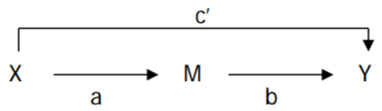
Social_Capital Mediation Analysis

Wesley_Tao wt2271@columbia.edu 9176558355

2018/5/12

Mediation Analysis



Treatment Variable	Mediation variable	Response Variable
X	M	Y
Jim Crow	Gini Coeficient	Socail Captial index

---

1. Preprocessing and Data Engineering

```
library("xlsx")
library(dplyr)
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 3.4.4
```

```
library(ggplot2)
library("mediation")
```

```
## Warning: package 'mediation' was built under R version 3.4.4
```

```
## Warning: package 'sandwich' was built under R version 3.4.4
```

```
library(dplyr)
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.4.4
```

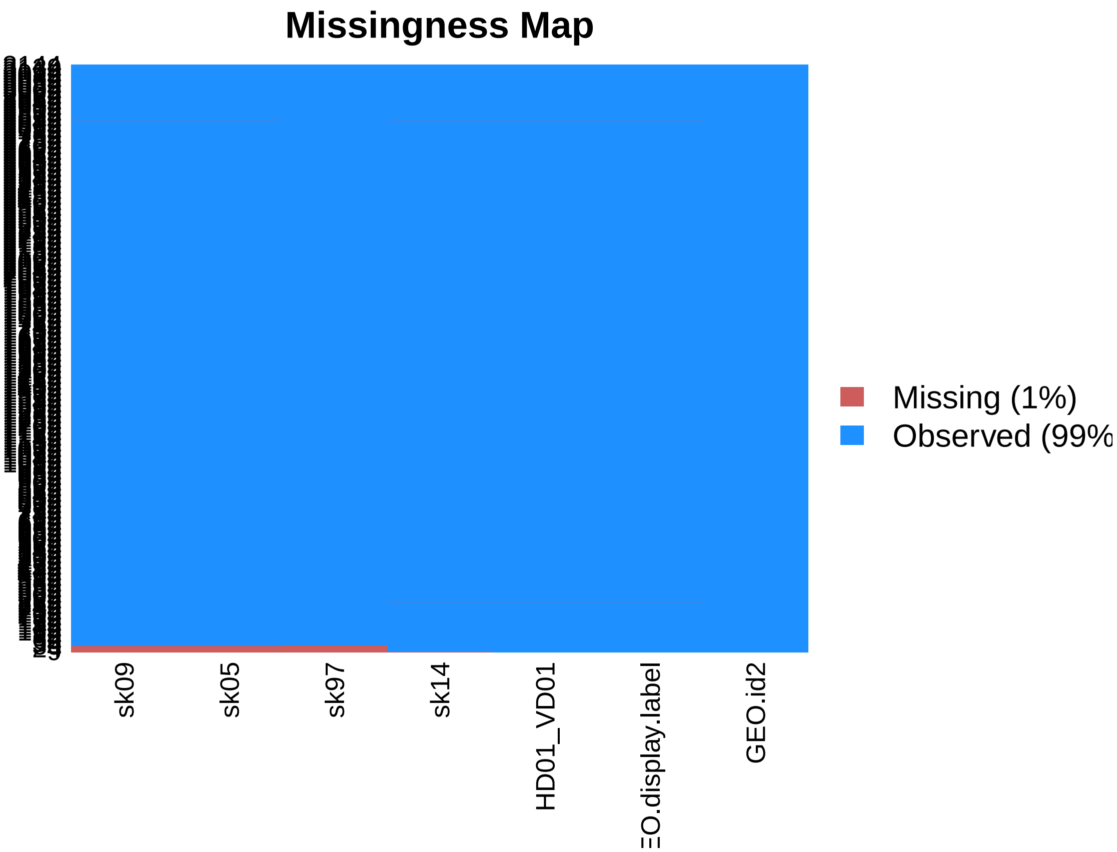
1.1 load and merge

```
# load the data
table_1997<-read.xlsx("../data/social capital 1997-2014.xlsx", 1)
table_2005<-read.xlsx("../data/social capital 1997-2014.xlsx", 2)
table_2009<-read.xlsx("../data/social capital 1997-2014.xlsx", 3)
table_2014<-read.xlsx("../data/social capital 1997-2014.xlsx", 4)
# change the header so we could align dataframe
names(table_2014)<-c("fips","areaname","sk14")

m.1<-table_1997 %>%
  full_join(table_2005, by = c("fips","areaname")) %>%
  dplyr::select(everything())

m.2<-table_2009 %>%
  full_join(table_2014, by = c("fips")) %>%
  dplyr::select(fips,sk09,sk14)

merged<-m.1 %>%
  full_join(m.2, by = c("fips"))%>%
  dplyr::select(everything())

all_content = readLines("../data/Gini coefficient 2010-2014.csv")
skip_second = all_content[-2]
Gini_coe    = read.csv(textConnection(skip_second), header = TRUE, stringsAsFactors = FALSE)

names(merged)[1]<-c("GEO.id2")
merged<-merged %>%
  full_join(Gini_coe, by = c("GEO.id2")) %>%
  dplyr::select(GEO.id2, sk97, sk05, sk09, sk14,GEO.display.label,HD01_VD01)
summary(merged)
```

```
##     GEO.id2           sk97              sk05              sk09
##  Min.   : 1001   Min.   :-4.3107   Min.   :-3.9094   Min.   :-3.9252
##  1st Qu.:18177   1st Qu.:-0.9961   1st Qu.:-0.9364   1st Qu.:-0.8347
##  Median :29176   Median :-0.2337   Median :-0.2259   Median :-0.2204
##  Mean   :30385   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:45082   3rd Qu.: 0.7578   3rd Qu.: 0.7022   3rd Qu.: 0.5265
##  Max.   :56045   Max.   : 8.2406   Max.   :14.2963   Max.   :17.4405
##                  NA's   :36        NA's   :36        NA's   :36
##       sk14           GEO.display.label    HD01_VD01
##  Min.   :-3.183280   Length:3144        Min.   :0.3346
##  1st Qu.:-0.756780   Class  :character  1st Qu.:0.4176
##  Median :-0.226120   Mode   :character  Median :0.4376
##  Mean   :-0.000003                      Mean   :0.4402
##  3rd Qu.: 0.477669                      3rd Qu.:0.4609
##  Max.   :21.808830                      Max.   :0.6519
##  NA's   :3                              NA's   :2
```

