# DID Homework

*Jennifer Hill, Ray Lu & Zarni Htet*

## Objective

The goal of this exercise is to simulate data that may or may not satisfy the assumptions of a difference in differences design.

### PART A: DID

The setting here is similar to the last assignment (RDD).

You will simulate hypothetical data collected on women who gave birth at any one of several hospitals in disadvantaged neighborhoods in New York City in 2010. This time we are envisioning a government policy that makes available job training for teen mothers. This program is only available for women in households with income below $20,000 at the time they gave birth. The general question of interest is whether this program increases annual income 3 years later. You will generate data for a sample of 1000 individuals. For this assignment we will make the unrealistic assumption that everyone who is eligible for the program participates and no one participates who is not eligible.

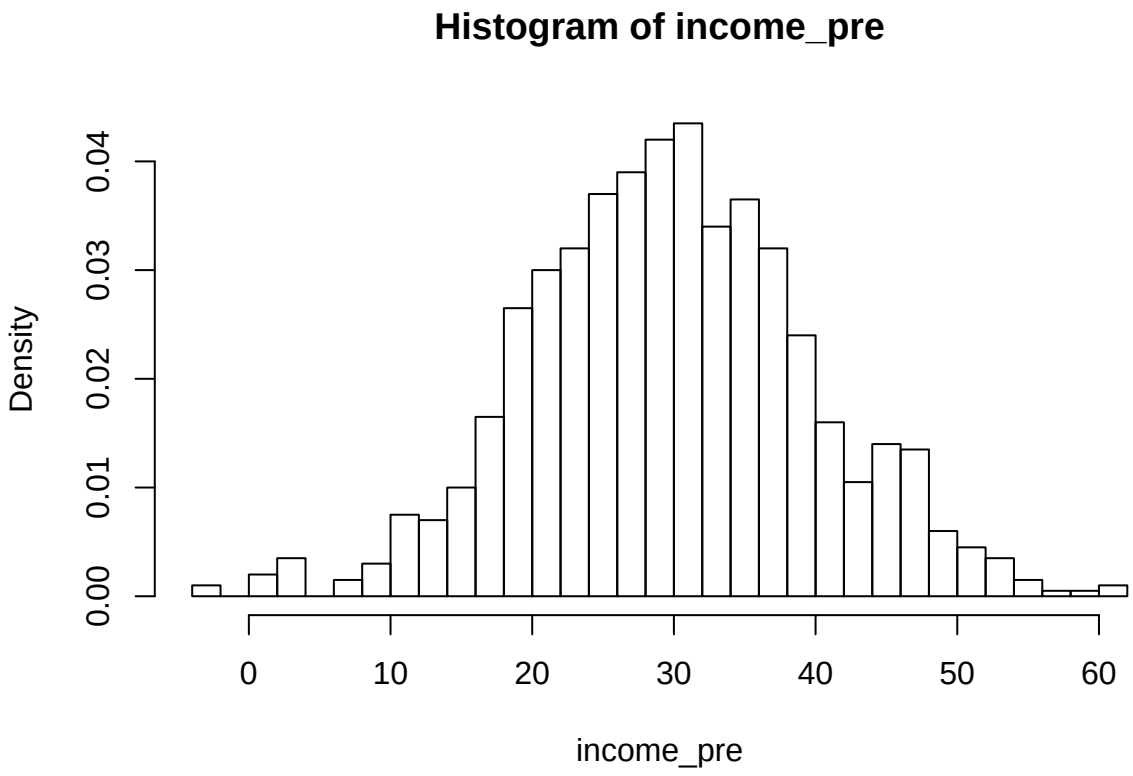
#### Question 1. God role: simulate income.

Simulate the "assignment variable" (sometimes referred to as the "running variable", "forcing variable", or "rating"), income, in units of thousands of dollars. Use the following model:

$$X \sim N(30, 100)$$

Then plot using a histogram.

```
set.seed(1234)
N = 1000
income_pre <- rnorm(N, 30, 10)
hist(income_pre,nclass=30,freq=FALSE)
```




**Question 2. Policy maker role: Assign eligibility indicator.**

Create an indicator for program eligibility for this sample. Call this variable "eligible". (You can use the same code as the previous assignment.)

```
eligible <- numeric(N)
eligible[income_pre<20] <- 1
```

