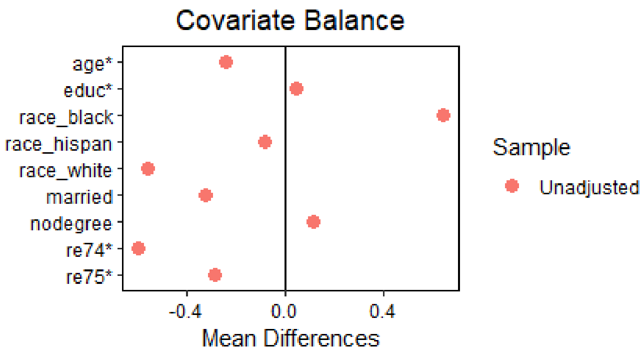
# Impact Evaluation with R

## Assignment 2

**R Script**

**Part 1**

```
library(cobalt)
data(lalonde, package = "MatchIt")

install.packages("causaldata")
library(causaldata)

l1 <- lalonde

love.plot(treat  ~ age + educ + race + married + nodegree + re74 + re75, data = l1, stars="std")
mod_la1 <- lm(re78  ~ treat, data = l1)
mod_la2 <- lm(re78  ~ treat + age + educ + race + married +
            nodegree + re74 + re75, data = l1)
library(texreg)
screenreg(list(mod_la1, mod_la2))
```




```
library(causaldata)
library(cem)
library(MatchIt)

# Breaks for educ, age, and race using cutpoints code

table(l1$educ)
table(l1$age)

educut <- c(0, 6.5, 8.5, 12.5, 16)
agecut <- c(16, 22, 28, 34, 40, 46, 55)

# New dataset with matched pairs

mat1 <- cem(treatment = "treat", data = l1, drop = "re78",
        cutpoints = list(educ = educut, age = agecut))
mat1
#          G0  G1
#All       429 185
#Matched   116  96
#Unmatched 313  89

# Estimated effect

m_ate <- att(mat1, re78 ~ treat, data = l1)
summary(m_ate)
```

# Impact Evaluation with R

## Assignment 2

**Linear regression model estimated on matched data only**

**Coefficients:**

```
         Estimate Std. Error t value   p-value
(Intercept)  4910.55     644.58  7.6183 8.667e-13 ***
treat        1638.10     957.87  1.7101   0.08872 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m_ate2 <- att(mat1, re78 ~ treat + age + educ + race + married +
              nodegree + re74 + re75, data = l1)
summary(m_ate2)
```

**Treatment effect estimation for data:**

```
          G0  G1
All       429 185
Matched   116  96
Unmatched 313  89
```

**Linear regression model estimated on matched data only**

```
Coefficients:
Estimate  Std. Error t value p-value
(Intercept) -8458.88346  5173.37751 -1.6351 0.10359
treat        1778.06402   944.91127  1.8817 0.06131 .
age           101.09146    71.67635  1.4104 0.15996
educ          881.15966   388.83549  2.2662 0.02450 *
racehispan    666.44976  2732.38587  0.2439 0.80755
racewhite     600.60117  1655.58616  0.3628 0.71715
married     -1807.37453  2370.67861 -0.7624 0.44672
nodegree     1808.08457  1530.99012  1.1810 0.23900
re74           -0.36963     0.45694 -0.8089 0.41952
re75            1.20956     0.52416  2.3076 0.02203 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Genetic Matching (w/out and w/ covariates)

install.packages("rgenoud")

set.seed(123)
match.l1 <- matchit(treat ~ age + educ + race + married +
                    nodegree + re74 + re75,
                  data = l1, method = "genetic",
                  replace = FALSE, pop.size = 50, print = 0) #, caliper = 0.4)
match.l1
```

