)                      (0.208)
##
## c                                 10.024***
##                                    (0.017)
##
## Constant                           0.497*                     103.557***
##                                    (0.278)                      (4.168)
##
## --------------------------------------------------------------------------------
## Observations                        1,000                        1,000
## R2                                  0.998                        0.308
## Adjusted R2                         0.998                        0.307
## Residual Std. Error           0.488 (df = 997)             9.289 (df = 998)
## F Statistic             260,771.500*** (df = 2; 997) 443.169*** (df = 1; 998)
## ================================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

```r
# Okay, so let's start working toward IV reg. Let's do the first stage regression and use its predictio
xHat <- lm(x ~ z)$fitted.values
```

```r
ols_corrected <- lm(y ~ xHat)
stargazer(ols_true,ols_endog,ols_corrected,
          type="text",
          title="True vs. Endogenous vs. Instrumented",
          column.labels = c("True","Endogenous","Manual"))
```

```
##
## True vs. Endogenous vs. Instrumented
## ================================================================================================
##                                                Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                        y
##                                  True                 Endogenous                Manual
##                                  (1)                    (2)                      (3)
## ------------------------------------------------------------------------------------------------
## x                              2.006***               4.378***
##                                (0.012)                (0.208)
##
## c                             10.024***
##                                (0.017)
##
## xHat                                                                           1.771***
##                                                                                (0.354)
##
## Constant                        0.497*               103.557***              155.674***
##                                (0.278)                (4.168)                  (7.094)
##
## ------------------------------------------------------------------------------------------------
## Observations                    1,000                  1,000                    1,000
## R2                              0.998                  0.308                    0.024
## Adjusted R2                     0.998                  0.307                    0.023
## Residual Std. Error       0.488 (df = 997)       9.289 (df = 998)        11.025 (df = 998)
## F Statistic         260,771.500*** (df = 2; 997) 443.169*** (df = 1; 998) 24.967*** (df = 1; 998)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

```r
# Note that the beta is correctly estimated but the standard errors are not if we use this approach.
# The ivreg package will calculate not only this beta, but the right standard errors.
ivreg <- ivreg(formula=y ~ x | z)
stargazer(ols_true,ols_endog,ols_corrected,ivreg,
          type="text",
          title="True vs. Endogenous vs. Manual vs. Instrumented",
          column.labels = c("True","Endogenous","Manual","IV"))
```

```
##
## True vs. Endogenous vs. Manual vs. Instrumented
## ================================================================================
##                                                Dependent variable:
##                     ---------------------------------------------------------
##                                                        y
##                                                       OLS
##
```

```
##                                True                    Endogenous              Manual
##                                (1)                       (2)                   (3)
## ------------------------------------------------------------------------------------------
## x                            2.006***                  4.378***
##                              (0.012)                   (0.208)
##
## c                           10.024***
##                              (0.017)
##
## xHat                                                                          1.771***
##                                                                               (0.354)
##
## Constant                      0.497*                  103.557***            155.674***
##                              (0.278)                   (4.168)               (7.094)
##
## ------------------------------------------------------------------------------------------
## Observations                  1,000                     1,000                 1,000
## R2                            0.998                     0.308                 0.024
## Adjusted R2                   0.998                     0.307                 0.023
## Residual Std. Error       0.488 (df = 997)          9.289 (df = 998)      11.025 (df = 998)    9.9
## F Statistic         260,771.500*** (df = 2; 997) 443.169*** (df = 1; 998) 24.967*** (df = 1; 998)
## ==========================================================================================
## Note:                                                                    *p<0.1; **p<0
```

```
summary(ivreg,diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = y ~ x | z)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.428e+01 -6.687e+00  7.278e-04  6.892e+00  3.502e+01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 155.6743     6.4297  24.212  < 2e-16 ***
## x             1.7707     0.3212   5.513  4.5e-08 ***
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1 998     940.1  <2e-16 ***
## Wu-Hausman         1 997     173.7  <2e-16 ***
## Sargan             0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.993 on 998 degrees of freedom
## Multiple R-Squared: 0.1985,  Adjusted R-squared: 0.1977
## Wald test: 30.39 on 1 and 998 DF,  p-value: 4.502e-08
```

```
# If you have multiple endogenous and instrumental variables in a single regression, you can tell R exp
# This syntax means "remove" x and include z.
```

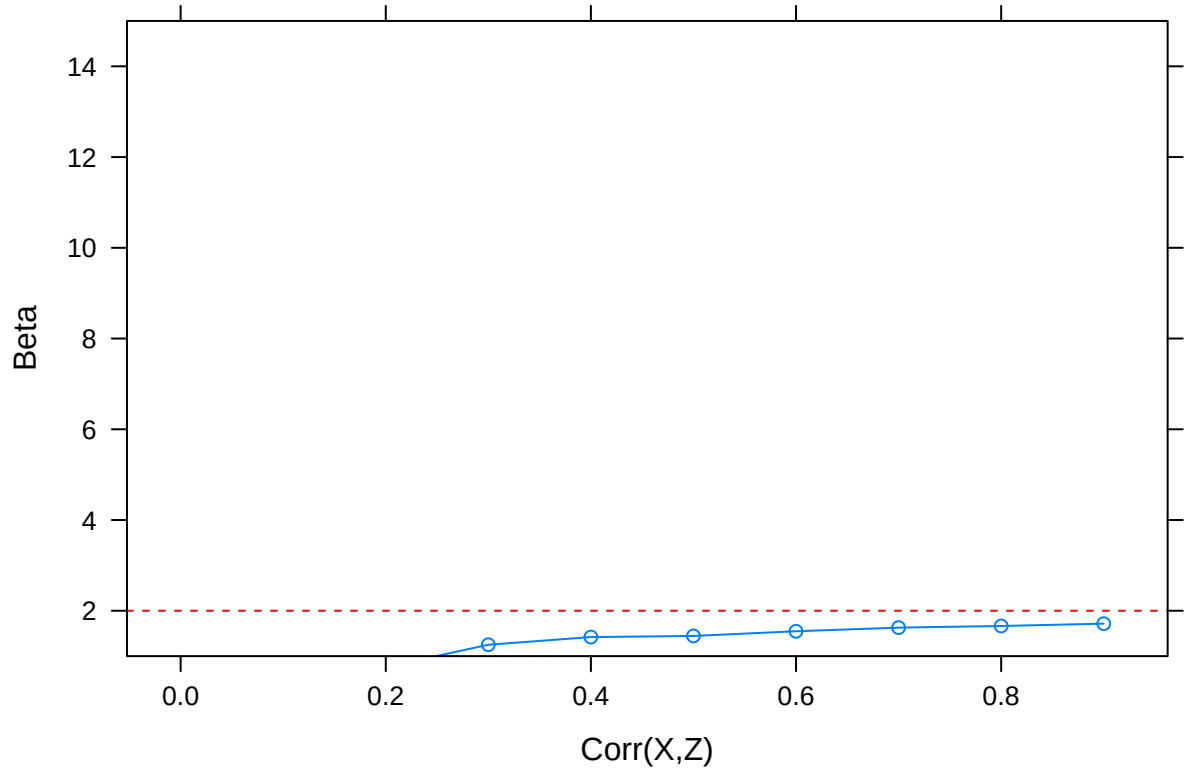
```r
# Any variables not mentioned after the pipe instrument for themselves (perfect predictors).
#ivreg <- ivreg(formula=y ~ x | .-x + z)

# Okay, now let's see what happens if we use instruments that are too weak in their association with x.
x1 <- xStar + 0.9*z
x2 <- xStar + 0.8*z
x3 <- xStar + 0.7*z
x4 <- xStar + 0.6*z
x5 <- xStar + 0.5*z
x6 <- xStar + 0.4*z
x7 <- xStar + 0.3*z
x8 <- xStar + 0.2*z
x9 <- xStar + 0.1*z
x10 <- xStar + 0.01*z

