# ✪ Psychological Statistics ✪

# Week 10: *Nonlinear Regression, Partial Correlation, & Mediation Analysis*


- Edited by Prof. **Changwei Wu**
- Graduate Institute of Mind, Brain and Consciousness (***GIMBC***), Taipei Medical University


In [1]:
```
### [ Setup the working directory ]

setwd("/Users/wesley/[Course]/Python/R_Script")
getwd()
```

'/Users/wesley/[Course]/Python/R_Script'

In [3]:
```
### [ Loading the required libraries ]

library("dplyr")
library("rstatix")
library("nlstools")
library("ppcor")
library("lavaan")
library("mediation")
library("gvlma")
```

---

## (1) Nonlinear Regression: Nonlinear Least Squares (NLS)

---

> ### *[ Hypothesis ]* The bone lenth of jaw is a function of age in deer. Theory indicates that the relationship is an asymptotic exponential with 3 parameters. Please check whether this dataset fits this expectation. *(2-tailed)*
>
> - Null hypothesis ***$H_0$*** **:** The data does not fit the asymptotic exponential relation.
> - Alternative hyp. ***$H_1$*** **:** The data fit the asymptotic exponential relation.

In [4]:
```
### Data Loading ###

deer <- read.csv("jaws.csv")
attach(deer)
```

In [5]:
```r
### Nonlinear modeling (1) ### Y = a-b*exp^(-c*X))

model1 <- nls(bone ~ a-b*exp(-c*age), start = list(a=120,b=110,c=0.064))

summary(model1)
```

```
Formula: bone ~ a - b * exp(-c * age)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a 115.2528     2.9139   39.55  < 2e-16 ***
b 118.6875     7.8925   15.04  < 2e-16 ***
c   0.1235     0.0171    7.22 2.44e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.21 on 51 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 2.391e-06
```

In [7]:
```r
### Nonlinear modeling (2) ### Y = d*(1-exp^(-e*X))

model2 <- nls(bone ~ d*(1-exp(-e*age)), start = list(d=120,e=0.064))

summary(model2)
AIC(model1,model2)
```

```
Formula: bone ~ d * (1 - exp(-e * age))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
d 115.58056    2.84365  40.645  < 2e-16 ***
e   0.11882    0.01233   9.635 3.69e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.1 on 52 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 1.369e-06
```

A data.frame: 2 × 2

| | df | AIC |
|---|---|---|
| | <dbl> | <dbl> |
| **model1** | 4 | 436.8894 |
| **model2** | 3 | 435.0823 |

In [8]:

```r
### Calculation of Effect Size (R-squared) ###

null.model <- lm(bone ~ 1) # calculate the total SS from null model

summary.aov(null.model)

summary(model2) # compare the SS_residuals

### SST = SSR + SSE ###

SST = 59008
MSE = (13.1)^2
SSE = 52 * MSE

SSR = SST-SSE
(R_squared <- SSR/SST)
```

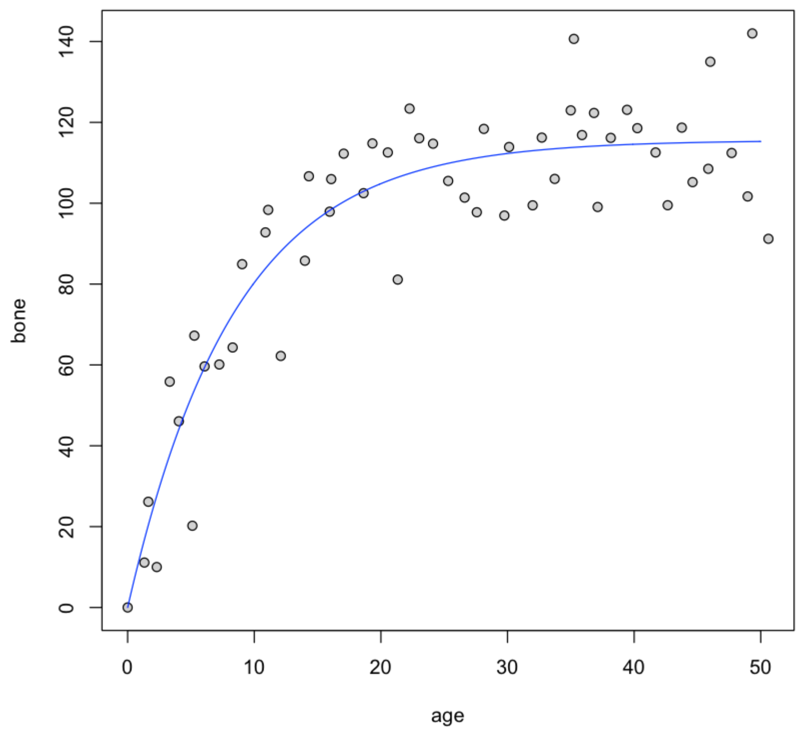
```
            Df Sum Sq Mean Sq F value Pr(>F)
Residuals   53  59008    1113


Formula: bone ~ d * (1 - exp(-e * age))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
d 115.58056    2.84365  40.645  < 2e-16 ***
e   0.11882    0.01233   9.635 3.69e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.1 on 52 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 1.369e-06
```

