Final Exam Prep Sessions 9 - 10

MSBA 6440: Causal Inference via Experimentation

Danny Moncada (monca016)

April 27, 2020

Session 9: Selection Bias, Measurement Error (selection)

Motivating Example 1
Effect of Education on Women's Wages

$$y_i = x_i\beta + \epsilon_i$$

where y_i is women i's wage and x_i is education. The selection problem is the sample consists only of women who choose to work.

Selection equation for entering the labor market might be:

$$U_i = w_i\gamma + u_i$$

- Where U_i is the utility of women i entering the labor market and w_i is a vector of factors that influence a women's decision to work.
- There is a selection issue if u_i is correlated with ε_i. We don't observe U_i. We observe $Z_i = 1$, if the women enters the workforce.

Outcome (Y) Wages	Selection Variable (Z) Labor Force Participation (lfp)	Independent Variable (X) Education
w_1	1	x_1
w_2	1	x_2
0	0	x_3
0	0	x_4

Selection model:

$$Probit(lfp_i) = \gamma * x_i + v_i$$

Outcome model:

$$OLS(Wages_i) = \beta * x_i + \beta_\lambda * IMR_i + \epsilon_i$$

Heckman Model, Sample Selection Model

- Selection Model: $z_i^* = w_i\gamma + u_i$

$$z_i = \begin{cases} 1 & \text{if } z_i^* > 0 \\ 0 & \text{if } z_i^* \leq 0 \end{cases}$$

- Outcome Model: $y_i = \begin{cases} x_i\beta + \epsilon_i & \text{if } z_i^* > 0 \\ - & \text{if } z_i^* \leq 0 \end{cases}$

- Assumptions:

$$u_i \sim N(0,1)$$
$$\epsilon_i \sim N(0,\sigma^2)$$
$$\text{corr}(u_i, \epsilon_i) = \rho$$

Selection model: if z_i^* is greater than 0, then you will participate, otherwise you will not.

Outcome model: if your z_i^* is greater than 0, then we will observe you, otherwise we ill not.

Assumptions: There are two error terms, and they have a normal distribution/standard deviation.

Estimation: Heckman's Two-Step Procedure

- Step 1: Estimate the selection (probit) equation to estimate γ. For each observation in the selected sample, compute $\hat{\lambda}_i = \frac{\phi(w_i\hat{\gamma})}{\Phi(w_i\hat{\gamma})}$ (the inverse Mills ratio, IMR).
  - We need at least one variable that affects selection but does not influence the outcome, for identification purposes.

- Step 2: Estimate β and $\beta_\lambda = \rho\sigma_\epsilon$ by OLS of y on x and $\hat{\lambda}$.
  - If the coefficient of the IMR in the outcome equation is significant, there is selection bias.

- We need one variable that affects selection but does not influence the outcome. This is the instrument variable that affects the selection or influences the outcome only through the selection variable.
- If the coefficient of IMR is not significant, then you can argue that selection bias is not an issue.

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

MROZ <-read.csv("MROZ.csv")

MROZ$kids <- (MROZ$kidslt6 + MROZ$kidsge6) # Count number of total kids

# Female labor supply (lfp = labour force participation)

## Outcome equations without correcting for selection
# I() means "as-is" -- do calculation in parentheses then use as variable

## Comparison of linear regression and selection model

outcome1 <- lm(wage ~ educ, data = MROZ)
summary(outcome1)
```

```
##
## Call:
## lm(formula = wage ~ educ, data = MROZ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6797 -1.6658 -0.4556  0.8794 21.1487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.09237    0.84829  -2.467    0.014 *
## educ         0.49531    0.06595   7.511 3.49e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.114 on 426 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.1169, Adjusted R-squared:  0.1149
## F-statistic: 56.41 on 1 and 426 DF,  p-value: 3.486e-13
```

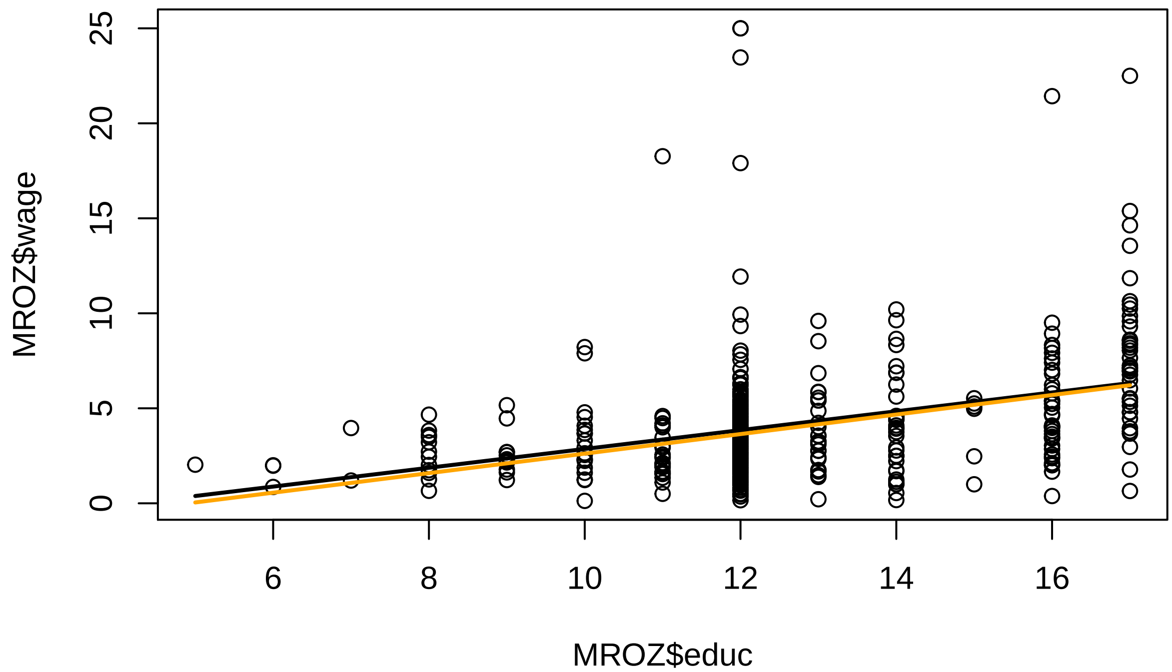
```r
# Education has significant relationship with wages

selection1 <- selection(selection = lfp ~ age + I(age^2) + faminc + kidslt6 + educ, # labor force part.
                        # Family chars MIGHT influence participation in labor force, but have NO affect
                        # wages EXCEPT for influencing their participation in labor force.
                        outcome = wage ~ educ,
                        data = MROZ, method = "2step") # 2 step Heckman

summary(selection1)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 753 observations (325 censored and 428 observed)
## 11 free parameters (df = 743)
## Probit selection equation:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -1.399e-01  1.514e+00  -0.092    0.926
## age         -1.174e-02  6.876e-02  -0.171    0.864
## I(age^2)    -2.567e-04  7.808e-04  -0.329    0.742
## faminc       3.233e-06  4.297e-06   0.752    0.452
## kidslt6     -8.531e-01  1.144e-01  -7.457 2.47e-13 ***
## educ         1.166e-01  2.365e-02   4.931 1.01e-06 ***
## Outcome equation:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.52489    1.30609  -1.933   0.0536 .
## educ         0.51403    0.07869   6.532  1.2e-10 ***
## Multiple R-Squared:0.1173,   Adjusted R-Squared:0.1132
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio   0.3149     0.7235   0.435    0.663
## sigma           3.1151         NA      NA       NA
## rho             0.1011         NA      NA       NA
## --------------------------------------------
```

- Predict if someone will participate in labor force; here, education and whether they have kids **does** affect their participation.
- Now our coefficients increase and become more significant
- invMillsRatio is not significant; therefore, selection bias is not statistically significant and thus not a problem.