1.2 treat missing data

```
missmap(merged)
```



```
merged<-na.omit(merged) # remove NA
```

Since the missing values only account for a small proportion of the dataset. And it is not correlated with Y (dependent) and X variables (independent) We could safely delete them.

1.3 Jim_crow

```
n_row   <-nrow(merged)
list_1  <-strsplit(merged$GEO.display.label,split=",")
State.County      <-data.frame(matrix(unlist(list_1),nrow=n_row,byrow=T))
colnames(State.County)<-c("County","State")
merged<-cbind(merged,State.County)

merged$State<-as.character(merged$State)
merged$State<-substr(merged$State,2,nchar(merged$State))
# add new independent variable
Jim_Crow_States_list<-unlist(strsplit(" Alabama, Arizona, Arkansas, Delaware, Florida, Georgia, Kansas,
Jim_Crow_States_list<-substr(Jim_Crow_States_list,2,nchar(Jim_Crow_States_list))
Jim_Crow_States_list
```

```
##  [1] "Alabama"        "Arizona"        "Arkansas"       "Delaware"
##  [5] "Florida"        "Georgia"        "Kansas"         "Kentucky"
##  [9] "Louisiana"      "Maryland"       "Mississippi"    "Missouri"
## [13] "New Mexico"     "North Carolina" "Oklahoma"       "South Carolina"
## [17] "Tennessee"      "Texas"          "Virginia"       "West Virginia"
## [21] "Wyoming"
```

```
# add dummy variables  Jim_Crow
"Alabama" %in% Jim_Crow_States_list
```

```
## [1] TRUE
```

```
merged$Jim_Crow<-(merged$State  %in% Jim_Crow_States_list)
```