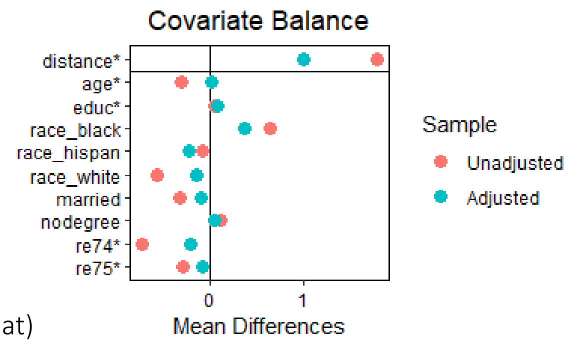
# Impact Evaluation with R

## Assignment 2

**A matchit object**

- method: 1:1 genetic matching without replacement
- distance: Propensity score
  - estimated with logistic regression
- number of obs.: 614 (original), 370 (matched)
- target estimand: ATT
- covariates: age, educ, race, married, nodegree, re74, re75

```
love.plot(match.l1, stars ="std") #to see adjusted covariate balance
match_dat <- match.data(match.l1)

# w/out covariates
mod_la_match1 <- lm(re78  ~ treat, data = match_dat)
screenreg(mod_la_match1)

#w/ covariates
mod_la_match2 <- lm(re78  ~ treat + age + educ + race +
            married + nodegree + re74 + re75, data = match_dat)
screenreg(list(mod_la_match1, mod_la_match2))
```




```
# Propensity Score Matching

library(tidyverse)
library(haven)

logit_l1 <- glm(treat ~ age +  educ + race +
          married + nodegree + re74 + re75, family = binomial(link = "logit"),
         data = l1)
```

**Call:**

```
glm(formula = treat ~ age + educ + race + married + nodegree +
    re74 + re75, family = binomial(link = "logit"), data = l1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7645  -0.4736  -0.2862   0.7508   2.7169

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.663e+00  9.709e-01  -1.713  0.08668 .
age          1.578e-02  1.358e-02   1.162  0.24521
educ         1.613e-01  6.513e-02   2.477  0.01325 *
racehispan  -2.082e+00  3.672e-01  -5.669 1.44e-08 ***
racewhite   -3.065e+00  2.865e-01 -10.699  < 2e-16 ***
married     -8.321e-01  2.903e-01  -2.866  0.00415 **
nodegree     7.073e-01  3.377e-01   2.095  0.03620 *
re74        -7.178e-05  2.875e-05  -2.497  0.01253 *
```

```
    re75         5.345e-05  4.635e-05   1.153  0.24884
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

l1control <- l1 %>%
  mutate(pscore = logit_l1$fitted.values)

# mean pscore
pscore_control <- l1control %>%
  filter(treat == 0) %>%
  pull(pscore) %>%
  mean()

pscore_treated <- l1control %>%
  filter(treat == 1) %>%
  pull(pscore) %>%
  mean()

summary(logit_l1)

m.nn     <- matchit(treat~ age     + educ   + race + nodegree          + re74   + re75,  data = l1,
           method= "nearest",      ratio      = 1)

# estimating treatment effect using PSM

mdata   = match.data(m.nn)
names(mdata)

match.data <- mdata

avg.income78.treated=weighted.mean(mdata$re78[mdata$treat == 1],
                           mdata$weights[mdata$treat==1])
avg.income78.control=weighted.mean(mdata$re78[mdata$treat==0],
                           mdata$weights[mdata$treat==0])
avg.income78.treated
avg.income78.control
avg.income78.treated- avg.income78.control

# estimating effect using regression --> get same result as method above of +278
lm_treat1 <- lm(re78 ~ treat, data = match.data)
screenreg(lm_treat1)
```