```r
plot(MROZ$wage ~ MROZ$educ)
curve(outcome1$coeff[1] + outcome1$coeff[2]*x, col="black", lwd="2", add=TRUE) # OLS regression
curve(selection1$coeff[7] + selection1$coeff[8]*x, col="orange", lwd="2", add=TRUE) # Heckman model
```




```r
## A more complete model comparison

outcome2 <- lm(wage ~ exper + I( exper^2 ) + educ + city, data = MROZ)
summary(outcome2)
```

```
##
## Call:
## lm(formula = wage ~ exper + I(exper^2) + educ + city, data = MROZ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6021 -1.6012 -0.4787  0.8950 21.2762
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.5609920  0.9288390  -2.757  0.00608 **
## exper        0.0324982  0.0615864   0.528  0.59800
## I(exper^2)  -0.0002602  0.0018378  -0.142  0.88747
## educ         0.4809623  0.0668679   7.193 2.91e-12 ***
## city         0.4492741  0.3177735   1.414  0.15815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.111 on 423 degrees of freedom
##   (325 observations deleted due to missingness)
## Multiple R-squared:  0.1248, Adjusted R-squared:  0.1165
```

```
## F-statistic: 15.08 on 4 and 423 DF,  p-value: 1.569e-11
```

```r
## Correcting for selection

selection.twostep2 <- selection(selection = lfp ~ age + I(age^2) + faminc + kidslt6 + educ,
                                outcome = wage ~ exper + I(exper^2) + educ + city,
                                data = MROZ, method = "2step")
summary(selection.twostep2)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 753 observations (325 censored and 428 observed)
## 14 free parameters (df = 740)
## Probit selection equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.399e-01  1.514e+00  -0.092    0.926
## age         -1.174e-02  6.876e-02  -0.171    0.864
## I(age^2)    -2.567e-04  7.808e-04  -0.329    0.742
## faminc       3.233e-06  4.297e-06   0.752    0.452
## kidslt6     -8.531e-01  1.144e-01  -7.457 2.48e-13 ***
## educ         1.166e-01  2.365e-02   4.931 1.01e-06 ***
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.7413454  1.3679742  -2.004   0.0454 *
## exper        0.0334859  0.0614715   0.545   0.5861
## I(exper^2)  -0.0003096  0.0018477  -0.168   0.8670
## educ         0.4887549  0.0795133   6.147 1.29e-09 ***
## city         0.4467138  0.3162288   1.413   0.1582
## Multiple R-Squared:0.1248,   Adjusted R-Squared:0.1145
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio  0.13220    0.73970   0.179    0.858
## sigma          3.09469         NA      NA       NA
## rho            0.04272         NA      NA       NA
## --------------------------------------------
```

```r
# Still not significant!