# Now let's simulate the data-generating process to recover y,
# a function of observed x, its confounder and an error term.
# Here, the true marginal effect of x on y is 2.
y1 <- 1 + 2*x1 + 10*c + rnorm(1000, 0, 0.5)
y2 <- 1 + 2*x2 + 10*c + rnorm(1000, 0, 0.5)
y3 <- 1 + 2*x3 + 10*c + rnorm(1000, 0, 0.5)
y4 <- 1 + 2*x4 + 10*c + rnorm(1000, 0, 0.5)
y5 <- 1 + 2*x5 + 10*c + rnorm(1000, 0, 0.5)
y6 <- 1 + 2*x6 + 10*c + rnorm(1000, 0, 0.5)
y7 <- 1 + 2*x7 + 10*c + rnorm(1000, 0, 0.5)
y8 <- 1 + 2*x8 + 10*c + rnorm(1000, 0, 0.5)
y9 <- 1 + 2*x9 + 10*c + rnorm(1000, 0, 0.5)
y10 <- 1 + 2*x10 + 10*c + rnorm(1000, 0, 0.5)

ivreg_weak1 <- ivreg(formula=y1 ~ x1 | z)
ivreg_weak2 <- ivreg(formula=y2 ~ x2 | z)
ivreg_weak3 <- ivreg(formula=y3 ~ x3 | z)
ivreg_weak4 <- ivreg(formula=y4 ~ x4 | z)
ivreg_weak5 <- ivreg(formula=y5 ~ x5 | z)
ivreg_weak6 <- ivreg(formula=y6 ~ x6 | z)
ivreg_weak7 <- ivreg(formula=y7 ~ x7 | z)
ivreg_weak8 <- ivreg(formula=y8 ~ x8 | z)
ivreg_weak9 <- ivreg(formula=y9 ~ x9 | z)
ivreg_weak10 <- ivreg(formula=y10 ~ x10 | z)

# The weaker our instrument, the less accurate our final estimate of X's effect becomes.
stargazer(ivreg_weak1,ivreg_weak2,
          ivreg_weak3,ivreg_weak4,
          ivreg_weak5,ivreg_weak6,
          ivreg_weak7,ivreg_weak8,
          ivreg_weak9,ivreg_weak10,
          type="text")
```

```
##
## =============================================================================================
##                                                                              Dependent variable:
##                                    ------------------------------------------------------------
##                                        y1          y2          y3          y4          y5          y6
```

```
##                                   (1)        (2)        (3)        (4)        (5)        (6)
## ------------------------------------------------------------------------------------------
## x1                              1.716***
##                                 (0.356)
##
## x2                                         1.664***
##                                            (0.403)
##
## x3                                                    1.628***
##                                                       (0.460)
##
## x4                                                               1.548***
##                                                                  (0.540)
##
## x5                                                                          1.444**
##                                                                             (0.650)
##
## x6                                                                                     1.420*
##                                                                                       (0.816)
##
## x7                                                                                              1
##                                                                                                (1
##
## x8
##
##
## x9
##
##
## x10
##
##
## Constant                       156.800*** 157.814*** 158.557*** 160.108*** 162.209*** 162.719*** 166
##                                 (7.128)    (8.064)    (9.209)    (10.815)   (13.007)   (16.329)   (2
##
## ------------------------------------------------------------------------------------------
## Observations                     1,000      1,000      1,000      1,000      1,000      1,000      1
## R2                               0.191      0.181      0.176      0.166      0.154      0.150      0
## Adjusted R2                      0.190      0.180      0.175      0.165      0.153      0.149      0
## Residual Std. Error (df = 998)   9.972      10.030     10.024     10.093     10.118     10.165     1
## ==========================================================================================
## Note:
```

```r
# Let's plot it for interests sake...
# First we pull out all the beta estimates, and we make a vector of the correlations we used (strength
betas <- rep(NA,10)
for (i in 1:10){
  betas[i] <- get(paste0('ivreg_weak',i))$coefficients[2]
}
weakness <- c(.9,.8,.7,.6,.5,.4,.3,.2,.1,.01)

# Now let's plot our recovered betas, against their strength, and include a ref line for true value of

p <- xyplot(betas~weakness,
```

```
            xlab="Corr(X,Z)",ylab="Beta",ylim=c(1,15),type="b")
update(p, panel=function(...){
  panel.xyplot(...)
  panel.abline(h=2,lty=2,col="red")
} )
```



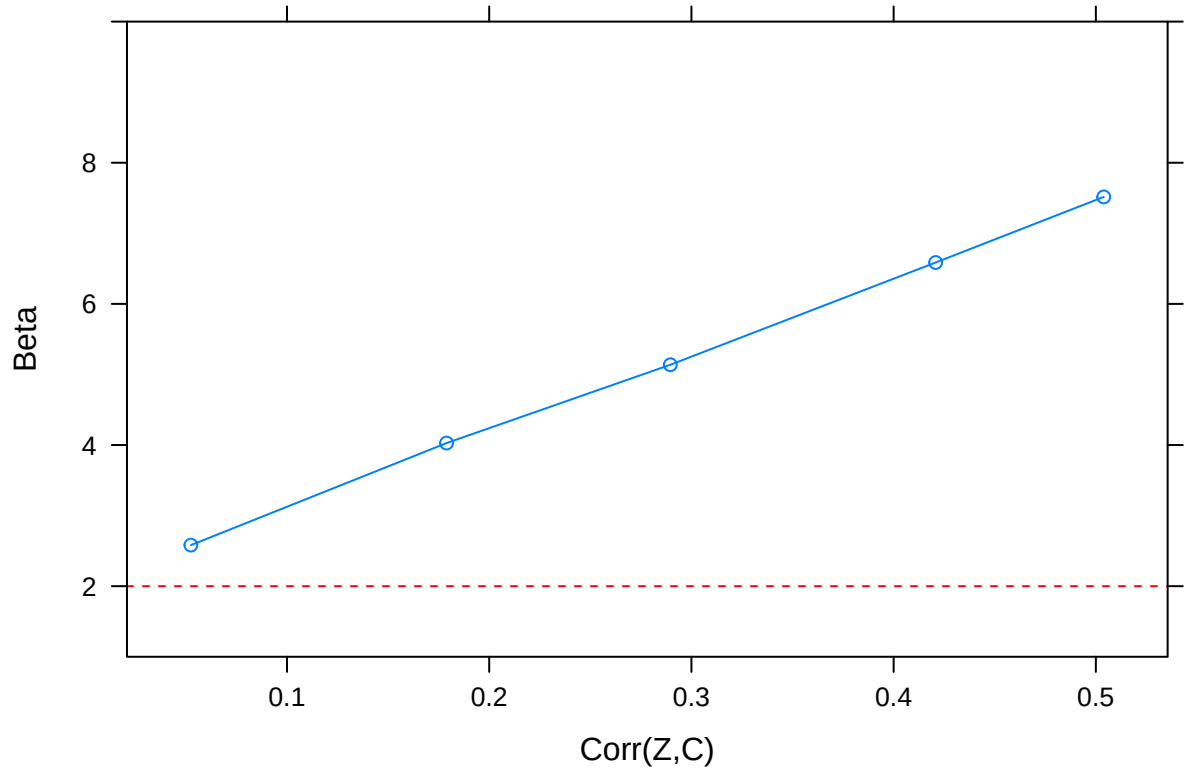

