prob set 4 Baixi Zhang

2025-11-18

Part 1

1.

A confounder is a variable that influences your independent variable X and dependent variable Y at the same time. We should control it in our model otherwise it will open a backdoor path causing bias in estimating relationship between X and Y.

A collider is a variable that being influenced by your independent variable X and dependent variable Y at the same time. We should not include the collider into our model otherwise it may create bias.

2.

In the original DAG with a collider, the collider blocks the path between X and Y, so the two variables are independent. However, conditioning on a collider opens the path between X and Y, causing spurious correlation.

3.

Statistical summaries or correlations cannot identified the causal mechanisms and causal directions between variables. Thus, it may create bias if adding controls solely relied on statistics. For example, collider, confounder, mediator and M-bias variables may have identical statistical estimates. We should use theory, substantive knowledge and DAG to assist us deciding controls.

4.

"Kitchen sink regression" refers to that adding all available variables into multivariate regression models, and then select them by p-value or model information. The problems including: ignoring the causal directionality, lacking of meaningful causal interpretation, inflated alpha error rate, overfitting and model instability, and ruling out professional knowledge.

5.

Backdoor path refers to a non-causal path connecting X and Y through another variable Z. It leads to a spurious association between X and Y even if there's no causal relationship between these two. By including the variable Z in backdoor path into the model, Z is holding constant and the backdoor path is blocked. In this case the bias caused by Z is excluded from the coefficient, and the coefficient estimate approaches the unbiased one.

Part 2

Treatment variable (X) : Intensity of populist rhetoric from incumbent

Outcome variable (Y): Democratic support of public

Confounder (C): Radical index of voters

Mediator (M): Backsliding events executed by incumbent

Collider (L): Political engagement of public

Instrument (I): Community note function in social media

Exogenous on Y (E): Education level

```
n <- 1000
set.seed(8964) #enable replication

## Generate exogenous var.

# Confounder: Radical index of voters
ideo_spec <- rnorm(n, mean = 0,sd = 1)

#Instrument: Community note function
comm_notes <- sample(c(0, 1), n, replace = TRUE, prob = c(0.5, 0.5))

#Exogenous on Y: Education level
edu <- rnorm(n, mean = 0,sd = 1)

#generate remaining var. as linear function

#Treatment variable: Intensity of populist rhetoric from incumbent
popu_rheto <- 0.7*ideo_spec + (-0.7)*comm_notes + rnorm(n)

#Mediator: Backsliding events  executed by incumbent
backsl_event <- 0.6*popu_rheto + rnorm(n)

#Outcome variable: Democratic support of public
dem_sup <- (-0.7)*ideo_spec + 0.6*edu + (-0.4)*popu_rheto + (-0.6) * backsl_event + rnorm(n)

#Collider: Political engagement of public
pol_engage <- 0.5*popu_rheto + 0.5*dem_sup + rnorm(n)
```

1.

```
#direct effect model
model1 <- lm(dem_sup ~ popu_rheto + ideo_spec + backsl_event)

summary(model1)
```

```
##
## Call:
## lm(formula = dem_sup ~ popu_rheto + ideo_spec + backsl_event)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9511 -0.7651  0.0480  0.7451  4.1743
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.004698   0.037995   0.124    0.902
## popu_rheto   -0.454912   0.041156 -11.053   <2e-16 ***
## ideo_spec    -0.702222   0.043248 -16.237   <2e-16 ***
```

```
## backsl_event -0.551026   0.034937 -15.772   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.128 on 996 degrees of freedom
## Multiple R-squared:  0.6658, Adjusted R-squared:  0.6648
## F-statistic: 661.4 on 3 and 996 DF,  p-value: < 2.2e-16
```

In the model estimating direct effect between X and Y, we need to include the mediator, the confounder and the exogenous variable on Y. By controlling the confounder, the backdoor path is blocked, so that we can have unbiased estimate. By controlling the mediator, the effect of mediator is excluded and the direct effect is isolated.

2.

```
#total effect model
model2 <- lm(dem_sup ~ popu_rheto + ideo_spec)

summary(model2)
```

```
##
## Call:
## lm(formula = dem_sup ~ popu_rheto + ideo_spec)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8957 -0.8554  0.0125  0.8218  4.1642
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01847    0.04242  -0.435    0.663
## popu_rheto  -0.82356    0.03785 -21.758   <2e-16 ***
## ideo_spec   -0.66060    0.04823 -13.696   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 997 degrees of freedom
## Multiple R-squared:  0.5823, Adjusted R-squared:  0.5815
## F-statistic: 695.1 on 2 and 997 DF,  p-value: < 2.2e-16
```

For identifying the total effect, we just need to exclude the mediator from the model. Therefore the effect of mediator will be included in the coefficient of X.