selection.mle <- selection(selection = lfp ~ age + I(age^2) + faminc + kids + educ,
                           outcome = wage ~ exper + I(exper^2) + educ + city,
                           data = MROZ, method = "mle") # Maximum likelihood estimation
summary(selection.mle)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 3 iterations
## Return code 2: successive function values within tolerance limit
## Log-Likelihood: -1579.498
## 753 observations (325 censored and 428 observed)
## 13 free parameters (df = 740)
## Probit selection equation:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.709e+00  1.399e+00  -2.652 0.008183 **
## age          1.649e-01  6.484e-02   2.543 0.011182 *
## I(age^2)    -2.189e-03  7.541e-04  -2.903 0.003808 **
## faminc       4.581e-06  4.525e-06   1.012 0.311667
## kids        -1.507e-01  3.830e-02  -3.935  9.1e-05 ***
## educ         9.061e-02  2.341e-02   3.870 0.000118 ***
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.2332665  1.3302676  -1.679   0.0936 .
## exper        0.0291691  0.0620275   0.470   0.6383
## I(exper^2)  -0.0001513  0.0018553  -0.082   0.9350
## educ         0.4679380  0.0766012   6.109 1.62e-09 ***
## city         0.4467800  0.3160013   1.414   0.1578
##    Error terms:
##       Estimate Std. Error t value Pr(>|t|)
## sigma  3.09755    0.10907  28.400   <2e-16 ***
## rho   -0.07081    0.20547  -0.345     0.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

```r
## Heckman model selection "by hand" ##

seleqn1 <- glm(lfp ~ age + I(age^2) + faminc + kidslt6 + educ, family=binomial(link="probit"),
               data=MROZ)
summary(seleqn1)
```

```
##
## Call:
## glm(formula = lfp ~ age + I(age^2) + faminc + kidslt6 + educ,
##     family = binomial(link = "probit"), data = MROZ)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0359  -1.1386   0.6860   0.9789   2.1831
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.399e-01  1.507e+00  -0.093    0.926
## age         -1.174e-02  6.852e-02  -0.171    0.864
## I(age^2)    -2.567e-04  7.784e-04  -0.330    0.742
## faminc       3.233e-06  4.353e-06   0.743    0.458
## kidslt6     -8.531e-01  1.149e-01  -7.425 1.13e-13 ***
## educ         1.166e-01  2.367e-02   4.926 8.38e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  931.42  on 747  degrees of freedom
## AIC: 943.42
##
```

```
## Number of Fisher Scoring iterations: 4
```

```r
## Calculate inverse Mills ratio by hand ##

MROZ$IMR <- dnorm(seleqn1$linear.predictors)/pnorm(seleqn1$linear.predictors)

## Outcome equation correcting for selection ##

outeqn1 <- lm(wage ~ exper + I(exper^2) + educ + city + IMR, data=MROZ, subset=(lfp==1))
summary(outeqn1)
```

```
##
## Call:
## lm(formula = wage ~ exper + I(exper^2) + educ + city + IMR, data = MROZ,
##     subset = (lfp == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6074 -1.6048 -0.4736  0.8876 21.2940
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.7413490  1.3773664  -1.990   0.0472 *
## exper        0.0334859  0.0619076   0.541   0.5889
## I(exper^2)  -0.0003096  0.0018608  -0.166   0.8679
## educ         0.4887551  0.0800561   6.105 2.33e-09 ***
## city         0.4467137  0.3184647   1.403   0.1614
## IMR          0.1322070  0.7448157   0.178   0.8592
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.115 on 422 degrees of freedom
## Multiple R-squared:  0.1248, Adjusted R-squared:  0.1145
## F-statistic: 12.04 on 5 and 422 DF,  p-value: 6.495e-11
```

```r
## compare to selection package -- coefficients right, se's wrong

summary(selection.twostep2)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 753 observations (325 censored and 428 observed)
## 14 free parameters (df = 740)
## Probit selection equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.399e-01  1.514e+00  -0.092    0.926
## age         -1.174e-02  6.876e-02  -0.171    0.864
## I(age^2)    -2.567e-04  7.808e-04  -0.329    0.742
## faminc       3.233e-06  4.297e-06   0.752    0.452
## kidslt6     -8.531e-01  1.144e-01  -7.457 2.48e-13 ***
## educ         1.166e-01  2.365e-02   4.931 1.01e-06 ***
## Outcome equation:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.7413454  1.3679742  -2.004   0.0454 *
## exper        0.0334859  0.0614715   0.545   0.5861
## I(exper^2)  -0.0003096  0.0018477  -0.168   0.8670
## educ         0.4887549  0.0795133   6.147 1.29e-09 ***
## city         0.4467138  0.3162288   1.413   0.1582
## Multiple R-Squared:0.1248,   Adjusted R-Squared:0.1145
##    Error terms:
##                Estimate Std. Error t value Pr(>|t|)
## invMillsRatio  0.13220    0.73970   0.179    0.858
## sigma          3.09469         NA      NA       NA
## rho            0.04272         NA      NA       NA
## --------------------------------------------
```

```r
stargazer(outeqn1,selection.twostep2,type="text",title="Heckman Two-step vs.Heckman by Hand",
          column.labels = c("Heckman By Hand","Heckman Command"))
```

```
##
## Heckman Two-step vs.Heckman by Hand
## ===========================================================
##                               Dependent variable:
##                       -------------------------------------
##                                      wage
##                              OLS              selection
##                        Heckman By Hand     Heckman Command
##                              (1)                (2)
## -----------------------------------------------------------
## exper                      0.033              0.033
##                           (0.062)            (0.061)
##
## I(exper2)                 -0.0003            -0.0003
##                           (0.002)            (0.002)
##
## educ                      0.489***           0.489***
##                           (0.080)            (0.080)
##
## city                       0.447              0.447
##                           (0.318)            (0.316)
##
## IMR                        0.132
##                           (0.745)
##
## Constant                  -2.741**           -2.741**
##                           (1.377)            (1.368)
##
## -----------------------------------------------------------
## Observations                428                753
## R2                         0.125
## Adjusted R2                0.114
## rho                                           0.043
## Inverse Mills Ratio                       0.132 (0.740)
## Residual Std. Error    3.115 (df = 422)
## F Statistic         12.040*** (df = 5; 422)
## ===========================================================
```

```
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

Classical Measurement Error in Independent Variable

- $B_{Yt\,X} = COV(Y_t, X)/\ VAR(X) \ < B_{YtXt} = COV(Y_t, X_t)/\ VAR(X_t)$
- Thus, measurement error in the independent variable produces a downward bias in the bivariate regression coefficient.
- The slope is attenuated by the reliability of the measure of the independent variable.
- Reliability of X = $VAR\ (X_t)\ /\ VAR\ (X)$

Classical Measurement Error in Dependent Variable

- $[B_{YXt} = COV(Y, X_t)/\ VAR(X_t)] \ = \ [B_{YtXt} = COV(Y_t, X_t)/\ VAR(X_t)]$
- Thus, random error in the dependent variable does not bias the slope coefficient.
- However, the standard error of the slope coefficient goes up. As the variance in Y goes up, R^2 goes down and the standard error of the slope coefficient goes up.
- When there are more than one independent variable, random measurement error can cause the coefficients to be biased upward or downward.

Classical Measurement Error

- Get better data. Get multiple indicators and check reliability.
- Given that measurement error in the dependent variables is more innocuous (does not bias the slope coefficient), we can run the reverse regression.
- The inverse of the slope of the reverse regression (g) and the slope of the standard regression (b) will bracket the true estimate. The bracketing result extends to multiple regression.
- $b/g = R^2$, so if R^2 is high b and g will be close.

- Reverse regression: use Y as your independent variable, and the slope of the reverse regression/linear regression will BRACKET the true relationship.
- The tru estimate will fall between b/g.

Classical Measurement Error

- Instrument variables can address measurement error, if instruments are correlated with X_t but not the measurement error.
- However, weak instruments will make the mis-measurement problem worse.
- If independent and dependent variables are mis-measured, the regression coefficient is biased downward, and the standard error is increased.

* There is no systematic bias in our measure; if this is true, then whatever treatment effect we are getting is understated because we are measuring X with error.

```
# Author: Gordon Burtch and Gautam Ray
# Course: MSBA 6440
# Session: Selection and Measurement Error
# Topic: Measurement Error
```

```r
# X and Y have classical measurement error. The true value are Xt and Yt, but they are meas-
# ured with error.
# X is measured as true X (Xt) plus error (ex)
# Y is measured as true Y (Yt) plus error (ey)
# The mean of Yt, Xt, ey and ex is (10, 7, 0, 0)
# The standard deviation of Yt, Xt, ey, and ex is (4, 8, 3, 6)
# The correlation of Yt and Xt is 0.7; Yt and Xt are uncorrelated with ey and ex; and ey and
# ex are uncorrelated with each other.

set.seed(1234)


Yt_Xt_ey_ex <- (mvrnorm(10000, c(10, 7, 0, 0), matrix(c(16, 22.4, 0.0, 0.0, 22.4, 64,0, 0, 0,
                                                         0, 9,0, 0, 0, 0, 36), ncol = 4)))
Yt <- Yt_Xt_ey_ex[,1]
Xt <- Yt_Xt_ey_ex[,2]
ey <- Yt_Xt_ey_ex[,3]
ex <- Yt_Xt_ey_ex[,4]

Y <- Yt + ey
X <- Xt + ex


# Check everything worked as expected.

cov(Yt_Xt_ey_ex) # co-variance table
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,] 15.8500152 21.7983191 -0.1185045  0.3525733
## [2,] 21.7983191 62.4120430 -0.3538506  0.4489615
## [3,] -0.1185045 -0.3538506  8.9843536 -0.0209746
## [4,]  0.3525733  0.4489615 -0.0209746 36.4322839
```

```r
cor(Yt_Xt_ey_ex)
```

```
##              [,1]         [,2]         [,3]         [,4]
## [1,]  1.000000000  0.693064996 -0.009930629  0.014672070
## [2,]  0.693064996  1.000000000 -0.014943156  0.009415246
## [3,] -0.009930629 -0.014943156  1.000000000 -0.001159330
## [4,]  0.014672070  0.009415246 -0.001159330  1.000000000
```

```r
sd(ey)
```

```
## [1] 2.997391
```

```r
sd(ex)
```

```
## [1] 6.035916
```

```r
sd(Yt)
```

```
## [1] 3.981208
```

```r
sd(Xt)
```

```
## [1] 7.900129
```

```r
mean (Yt)
```

```
## [1] 10.04655
```

```r
mean(Xt)
```

```
## [1] 7.037755
```

```r
mean(ey)
```

```
## [1] 0.01225815
```

```r
mean(ex)
```

```
## [1] -0.04648214
```

```r
#1. Measurement error in X, underestimates the effect of X on Y. The reliability of mismea-
# surement is the magnitude of the mismeasurement.

summary(lm(Yt~Xt)) # true regression
```

```
##
## Call:
## lm(formula = Yt ~ Xt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0613  -1.9560   0.0189   1.9343  10.9935
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.588506   0.038439  197.42   <2e-16 ***
## Xt          0.349265   0.003633   96.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.87 on 9998 degrees of freedom
## Multiple R-squared:  0.4803, Adjusted R-squared:  0.4803
## F-statistic:  9241 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
# 0.349 is true coefficient

summary(lm(Yt~X)) # faulty regression
```

```
##
## Call:
## lm(formula = Yt ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0113  -2.2426  -0.0154   2.2387  13.9810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.493914   0.040360  210.45   <2e-16 ***
## X           0.222081   0.003311   67.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.306 on 9998 degrees of freedom
## Multiple R-squared:  0.3104, Adjusted R-squared:  0.3103
## F-statistic:  4500 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
# 0.22 is understated coefficient

Reliability <- var(Xt)/var(X)

Reliability
```

```
## [1] 0.6257333
```

```r
summary(lm(Yt~X))$coefficients[2,1]/ summary(lm(Yt~Xt))$coefficients[2,1]
```

```
## [1] 0.6358541
```

```r
# The ratio of the coefficient from faulty regression compared to the true regression is
# around the same as the reliability number.