```
# Okay, now let's see what happens as we violate exclusion
# That is, as we allow z to be correlated to an increasing degree with the confounders in the error ter
# To make this work, we now need to draw all three variables from a joint distribution (good x, z and c
x_C_Z_1 <- mvrnorm(1000, c(20, 15, 10), matrix(c(1,0.5,0.9,0.5,1,.1,.9,.1,1), 3, 3))
x_C_Z_2 <- mvrnorm(1000, c(20, 15, 10), matrix(c(1,0.5,0.9,0.5,1,.2,.9,.2,1), 3, 3))
x_C_Z_3 <- mvrnorm(1000, c(20, 15, 10), matrix(c(1,0.5,0.9,0.5,1,.3,.9,.3,1), 3, 3))
x_C_Z_4 <- mvrnorm(1000, c(20, 15, 10), matrix(c(1,0.5,0.9,0.5,1,.4,.9,.4,1), 3, 3))
x_C_Z_5 <- mvrnorm(1000, c(20, 15, 10), matrix(c(1,0.5,0.9,0.5,1,.5,.9,.5,1), 3, 3))

x11 <- x_C_Z_1[, 1]
x12 <- x_C_Z_2[, 1]
x13 <- x_C_Z_3[, 1]
x14 <- x_C_Z_4[, 1]
x15 <- x_C_Z_5[, 1]
c1 <- x_C_Z_1[, 2]
c2 <- x_C_Z_2[, 2]
c3 <- x_C_Z_3[, 2]
c4 <- x_C_Z_4[, 2]
c5 <- x_C_Z_5[, 2]
```

```r
z1 <- x_C_Z_1[, 3]
z2 <- x_C_Z_2[, 3]
z3 <- x_C_Z_3[, 3]
z4 <- x_C_Z_4[, 3]
z5 <- x_C_Z_5[, 3]

# What are we doing here? Making versions of z that are increasingly correlated with c.
# Let's store those correlations for our plot later.
exclusion <- rep(NA,5)
for (i in 1:5){
  exclusion[i] <- cor(get(paste0("c",i)),get(paste0("z",i)))
}

# Okay, now let's simulate our Y's
y11 <- 1 + 2*x11 + 10*c1 + rnorm(1000, 0, 0.5)
y12 <- 1 + 2*x12 + 10*c2 + rnorm(1000, 0, 0.5)
y13 <- 1 + 2*x13 + 10*c3 + rnorm(1000, 0, 0.5)
y14 <- 1 + 2*x14 + 10*c4 + rnorm(1000, 0, 0.5)
y15 <- 1 + 2*x15 + 10*c5 + rnorm(1000, 0, 0.5)

ivreg_endog1 <- ivreg(formula=y11 ~ x11 | z1)
ivreg_endog2 <- ivreg(formula=y12 ~ x12 | z2)
ivreg_endog3 <- ivreg(formula=y13 ~ x13 | z3)
ivreg_endog4 <- ivreg(formula=y14 ~ x14 | z4)
ivreg_endog5 <- ivreg(formula=y15 ~ x15 | z5)

stargazer(ivreg_endog1,ivreg_endog2,
          ivreg_endog3,ivreg_endog4,
          ivreg_endog5,type="text")
```

```
##
## ================================================================================
##                                          Dependent variable:
##                                 ---------------------------------------------
##                                    y11        y12        y13        y14       y15
##                                    (1)        (2)        (3)        (4)       (5)
## --------------------------------------------------------------------------------
## x11                              2.582***
##                                  (0.342)
##
## x12                                        4.028***
##                                            (0.335)
##
## x13                                                   5.138***
##                                                       (0.304)
##
## x14                                                              6.587***
##                                                                  (0.297)
##
## x15                                                                        7.516***
##                                                                            (0.302)
##
## Constant                        139.479*** 110.714*** 88.426*** 58.912*** 41.136***
```

```
##                                 (6.851)    (6.724)    (6.106)    (5.938)    (6.038)
##
## ------------------------------------------------------------------------------------
## Observations                     1,000      1,000      1,000      1,000      1,000
## R2                               0.224      0.307      0.364      0.414      0.390
## Adjusted R2                      0.223      0.306      0.363      0.414      0.390
## Residual Std. Error (df = 998)   9.611      9.354      8.916      8.678      8.909
## ====================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

```r
# Let's plot it again...
# As you can see, as Z becomes less "excluded" we see it yields worse and worse estimates of X's margin
# Effect on Y.
betas <- rep(NA,5)
for (i in 1:5){
  betas[i] <- get(paste0('ivreg_endog',i))$coefficients[2]
}
p <- xyplot(betas~exclusion,xlab="Corr(Z,C)",ylab="Beta",ylim=c(1,10),type="b")
update(p, panel=function(...){
  panel.xyplot(...)
  panel.abline(h=2,lty=2,col="red")
} )
```




```r
# Okay let's do a real example here...
# This dataset is state-level data on cigarette sales, prices
```

```r
# and taxes. Taxes are used as an instrument for prices here.
data("CigarettesSW")
sales <- lm(log(packs) ~ log(price) + year + state, data=CigarettesSW)
sales_iv <- ivreg(log(packs) ~ log(price) + year + state | .-log(price) + tax, data = CigarettesSW)
stargazer(sales,sales_iv,
          title="OLS vs. IV",
          type="text",
          column.labels = c("OLS","IV"),
          omit=c("state","year"))
```

```
##
## OLS vs. IV
## ===============================================================
##                                       Dependent variable:
##                                   -----------------------------
##                                          log(packs)
##                                     OLS           instrumental
##                                                     variable
##                                     OLS               IV
##                                     (1)               (2)
## ---------------------------------------------------------------
## log(price)                       -1.085***         -1.380***
##                                   (0.151)           (0.192)
##
## Constant                         9.763***          11.108***
##                                   (0.689)           (0.879)
##
## ---------------------------------------------------------------
## Observations                        96                96
## R2                                 0.966             0.963
## Adjusted R2                        0.929             0.923
## Residual Std. Error (df = 46)      0.065             0.068
## F Statistic                 26.306*** (df = 49; 46)
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

```r
# Setting "diagnostics = TRUE" let's us assess a hausman test, weak IV stats and overidentifying tests
summary(sales_iv,diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = log(packs) ~ log(price) + year + state | . -
##     log(price) + tax, data = CigarettesSW)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.068e-01 -3.755e-02 -2.665e-15  3.755e-02  1.068e-01
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.107617   0.878778  12.640  < 2e-16 ***
## log(price)  -1.379516   0.192287  -7.174 5.00e-09 ***
## year1995     0.528352   0.109568   4.822 1.59e-05 ***
```