0.848771014099783

In [9]:
```
### Visualization ###

nlstools::confint2(model2)

par(mfrow=c(1,1))
plot(age, bone, pch=21,bg="lightgrey")
av <- seq(0,50,0.1)
cv <- predict(model2,list(age=av))
lines(av,cv,col="blue")
```

A matrix: 2 × 2 of type dbl

| | 2.5 % | 97.5 % |
|---|---|---|
| **d** | 109.87435953 | 121.2867506 |
| **e** | 0.09407099 | 0.1435604 |




## ~ *Report* ~

- **The asymptotic exponential model y= a(1-exp(-bx)) fits the bone-length data well (n=54, *p* < 0.001), explaining 84.9% of the total variation with a = 115.58±2.84 and b = 0.12±0.01.**

In [ ]:
```
#--- generalized additive models GAM ---#
### When we know it is nonlinear without theoretical model ###

hump <- read.csv("hump.csv")
attach(hump)

model <- mgcv::gam(y~s(x))

plot(model,col="blue")
points(x,y-mean(y),pch=21,bg="yellow")

summary(model)
```

# (2) Partial Correlation

***[ Scenario →] We would like to estimate the relations of math aptitude (SAT-Q) and math achievement (CLEP) on the mathematic performance (GPA) among the first-year undergraduate students.***

- Null hypothesis $H_0$ **:** No relationship between GPA and CLEP → $r = 0$
- Alternative hyp. $H_1$ **:** There is (linear) relationship between GPA and CLEP → $r \neq 0$

In [10]:
```
### [ Step.1 ] Data Loading

# → Loading the dataset
examData <- read.csv("Aptitude_Achievement.csv",  header = TRUE)
examData <- examData[,2:4]
#examData %>% select(-Person)
examData %>% head(4)
```

A data.frame: 4 × 3

| | SAT.Q | CLEP | GPA |
|---|---|---|---|
| | <int> | <int> | <dbl> |
| **1** | 500 | 30 | 2.8 |
| **2** | 550 | 32 | 3.0 |
| **3** | 450 | 28 | 2.9 |
| **4** | 400 | 25 | 2.8 |

In [11]:
```
### [ Step.2 ] Check assumptions

#----- (a) Outliers -----#
examData %>% identify_outliers(SAT.Q)
examData %>% identify_outliers(CLEP)
examData %>% identify_outliers(GPA)

#----- (b) Normality -----#
examData %>% shapiro_test(SAT.Q)
examData %>% shapiro_test(CLEP)
examData %>% shapiro_test(GPA)
```

A data.frame: 0 × 5

| SAT.Q | CLEP | GPA | is.outlier | is.extreme |
|---|---|---|---|---|
| <int> | <int> | <dbl> | <lgl> | <lgl> |

A data.frame: 0 × 5

| SAT.Q | CLEP | GPA | is.outlier | is.extreme |
|---|---|---|---|---|
| <int> | <int> | <dbl> | <lgl> | <lgl> |