#2. Measurement error in Y, does not influence the coefficient of X, but exaggerartes the
# standard error of the regression coefficient.

summary(lm(Yt~Xt))
```

```
##
## Call:
## lm(formula = Yt ~ Xt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0613  -1.9560   0.0189   1.9343  10.9935
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.588506   0.038439  197.42   <2e-16 ***
## Xt          0.349265   0.003633   96.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.87 on 9998 degrees of freedom
## Multiple R-squared:  0.4803, Adjusted R-squared:  0.4803
## F-statistic:  9241 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
summary(lm(Y~Xt))
```

```
##
## Call:
## lm(formula = Y ~ Xt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4804  -2.7805   0.0196   2.7619  16.3189
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.640666   0.055594  137.44   <2e-16 ***
## Xt          0.343595   0.005255   65.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.151 on 9998 degrees of freedom
## Multiple R-squared:  0.2996, Adjusted R-squared:  0.2995
## F-statistic:  4276 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
# 0.005 standard error is higher

#3. Given that the measurement error in Y is more innocuous than the measurement error in
# X, we might run the reverse regression.
# The coefficient of the regular regression and the inverse of the coefficient of the rev-
# erse regression, bracket the true coefficient.

regular_reg <- (lm(Yt~X))
b <- summary(regular_reg)$coefficients[2,1]


reverse_reg <- (lm(X~Yt))
reverse_reg_coeff <- summary(reverse_reg)$coefficients[2,1]

g <- 1/(summary(reverse_reg)$coefficients[2,1])

b/g # bracket the true estimate
```

```
## [1] 0.3103656
```

```r
# 0.31 is R-squared, bracketing result extends to multiple regression

summary((lm(Yt~X)))
```

```
##
## Call:
## lm(formula = Yt ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0113  -2.2426  -0.0154   2.2387  13.9810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.493914   0.040360  210.45   <2e-16 ***
## X           0.222081   0.003311   67.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.306 on 9998 degrees of freedom
## Multiple R-squared:  0.3104, Adjusted R-squared:  0.3103
## F-statistic:  4500 on 1 and 9998 DF,  p-value: < 2.2e-16
```

```r
#4. If there is a good instrument for Xt, then the true estimate can be recovered.
# Lets say that Z is a good instrument for Xt. Like Xt, Z has a mean of 7 and a sd of 8. Z
# has a correlation of 0.5 with Xt, and Z is uncorrelated with ey and ex.
# Z has a correlation of 0.35 with Yt which is the product of 0.7 and 0.5 i.e, the correl-
# ation between Yt and Xt and the correlation between Xt and Z.


Yt_Xt_ey_ex_Z <- (mvrnorm(10000, c(10, 7, 0, 0, 7), matrix(c(16, 22.4, 0.0, 0.0, 11.2, 22.4,
                    64,0, 0, 32, 0, 0, 9,0, 0, 0, 0, 0, 36, 0, 11.2, 32, 0, 0, 64), ncol = 5)))

Yt <- Yt_Xt_ey_ex_Z[,1]
Xt <- Yt_Xt_ey_ex_Z[,2]
ey <- Yt_Xt_ey_ex_Z[,3]
ex <- Yt_Xt_ey_ex_Z[,4]
Z <- Yt_Xt_ey_ex_Z[,5]

Y <- Yt + ey
X <- Xt + ex

cov(Yt_Xt_ey_ex_Z)
```

```
##            [,1]        [,2]        [,3]        [,4]       [,5]
## [1,] 16.0047135 22.63816128 -0.00282960 -0.37864749 11.2970827
## [2,] 22.6381613 64.48127066  0.02019033 -1.05799144 32.3030631
## [3,] -0.0028296  0.02019033  8.92867464 -0.08615143 -0.1550035
## [4,] -0.3786475 -1.05799144 -0.08615143 35.60204905 -0.9907565
## [5,] 11.2970827 32.30306312 -0.15500346 -0.99075646 64.8426342
```

```r
cor(Xt, Z)
```

```
## [1] 0.4995703
```

```r
cor(X, Z)
```

```
## [1] 0.3928657
```

```r
cor(Yt, Z)
```

```
## [1] 0.3506808
```

```r
cor(Y, Z)
```

```
## [1] 0.277137
```

```r
ols_true <- lm((Yt~Xt)) # true regression which should give us correctly exact estimate

ivreg1 <- ivreg(formula=Yt ~ X | Z)

ivreg2 <- ivreg(formula=Y ~ X | Z)

stargazer(ols_true,ivreg1, ivreg2,type="text",title="True vs.Instrumented",
          column.labels = c("True","IV1", "IV2"))
```

```
##
## True vs.Instrumented
## ===================================================================================
##                                                 Dependent variable:
##                                    ------------------------------------------------
##                                          Yt ~ Xt             Yt           Y
##                                            OLS        instrumental instrumental
##                                                          variable     variable
##                                           True             IV1          IV2
##                                            (1)             (2)          (3)
## -----------------------------------------------------------------------------------
## Xt                                      0.351***
##                                          (0.004)
##
## X                                                     0.361***     0.356***
##                                                        (0.009)      (0.012)
##
## Constant                                7.567***      7.516***     7.548***
##                                          (0.038)       (0.073)      (0.095)
##
## -----------------------------------------------------------------------------------
## Observations                             10,000         10,000       10,000
## R2                                        0.497          0.207        0.136
## Adjusted R2                               0.497          0.207        0.136
## Residual Std. Error (df = 9998)           2.839          3.563        4.641
## F Statistic                    9,862.679*** (df = 1; 9998)
## ===================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

A good instrument can solve measurement problems.

Session 10: Synthetic Control Method (synth)

An Original Application

- California Tobacco Law of 1988
  - ADH (2010) wanted to estimate the total effect of Proposition 99 on state-wide smoking rates.
  - Prop 99 introduced a $0.25 tax per cigarette pack, with revenue to be reinvested in anti-smoking education and healthcare initiatives.
  - This prop drove many follow-up local laws around banned smoking in restaurants, etc.
  - Goal was to estimate the total effect of these developments on smoking rates *in California*.

