Zhengyu Xiao_PS4

2025-09-05

1. A confounder is a random variable that have causal effect on both the treatment variable and the outcome variable. It is something that we should control in a regression or causal identification. A collider is a random variable that is caused by both the treatment variable and the outcome variable, we should not control it to make sure not openning new back door paths.
2. Conditioning on a collider can create bias by openning a spurious pathway or non-causal correlation that is probability unwanted in the causal identification.
3. Because statistical summaries or correlations alone cannot identify the causal relationship or logic as shown in a DAG. It only shows correlation not causal direction and we wont be able to tell which variables are colliders or confounders so we dont know who to control correctly.
4. A "kitchen sink" regression is trying to put every possible variable in the regression model. We have to consider the different causal pathways and existence of confounders and colliders, which is highly likely given putting a lot of variables in one model. We would not identify the correct causal relationship between the treatment and the outcome. In addition, we will have over-adjustment bias that possibly neglect the causal influence of mediators on the outcomes.
5. A backdoor path is any path between the treatment and the outcome that starts with an arrow pointing INTO the treatment($D \leftarrow Z \cdots \rightarrow Y$). It is when unobserved confounder is not controlled or when a collider is controlled such that we have a seperate pathway that might interfere with the real causal pathway and relationship we want to identify. Using mutiple regression, we can condition the confounders to block the backdoor pathways, allowing us to isolate the true causal effect.

#Part 2 Simulation 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}

```
install.packages("dagitty", 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 'dagitty' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\17985\AppData\Local\Temp\RtmpSm4jCt\downloaded_packages
```

```
install.packages("ggdag", 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 'ggdag' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\17985\AppData\Local\Temp\RtmpSm4jCt\downloaded_packages
```

```r
library(dagitty)
```

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

```r
library(ggdag)
```

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

```
##
## Attaching package: 'ggdag'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(dplyr)
```

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

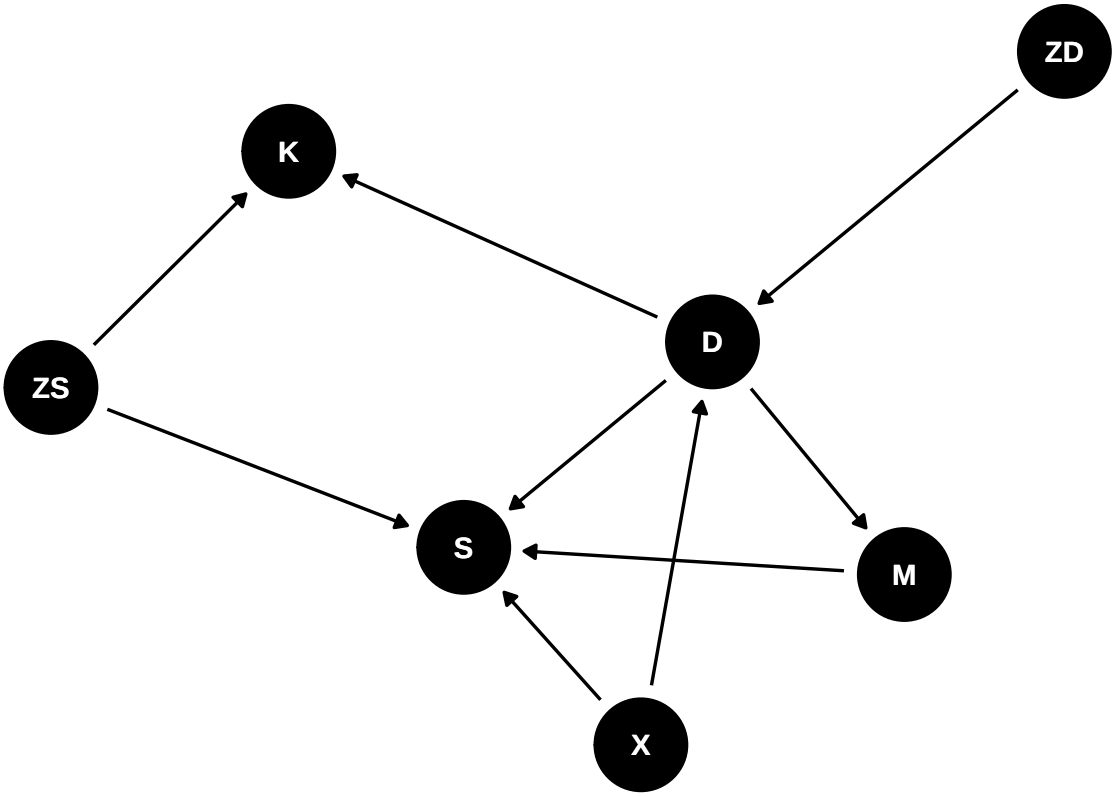
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```r
library(ggplot2)
theme_set (ggdag::theme_dag())
```