# Impact Evaluation with R

## Assignment 2

```
========================
           Model 1
------------------------
(Intercept)  6070.34 ***
           (531.30)
treat        278.80
           (751.38)
------------------------
R^2           0.00
Adj. R^2     -0.00
Num. obs.     370
========================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

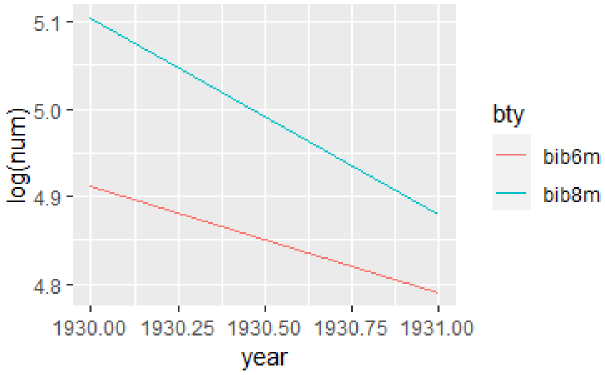
---

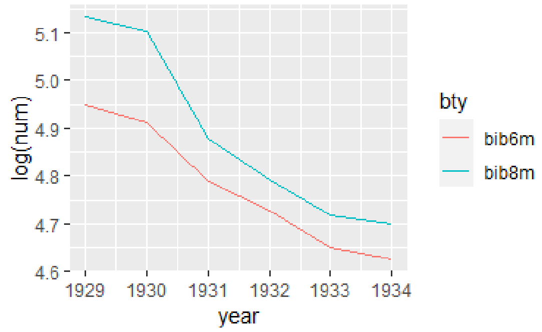
```
#Installing packages
install.packages("tidyverse")
library(tidyverse)
install.packages("tidyr")
library(tidyr)
install.packages("lubridate")
library(lubridate)
install.packages("dplyr")
library(dplyr)
install.packages("ggplot2")
library(ggplot2)

library(readr)
banks <- read_csv("C:/Users/Lg/Downloads/banks.csv")
View(banks)

#creating variables
date = banks$date
day = banks$day
month = banks$month
weekday = banks$weekday
year = banks$year
bib6 = banks$bib6
bio6 = banks$bio6
bib8 = banks$bib8
bio8 = banks$bio8

#Calculate the mean number of banks in business each year in the 6th and 8th districts
mean(bib6, trim = 0, na.rm = FALSE) #118.6193
mean(bio6, trim = 0, na.rm = FALSE) #117.1283
mean(bib8, trim = 0, na.rm = FALSE) #133.0367
mean(bio8, trim = 0, na.rm = FALSE) #131.1416
```

# Impact Evaluation with R

## Assignment 2

```
#stack or gather bib in one column
bankag = banks %>%
  group_by(year) %>%
  summarize(bib6m = mean(bib6),
        bib8m=mean(bib8))
head(bankag)

bankag2 = gather(bankag, "bty","num", 2:3)

bankag3 <- filter(bankag2,
          year == 1930 | year == 1931)

#graph with filter for years 1930 and 1931
ggplot(bankag3, aes(x = year, y=log(num), color=bty)) +
  geom_line()

#graph for years 1929 to 1934
ggplot(bankag2, aes(x = year, y=log(num), color=bty)) +
  geom_line()
```




Compared to the graph for years 1930 and 1931, the graph for 1929 - 1934 shows a decline in descension. The banks in the 8th district after 1931 are closing at a slower rate than what it was prior to the implementation of the treatment. Post treatment shows similar trends for the 6th and 8th district. The DiD for the two districts is (120 – 136) – (132 – 165) = -16 – (-33) = 17. In this final graph by using the DiD, we can say that it withstands the parallel trend assumption.

---

Part 3

```
install.packages("reshape2")
library(reshape2)
install.packages("tidyr")
library(tidyr)
install.packages("magrittr")
library(magrittr)
library(foreign)
library(tidyverse)

traffic1_long <- reshape(data = traffic1, idvar = "state",
              varying = c("admn90", "admn85", "open90", "open85", "dthrte90", "dthrte85", "speed90",
"speed85", "cdthrte", "cadmn", "copen", "cspeed"),
              v.name = "value",
              time=c("admn90", "admn85", "open90", "open85", "dthrte90", "dthrte85", "speed90",
"speed85", "cdthrte", "cadmn", "copen", "cspeed"),
              new.row.names = 1:1000,
              direction= c("long"))
```

# Impact Evaluation with R

## Assignment 2

traffic1_long1 = gather(traffic1, key = "state", value = "admn90", "admn85", "open90", "open85", "dthrte90", "dthrte85", "speed90", "speed85", "cdthrte", "cadmn", "copen", "cspeed", na.rm = FALSE, convert = FALSE, factor_key = FALSE)

#Estimate the difference-in-differences estimate - treatment effect, including interaction term