A data.frame: 0 × 5

| SAT.Q | CLEP | GPA | is.outlier | is.extreme |
|---|---|---|---|---|
| <int> | <int> | <dbl> | <lgl> | <lgl> |

A tibble: 1 × 3

| variable | statistic | p |
|---|---|---|
| <chr> | <dbl> | <dbl> |
| SAT.Q | 0.9656533 | 0.8478815 |

A tibble: 1 × 3

| variable | statistic | p |
|---|---|---|
| <chr> | <dbl> | <dbl> |
| CLEP | 0.9452356 | 0.6125938 |

A tibble: 1 × 3

| variable | statistic | p |
|---|---|---|
| <chr> | <dbl> | <dbl> |
| GPA | 0.9059681 | 0.2544281 |

In [12]:
```r
### [ Step.3 (0)] Partial correlation -- original Correlation matrix

(rmap <- cor(examData))

#examData %>% cor_test(GPA, CLEP, SAT.Q)
```

A matrix: 3 × 3 of type dbl

| | SAT.Q | CLEP | GPA |
|---|---|---|---|
| **SAT.Q** | 1.0000000 | 0.8745001 | 0.7180459 |
| **CLEP** | 0.8745001 | 1.0000000 | 0.8762720 |
| **GPA** | 0.7180459 | 0.8762720 | 1.0000000 |

### *[ Speculation 1 ]* The GPA scores has positive correlation with the math achievement (CLEP) after controlling the math aptitude (SAT-Q) among the first-year undergraduate students. *(1-tailed)*

- Null hypothesis $\boldsymbol{H_0}$ **:** No relationship between GPA and CLEP after controlling SATQ → $\boldsymbol{r_{GC \cdot S}} = 0$
- Alternative hyp. $\boldsymbol{H_1}$ **:** There is (linear) relationship between GPA and CLEP after controlling SATQ → $\boldsymbol{r_{GC \cdot S}} \neq 0$

In [13]:
```r
### [ Step.3 (1a)] Partial correlation -- calculated from equation & correlati

r_GC = rmap[2,3]
r_GS = rmap[1,3]
r_CS = rmap[1,2]

(partCor_GC.S <- (r_GC - r_GS*r_CS)/(sqrt(1-r_GS^2)*sqrt(1-r_CS^2)))
```

0.735659966483154

In [14]:
```r
### [ Step.3 (1b) ] Partial correlation -- Through ppcor

ppcor::pcor(examData)$estimate # export the partial correlation matrix

ppcor::pcor.test(examData$GPA, examData$CLEP, examData$SAT.Q)

# after controlling SAT.Q from both variables, the correlation is still signifi
# This is the first-order partial correlation by controlling single variable.
```

A matrix: 3 × 3 of type dbl

| | SAT.Q | CLEP | GPA |
|---|---|---|---|
| **SAT.Q** | 1.0000000 | 0.7314811 | -0.2064849 |
| **CLEP** | 0.7314811 | 1.0000000 | 0.7356600 |
| **GPA** | -0.2064849 | 0.7356600 | 1.0000000 |

A data.frame: 1 × 6

| estimate | p.value | statistic | n | gp | Method |
|---|---|---|---|---|---|
| <dbl> | <dbl> | <dbl> | <int> | <dbl> | <fct> |
| 0.73566 | 0.02387202 | 2.873508 | 10 | 1 | pearson |

In [15]:
```r
### [ Step.3 (1c) ] Partial correlation -- Through regression residuals

resG.S <- residuals(lm(GPA ~ SAT.Q, data = examData))  # regress out the effect
resC.S <- residuals(lm(CLEP ~ SAT.Q, data = examData)) # regress out the effect

cor.test(resG.S,resC.S) # Using the residuals to conduct the correlation analys
```

```
	Pearson's product-moment correlation

data:  resG.S and resC.S
t = 3.0719, df = 8, p-value = 0.0153
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1975258 0.9330883
sample estimates:
    cor
0.73566
```

## ~ *Report (Partial correlation)* ~

- **The GPA was correlated with CLEP after controlling SAT-Q ($r$ = 0.736, $p$ < 0.02).**