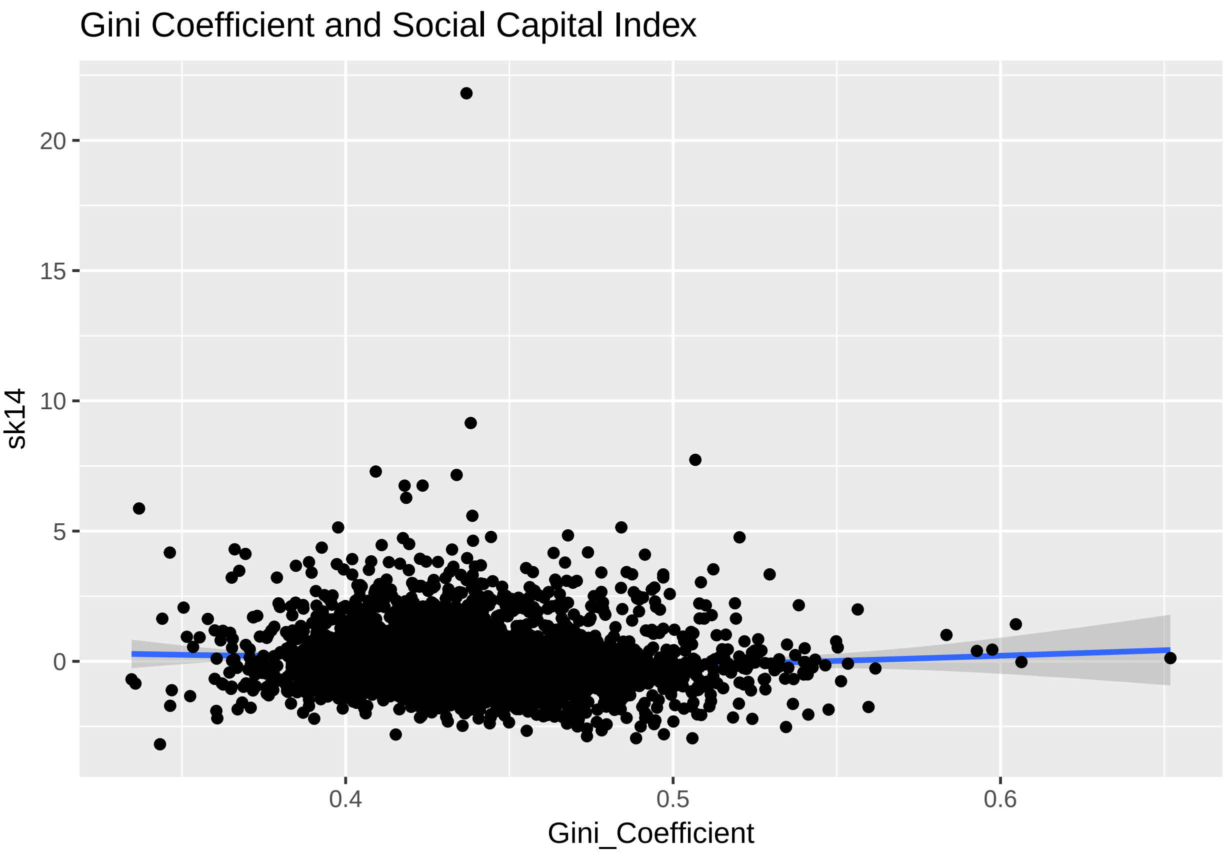
2 Structural Mediation Regression analysis

2.1 Step 1 Regression M —> Y

Gini coefficient 2010-2014 –> on Social Capital 2014 (Social Capital is dependent variable)

```
names(merged)[7]<-"Gini_Coefficient"

ggplot(data = merged)+
  geom_smooth(aes(x=Gini_Coefficient,y=sk14))+ geom_point(aes(x=Gini_Coefficient,y=sk14))+
  labs(title="Gini Coefficient and Social Capital Index")
```

```
## `geom_smooth()` using method = 'gam'
```



```
cat("the maximum value for social capital is ",merged$GEO.display.label[which.max(merged$sk14)])
```

```
## the maximum value for social capital is  Edgefield County, South Carolina
```

```
cat("\nThis value is around the center of the data. In statistics, it wouldn't affect the model estimat
```

```
##
## This value is around the center of the data. In statistics, it wouldn't affect the model estimate ve
```

