PS5

Zhengyu Xiao

2025-11-25

I will use the variables and regression design in my proposed research. Treatment variable: the pressure event intensity: $D_{i,t+k}$ Mediator: the Strategic Calculation of Local Officials/Bureaucrats: M_{it} Outcome: Semantic Compliance in Local Government Documents S_{it} Confounder: X_{it} Collider: K_{it} caused by D and outcome-only shock An independent variable that has an exogenous effect on the outcome variable: Z_{it}^{S} Instrument variable: Z_{it}^{D}

```
# Construct Variables
```

```
set.seed(123)

N <- 300
X_it <- rnorm(N, 0 ,1)
Z_D <- rnorm (N, 0 ,1)
Z_S <- rnorm (N, 0 ,1)

# Treatment (pressure event intensity)
e_D <- rnorm(N, 0, 1)
D   <- 0.80*X_it + 0.90*Z_D + e_D

# Mediator: strategic calculation
e_M <- rnorm(N, 0, 1)
M   <- 0.60*D + 0.50*X_it + e_M

# Outcome = direct(D) + via(M) + confounding(X) + exogenous outcome shock(Z_S)
e_S <- rnorm(N, 0, 1)
S   <- 0.50*D + 0.80*M + 0.40*X_it + 0.70*Z_S + e_S

# Collider: caused by D and Z_S
e_K <- rnorm(N, 0, 1)
K   <- 0.70*D + 0.70*Z_S + e_K

data <- data.frame(X_it, Z_D, Z_S, D, M, K, S)
```

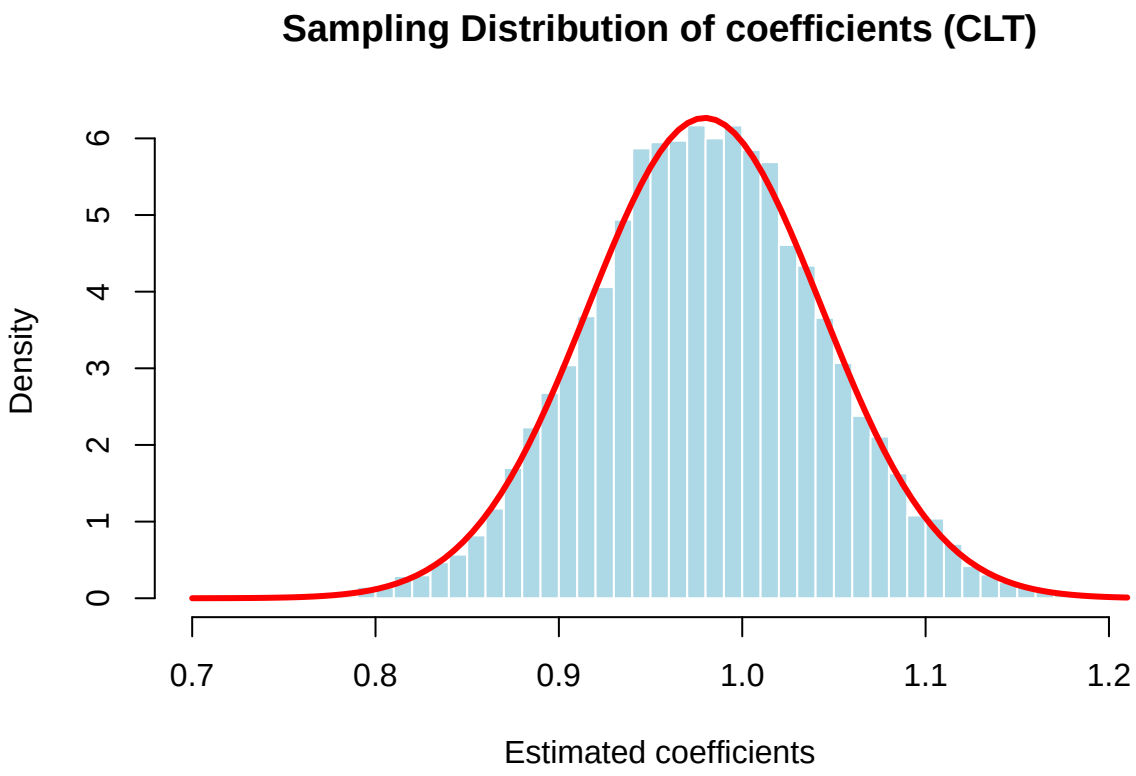
```
m_true <- lm(S ~ D + X_it, data = data)
summary(m_true)
```

```
##
## Call:
## lm(formula = S ~ D + X_it, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8277 -1.0275 -0.0480  0.9613  3.8894
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07675    0.08336   0.921    0.358
## D            1.04508    0.05851  17.861  < 2e-16 ***
## X_it         0.52991    0.09879   5.364 1.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.442 on 297 degrees of freedom
## Multiple R-squared:  0.6471, Adjusted R-squared:  0.6447
## F-statistic: 272.3 on 2 and 297 DF,  p-value: < 2.2e-16
```