Original Application

- The authors stress the difficulty in their setting of identifying *other* states that could serve as a suitable control for California (among the available states that did not implement a similar policy around the same time, N = 38 states). *Parallel trends assumption is violated!*

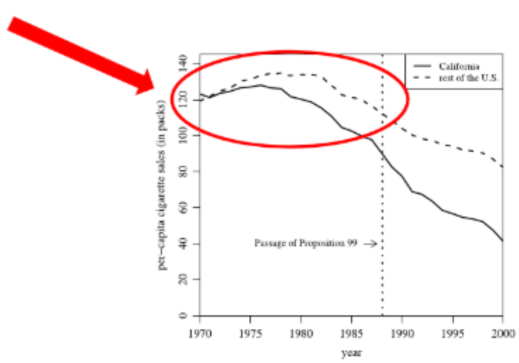


Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States.

* Dashed line is 38 states that didn't pass the law, form our control.

They Synthesize a Counterfactual

- Using pre-period data from other states, build a model that assigns weights to each control state, and arrives at a weighted average that closely resembles California smoking activity before the law was changed.

- Use the resulting model to synthesize what California *would have looked like* in post period (absent treatment).

- A feature of this approach is that you end up recovering weights which indicate how (dis)similar a given control unit is to the treated unit, in the pre period.

* Build a convex roll-up of states that look like CA in the pre period (a synthetic state).

Synthetic Control

Table 1. Cigarette sales predictor means

Variables	California Real	California Synthetic	Average of 38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.

* All of these are predictors of smoking rate

- Synthetic control looks closer to CA than average of 38 states

Synthetic Control

Journal of the American Statistical Association, June 2010

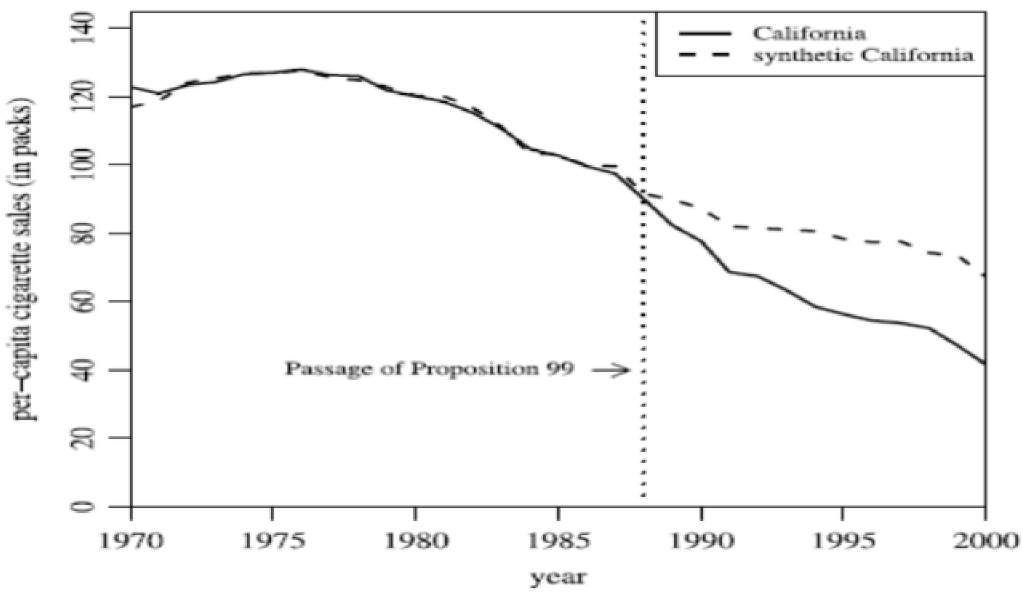


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

* Per capita sales

- We can use the dashed line to create a counter factual and determine what would have happened to CA if they didn't pass the law

Permutation Methods to compute "standard errors"

- Whether the effect estimated by the synthetic control for the unit affected by the intervention is large relative to the distribution of the effects estimated for the control units not exposed to the intervention.

- Iteratively apply the synthetic method to each state in the control pool and obtain a distribution of placebo effects. Compare the gap for California to the distribution of the placebo gaps.

- If the placebo studies create gaps of magnitude similar to the one estimated for California, then the analysis does not provide significant evidence of a negative effect of Proposition 99 on cigarette sales in California.

* The CA estimate should be **higher** than for the other states

- If the placebo shows that they are the same, then our analysis is no good - there is no effect nor is it statistically significant

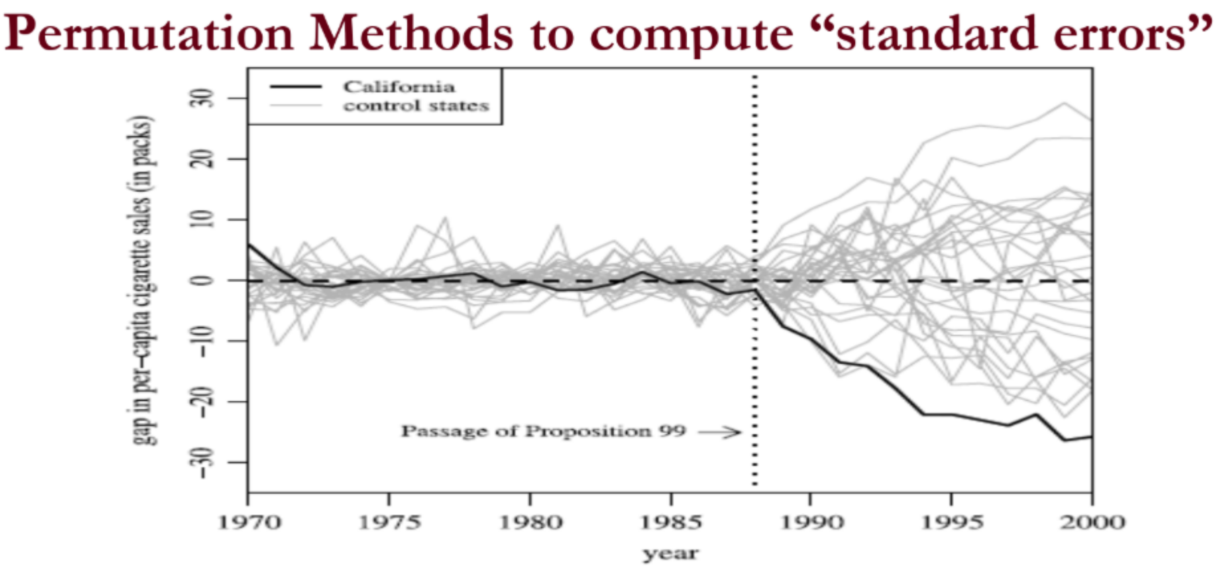


Figure 6. Per-capita cigarette sales gaps in California and placebo gaps in 29 control states (discards states with pre-Proposition 99 MSPE five times higher than California’s).