**Question 3: God role.**

For question 3 you will simulate income at 3 years post treatment in thousands of dollars. You will assume linear models for both E[Y(0) | X] and E[Y(1) | X]. The *expected* treatment effect for everyone should be 4 (in other words, E[Y(1) - Y(0) | X] should be 4 at all levels of X). The residual standard deviation of each potential outcome should be *2*.
a) You will simulate using the following model

$$Y(0) \sim N(6 + .3 * \text{income}_{\text{pre}}, 2^2) Y(1) \sim N(6 + .3 * \text{income}_{\text{pre}} + 4, 2^2)$$

b) You will save two datasets:
(1) fullA should have the forcing variable and both potential outcomes (2) obsA should have the forcing variable, the eligibility variable, and the observed outcome.

```
#set.seed(1234)
y0A <- rnorm(N, 6 + .3*income_pre, 2)
y1A <- rnorm(N, 6 + .3*income_pre + 4, 2)
#y0A <- 6 + .3*income_pre + rnorm(N, 0 , 2)
summary(y0A); summary(y1A)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.647  12.489  14.883  14.949  17.401  28.409
```

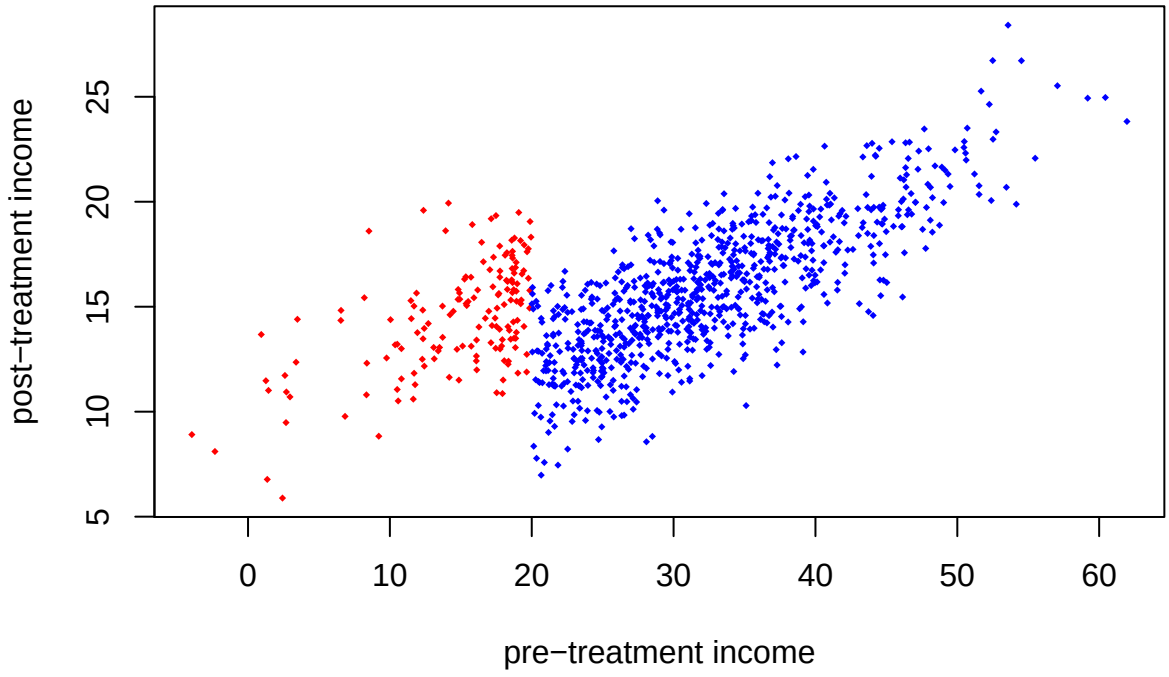
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.886  16.587  19.046  18.978  21.388  29.056
```