```
## stateAR      0.162404   0.068269   2.379 0.021573 *
## stateAZ     -0.026116   0.073058  -0.357 0.722376
## stateCA     -0.128743   0.075062  -1.715 0.093046 .
## stateCO     -0.134011   0.067687  -1.980 0.053721 .
## stateCT      0.224606   0.085641   2.623 0.011793 *
## stateDE      0.242414   0.067798   3.576 0.000835 ***
## stateFL      0.183090   0.073189   2.502 0.015984 *
## stateGA     -0.017651   0.067915  -0.260 0.796100
## stateIA      0.069703   0.069904   0.997 0.323916
## stateID     -0.119940   0.068841  -1.742 0.088142 .
## stateIL      0.101860   0.071709   1.420 0.162213
## stateIN      0.166653   0.068135   2.446 0.018333 *
## stateKS     -0.016880   0.067964  -0.248 0.804951
## stateKY      0.334593   0.071542   4.677 2.58e-05 ***
## stateLA      0.145811   0.068569   2.126 0.038860 *
## stateMA      0.140108   0.078244   1.791 0.079931 .
## stateMD     -0.075703   0.067854  -1.116 0.270358
## stateME      0.237828   0.072314   3.289 0.001934 **
## stateMI      0.246150   0.080069   3.074 0.003544 **
## stateMN      0.186660   0.079611   2.345 0.023416 *
## stateMO      0.127731   0.067728   1.886 0.065628 .
## stateMS      0.090566   0.068273   1.327 0.191217
## stateMT     -0.158804   0.067765  -2.343 0.023485 *
## stateNC      0.071140   0.071434   0.996 0.324515
## stateND     -0.003623   0.071412  -0.051 0.959758
## stateNE     -0.001406   0.069445  -0.020 0.983938
## stateNH      0.471720   0.067662   6.972 1.00e-08 ***
## stateNJ      0.110496   0.074559   1.482 0.145162
## stateNM     -0.281978   0.068497  -4.117 0.000158 ***
## stateNV      0.320003   0.076739   4.170 0.000133 ***
## stateNY      0.106181   0.078329   1.356 0.181853
## stateOH      0.114078   0.067700   1.685 0.098754 .
## stateOK      0.124564   0.067928   1.834 0.073162 .
## stateOR      0.057357   0.068822   0.833 0.408925
## statePA      0.095985   0.069629   1.379 0.174709
## stateRI      0.253760   0.074983   3.384 0.001468 **
## stateSC     -0.029784   0.069276  -0.430 0.669253
## stateSD     -0.058327   0.067647  -0.862 0.393040
## stateTN      0.184855   0.067821   2.726 0.009048 **
## stateTX      0.020787   0.072590   0.286 0.775883
## stateUT     -0.483985   0.070562  -6.859 1.48e-08 ***
## stateVA      0.050808   0.067873   0.749 0.457924
## stateVT      0.269097   0.068215   3.945 0.000271 ***
## stateWA      0.135134   0.092001   1.469 0.148687
## stateWI      0.161820   0.075760   2.136 0.038038 *
## stateWV      0.129671   0.068524   1.892 0.064749 .
## stateWY      0.044147   0.068171   0.648 0.520465
##
## Diagnostic tests:
##                  df1 df2 statistic  p-value
## Weak instruments   1  46    91.451 1.65e-12 ***
## Wu-Hausman         1  45     8.905  0.00458 **
## Sargan             0  NA        NA       NA
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06764 on 46 degrees of freedom
## Multiple R-Squared: 0.9627,  Adjusted R-squared: 0.9229
## Wald test: 24.36 on 49 and 46 DF,  p-value: < 2.2e-16
```

Session 8: Regression Discontinuity (rdrobust)

Regression Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment – Thistlewait and Campbell (1960)

- Research Question: Do the students who win merit scholarship awards are more likely to finish college?

- Y: Likelihood of Finishing College
- X: Score on the PSAT
- D: National Merit Scholarship Award based on X

- Estimate of the Treatment Effect: Jump in the relationship between PSAT score and college degree in the neighborhood of the award threshold.
- Rubin (1977): Treatment assignment based on a covariate.

* Measure the difference of the likelihood of students finishing college after getting a 79 on the PSAT and getting scholarship vs students who did not.

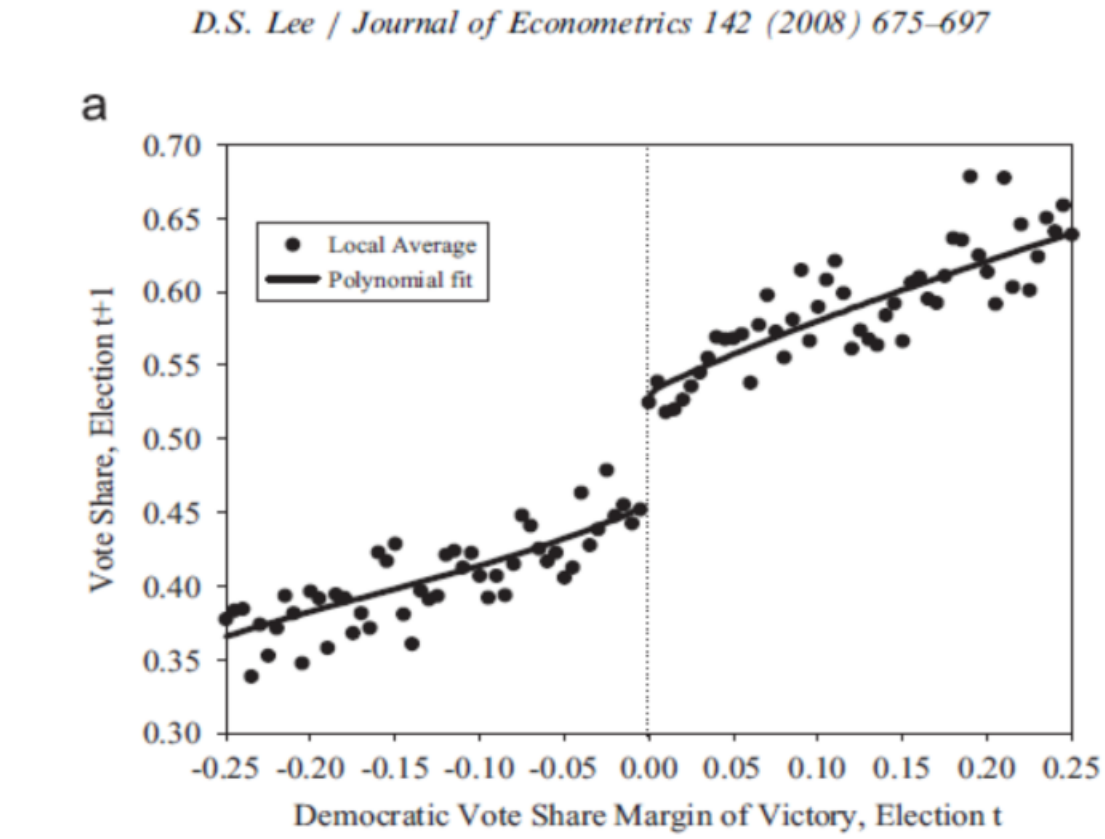


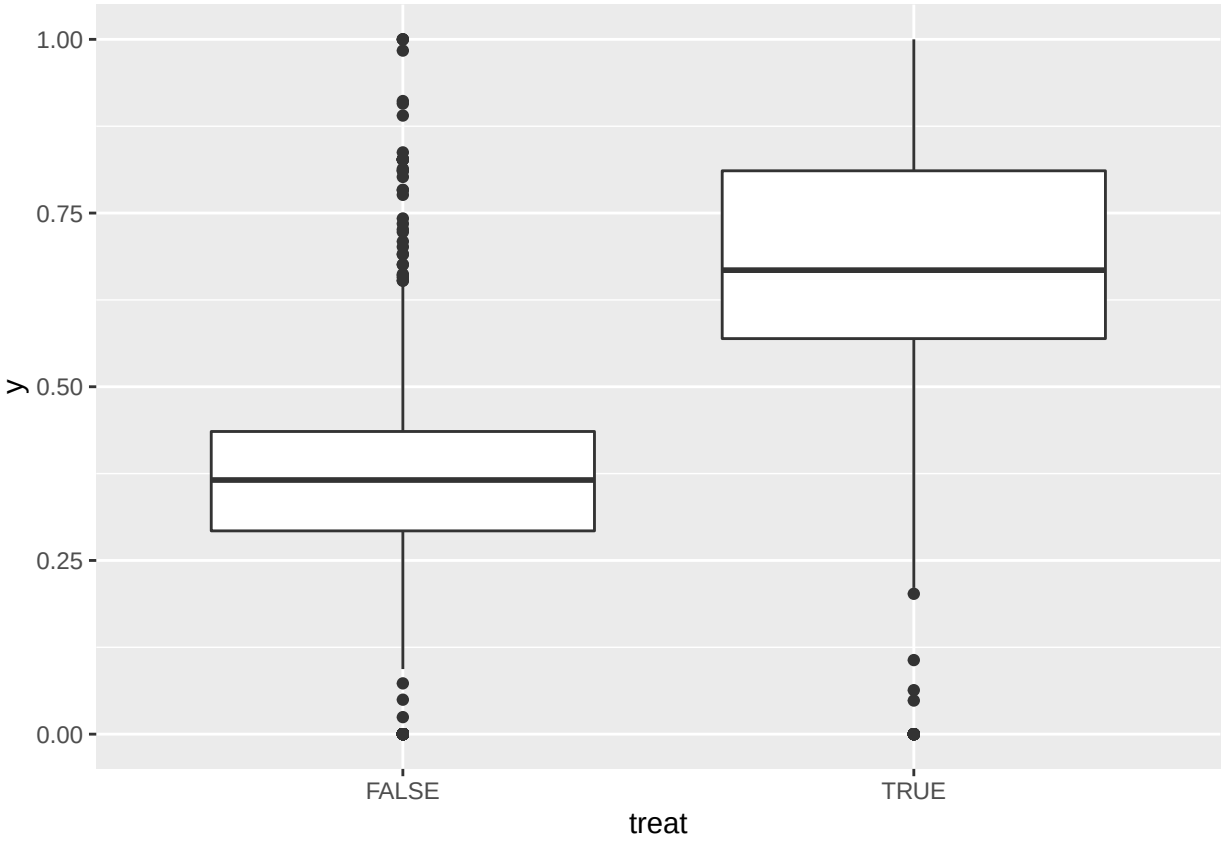
Figure 3: Marginal or the "jump" between margin of victory, at 50% threshold