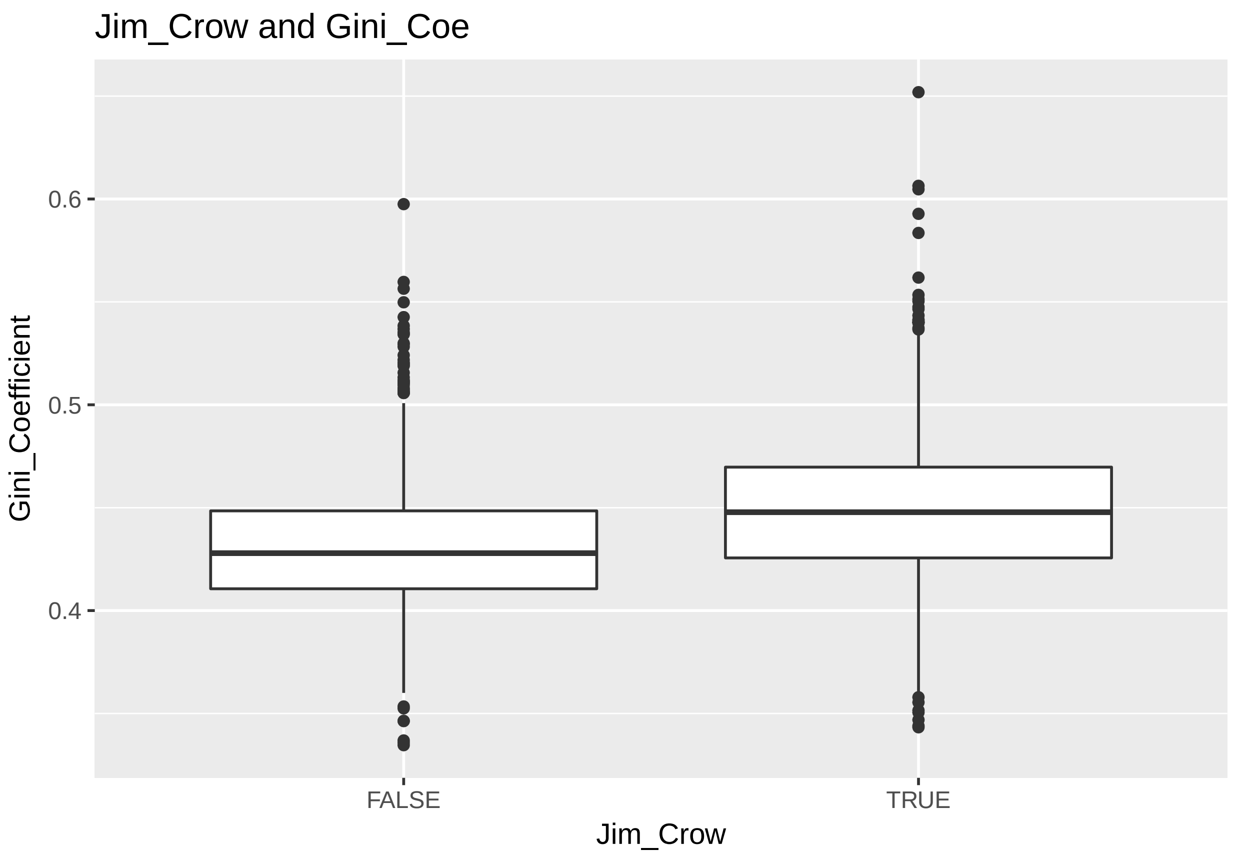
```
m1<-lm(sk14~Gini_Coefficient,data=merged)
summary(m1)
```

```
##
## Call:
## lm(formula = sk14 ~ Gini_Coefficient, data = merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5483 -0.7458 -0.2072  0.4641 21.7892
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.6318     0.2888   5.650 1.74e-08 ***
## Gini_Coefficient  -3.6900     0.6536  -5.646 1.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.255 on 3104 degrees of freedom
## Multiple R-squared:  0.01016,    Adjusted R-squared:  0.009846
## F-statistic: 31.88 on 1 and 3104 DF,  p-value: 1.791e-08
```