#a

```r
set.seed(123)

R <- 10000
coefficients <- numeric(R)

for (r in 1:R) {
  N <- 300
  X_it <- rnorm(N, 0 ,1)
  Z_D <- rnorm (N, 0 ,1)
  Z_S <- rnorm (N, 0 ,1)

  # Treatment (pressure event intensity)
  e_D <- rnorm(N, 0, 1)
  D   <- 0.80*X_it + 0.90*Z_D + e_D
  # Mediator: strategic calculation
  e_M <- rnorm(N, 0, 1)
  M   <- 0.60*D + 0.50*X_it + e_M

  # Outcome = direct(D) + via(M) + confounding(X) + exogenous outcome shock(Z_S)
  e_S <- rnorm(N, 0, 1)
  S   <- 0.50*D + 0.80*M + 0.40*X_it + 0.70*Z_S + e_S

  m1 <- lm(S ~ D + X_it)
  coefficients[r] <- coef(m1)["D"]
}

hist(coefficients, breaks = 40, freq = FALSE,
     main = "Sampling Distribution of coefficients (CLT)",
     xlab = "Estimated coefficients", col = "lightblue", border = "white")

curve(dnorm(x, mean(coefficients), sd(coefficients)),
      col = "red", lwd = 3, add = TRUE)
```



Under repeated sampling, the OLS estimator for the treatment variable is asymptotically Normal.