### *[ Speculation 2 ]* Someone argues that the SAT-Q will add notihing to the prediction of GPA once we know CLEP. Therefore, in this case, we will hold CLEP constant for the SAT-Q and see whether the SAT still has correlation with GPA (by regressing out the CLEP scores from SAT-Q, but not from GPA). *(1-tailed)*

- Null hypothesis $H_0$ **:** No relationship between GPA and SATQ, with controlling CLEP from SATQ → $r_{G(S \cdot C)} = 0$
- Alternative hyp. $H_1$ **:** There is (linear) relationship between between GPA and SATQ, with controlling CLEP from SATQ → $r_{G(S \cdot C)} \neq 0$

In [16]:
```
### [ Step.3 (2a) ] Semi-Partial correlation -- Calculated from equation

(rmap <- cor(examData))
#(r_GC = rmap[2,3])
#(r_GS = rmap[1,3])
#(r_CS = rmap[1,2])

(semipartCor_GS.C <- (r_GS - r_GC*r_CS)/(sqrt(1-r_CS^2)))
```

A matrix: 3 × 3 of type dbl

| | SAT.Q | CLEP | GPA |
|---|---|---|---|
| **SAT.Q** | 1.0000000 | 0.8745001 | 0.7180459 |
| **CLEP** | 0.8745001 | 1.0000000 | 0.8762720 |
| **GPA** | 0.7180459 | 0.8762720 | 1.0000000 |

-0.0994878645242558

In [17]:
```
### [ Step.3 (2b) ] Semi-Partial correlation -- Through ppcor

ppcor::spcor.test(examData$GPA, examData$SAT.Q, examData$CLEP)

# [!Attention!] the order of placing variables will change the results!!
```

A data.frame: 1 × 6

| estimate | p.value | statistic | n | gp | Method |
|---|---|---|---|---|---|
| <dbl> | <dbl> | <dbl> | <int> | <dbl> | <fct> |
| -0.09948786 | 0.7989893 | -0.2645326 | 10 | 1 | pearson |

In [19]:
```
### [ Step.3 (2c) ] Semi-Partial correlation -- Through regression residuals

# only regress out the effect of CLEP from SAT-Q
resS.C <- residuals(lm(SAT.Q ~ CLEP, data = examData))

cor.test(examData$GPA,resS.C)
```

```
        Pearson's product-moment correlation

data:  examData$GPA and resS.C
t = -0.2828, df = 8, p-value = 0.7845
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6861346  0.5655656
sample estimates:
        cor
-0.09948786
```

### *~ Report (Semi-partial correlation) ~*

- **The GPA was not associated with SAT-Q ($r$ = -0.099), when the SAT-Q was controlled by CLEP ( $p$ > 0.05).**

# (3) Mediation Analysis

### 3 steps of Regression

### *[ Hypothesis ]* The mediator (M) intervenes the relationship between X and Y. *(2-tailed)*

- Null hypothesis $H_0$ **:** No mediation effect was observed between X and Y (a*b = 0).
- Alternative hyp. $H_1$ **:** M is mediating the X-Y relationship (a*b ≠ 0).

In [20]:
```r
### [ Step.1 ] Data Loading

# → Loading the dataset
XMY <- read.csv("PlannedBehavior.csv", header = TRUE)
XMY %>% head(4)
```

A data.frame: 4 × 5

| | attitude | norms | control | intention | behavior |
|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| **1** | 2.31 | 2.31 | 2.03 | 2.50 | 2.62 |
| **2** | 4.66 | 4.01 | 3.63 | 3.99 | 3.64 |
| **3** | 3.85 | 3.56 | 4.20 | 4.35 | 3.83 |
| **4** | 4.24 | 2.25 | 2.84 | 1.51 | 2.25 |

In [21]:
```r
### [ Step.3 ] Analyses of multiple regression models

#--- Regression (1): Total effect, must be significantly different from 0 ---#
path_c <- lm(behavior ~ attitude, data = XMY)
summary(path_c)
# confint(path_c)
```

```
Call:
lm(formula = behavior ~ attitude, data = XMY)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.96792 -0.62906  0.08816  0.66248  1.81281 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.33552    0.22214  10.514  < 2e-16 ***
attitude     0.24126    0.06688   3.608 0.000392 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9087 on 197 degrees of freedom
Multiple R-squared:  0.06197,	Adjusted R-squared:  0.05721 
F-statistic: 13.02 on 1 and 197 DF,  p-value: 0.0003917
```