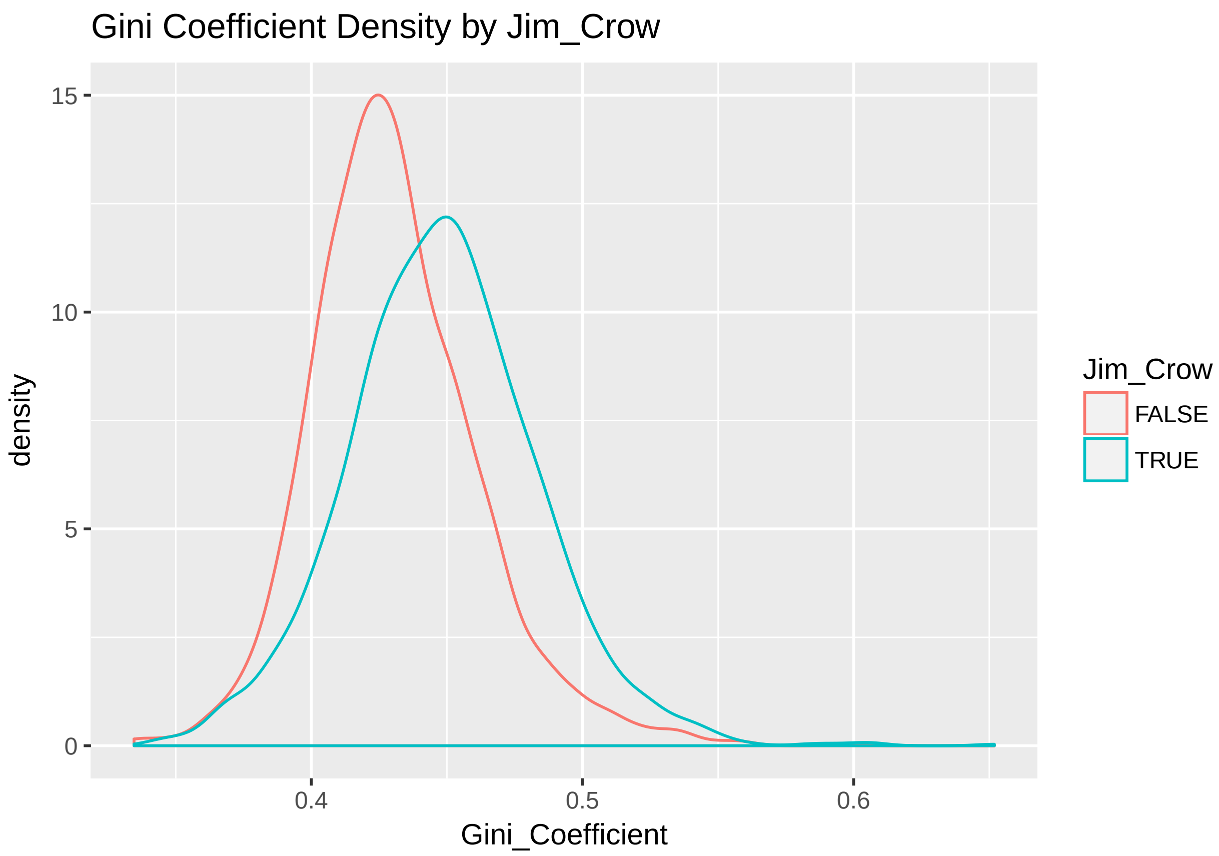
2.2 step2 X—->M

2. States with Jim Crow Law are = 1, states without Jim Crow Laws are = 0.

```
ggplot(data =merged)+
  geom_boxplot(aes(x=Jim_Crow,y=Gini_Coefficient))+
  labs(title="Jim_Crow and Gini_Coe ")
```



```r
ggplot(data=merged,aes(x=Gini_Coefficient,color=Jim_Crow))+
  geom_density()+
  labs(title="Gini Coefficient Density by Jim_Crow")
```



```
cat("summary statistics of gini coefficient by group")
```

```
## summary statistics of gini coefficient by group
```

```
describeBy(merged$Gini_Coefficient,list(jim_crow=merged$Jim_Crow))
```

```
##
##  Descriptive statistics by group
## jim_crow: FALSE
##    vars    n mean   sd median trimmed  mad  min max range skew kurtosis se
## X1    1 1395 0.43 0.03   0.43    0.43 0.03 0.33 0.6  0.26 0.67      1.6  0
## --------------------------------------------------------
## jim_crow: TRUE
##    vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
## X1    1 1711 0.45 0.04   0.45    0.45 0.03 0.34 0.65  0.31 0.32     1.23
##    se
## X1  0
```