```r
g <- dagitty("
dag {
  D -> M
  M -> S
  D -> S

  X -> D
  X -> S

  ZD -> D
  ZS -> S

  D -> K
  ZS -> K
}
")

coords <- tibble::tribble(
  ~name, ~x, ~y,
  "ZD",  0, 0.7,
  "X",   0, 0.3,
  "D",   1, 0.5,
  "M",   2, 0.7,
  "S",   3, 0.5,
  "ZS",  2, 0.1,
  "K",   1.7, 0.3
)
```

```
coordinates(g) <- coords

ggdag(g) +
  geom_dag_text()
```



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)
```

Fit a model that recovers the direct effect of the treatment on the outcome variable.

```
m_1 <- lm(S ~ D + M + X_it, data = data)
summary(m_1)
```

```
##
## Call:
## lm(formula = S ~ D + M + X_it, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9007 -0.7566 -0.0541  0.8060  2.9686
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06684    0.07214   0.927  0.35493
## D            0.57309    0.06910   8.294  3.9e-15 ***
## M            0.74885    0.07461  10.037  < 2e-16 ***
## X_it         0.26570    0.08944   2.971  0.00321 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.248 on 296 degrees of freedom
## Multiple R-squared:  0.7367, Adjusted R-squared:  0.734
## F-statistic:   276 on 3 and 296 DF,  p-value: < 2.2e-16
```

To recover the direct effect, we must block the mediated path by conditioning on the mediator M, and close the backdoor via the confounder X_it.

Fit a model that recovers the total e>ect of the treatment on the outcome variable

```
m_2 <- lm(S ~ D + X_it, data = data)
summary(m_2)
```

```
##
## 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
```

To recover the total effect, do not condition on the mediator M, but still control the confounder X_it in the model.

Controlling for the collider, the exogenous independent variable and the instrument.

```
m_3 <- lm(S ~ D + K + X_it, data = data)
m_4 <- lm(S ~ D + Z_S + X_it, data = data)
m_5 <- lm(S ~ D + Z_D + X_it, data = data)
summary(m_3)
```

```
##
## Call:
## lm(formula = S ~ D + K + X_it, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9369 -0.9555 -0.0248  1.0740  3.6319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07061    0.08031   0.879     0.38
## D            0.82891    0.07153  11.589  < 2e-16 ***
## K            0.32099    0.06540   4.908 1.52e-06 ***
## X_it         0.52342    0.09517   5.500 8.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.389 on 296 degrees of freedom
## Multiple R-squared:  0.6736, Adjusted R-squared:  0.6703
## F-statistic: 203.6 on 3 and 296 DF,  p-value: < 2.2e-16
```

```
summary(m_4)
```

```
##
## Call:
## lm(formula = S ~ D + Z_S + X_it, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4087 -0.7929 -0.0426  0.8438  3.9433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06339    0.07102   0.893    0.373
## D            1.04354    0.04984  20.938  < 2e-16 ***
## Z_S          0.73283    0.06883  10.648  < 2e-16 ***
```

```
## X_it         0.56725    0.08422   6.735 8.51e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.229 on 296 degrees of freedom
## Multiple R-squared:  0.7448, Adjusted R-squared:  0.7422
## F-statistic:   288 on 3 and 296 DF,  p-value: < 2.2e-16
```

```
summary(m_5)
```

```
##
## Call:
## lm(formula = S ~ D + Z_D + X_it, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0024 -1.0141 -0.0499  1.0123  3.6083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07523    0.08298   0.907 0.365363
## D            1.15712    0.08201  14.110  < 2e-16 ***
## Z_D         -0.22988    0.11845  -1.941 0.053249 .
## X_it         0.43012    0.11096   3.876 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.436 on 296 degrees of freedom
## Multiple R-squared:  0.6515, Adjusted R-squared:  0.648
## F-statistic: 184.5 on 3 and 296 DF,  p-value: < 2.2e-16
```

Adding the collider K lowered the estimated effect of D from roughly 0.98 to 0.83. Because K is a common effect of D and Z_S, conditioning on it opens a non-causal path between D and Z_S, inducing collider bias and attenuating the estimated effect.

Adding Z_S had almost no effect on the coefficient of D (from ~0.98 to ~1.04) because Z_S affects only the outcome and is independent of D. Controlling it improves precision but does not change causal identification.

Controlling Z_D inflated the estimated effect of D from $\approx$ 0.98 to 1.16. Because Z_D provides exogenous variation in D, conditioning on it removes the identifying variation and leaves only the endogenous component, biasing the OLS estimate. Instruments should never be added as control variables.

The selection of variables should be based on the estimand and the causal structure encoded in the DAG. The purpose of controlling variables is to block backdoor confounding paths between the treatment and outcome—not to describe "everything correlated with the outcome." Therefore, only variables that are true confounders should be included. According to the results, we should control only the confounder not the collider. And control for mediator only when we want to estimate the direct direct effect of the indepedent variable. Controlling for variables that affect only the outcome do not change the estimand and may be added only to improve precision. Do not control for variables that effect only the treatment as a covariate in the OLS. Only control it in a 2SLS design.