```
setwd("~/MSBA 2020 All Files/Spring 2020/MSBA 6440 - Causal Inference via Ecnmtrcs Exprmnt/Week 8 - Reg

# Author: Gordon Burtch and Gautam Ray
# Course: MSBA 6440
# Session: Regression Discontinuity
# Topic: RDD Example
# Lecture 8

# Dataset used by Lee (2008), which is a paper that talks about the "incumbency advantage"
# in US politics. If I hold a congressional seat right now, to what degree does that increase
# my party's chances of winning re-election?

# The discontinuity design here is based on vote share in the *last* election. Essentially, I
# am "assigned" to incumbency: if I won the last election, and not if not. Whether I win an election
# is based on a simple majority threshold (given the two-party system).  Usually it means I passed 50%
# of the vote. So, we're going to use that 50% cutoff in the last election to estimate the effect of
# "just barely" winning last time, on winning this time (versus losing).

HouseData <-read.csv("house.csv")

# Let's define our treatment variable (0 = equal proportion of vote in last election)
HouseData$treat <- (HouseData$x>0)
# Your vote difference is greater than 0, you won and are the incumbent.
```

```r
# Let's run the endogenous regression first... says 35% increase in vote share due to winning last
# election! Of course we know that's wrong...
ols <- lm(data=HouseData,y~treat)
# y is a function of the vote share in the current election as part of the treated group, those
# that were incumbents.
summary(ols)
```

```
##
## Call:
## lm(formula = y ~ treat, data = HouseData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69788 -0.10061 -0.00360  0.09631  0.65348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.346522   0.003201  108.25   <2e-16 ***
## treatTRUE   0.351358   0.004195   83.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1676 on 6556 degrees of freedom
## Multiple R-squared:  0.5169, Adjusted R-squared:  0.5168
## F-statistic:  7014 on 1 and 6556 DF,  p-value: < 2.2e-16
```

```r
## 0.35; 35% greater chance of winning, but not strictly because of incumbency, this is also part
# due to better campagin, better funds.  We need to whittle this down.

# What is this OLS actually estimating? It's a t-test comparing vote outcomes in current election
# between incumbents and non-incumbents.
ggplot(data=HouseData) + geom_boxplot(aes(y=y,x=treat))
```

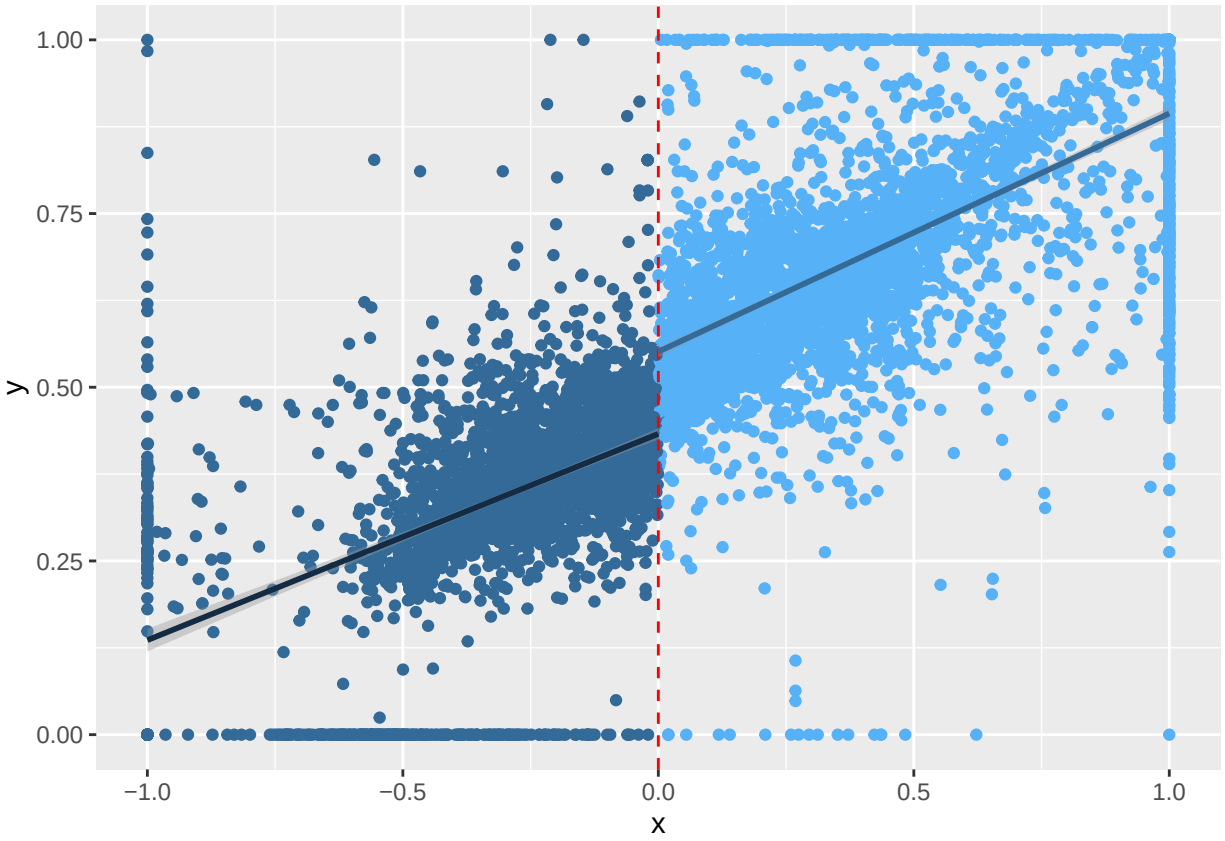