3. Regression Jim Crow Laws –> on Gini coefficient 2010-2014

```
med.fit<-lm(Gini_Coefficient~Jim_Crow,data=merged)
summary(med.fit)
```

```
##
## Call:
## lm(formula = Gini_Coefficient ~ Jim_Crow, data = merged)
##
## Residuals:
```

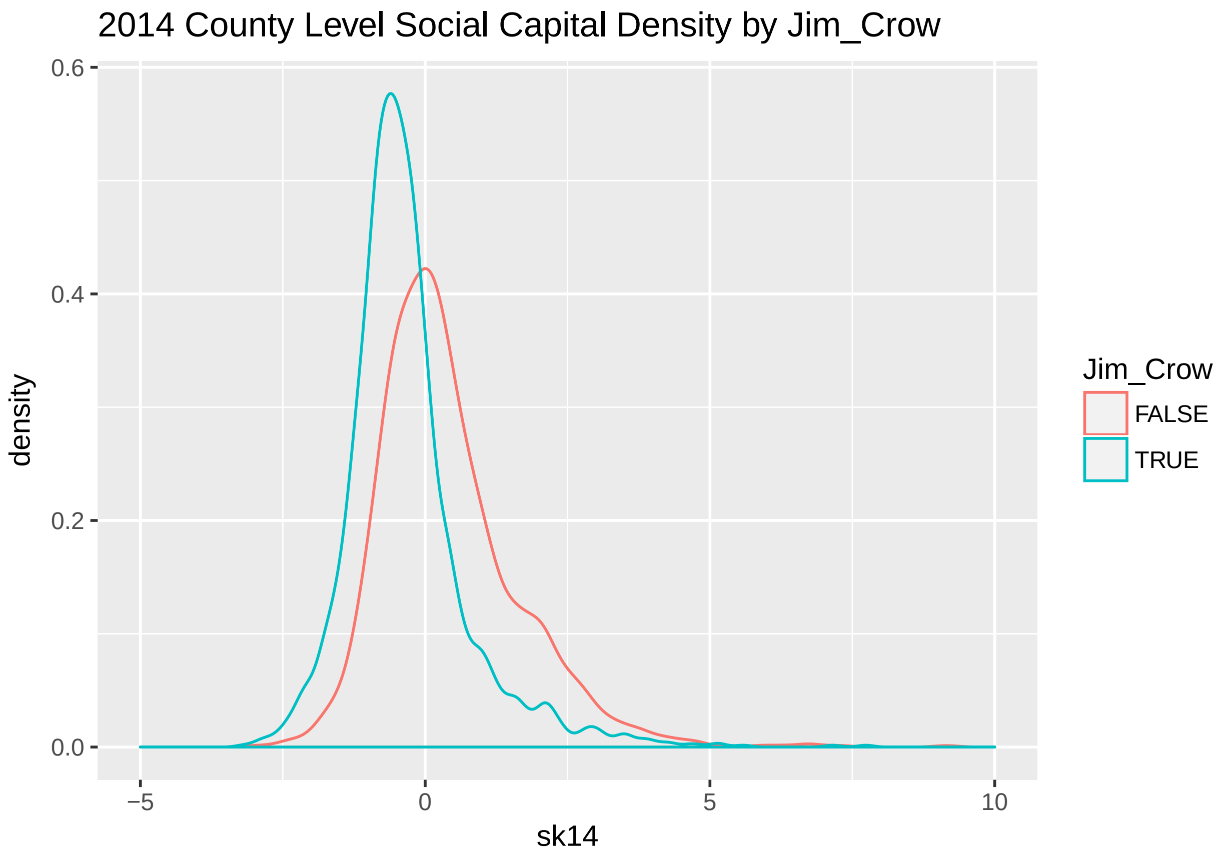
```
##       Min        1Q    Median        3Q       Max
## -0.105169 -0.021469 -0.001669  0.019711  0.203431
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4307495  0.0008921  482.87   <2e-16 ***
## Jim_CrowTRUE 0.0177197  0.0012019   14.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03332 on 3104 degrees of freedom
## Multiple R-squared:  0.06544,    Adjusted R-squared:  0.06514
## F-statistic: 217.4 on 1 and 3104 DF,  p-value: < 2.2e-16
```