```
#plot(income_pre,y1A)
 # save data
fullA <- data.frame(y0=y0A,y1=y1A,income_pre=income_pre)
y <- y0A*(1-eligible) + y1A*eligible
obsA = data.frame(income_post=y,income_pre=income_pre,eligible=eligible)
```

**Question 4. Researcher and god role. Plot your data!**

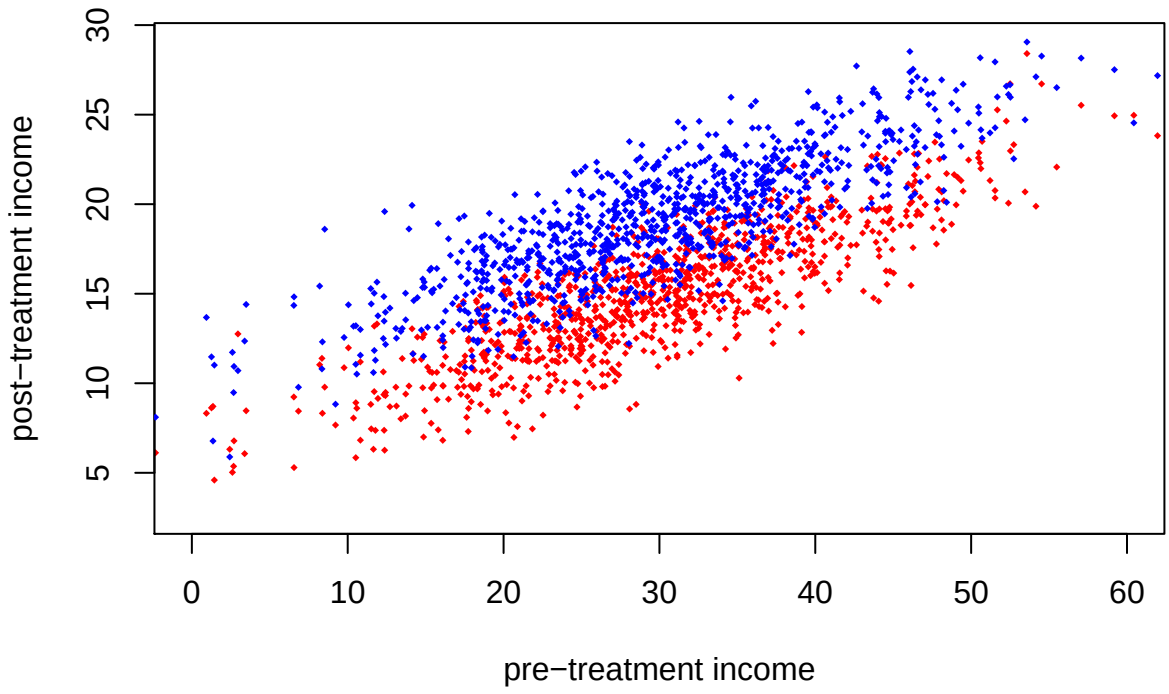
Make a scatter plots of pre-treatment income (x-axis) versus observed post-treatment income (y-axis). Plot eligible participants in red and non-eligible participants in blue.

```
par(mfrow=c(1,1))
plot(obsA$income_pre[obsA$eligible==1],obsA$income_post[obsA$eligible==1], xlab="pre-treatment income",
points(obsA$income_pre[obsA$eligible==0],obsA$income_post[obsA$eligible==0], pch=18,col="blue",cex=.5)
```




Now plot the full response surface.

```
par(mfrow=c(1,1))
# World A
plot(fullA$income_pre,fullA$y0,xlab="pre-treatment income", xlim=c(0,60), ylim=range(c(fullA$y0,fullA$y
points(fullA$income_pre, fullA$y1, pch=18,col="blue",cex=.5)
```

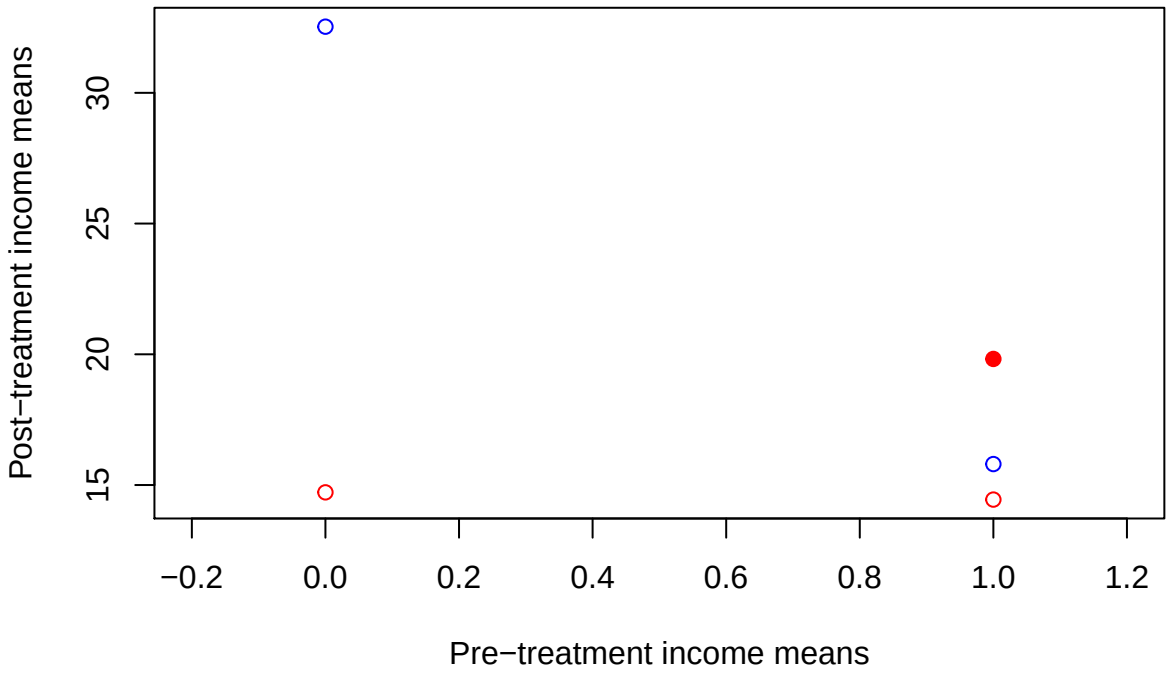



Now make a plot like the ones in the DID lecture with time on the x axis and income on the y axis. Plot observed means as open circles; blue corresponds to the control and red corresponds to the treatment. In the same figure plot

$$E[Y(1) \mid Z = 0]$$

using different red symbol (square, triangle, filled in circle).

