TEXAS TECH UNIVERSITY™

STATS CAMP

statscamp.org

# Lecture 2: Simple Mediation Analysis

Kyle M. Lang

*Institute for Measurement, Methodology, Analysis & Policy*
*Texas Tech University*
*Lubbock, TX*

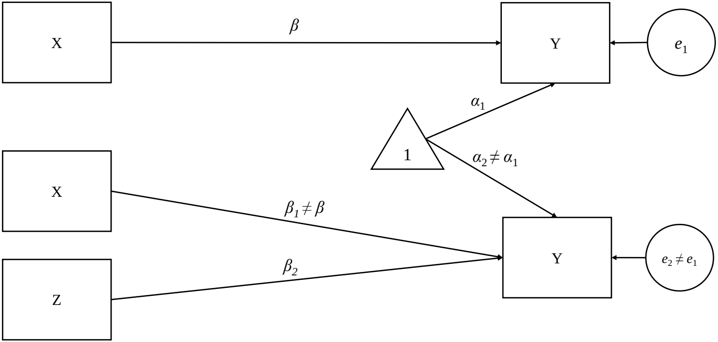
*2016 Stats Camp*

# Outline

- Little review of ordinary least squares (OLS) regression
- Baron and Kenny (1986) Causal Steps approach
- The Sobel (1982) Z test

# A Wee Bit o' Regression

$$Y = \alpha_1 + \beta + e_1 \tag{1}$$
$$Y = \alpha_2 + \beta_1 X + \beta_2 Z + e_2 \tag{2}$$



# Example

```r
## Fit model 1:
fit1 <- lm(y ~ x, data = dat1)
## Fit model 2:
fit2 <- lm(y ~ x + z, data = dat1)
## Look at the results:
summary(fit1)
```

```
Call:
lm(formula = y ~ x, data = dat1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.56967 -0.61039  0.00362  0.61399  2.75563

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03921    0.04201   0.933    0.351
x            0.32496    0.03951   8.224 1.72e-15 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05
        .   0.1         1
```

# Example

```
Residual standard error: 0.939 on 498 degrees of freedom
Multiple R^2:  0.1196, Adjusted R^2:  0.1178
F-statistic: 67.64 on 1 and 498 DF,  p-value: 1.724e-15
```

```
summary(fit2)
```

```
Call:
lm(formula = y ~ x + z, data = dat1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.72281 -0.57410 -0.05566  0.68002  2.75771

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04322    0.04046   1.068    0.286
x            0.24003    0.04034   5.950 5.06e-09 ***
z            0.26533    0.04188   6.335 5.32e-10 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05
        .    0.1        1
```

# Example

```
Residual standard error: 0.9042 on 497 degrees of freedom
Multiple R^2:  0.1854, Adjusted R^2:  0.1821
F-statistic: 56.54 on 2 and 497 DF,  p-value: < 2.2e-16
```

```r
## Get the coefficients:
coef(fit1)
```

```
(Intercept)           x
 0.03921385  0.32496211
```

```r
coef(fit2)
```

```
(Intercept)           x           z
 0.04321606  0.24002536  0.26532809
```

```r
## Get the standard errors:
sqrt(diag(vcov(fit1)))
```

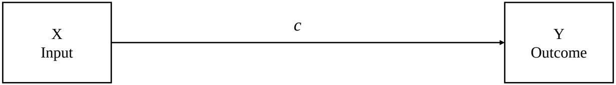
```
(Intercept)           x
 0.04201438  0.03951356
```

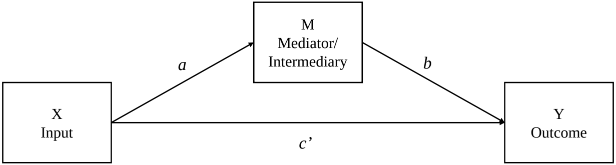
# Example

```r
sqrt(diag(vcov(fit2)))
```

```
(Intercept)           x           z
 0.04045992  0.04034027  0.04188304
```