```
# Let's try RDD now, where we condition on the relationship between y and x, to get at the effect
# right around the threshold.
# By using "all" the data we are implicitly using a maximum bandwidth (use all of the range of x
# around c) This says the local treatment effect is 0.11 (11% increase in vote outcome due to the
# discontinuity).
ols <- lm(data=HouseData,y~treat + x) # Control for how  much did the previous incumbent win by in
# previous election.
summary(ols)
```

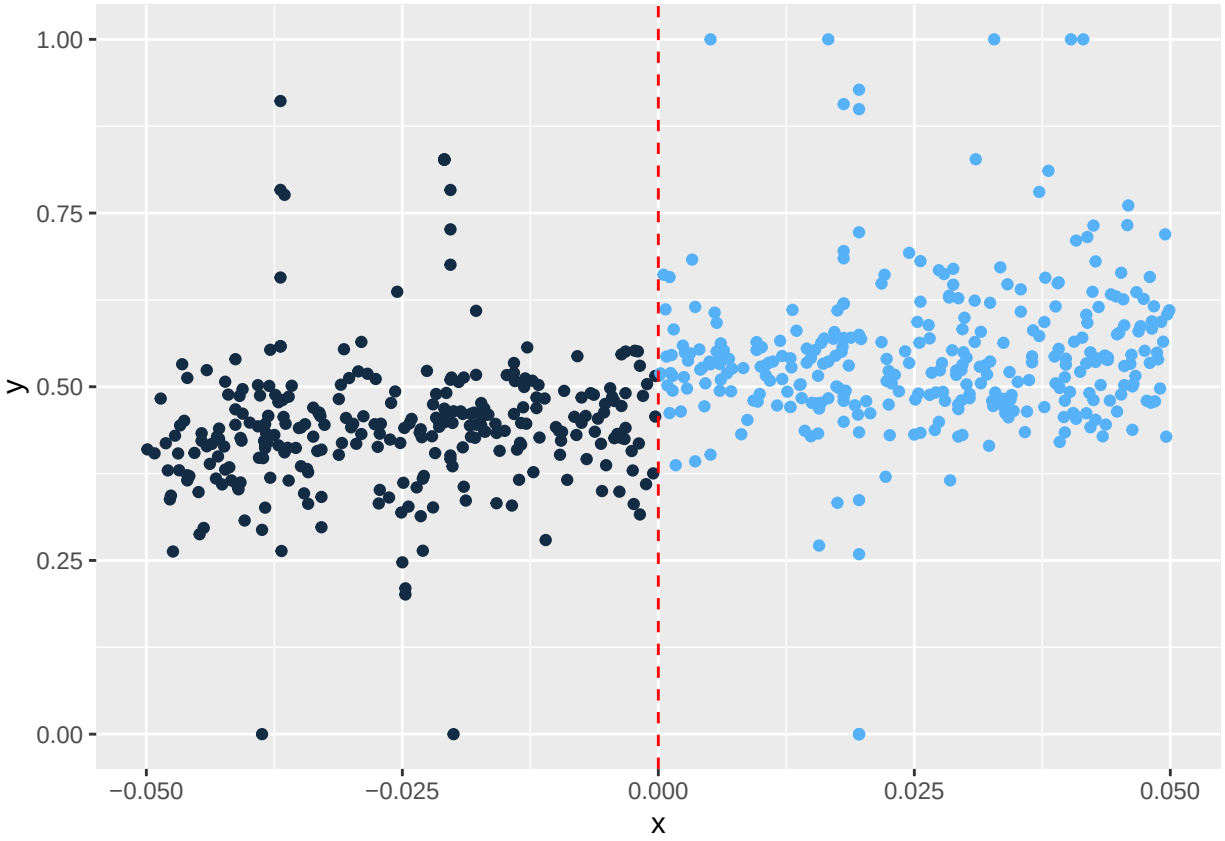
```
##
## Call:
## lm(formula = y ~ treat + x, data = HouseData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88700 -0.06324  0.00027  0.07082  0.88780
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.442736   0.003168  139.75   <2e-16 ***
## treatTRUE   0.113728   0.005528   20.57   <2e-16 ***
## x           0.330533   0.005989   55.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1385 on 6555 degrees of freedom
## Multiple R-squared:  0.6701, Adjusted R-squared:   0.67
## F-statistic:  6658 on 2 and 6555 DF,  p-value: < 2.2e-16
```

```
# 11% advantage to being an incumbent.

# Here's a plot of what we are estimating by running this regression.
ggplot(HouseData, aes(y=y,x=x)) + geom_point(aes(col=treat+1),show.legend = FALSE) + geom_vline(xinterc
  geom_smooth(aes(group=treat,col=as.numeric(treat)),method = "lm",show.legend=FALSE)
```



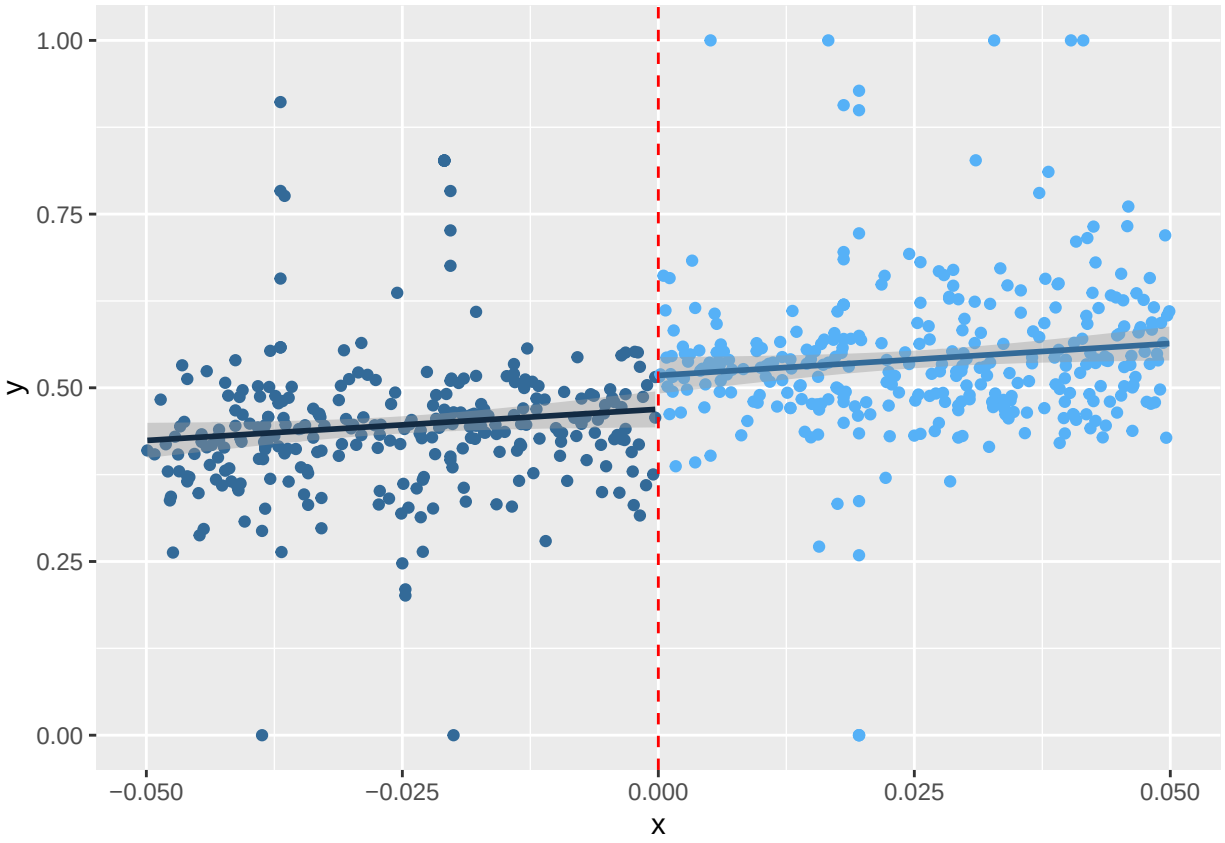

```
# But we don't really believe that incumbents and prior losers are comparable "generally",
# so we don't trust this bandwidth value,
# for the same reason we don't trust the OLS more generally...
# We *might* believe the comparison is fair right around the election win threshold, however.
# Here, we are zooming in to a 5% differential on either side of the cutoff, h = 0.05.
Pared_House <- HouseData[HouseData$x >= -0.05 & HouseData$x <= 0.05,] # narrow bandwidth to +/- 5%
ggplot(Pared_House, aes(y=y,x=x,col=as.numeric(treat))) + geom_point(show.legend = FALSE) + geom_vline(i
```




```
# Looks better...

# Okay let's run our RDD regression now. Our estimate falls to about 5% with this tighter bandwidth.
house_rdd <- lm(data=Pared_House,y ~ treat + x)
summary(house_rdd)
```

```
##
## Call:
## lm(formula = y ~ treat + x, data = Pared_House)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53590 -0.05496 -0.00756  0.03574  0.47729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.46942    0.01046  44.872  < 2e-16 ***
## treatTRUE    0.04865    0.01881   2.586  0.00995 **
## x            0.90988    0.31979   2.845  0.00459 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1113 on 607 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1625
## F-statistic: 60.09 on 2 and 607 DF,  p-value: < 2.2e-16
```

```
# Now incumbent only has 5% advantage compared to challenger.

ggplot(Pared_House, aes(y=y,x=x)) + geom_point(aes(col=treat+1),show.legend = FALSE) + geom_vline(xintei
  geom_smooth(aes(group=treat,col=as.numeric(treat)),method = "lm",show.legend=FALSE)
```




```
# Jump is 5% difference, which can be seen here.