```
y11m <- mean(obsA$income_post[obsA$eligible==1])
y10m <- mean(obsA$income_pre[obsA$eligible==1])
y01m <- mean(obsA$income_post[obsA$eligible==0])
y00m <- mean(obsA$income_pre[obsA$eligible==0])
y11e <- mean(fullA$y1[obsA$eligible==0])
plot(x=c(0,1),y=c(y10m,y11m), col="red", ylim=c(range(c(y11m,y10m,y01m,y00m))),
     ylab = "Post-treatment income means", xlab = "Pre-treatment income means", xlim = c(-.2,1.2))
points(x=c(0,1),y=c(y00m,y01m), col="blue")
points(x=1,y=y11e, col="red", pch=19)
```




Based on what you know about the DID assumptions, will the DID estimate of E[Y(1)-Y(0) | Z=0] be close to the truth? Why or why not?

** No! if you assume that the control trajectory represents what would have happened to the treated if they hadn't participated in the treatment then the E[Y(1) | Z=0] would be very different than what is predicted by the DID model thus throwing off the t.e. estimate. **

**Question 5. Researcher role. Estimate the treatment effect using all the data using two approaches:**

a) a regression discontinuity approach (linear model only)

```
obsA$income_preT <- obsA$income_pre - 20
# i
summary(glm(income_post ~ eligible + income_preT,data=obsA))$coef[2,1:2]
```

```
##   Estimate Std. Error
##  4.2762198  0.2290906
```

```
# or, ii
summary(glm(income_post ~ eligible*income_preT,data=obsA))$coef[2,1:2]
```

```
##   Estimate Std. Error
##  4.1569114  0.2626164
```

b) a DID approach

```
# could just use the four means
d1 <- y11m - y10m
d0 <- y01m - y00m
d1 - d0
```

```
## [1] 16.45288
```

```
#
#  or could use a regression approach
```

```r
obsA$income_change = obsA$income_post - obsA$income_pre
summary(glm(income_change ~ eligible, data=obsA))$coef[2,1:2]
```

```
##   Estimate Std. Error
## 16.4528762  0.4830459
```

**Question 6. Researcher and god roles. Thinking about the assumptions?**

Can you think of a way of altering the simulation setup so that the DID assumptions would hold? Make this change, rerun (3) and (5) above, and comment on what you find.

```r
set.seed(12345)
y0B <- rnorm(N, 6 + 1*income_pre,     2)
y1B <- rnorm(N, 6 + 1*income_pre + 4, 2)
summary(y0B); summary(y1B)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.036  28.937  35.812  35.826  42.491  68.381
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.036  32.698  40.016  39.673  46.377  71.766
```

```r
# save data
fullB= data.frame(y0=y0B,y1=y1B,income_pre=income_pre)
yB = y0B*(1-eligible) + y1B*eligible
obsB = data.frame(income_post=yB,income_pre=income_pre,eligible=eligible)
```

```r
obsB$income_preT <- obsB$income_pre - 20
# i
summary(glm(income_post ~ eligible + income_preT,data=obsB))$coef[2,1:2]
```

```
##   Estimate Std. Error
##  3.7970102  0.2282643
```

```r
# ii
obsB$income_change = obsB$income_post - obsB$income_pre
summary(glm(income_change ~ eligible, data=obsB))$coef[2,1:2]
```

```
##   Estimate Std. Error
##   3.698876   0.173414
```

** In the original simulation even though the regression coefficient on pretreatment income was the same across groups that just means that the percentage change over time is the same. That's not what the DID assumption requires. When we change the coefficient on pre-treatment income to be 1 we implicitly satisfy the assumption that the change over time (represented as a difference in means not a percentage!!) is the same across groups. Thus now both approaches yield good estimates. **