# Path Diagrams



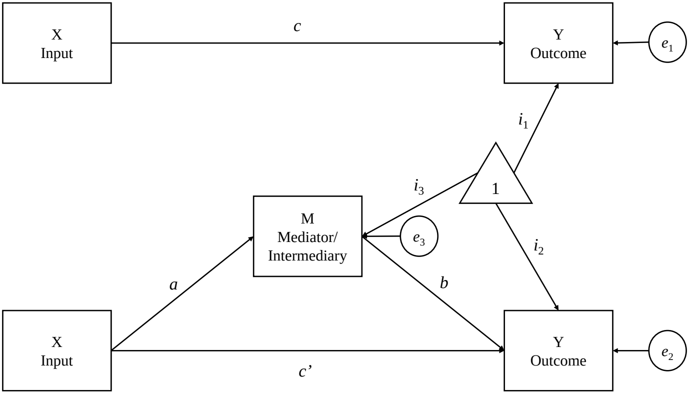
# Necessary Equations

To get all the pieces of the preceding diagram, we'll need to fit three equations.

$$Y = i_1 + cX + e_1 \tag{3}$$
$$Y = i_2 + c'X + bM + e_2 \tag{4}$$
$$M = i_3 + aX + e_3 \tag{5}$$

- Equation 3 gives us the total effect ($c$).
- Equation 4 gives us the direct effect ($c'$) and the partialled effect of the mediator on the outcome ($b$).
- Equation 5 gives us the effect of the input on the outcome ($a$).

# More Complex Path Diagram



# Two Measures of Indirect Effect

Indirect effects can be quantified in two different ways:

$$IE_{diff} = c - c' \tag{6}$$

$$IE_{prod} = a \cdot b \tag{7}$$

$IE_{diff}$ and $IE_{prod}$ are equivalent in simple mediation.

- Both give us information about the proportion of the total effect that is transmitted through the intermediary variable.
- $IE_{prod}$ provides a more direct representation of the actual pathway we're interested in testing.
- $IE_{diff}$ gets at our desired hypothesis indirectly.

# The *Causal Steps Approach*

Baron and Kenny (1986, p. 1176) describe three/four conditions as being sufficient to demonstrate statistical "mediation."

1. Variations in levels of the independent variable significantly account for variations in the presumed mediator (i.e., Path $a$).
   - Need a significant $a$ path.
2. Variations in the mediator significantly account for variations in the dependent variable (i.e., Path $b$).
   - Need a significant $b$ path.
3. When Paths $a$ and $b$ are controlled, a previously significant relation between the independent and dependent variables is no longer significant.
   - Need a significant total effect
   - The direct effect must be "less" than the total effect

# Example

```r
dat1 <- readRDS("../data/adamsKlpsScaleScore.rds")
## Check pre-conditions:
mod1 <- lm(policy ~ polAffil, data = dat1)
mod2 <- lm(policy ~ sysRac, data = dat1)
mod3 <- lm(sysRac ~ polAffil, data = dat1)
## Partial out the mediator's effect:
mod4 <- lm(policy ~ sysRac + polAffil, data = dat1)
summary(mod1)
```

```
Call:
lm(formula = policy ~ polAffil, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7357 -0.8254  0.0643  0.6827  3.2481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.71516    0.35648   7.617 3.32e-11 ***
polAffil     0.23675    0.07775   3.045   0.0031 **
---
```

# Example

```
Signif. codes:  0  ***  0.001  **  0.01  *  0.05
       .   0.1       1

Residual standard error: 1.134 on 85 degrees of freedom
Multiple R^2:  0.09836,  Adjusted R^2:  0.08775
F-statistic: 9.273 on 1 and 85 DF,  p-value: 0.003096
```

```
summary(mod2)
```

```
Call:
lm(formula = policy ~ sysRac, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7700 -0.5593  0.0255  0.6277  3.6835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9295     0.3896   2.386   0.0193 *
sysRac        0.7557     0.1014   7.450 7.14e-11 ***
---
```

# Example

```
Signif. codes:  0  ***  0.001  **  0.01  *  0.05
      .  0.1      1

Residual standard error: 0.9286 on 85 degrees of freedom
Multiple R^2:  0.395,  Adjusted R^2:  0.3879
F-statistic:  55.5 on 1 and 85 DF,  p-value: 7.145e-11
```

```
summary(mod3)
```

```
Call:
lm(formula = sysRac ~ polAffil, data = dat1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.44714 -0.50502  0.05286  0.54498  2.25286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.60605    0.28489   9.147 2.72e-14 ***
polAffil     0.25685    0.06213   4.134 8.34e-05 ***
---
```