# Lets try an interaction to see if the slopes are different
house_rdd_int <- lm(data=Pared_House,y ~ treat*x)
# We want to see if the slop AFTER the treatment effects is the same as the effect before,
# so we add an interaction term to see what happens.
summary(house_rdd_int)
```

```
##
## Call:
## lm(formula = y ~ treat * x, data = Pared_House)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53585 -0.05515 -0.00759  0.03566  0.47747
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.46914    0.01387  33.824   <2e-16 ***
```

```
## treatTRUE    0.04870    0.01890   2.577   0.0102 *
## x            0.89896    0.47951   1.875   0.0613 .
## treatTRUE:x  0.01970    0.64394   0.031   0.9756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1114 on 606 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1611
## F-statistic:    40 on 3 and 606 DF,  p-value: < 2.2e-16
```

```
# Slope to the left and right are the same


# Lets try a square term to see if there is any curvilinearity

Pared_House$x_sq <- Pared_House$x*Pared_House$x
house_rdd_sq <- lm(data=Pared_House,y ~ treat + x + x_sq)
summary(house_rdd_sq)
```

```
##
## Call:
## lm(formula = y ~ treat + x + x_sq, data = Pared_House)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53600 -0.05504 -0.00757  0.03580  0.47714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.46962    0.01216  38.631  < 2e-16 ***
## treatTRUE    0.04860    0.01890   2.571  0.01037 *
## x            0.91112    0.32233   2.827  0.00486 **
## x_sq        -0.20117    6.21873  -0.032  0.97420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1114 on 606 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1611
## F-statistic:    40 on 3 and 606 DF,  p-value: < 2.2e-16
```

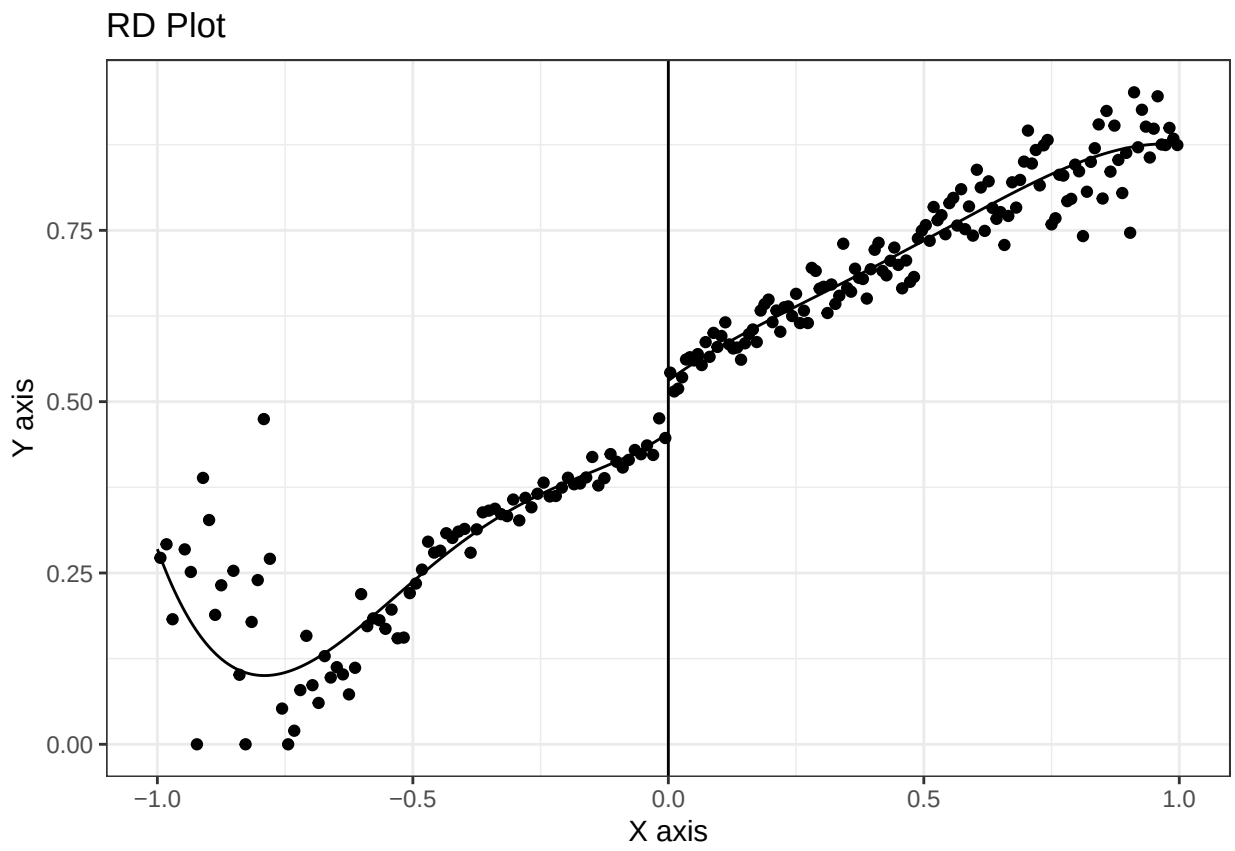
```
# Not significant.

# These days, we don't implement it all manually.
# We use packages that implement algorithms that choose bandwidth, specification and other things
# for us based on statistics... We probably want to use a weighting function, for example (further
# away from cutoff, we down-weight you) in tandem with the optimally chosen band-width, etc.
# rdrobust() chooses everything for you, based on some cross-validation, etc.
# This says that we are still over-doing it! A more accurate estimate of the effect is actually about
# just 6%.
House_Robust_RDD <- rdrobust(HouseData$y,HouseData$x,c=0)
```

```
## [1] "Mass points detected in the running variable."
```

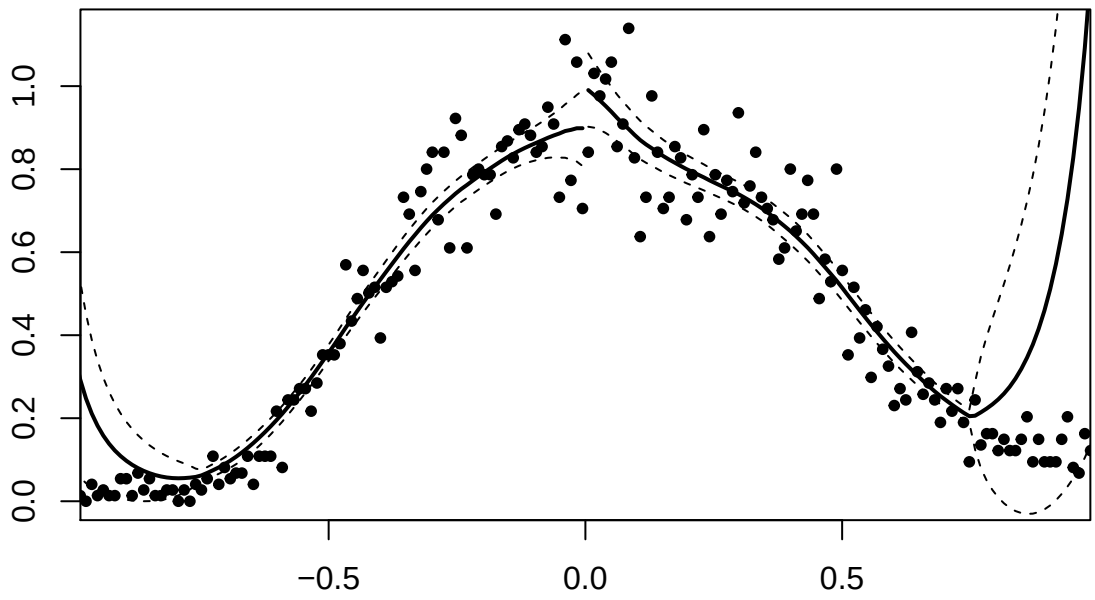
```r
summary(House_Robust_RDD)
```