#b:Bootstrap simulation

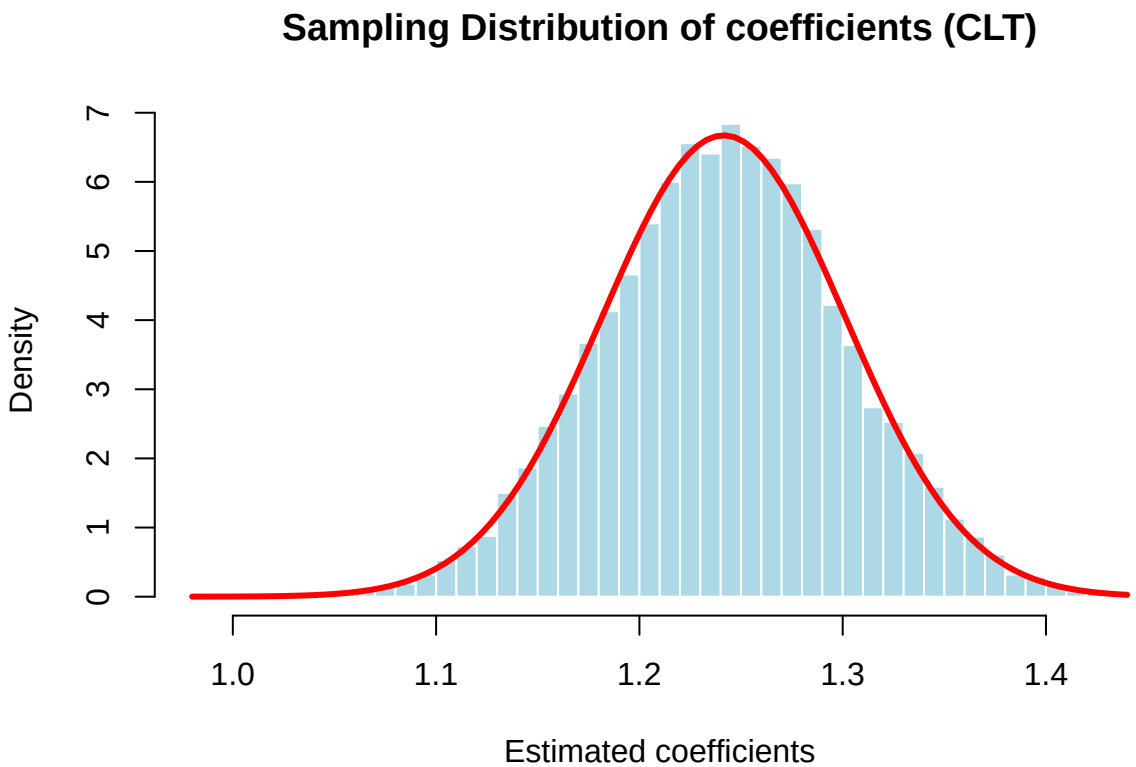
```
set.seed(123)
R <- 10000
boot_coef <- numeric(R)

for (r in 1:R) {
  boot_index <- sample(1:nrow(data), replace = TRUE)
  boot_data  <- data[boot_index, ]

  m2 <- lm(S ~ D + X_it, data = boot_data)

  boot_coef[r] <- coef(m2)["D"]
}

# Bootstrapped standard error:
boot_se <- sd(boot_coef)
boot_se
```

```
## [1] 0.05864912
```

#c: Omits the confounding variable

```
set.seed(123)

R <- 10000
coefficients <- numeric(R)

for (r in 1:R) {
  N <- 300
```

```r
  X_it <- rnorm(N, 0 ,1)
  Z_D <- rnorm (N, 0 ,1)
  Z_S <- rnorm (N, 0 ,1)

  # Treatment (pressure event intensity)
  e_D <- rnorm(N, 0, 1)
  D   <- 0.80*X_it + 0.90*Z_D + e_D
  # Mediator: strategic calculation
  e_M <- rnorm(N, 0, 1)
  M   <- 0.60*D + 0.50*X_it + e_M

  # Outcome = direct(D) + via(M) + confounding(X) + exogenous outcome shock(Z_S)
  e_S <- rnorm(N, 0, 1)
  S   <- 0.50*D + 0.80*M + 0.40*X_it + 0.70*Z_S + e_S

  m3 <- lm(S ~ D)
  coefficients[r] <- coef(m3)["D"]
}

hist(coefficients, breaks = 40, freq = FALSE,
     main = "Sampling Distribution of coefficients (CLT)",
     xlab = "Estimated coefficients", col = "lightblue", border = "white")

curve(dnorm(x, mean(coefficients), sd(coefficients)),
      col = "red", lwd = 3, add = TRUE)
```



When I omit the confounder , the sampling distribution of the treatment coefficient shifts substantially, from being centered around the true total effect (~0.98) to being centered near ~1.22. Although the distribution remains approximately Normal due to the CLT, it is centered at the wrong value, reflecting omitted variable bias from the backdoor path. This demonstrates that standard hypothesis tests and confidence intervals can

look well even when the underlying estimator is biased, meaning that inference is only valid if the causal model is correctly specified.