# Example

```
Signif. codes:  0  ***  0.001  **  0.01  *  0.05
        .  0.1       1

Residual standard error: 0.906 on 85 degrees of freedom
Multiple R^2:  0.1674, Adjusted R^2:  0.1576
F-statistic: 17.09 on 1 and 85 DF,  p-value: 8.336e-05
```

```
summary(mod4)
```

```
Call:
lm(formula = policy ~ sysRac + polAffil, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7156 -0.6043  0.0262  0.6474  3.7992

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.83266    0.41246   2.019   0.0467 *
sysRac       0.72236    0.11148   6.480 5.93e-09 ***
polAffil     0.05121    0.06998   0.732   0.4663
---
```

# Example

```
Signif. codes:  0     ***     0.001     **     0.01     *     0.05
        .     0.1           1

Residual standard error: 0.9312 on 84 degrees of freedom
Multiple R^2:  0.3989, Adjusted R^2:  0.3845
F-statistic: 27.87 on 2 and 84 DF,  p-value: 5.211e-10
```

# Example

```
## Extract important parameter estimates:
a <- coef(mod3)["polAffil"]
b <- coef(mod4)["sysRac"]
c <- coef(mod1)["polAffil"]
cPrime <- coef(mod4)["polAffil"]
## Compute indirect effects:
ieDiff <- c - cPrime
ieProd <- a * b
ieDiff
```

```
 polAffil
0.1855374
```

```
ieProd
```

```
 polAffil
0.1855374
```

# Sobel's Z

In the previous example, do we have a *significant* indirect effect?

- The direct effect is substantially smaller than the total effect, but is the difference statistically significant?
- Sobel (1982) developed an asymptotic standard error for $IE_{prod}$ that we can use to assess this hypothesis.

$$SE_{sobel} = \sqrt{a^2 \cdot SE_b^2 + b^2 \cdot SE_a^2} \tag{8}$$

$$Z_{sobel} = \frac{ab}{SE_{sobel}} \tag{9}$$

$$95\% CI_{sobel} = ab \pm 1.96 \cdot SE_{sobel} \tag{10}$$

# Example

```
## Calculate Sobel's Z:
seA <- sqrt(diag(vcov(mod3)))["polAffil"]
seB <- sqrt(diag(vcov(mod4)))["sysRac"]
sobelSE <- sqrt(b^2 * seA^2 + a^2 * seB^2)
sobelZ <- ieProd / sobelSE
sobelZ
```

```
polAffil
 3.48501
```

```
sobelP <- 2 * pnorm(sobelZ, lower = FALSE)
sobelP
```

```
    polAffil
0.0004921178
```

```
sobelUB <- ieProd + 1.96 * sobelSE
sobelLB <- ieProd - 1.96 * sobelSE
## 95% Sobel CI:
c(sobelLB, sobelUB)
```

```
  polAffil   polAffil
0.08118957 0.28988525
```

# Alternative formulations

There are, at least, two alternative formulation of the Sobel SE.

- The first is due to Aroian (1947):

$$SE_{aroian} = \sqrt{a^2 \cdot SE_b^2 + b^2 \cdot SE_a^2 + SE_a^2 \cdot SE_b^2} \tag{11}$$

- The other is due to Goodman (1960):

$$SE_{goodman} = \sqrt{a^2 \cdot SE_b^2 + b^2 \cdot SE_a^2 - SE_a^2 \cdot SE_b^2} \tag{12}$$

- The Goodman formulation is unbiased, but can lead to negative estimated SEs.
- The Aroian formulation is recommended since it does not assume that $SE_a^2 \cdot SE_b^2$ asymptotically vanishes.
- The Aroian and Sobel versions will probably perform equivalently with $N \geq 50$.

NOTE: The information on this slide was drawn from `http://quantpsy.org/sobel/sobel.htm`

# Example

```
## Calculate Aroian's Z:
aroianSE <- sqrt(b^2 * seA^2 + a^2 * seB^2 + seA^2 * seB^2)
aroianZ <- ieProd / aroianSE
aroianZ
```

```
polAffil
3.455885
```

```
aroianP <- 2 * pnorm(aroianZ, lower = FALSE)
aroianP
```

```
    polAffil
0.0005484897
```

```
aroianUB <- ieProd + 1.96 * aroianSE
aroianLB <- ieProd - 1.96 * aroianSE
## 95% Aroian CI:
c(aroianLB, aroianUB)
```

```
  polAffil   polAffil
0.08031014 0.29076468
```