In [22]:
```
### [ Step.3 ] Analyses of multiple regression models

#--- Regression (2): Path A ---#
path_a <- lm(intention ~ attitude, data = XMY)
summary(path_a)
# confint(path_a)
```

```
Call:
lm(formula = intention ~ attitude, data = XMY)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.00320 -0.54732 -0.07735  0.59008  1.94006 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.46083    0.19393   7.533 1.76e-12 ***
attitude     0.48405    0.05838   8.291 1.73e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7933 on 197 degrees of freedom
Multiple R-squared:  0.2587,	Adjusted R-squared:  0.2549 
F-statistic: 68.74 on 1 and 197 DF,  p-value: 1.728e-14
```

In [23]:
```
### [ Step.3 ] Analyses of multiple regression models

#--- Regression (3): Path B & C' ---#
path_b <- lm(behavior ~ intention + attitude, data = XMY)
summary(path_b)
# confint(path_b)
```

```
Call:
lm(formula = behavior ~ intention + attitude, data = XMY)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.01916 -0.57280  0.06326  0.63735  1.73810 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.69617    0.23358   7.262 8.79e-12 ***
intention    0.43767    0.07561   5.788 2.78e-08 ***
attitude     0.02941    0.07196   0.409    0.683    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8419 on 196 degrees of freedom
Multiple R-squared:  0.1989,	Adjusted R-squared:  0.1907 
F-statistic: 24.33 on 2 and 196 DF,  p-value: 3.636e-10
```

In [24]:
```
### [ Step.2 ] Check assumptions if using Sobel Test

library(gvlma)

#----- Check the assumptions of Path A & B -----#
gvlma(path_a)
```

```
Call:
lm(formula = intention ~ attitude, data = XMY)

Coefficients:
(Intercept)     attitude
     1.4608       0.4841


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = path_a)

                    Value p-value                   Decision
Global Stat        7.2598 0.12278    Assumptions acceptable.
Skewness           1.2650 0.26070    Assumptions acceptable.
Kurtosis           1.5705 0.21014    Assumptions acceptable.
Link Function      4.1010 0.04286 Assumptions NOT satisfied!
Heteroscedasticity 0.3233 0.56961    Assumptions acceptable.
```

In [25]:
```
### [ Step.2 ] Check assumptions if using Sobel Test

#----- Check the assumptions of Path A & B -----#
gvlma(path_b)
```

```
Call:
lm(formula = behavior ~ intention + attitude, data = XMY)

Coefficients:
(Intercept)    intention     attitude
    1.69617      0.43767      0.02941


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = path_b)

                     Value p-value                Decision
Global Stat        4.73471 0.31562 Assumptions acceptable.
Skewness           0.87538 0.34947 Assumptions acceptable.
Kurtosis           3.72017 0.05376 Assumptions acceptable.
Link Function      0.09124 0.76260 Assumptions acceptable.
Heteroscedasticity 0.04792 0.82672 Assumptions acceptable.
```