* CA is an outlier from any other state that did not have the treatment

```r
setwd("~/MSBA 2020 All Files/Spring 2020/MSBA 6440 - Causal Inference via Ecnmtrcs Exprmnt/Week 10 - Sy

#Change the read-in line to wherever your saved version of the fracking data csv file lives
#Note: your panel unit 'names' variable must be a character / string, not a factor, or it won't work.
fracking.data = read.csv("fracking.csv",stringsAsFactors=FALSE)
head(fracking.data)
```

```
##   id panel.id year     state        Y res.share  edu pop.dense
## 1  1        2 1990 Wisconsin 21.96667        NA 21.5        NA
## 2  2        2 1991 Wisconsin 21.45000 0.2534060 24.1        NA
## 3  3        2 1992 Wisconsin 21.23333 0.2512521 23.8        NA
## 4  4        2 1993 Wisconsin 21.33333 0.2489048 21.6        NA
## 5  5        2 1994 Wisconsin 20.51667 0.2462865 23.9    85.025
## 6  6        2 1995 Wisconsin 20.96667 0.2434099 22.9    85.975
```

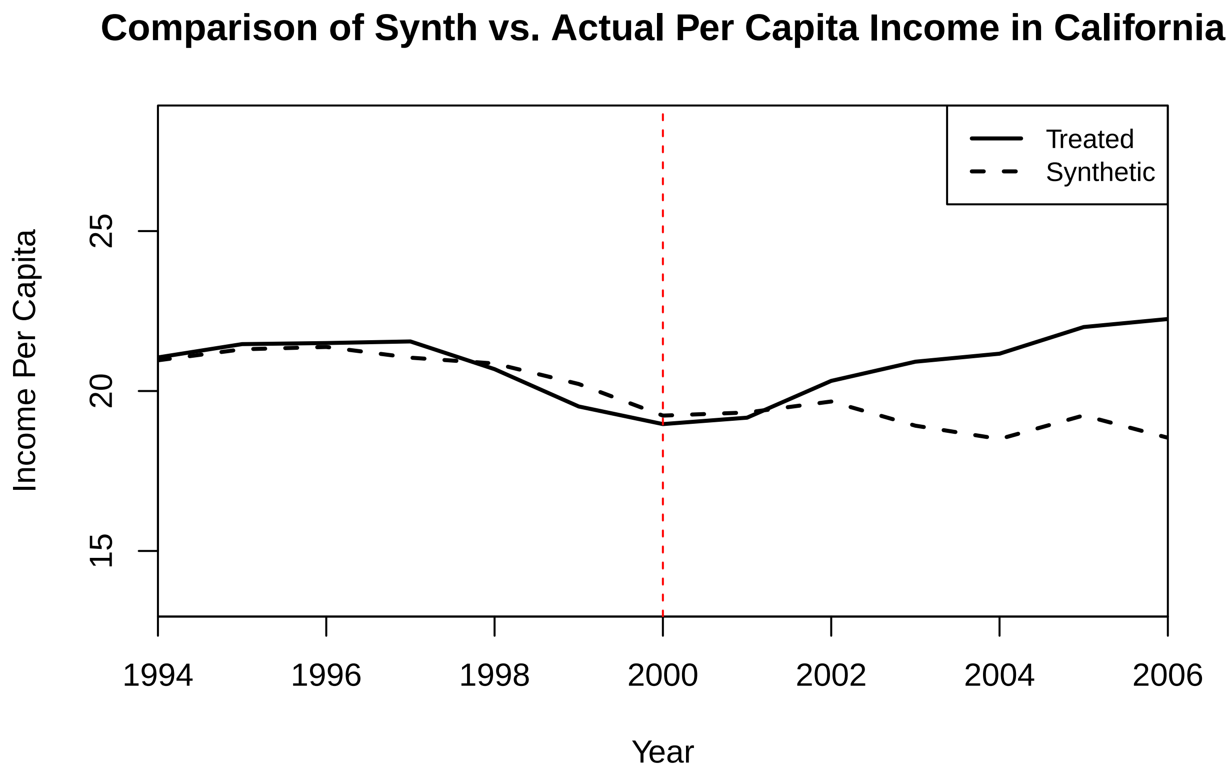
```r
fracking.data$treated = (fracking.data$state=="California")
ggplot(fracking.data, aes(x=year,y=Y,group=state)) +
  geom_line(aes(color=treated,linetype=!(treated))) +
  geom_vline(xintercept=2000,linetype="dashed")
```


```r
#Let's drop the ID column.
fracking.data = fracking.data[,-c(1)]

# your outcome variable *must* be named Y for Synth to accept it (bad coding practices in
# here I suspect)
dataprep.out=
   dataprep(foo = fracking.data,
   predictors = c("res.share", "edu", "pop.dense"),
   predictors.op = "mean",
   dependent = "Y",
   unit.variable = "panel.id",
   time.variable = "year",

   #Any pre-period X's we want to include using different aggregation function, other than
   # mean, or different time windows, specific years vs. all years, we enter here.
   special.predictors = list(list("Y", 1999, "mean"),list("Y", 1995, "mean"),list("Y", 1990, "mean")),

   #which panel is treated?
   treatment.identifier = 7,

   #which panels are we using to construct the synthetic control?
   controls.identifier = c(29, 2, 13, 17, 32, 38),

   #what is the pre-treatment time period?
   time.predictors.prior = c(1994:1999),
```

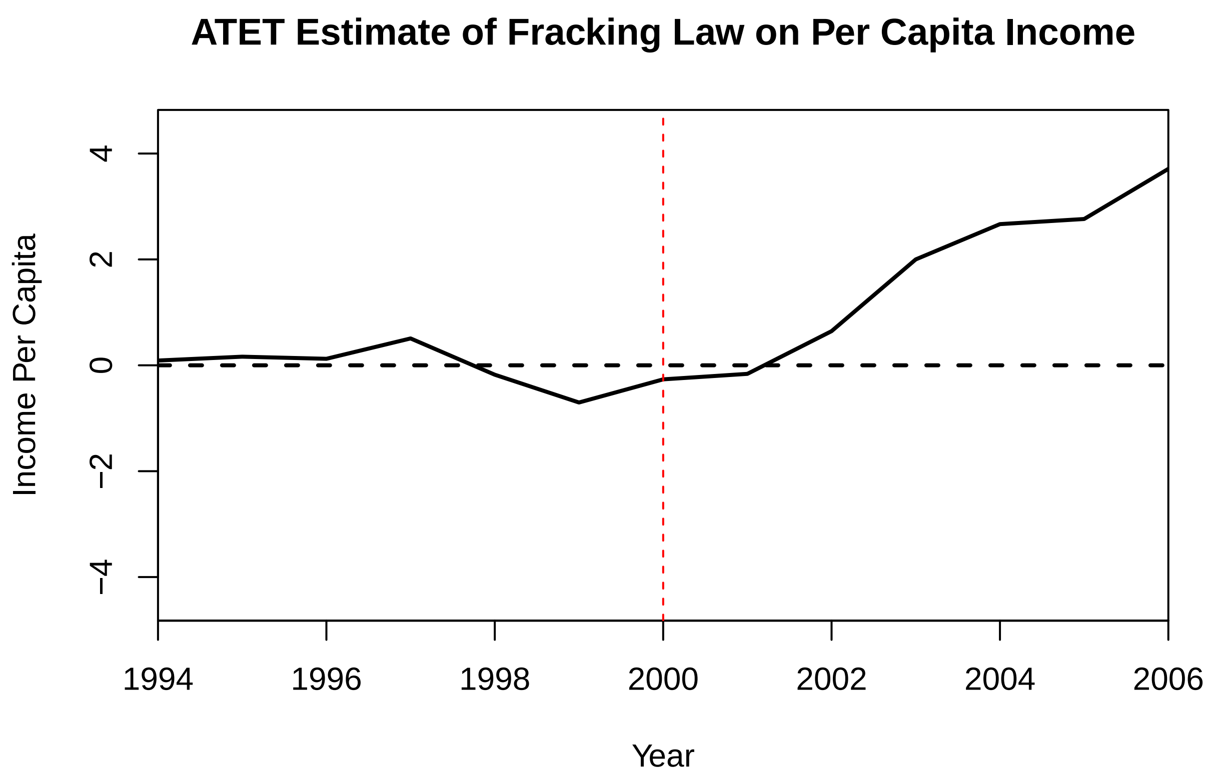
```r
   time.optimize.ssr = c(1994:1999),

   #name of panel units
   unit.names.variable = "state",

   #time period to generate the plot for.
   time.plot = 1994:2006)

synth.out = synth(dataprep.out)
```

```
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.1387035
##
## solution.v:
##  2.59612e-05 0.001955033 0.5012642 0.002919795 0.0005281146 0.4933069
##
## solution.w:
##  0.2574452 0.01879814 3.48127e-05 0.1457779 0.4939386 0.08400536
```