2.3 step3 X—->Y

4. Regression Jim Crow Laws –> social capital 1997, 2005, 2009, 2014 (do for each year separately)

```
ggplot(data=merged,aes(x=sk14,color=Jim_Crow))+
  geom_density()+
  labs(title="2014 County Level Social Capital Density by Jim_Crow")+
  xlim(-5, 10)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
```



```r
cat("summary statistics of 2014 social capital index by group")
```

```
## summary statistics of 2014 social capital index by group
```

```r
describeBy(merged$sk14,list(jim_crow=merged$Jim_Crow))
```

```
##
##  Descriptive statistics by group
## jim_crow: FALSE
##    vars    n mean   sd median trimmed mad   min  max range skew kurtosis
## X1    1 1395  0.4 1.23   0.19    0.29   1 -2.95 9.15  12.1 1.35     4.15
##      se
## X1 0.03
## --------------------------------------------------------
## jim_crow: TRUE
##    vars    n  mean   sd median trimmed  mad   min   max range skew
## X1    1 1711 -0.32 1.19  -0.48   -0.44 0.69 -3.18 21.81 24.99 4.96
##    kurtosis   se
## X1    72.25 0.03
```

year 1999

```r
summary(lm(sk97~Jim_Crow,data=merged))
```

```
##
## Call:
## lm(formula = sk97 ~ Jim_Crow, data = merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7466 -0.8196 -0.2151  0.5571  7.5504
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.69023    0.03463   19.93   <2e-16 ***
## Jim_CrowTRUE -1.25434    0.04665  -26.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.293 on 3104 degrees of freedom
## Multiple R-squared:  0.1889, Adjusted R-squared:  0.1886
## F-statistic: 722.9 on 1 and 3104 DF,  p-value: < 2.2e-16
```

year 2005

```r
summary(lm(sk05~Jim_Crow,data=merged))
```

```
##
## Call:
## lm(formula = sk05 ~ Jim_Crow, data = merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6864 -0.7663 -0.1879  0.5501 14.8031
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.61999    0.03411   18.18   <2e-16 ***
## Jim_CrowTRUE -1.12685    0.04596  -24.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.274 on 3104 degrees of freedom
## Multiple R-squared:  0.1623, Adjusted R-squared:  0.162 
## F-statistic: 601.2 on 1 and 3104 DF,  p-value: < 2.2e-16
```

year 2009

```
summary(lm(sk09~Jim_Crow,data=merged))
```

```
## 
## Call:
## lm(formula = sk09 ~ Jim_Crow, data = merged)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5520 -0.7559 -0.2011  0.4926 17.8138 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.45907    0.03408   13.47   <2e-16 ***
## Jim_CrowTRUE -0.83234    0.04591  -18.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.273 on 3104 degrees of freedom
## Multiple R-squared:  0.09574,    Adjusted R-squared:  0.09544 
## F-statistic: 328.6 on 1 and 3104 DF,  p-value: < 2.2e-16
```

year 2014

```
summary(lm(sk14~Jim_Crow,data=merged))
```

```
## 
## Call:
## lm(formula = sk14 ~ Jim_Crow, data = merged)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.3573 -0.6947 -0.1905  0.4149 22.1271 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.40433    0.03237   12.49   <2e-16 ***
## Jim_CrowTRUE -0.72256    0.04362  -16.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.209 on 3104 degrees of freedom
## Multiple R-squared:  0.08123,    Adjusted R-squared:  0.08094
## F-statistic: 274.4 on 1 and 3104 DF,  p-value: < 2.2e-16
```

step 4 all together

5. Regression Jim Crow Laws + Gini Coefficient –> Social Capital (2014)

```
out.fit<-lm(sk14~Jim_Crow+Gini_Coefficient,data=merged)
summary(out.fit)
```

```
##
## Call:
## lm(formula = sk14 ~ Jim_Crow + Gini_Coefficient, data = merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2751 -0.7033 -0.1946  0.4299 22.1144
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.8751     0.2823   3.100  0.00196 **
## Jim_CrowTRUE       -0.7032     0.0451 -15.590  < 2e-16 ***
## Gini_Coefficient  -1.0930     0.6512  -1.679  0.09333 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.209 on 3103 degrees of freedom
## Multiple R-squared:  0.08207,    Adjusted R-squared:  0.08148
## F-statistic: 138.7 on 2 and 3103 DF,  p-value: < 2.2e-16
```