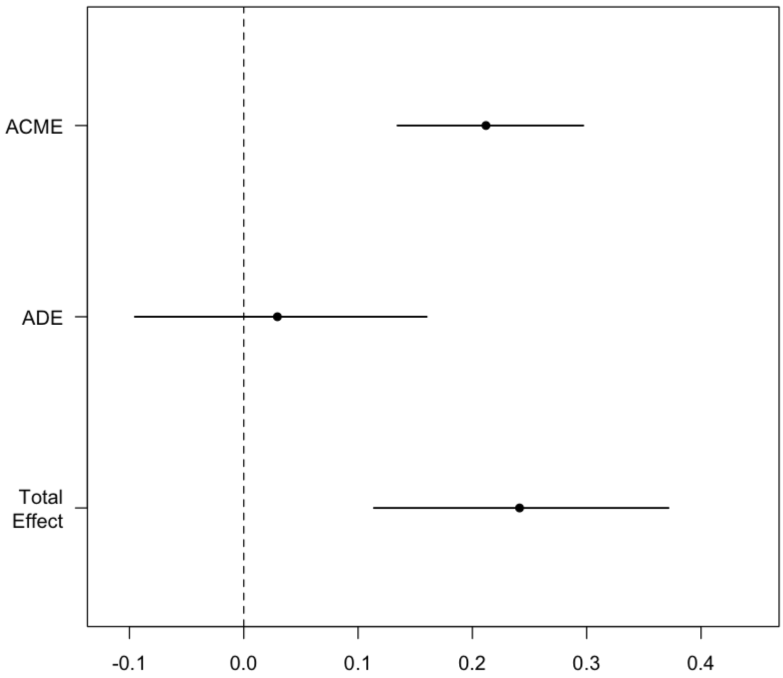
In [26]:
```
### [ Step.4 ] Effect size: Product of Coefficients (Indirect effect)

#---- Compute the indirect effect ----#

(indirect_eff <- path_a$coefficients[2] * path_b$coefficients[2])

#library(multilevel)
#sobel(XMY$X, XMY$M, XMY$Y)
```

**attitude:** 0.21185243108384

In [27]:
```
### [ Step.4 ] Effect size: Proportion Mediated

library(mediation)

set.seed(123)

boot.mediate <- mediate(path_a, path_b, sims=5000, treat='attitude', mediator='
summary(boot.mediate)

# ACME(Average Causal Mediation Effects) = Indirect Effect = a*b
# ADE(Average Direct Effects) = c'
# Total Effect = c
# Proportion Mediated = ACME/TotalEffect
```

```
Running nonparametric bootstrap



Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile Method

               Estimate 95% CI Lower 95% CI Upper p-value    
ACME             0.2119       0.1344         0.30  <2e-16 ***
ADE              0.0294      -0.0953         0.16  0.6440    
Total Effect     0.2413       0.1138         0.37  0.0004 ***
Prop. Mediated   0.8781       0.5200         1.69  0.0004 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sample Size Used: 199


Simulations: 5000
```

In [28]:
```
### [ Step.5 ] Visualization of mediation analysis

plot(boot.mediate)
```




## ~ *Report* ~

- **Based on the 199 observations, the Intention acts as a complete mediator on the relationship between Attitude and Planned behavior.** (indirect effect = 0.212, C.I.[0.132, 0.30], $p < 0.001$)

---

---

*** The follow up is the same mediation analysis using SEM strategy. ***

## (Optional) Structural Equation Modeling (SEM)

## *[ Hypothesis ]* The mediator (M) intervenes the relationship between X and Y. *(2-tailed)*

- Null hypothesis $H_0$ **:** No mediation effect was observed between X and Y (a*b = 0).
- Alternative hyp. $H_1$ **:** M is mediating the X-Y relationship (a*b ≠ 0).

In [ ]:
```
### [ Step 2b.1 ] Data Loading

# → Loading the dataset
XMY <- read.csv("PlannedBehavior.csv", header = TRUE)
XMY %>% head(4)
```

In [ ]:
```
### [ Step 2b.3 ] SEM analysis

library(lavaan)

#--- set up the model ---#
# set up the structural equation of mediation analysis

Behav_model <- "
# Path c' (direct effect)
behavior ~ c * attitude

# Path a
intention ~ a * attitude

# Path b
behavior ~ b * intention

# Indirect Effect (a*b): Sobel test (delta method)
ab := a * b
"
```

In [ ]:
```
### [ Step 2b.3 ] SEM analysis

set.seed(123)

#--- fit/estimate the model ---#

BehavMediat <- sem(Behav_model, data=XMY, bootstrap=5000, se="bootstrap")

# summarize the results
summary(BehavMediat, fit.measures=FALSE, rsquare=TRUE)
```

In [ ]:
```
### [ Step 2b.4 ] Effect size

#--- Calculate Proportion Mediated (= a*b/c) ---#

(Mediat.par <- parameterEstimates(BehavMediat, ci=TRUE, level=0.95, boot.ci.typ

(Proportion_Mediated <- Mediat.par$est[7]/path_c$coefficients[2])
```