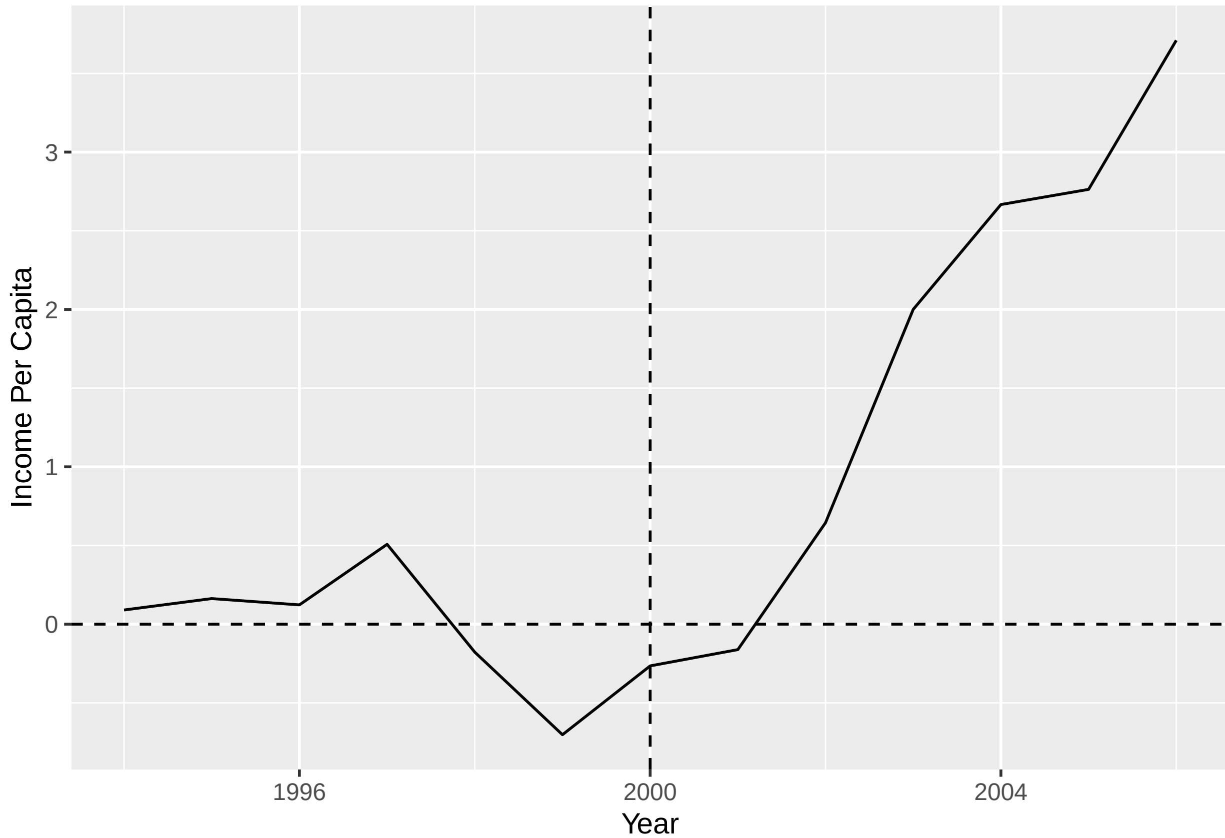
```r
# Two native plotting functions.
# Path.plot() plots the synthetic against the actual treated unit data.
path.plot(dataprep.res = dataprep.out, synth.res = synth.out,Xlab="Year",
          Ylab="Income Per Capita",
          Main="Comparison of Synth vs. Actual Per Capita Income in California")
abline(v=2000,lty=2,col="red")
```




```r
# Gaps.plot() shows the deviation between the synthetic and the actual over time.
gaps.plot(dataprep.res = dataprep.out, synth.res = synth.out,Xlab="Year",
          Ylab="Income Per Capita",Main="ATET Estimate of Fracking Law on Per Capita Income")
abline(v=2000,lty=2,col="red")
```




```r
controls <- c(29, 2, 13, 17, 32, 38)

# We can pull out the data from the result, to make our own nicer plots in ggplot of course
plot.df = data.frame(dataprep.out$Y0plot%*%synth.out$solution.w)
years = as.numeric(row.names(plot.df))
plot.df = data.frame(y=fracking.data$Y[fracking.data$state=='California' &
                              fracking.data$year %in% years]) - data.frame(y=plot.df$w.weight)
plot.df$years <- years
plot.df$state <- "California"
ggplot(plot.df,aes(y=y,x=years)) +
  geom_line() +
  geom_hline(yintercept=0,linetype="dashed") +
  geom_vline(xintercept=2000,linetype="dashed") + xlab("Year") + ylab("Income Per Capita")
```




```
# Okay, let's simulate a null distribution
# We'll run synthetic control on each of the untreated units, using the other units as
# controls (we exclude the treated unit from the control set in each placebo run).
for (i in 1:length(controls)){
  controls_temp <- controls[!controls %in% controls[i]]
  #your outcome variable *must* be named Y for Synth to accept it (bad coding practices in
  # here I suspect)
  dataprep.out.placebo=
    dataprep(foo = fracking.data,
             predictors = c("res.share", "edu", "pop.dense"),
             predictors.op = "mean",
             dependent = "Y",
             unit.variable = "panel.id",
             time.variable = "year",

             #Any pre-period X's we want to include using different aggregation function,
             # other than mean, or different
             # time windows, specific years vs. all years, we enter here.
             special.predictors = list(list("Y", 1999, "mean"),
                                       list("Y", 1995, "mean"),
                                       list("Y", 1990, "mean")),

             # which panel is treated?
             treatment.identifier = controls[i],

             # which panels are we using to construct the synthetic control?
```

```r
            controls.identifier = controls_temp,

            # what is the pre-treatment time period?
            time.predictors.prior = c(1994:1999),

            time.optimize.ssr = c(1994:1999),

            # name of panel units
            unit.names.variable = "state",

            # time period to generate the plot for.
            time.plot = 1994:2006)

  synth.out.placebo = synth(dataprep.out.placebo)
  plot.df.temp <- data.frame(dataprep.out.placebo$Y0plot%*%synth.out.placebo$solution.w)
  years = as.numeric(row.names(plot.df.temp))
  plot.df.update <- data.frame(y=fracking.data$Y[fracking.data$panel.id==controls[i] &
                      fracking.data$year %in% years]) - data.frame(y=plot.df.temp$w.weight)
  plot.df.update$years <- years
  plot.df.update$state <- unique(fracking.data[fracking.data$panel.id==controls[i],]$state)
  plot.df <- rbind(plot.df, plot.df.update)
}
```