# Example

```
## Calculate Goodman's Z:
goodSE <- sqrt(b^2 * seA^2 + a^2 * seB^2 - seA^2 * seB^2)
goodZ <- ieProd / goodSE
goodZ
```

```
polAffil
3.514885
```

```
goodP <- 2 * pnorm(goodZ, lower = FALSE)
goodP
```

```
    polAffil
0.0004399441
```

```
goodUB <- ieProd + 1.96 * goodSE
goodLB <- ieProd - 1.96 * goodSE
## 95% Goodman CI:
c(goodLB, goodUB)
```

```
  polAffil   polAffil
0.08207647 0.28899835
```

# Compare Results

All three formulations give similar answers for this problem:

| | Z-Stat | P-Value | 95% CI LB | 95% CI UB |
|---|---|---|---|---|
| Sobel | 3.485 | 0.000 | 0.081 | 0.290 |
| Aroian | 3.456 | 0.001 | 0.080 | 0.291 |
| Goodman | 3.515 | 0.000 | 0.082 | 0.289 |

Table: Mediation Test Results Using Three SE Formulations

# Example

```
## Check preconditions:
mod1 <- lm(policy ~ revDisc, data = dat1)
mod2 <- lm(sysRac ~ revDisc, data = dat1)
mod3 <- lm(policy ~ sysRac, data = dat1)
summary(mod1)
```

```
Call:
lm(formula = policy ~ revDisc, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6840 -0.8584  0.1331  0.7941  3.2112

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.22306    0.29023  14.551   <2e-16 ***
revDisc     -0.13904    0.07465  -1.862    0.066 .
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05
        .   0.1         1

Residual standard error: 1.17 on 85 degrees of freedom
```

```
Multiple R^2:  0.03921,  Adjusted R^2:  0.02791
F-statistic: 3.469 on 1 and 85 DF,  p-value: 0.06599
```

```
summary(mod2)
```

```
Call:
lm(formula = sysRac ~ revDisc, data = dat1)

Residuals:
    Min       1Q   Median       3Q      Max
-2.2292 -0.6877  0.0492  0.5861  2.3708

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.15540    0.24043   17.28   <2e-16 ***
revDisc     -0.12615    0.06184   -2.04   0.0445 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05
        .  0.1       1

Residual standard error: 0.9695 on 85 degrees of freedom
Multiple R^2:  0.04667,  Adjusted R^2:  0.03545
```

# Example

```
F-statistic: 4.161 on 1 and 85 DF,  p-value: 0.04447
```

```
summary(mod3)
```

```
Call:
lm(formula = policy ~ sysRac, data = dat1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7700 -0.5593  0.0255  0.6277  3.6835

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9295     0.3896   2.386   0.0193 *
sysRac        0.7557     0.1014   7.450 7.14e-11 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05
        .   0.1        1

Residual standard error: 0.9286 on 85 degrees of freedom
Multiple R^2:  0.395,  Adjusted R^2:  0.3879
F-statistic:  55.5 on 1 and 85 DF,  p-value: 7.145e-11
```

# Example

```
## Fit partial model:
mod4 <- lm(policy ~ revDisc + sysRac, data = dat1)
## Extract parameter estimates:
a <- coef(mod2)["revDisc"]
b <- coef(mod4)["sysRac"]
seA <- sqrt(diag(vcov(mod2)))["revDisc"]
seB <- sqrt(diag(vcov(mod4)))["sysRac"]
sobelSE <- sqrt(a^2 * seB^2 + b^2 * seA^2)
sobelZ <- (a * b) / sobelSE
sobelZ
```

```
  revDisc
-1.960368
```

```
sobelP <- 2 * pnorm(sobelZ, lower = TRUE)
sobelP
```

```
  revDisc
0.0499528
```

# Practice

```
mod1 <- lm(y ~ x, data = dat1)
mod2 <- lm(m ~ x, data = dat1)
mod3 <- lm(y ~ m + x, data = dat1)
round(summary(mod1)$coef, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.033      0.038  -0.880    0.379
x              0.475      0.038  12.455    0.000
```

```
round(summary(mod2)$coef, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.044       0.04  -1.104     0.27
x              0.504       0.04  12.442     0.00
```

```
round(summary(mod3)$coef, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.021      0.036  -0.588    0.557
m              0.270      0.041   6.663    0.000
x              0.339      0.042   8.097    0.000
```

# References


Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics*, 265–271.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, *55*(292), 708–713.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*(1982), 290–312.