```
## Call: rdrobust
##
## Number of Obs.                 6558
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                2740        3818
## Eff. Number of Obs.            786         816
## Order est. (p)                   1           1
## Order bias  (q)                  2           2
## BW est. (h)                  0.135       0.135
## BW bias (b)                  0.240       0.240
## rho (h/b)                    0.564       0.564
## Unique Obs.                   2108        2581
##
## ===============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ===============================================================================
##   Conventional     0.064     0.011     5.806     0.000     [0.042 , 0.085]
##         Robust         -         -     4.731     0.000     [0.035 , 0.084]
## ===============================================================================
```

```r
# 6% advantage for being an incumbent, as compared to results from earlier in the analysis.
rdplot(HouseData$y,HouseData$x)
```




```
# It can't "fix" self-selection, however, so let's again run a density check around the cut-point to
# evaluate self-selection ("sorting") the number it spits out is the p-value associated with the non-
# parametric test of density differences around the threshold.
# In this case, the p-value is ~0.19, which is fairly far away from being a problem (no evidence of sor
DCdensity(HouseData$x,0,plot=TRUE)
```




```
## [1] 0.1952357
```

```
# bins the data set, and centers it around cut off point to demonstrate data is unbalanced.
```

- We cannot reject the null hypothesis that the density is even on both sides of the cutoff which is 0.
- This is the p-value that comes out of the density test; since it is so low, there is likely no self selection happning in the data set.
- If it WAS significant, then there is sorting happening, and the assumption of regression discontuinity is FALSE.

How to Model / Estimate RDD

Model the Dummy (Above Threshold), as well as Distance from Threshold.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 (X_i - X_c) + e_i$$

- Here, Z is the threshold dummy (1 if above, 0 if below), X is the assignment variable, and X_c is the cutoff value of X that defines the threshold (Note: $X_i - X_c$ captures 'distance' from the cutoff).

- **Interpretation:** coefficient on Z captures mean difference in Y between treated and untreated when X_i - $X_c = 0$, i.e., right at the threshold!

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 (X_i - X_c) + e_i$$

Bandwidth Choice

Choosing What Data to Enter Into Your Regression

- We need to select a bandwidth, h, which is the range around the cutoff point of X that we consider. Lowering h yields a more trustworthy estimate of the treatment (less bias), but it comes at the cost of power (wider standard errors).

- **Good News:** Algorithms proposed in recent years for optimally selecting the bandwidth (and R implements them).

Assumptions / Requirements

1. The assignment rule needs to be arbitrary! For example, this means that it must not be something units actively 'manage' (self-selection).
   a) For example, a majority voting outcome *may* serve as a good regression discontinuity design. However, *not if candidates are stuffing the ballot box to win.*

2. The outcome variable (Y) must be a continuous, smooth function of X, particularly around the threshold, and that function must be specified correctly, e.g., linear, quadratic, cubic, ...

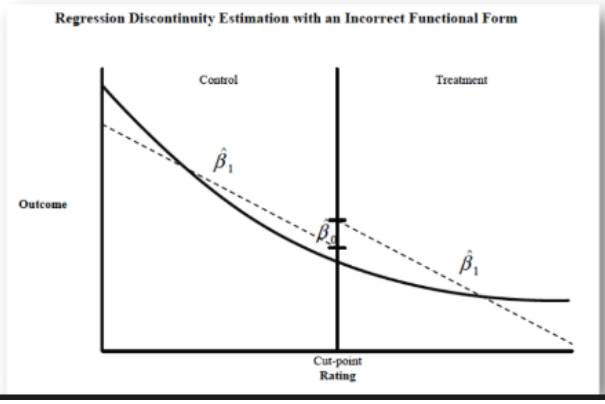
We can examine these assumptions visually, typically by creating a scatterplot of Y on X and fit trend lines around the cutoff.

* If the treatment effect is known then this is not valid; for example, if you know the cutoff for PSAT is an 80, and as a student you go for the lowest possible score required, then this student should **not** be included at the cutoff because you *could* have done better and chose not to.

When Assumptions are Violated

Model Misspecification:

2. **Recall that difference in means right at the cutoff is our treatment effect.** Model misspecification can influence our estimate of that treatment effect!



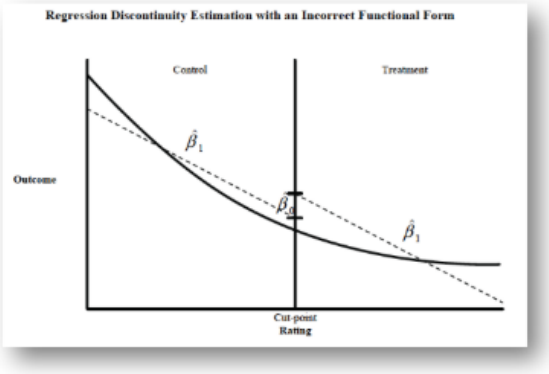
* Y is not linear here; y as a function of x **must** be continous.

When Assumptions are Violated

Model Misspecification:

- Modeled correctly, we would want to include non-linear terms here.

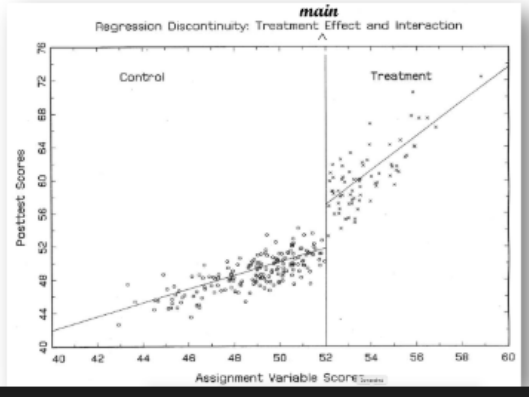
$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2(X_i - X_c) + \beta_3(X_i - X_c)^2 + \varepsilon_i$$



When Assumptions are Violated

Model Misspecification:

- It may also be that the effect of assignment depends on the value of X, generally! Notice how treatment not only causes a jump, it also causes the slope associated with $(X_i - X_c)$ to change?

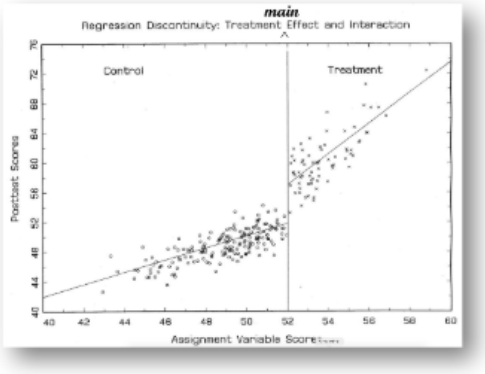


When Assumptions are Violated

Model Misspecification:

- To account for different slopes on the two sides of threshold, add an interaction term between Z and $(X-Xc)$.

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2(X_i - X_c) + \beta_3 Z_i (X_i - X_c) + \varepsilon_i$$



Sometimes RDDs are "Fuzzy" (FRDD)

When Assignment Rule is Loosely Applied:

- E.g., we make exceptions to the rule. This can still be used, the assumption required is that exceeding the cutoff probabilistically increases your chances of assignment to treatment.

- The idea here is quite similar to ITT in experiments (the policy says we intended to treat you and not you, but maybe we have defiers and compliers). The presence of "exceptions" will mute the treatment effect.

- ***With Fuzzy RDD, we use an indicator of exceeding threshold as an instrument for assignment to treatment.***

Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach – van der Klaauw (2002)

- Y: Acceptance of Admission Offers
- X: Score based on SAT, GPA
- D: Financial Aid Offer based on SAT, GPA, Class Rank, etc.
- Fuzzy Cut-off: Financial Aid Officers use discretion to offer aid based on essay quality, extracurricular activities, ethnicity, family income, quality of recommendation letters.

- Instrument Variable: X Score cut offs used to determine financial aid eligibility.