```
------
install.packages("texreg")
library(texreg)
#OLS w/ deaths and open container laws in 1985 and 1990

mod85_open <- lm(dthrte85~open85, data=traffic1)
screenreg(mod85_open)

mod90_open <- lm(dthrte90~open90, data=traffic1)
screenreg(mod90_open)
screenreg(list(mod85_open, mod90_open))

#DiD for open models

DD_model_open <- lm(cdthrte ~ copen, data = traffic1)
screenreg(DD_model_open)

#OLS w/ deaths and admin laws in 1985 and 1990

mod85_admn <- lm(dthrte85~admn85, data=traffic1)

mod90_admn <- lm(dthrte90~admn90, data=traffic1)
screenreg(list(mod85_admn, mod90_admn))

#DiD for admn models
```

```
=================================
            Model 1    Model 2
---------------------------------
(Intercept)   2.61 ***   2.11 ***
             (0.11)     (0.11)
admn85        0.23
                 (0.17)
admn90                   0.07
             (0.15)
---------------------------------
R^2          0.03       0.00
Adj. R^2      0.02      -0.02
Num. obs.    51         51
=================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

# Impact Evaluation with R

## Assignment 2

Model 1 is the OLS regression with deaths in 1985. Model 2 is the regression for 1990. There is a positive .23 unit increase in traffic deaths whereas, in model 2 there is a 0.07 unit increase. .23 - .07 = .16.

```
DD_model_admn <- lm(cdthrte ~ cadmn, data = traffic1)

screenreg(list(mod85_admn, mod90_admn, DD_model_admn))
```

```
============================================
             Model 1    Model 2    Model 3
--------------------------------------------
(Intercept)   2.61 ***   2.11 ***  -0.52 ***
             (0.11)     (0.11)     (0.05)
admn85        0.23
             (0.17)
admn90                   0.07
                        (0.15)
cadmn                              -0.18
                                   (0.12)
--------------------------------------------
R^2           0.03       0.00       0.04
Adj. R^2      0.02      -0.02       0.02
Num. obs.    51         51         51
============================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

```
#Estimate ATE for both laws

DD_model_both <- lm(cdthrte ~ copen + cadmn, data = traffic1)
screenreg(DD_model_both)
```

```
======================
             Model 1
----------------------
(Intercept)  -0.50 ***
             (0.05)
copen        -0.42 *
             (0.21)
cadmn        -0.15
             (0.12)
----------------------
R^2           0.12
Adj. R^2      0.08
Num. obs.    51
======================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

# Impact Evaluation with R

## Assignment 2

Question 1:

Some of the issues that arise when using this cross-sectional data is that states vary in culture, attitude, location, etc, which may have unobserved omitted variables that cannot be controlled for. We also cannot derive counterfactual outcomes from single cross-sectional data as it creates randomization uncertainty. Cross sectional data cannot control for time invariant unobserved heterogeneity like panel and pooled data.

Question 2:

Difference-in-differences would be the method of choice for having data with two periods. This method allows us to utilize the data at two points to establish trends prior to instilling the treatment as a baseline to understand how the two groups relate to each other. Then once the treatment has been applied, we can predict what the counterfactual would have been in the instance that the treatment was not applied, and the difference between the counterfactual and the outcome is the treatment effect. Furthermore, difference-in-differences does not require that the states be alike, due to being able to establish the trends from the two-period data. Controlling for covariates before and after the treatment to avoid bias as a change in one covariate might have a significant effect on the treatment or control group.

Question 5:

We cannot completely prove the parallel assumption only provide the evidence to support it. Pre-trends are the typical way, looking at trends in treatment and control group in the years before the treatment. But more should be done in terms of arguing why it should hold at the time of treatment. Some would show that after the implementation of the treatment, the trend became closer to the control group.