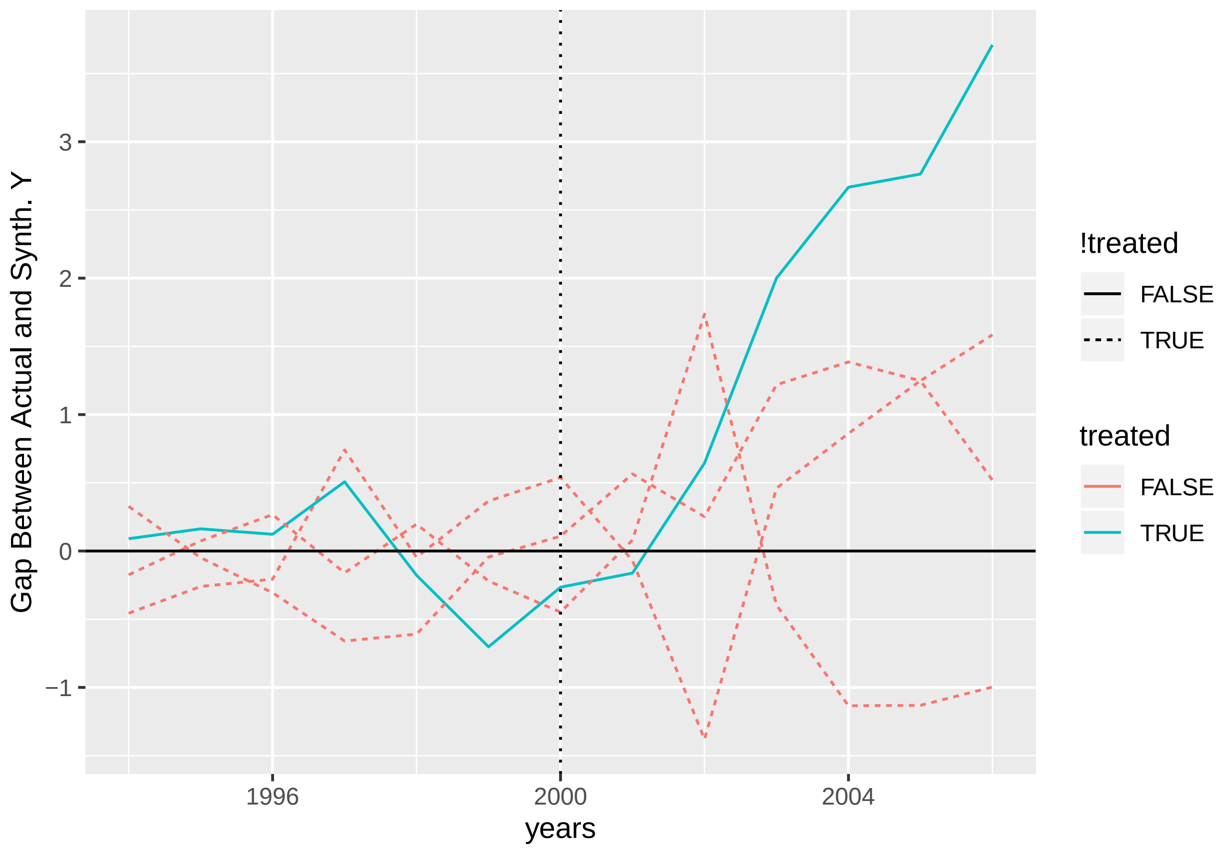
```
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.7891398
##
## solution.v:
##  0.1953372 0.1280436 0.5913547 0.007033043 0.07719718 0.001034261
##
## solution.w:
##  5.6635e-06 0.364165 0.1109425 0.0001380115 0.5247488
##
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
```

```
## ****************
##
## MSPE (LOSS V): 0.03662159
##
## solution.v:
##  0.008192141 0.2159541 0.2063148 0.1875261 0.1993554 0.1826575
##
## solution.w:
##  0.05248441 0.4598678 7.2708e-06 0.4876399 6.843e-07
##
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 57.62437
##
## solution.v:
##  0.01105524 0.03814458 0.01982471 0.2023291 0.311199 0.4174474
##
## solution.w:
##  5.242e-07 9.46e-08 0.9999988 5.234e-07 2.4e-08
##
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.1685874
##
## solution.v:
##  0.0136038 0.003942 1.7308e-06 0.5238024 0.4373643 0.02128577
##
## solution.w:
##  0.01719944 0.002975143 0.1419645 0.1980392 0.6398217
##
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
```

```
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.1673348
##
## solution.v:
##  0.0007177705 0.006915393 0.001917603 0.6964686 0.04423186 0.2497487
##
## solution.w:
##  0.9211801 0.07879777 1.796e-07 1.21924e-05 9.7362e-06
##
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 1.700606
##
## solution.v:
##  0.0003706307 0.004005661 0.03501926 0.2383374 0.3876867 0.3345803
##
## solution.w:
##  2.9e-09 4.4e-09 0.9999995 4.315e-07 2.58e-08
```

```r
plot.df$treated <- (plot.df$state=="California")

# Let's plot the diffs associated with each control state.
ggplot(plot.df,aes(y=y,x=years,group=state)) +
  geom_line(aes(color=treated,linetype=!treated)) +
  geom_vline(xintercept=2000,linetype="dotted") +
  geom_hline(yintercept=0)
```


```r
# Our syntheses for ID, OR and IL are all terrible in the pre period.
# I can remove them here for now, but you'd want to tweak the inputs to try to get a better
# MSPE for those three.
ggplot(plot.df[plot.df$state!="Idaho" & plot.df$state!="Oregon" & plot.df$state!="Illinois",],
       aes(y=y,x=years,group=state)) +
  geom_line(aes(color=treated,linetype=!treated)) +
  geom_vline(xintercept=2000,linetype="dotted") +
  geom_hline(yintercept=0) +
  ylab("Gap Between Actual and Synth. Y")
```




```r
# I can also recover my cumulative alpha (the ATT) for CA and all placebo estimates.
# by summing over the gaps in the post period.
# If I exclude the 3 poorly synthesized states, CA is the biggest effect in the distribution.
# This is a sparse null distribution, but technically empirical p-value = 0.000.
post.treats <- plot.df[plot.df$year>=2000,]
alphas <- aggregate(post.treats[-c(2:3)], by=list(post.treats$state),FUN=sum)
View(alphas[alphas$Group.1!="Idaho" & alphas$Group.1!="Oregon" & alphas$Group.1!="Illinois",])
```

Conclusion

- SC is a method to evaluate the causal effect of shocks / policies
- SC builds upon the setting of the standard DD model, but makes two changes:
  - Synthetic Control allows for time-varying individual-specific heterogeneity
  - Synthetic Control takes a serious, data driven approach to forming counterfactuals / selecting the control group

- The benefits of Synthetic Control come with costs:
  - Large scale (asymptotic) inference cannot be conducted on Synthetic Control estimators
  - Instead, SC uses Permutation Methods to compute "standard errors"

* empirical p-values