Mediation Analysis

6. Mediation analysis for Jim Crow Laws + Gini Coefficient –> Social Capital (2014)

```
med.out<-mediation::mediate(med.fit,out.fit,treat = "Jim_Crow",mediator = "Gini_Coefficient")
summary(med.out)
```

```
##
## Causal Mediation Analysis
##
## Quasi-Bayesian Confidence Intervals
##
##                 Estimate 95% CI Lower 95% CI Upper p-value
## ACME            -0.01951     -0.04166         0.00   0.086 .
## ADE             -0.70447     -0.79332        -0.62  <2e-16 ***
## Total Effect    -0.72399     -0.80843        -0.64  <2e-16 ***
## Prop. Mediated  0.02643     -0.00315         0.06   0.086 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sample Size Used: 3106
##
##
```

```
## Simulations: 1000
```

This method is using simulation samples to construct the confidence interval for direct effect and mediate effect.

ADE stands for average direct effect $DirectEffect = Y_i(1, M_i(t)) - Y_i(0, M_i(t)))$

ACME means the average causal mediation effects $MediateEffect = Y_i(t, M_i(1)) - Y_i(t, M_i(0)))$

$ Total Effect = Mediation Effect + Direct Effect$ prop.mediated stands for the proportion of meidation effect. This is a ratio $ prop.mediate = mediate.effect/direct.effect$

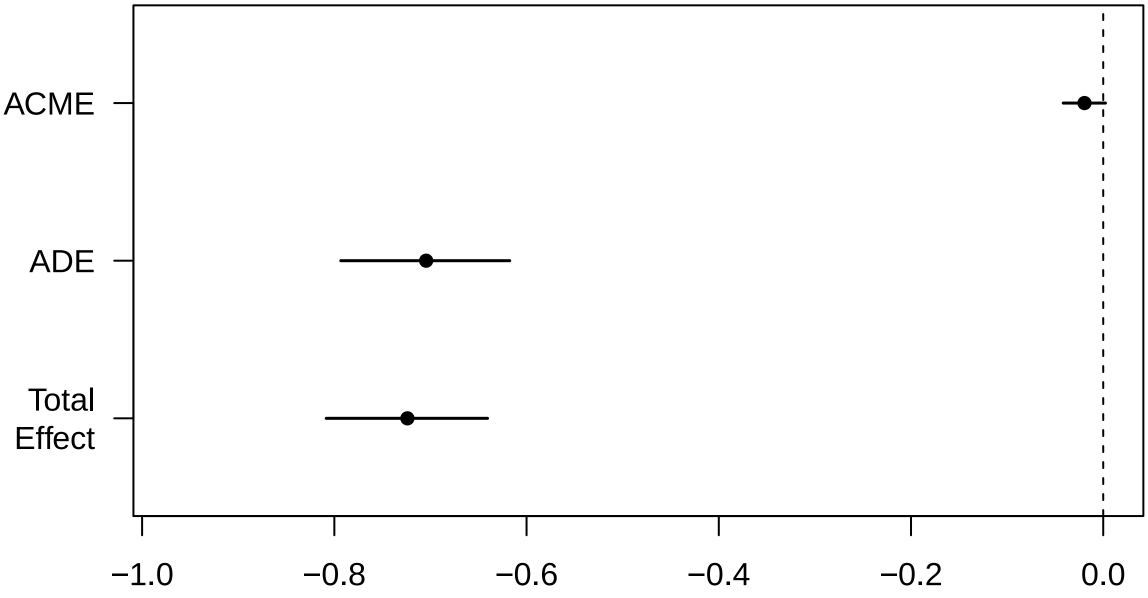
```r
cat("Prop Mediated is a ratio of two estimates, which are known to have a very high variance especially
```

```
## Prop Mediated is a ratio of two estimates, which are known to have a very high variance especially wh
```

```r
cat(" I would focus on the point estimate of this quantity rather than its CI.  The most important thin
```

```
##  I would focus on the point estimate of this quantity rather than its CI.  The most important thing i
```

```r
plot(med.out)
```


ACME
ADE
Total
Effect