3.

The discussion below will based on direct effect model.

```
#control for the collider
model3 <- lm(dem_sup ~ popu_rheto + ideo_spec + backsl_event + pol_engage)

summary(model3)
```

```
##
## Call:
## lm(formula = dem_sup ~ popu_rheto + ideo_spec + backsl_event +
##     pol_engage)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5072 -0.6149 -0.0055  0.6361  2.9994
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.005276   0.032252  -0.164     0.87
## popu_rheto   -0.632118   0.036072 -17.524   <2e-16 ***
## ideo_spec    -0.495532   0.038178 -12.979   <2e-16 ***
## backsl_event -0.388095   0.030786 -12.606   <2e-16 ***
## pol_engage    0.516211   0.026220  19.688   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9572 on 995 degrees of freedom
## Multiple R-squared:  0.7595, Adjusted R-squared:  0.7585
## F-statistic: 785.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

Comparing with coefficient in model 1 (-0.45) and the true effect (-0.4), the coefficient of populist rhetoric increased to -0.63, which is significantly biased because a new causal path is opened by controlling the collider.

```
#control for the exogenous Y var.
model4 <- lm(dem_sup ~ popu_rheto + ideo_spec + backsl_event + edu)

summary(model4)
```

```
##
## Call:
## lm(formula = dem_sup ~ popu_rheto + ideo_spec + backsl_event +
##     edu)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1005 -0.6257  0.0017  0.6847  3.3732
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.01101    0.03319  -0.332     0.74
## popu_rheto   -0.44196    0.03595 -12.294   <2e-16 ***
## ideo_spec    -0.70497    0.03777 -18.665   <2e-16 ***
## backsl_event -0.57099    0.03053 -18.701   <2e-16 ***
## edu           0.57517    0.03262  17.632   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9849 on 995 degrees of freedom
## Multiple R-squared:  0.7454, Adjusted R-squared:  0.7443
## F-statistic: 728.1 on 4 and 995 DF,  p-value: < 2.2e-16
```

By controlling the exogenous variable, although the coefficient only changed marginally, the residual standard error decreased from 1.12 to 0.98, and the standard error of X coefficient decreased from 0.041 to 0.035, which indicates the estimate is more precise.

```
#control for the instrument
model5 <- lm(dem_sup ~ popu_rheto + ideo_spec + backsl_event + comm_notes)
```

```r
summary(model5)
```

```
##
## Call:
## lm(formula = dem_sup ~ popu_rheto + ideo_spec + backsl_event +
##     comm_notes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9737 -0.7807  0.0415  0.7265  4.1033
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.05021    0.04926   1.019    0.308
## popu_rheto   -0.47617    0.04367 -10.905   <2e-16 ***
## ideo_spec    -0.69026    0.04400 -15.686   <2e-16 ***
## backsl_event -0.54945    0.03493 -15.728   <2e-16 ***
## comm_notes   -0.11205    0.07725  -1.450    0.147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.127 on 995 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6652
## F-statistic: 497.2 on 4 and 995 DF,  p-value: < 2.2e-16
```

```r
#control for the instrument but not controlling confounder
model6 <- lm(dem_sup ~ popu_rheto  + backsl_event + comm_notes)

summary(model6)
```

```
##
## Call:
## lm(formula = dem_sup ~ popu_rheto + backsl_event + comm_notes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0039 -0.8608  0.0347  0.8186  4.5738
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.03072    0.05497   0.559    0.576
## popu_rheto   -0.82288    0.04204 -19.575  < 2e-16 ***
## backsl_event -0.51343    0.03891 -13.195  < 2e-16 ***
## comm_notes   -0.33917    0.08471  -4.004 6.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.258 on 996 degrees of freedom
## Multiple R-squared:  0.584,  Adjusted R-squared:  0.5828
## F-statistic: 466.2 on 3 and 996 DF,  p-value: < 2.2e-16
```

If we control the instrumental variable, we actually exclude the part of clean and good variation of X, which will lead to bias in X coefficient. However in our model the confounder has been controlled, so the backdoor path is blocked. We need to stop controlling the confounder to see the significant bias. In model 5, the X coefficient -0.82 is significantly biased because it actually shows the variation caused by $ideo_spec$.

4.

Given the reading and simulation results, we should base on our theory and causal argument to choose which variables to include in a model, instead of base on p-value or add everything we have. In the case of my study, since I want to examine the direct effect, I should include confounder, mediator and exogenous variable on Y.