Part2 Data Analysis: Continue using my simulated data in the research proposal

a

I split the pressure event intensity in two groups: high and low pressure by median, which creates a binary variable.

```
data$D_hi <- as.integer(data$D >= median(data$D))

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
stats <- data %>%
  group_by(D_hi) %>%
  summarise(n = n(),
            mean_S = mean(S),
            sd_S = sd(S), .groups = "drop")
stats
```

```
## # A tibble: 2 x 4
##    D_hi     n mean_S  sd_S
##   <int> <int>  <dbl> <dbl>
## 1     0   150  -1.39  1.80
## 2     1   150   1.71  1.91
```

```
diff_means <- with(stats, mean_S[D_hi==1] - mean_S[D_hi==0])
diff_means
```

```
## [1] 3.10585
```

```
tt <- t.test(S ~ D_hi, data = data, var.equal = FALSE, alternative = "two.sided")
tt
```

```
##
##  Welch Two Sample t-test
##
## data:  S by D_hi
## t = -14.485, df = 296.92, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -3.527818 -2.683882
## sample estimates:
## mean in group 0 mean in group 1
```

```
##       -1.392142        1.713708
```

```
install.packages("effsize", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/17985/AppData/Local/R/win-library/4.5'
## (as 'lib' is unspecified)
```

```
## package 'effsize' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\17985\AppData\Local\Temp\RtmpKk1OTM\downloaded_packages
```

```
library(effsize)
```

```
## Warning: package 'effsize' was built under R version 4.5.2
```

```
d_eff <- cohen.d(S ~ D_hi, data = data, hedges.correction = TRUE)
```

```
## Warning in cohen.d.formula(S ~ D_hi, data = data, hedges.correction = TRUE):
## Cohercing rhs of formula to factor
```

```
d_eff
```

```
##
## Hedges's g
##
## g estimate: -1.668388 (large)
## 95 percent confidence interval:
##     lower     upper
## -1.931551 -1.405225
```

Interpretation: The estimated difference in means is 3.106, with high-pressure units exhibiting substantially higher semantic-weakening scores. The test statistic is large in magnitude (t=-14.49) and highly statistically significant, and the 95% confidence interval for the difference in means (-3.53,-2.68) excludes zero. Statistically, this indicates a clear difference in average outcomes between the two groups. Substantively, given that higher values of S correspond to weaker compliance with central policy language, this finding implies that months following stronger political pressure events are associated with markedly more strategic, less compliant bureaucratic drafting behavior. To assess the magnitude of this effect, I also computed Hedges's g, a bias-corrected version of Cohen's d suitable for unequal variances. The effect size is extremely large (g=1.67), with a confidence interval from -1.41 to -1.93, suggesting that the difference is not only statistically significant but also substantively substantial—roughly equivalent to comparing two markedly different compliance regimes.

#b

```
m_true <- lm(S ~ D + X_it, data = data)
summary(m_true)
```

```
##
## Call:
## lm(formula = S ~ D + X_it, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8277 -1.0275 -0.0480  0.9613  3.8894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07675    0.08336   0.921    0.358
```

```
## D            1.04508    0.05851  17.861  < 2e-16 ***
## X_it         0.52991    0.09879   5.364 1.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.442 on 297 degrees of freedom
## Multiple R-squared:  0.6471, Adjusted R-squared:  0.6447
## F-statistic: 272.3 on 2 and 297 DF,  p-value: < 2.2e-16
```

Interpretation: The coefficient of 1.045 indicates that a one-unit increase in pressure intensity is associated with a 1.045-unit increase in the semantic-weakening score. holding the confounder X_it constant. Substantively, this means stronger political pressure leads to noticeably weaker, more strategically diluted policy language.

The standard error of 0.0585 measures the uncertainty of the coefficient estimate. It measures how far the estimator tends to fall from the center of its sampling distribution. The small value implies the estimate is precise.

The t-value of 17.86 is the ratio of the coefficient to its standard error. It means the estimated effect is over 17 standard errors away from zero. This is extremely strong evidence against the null hypothesis.

The p-value is nearly zero. Statistically, this implies a highly significant effect of pressure intensity on semantic weakening.