Assignment 2

Gao Haochun A0194525Y Ge Siqi A0194550A Wang Pei A0194486M
Wei Yifei A0203451W

3/4/2021

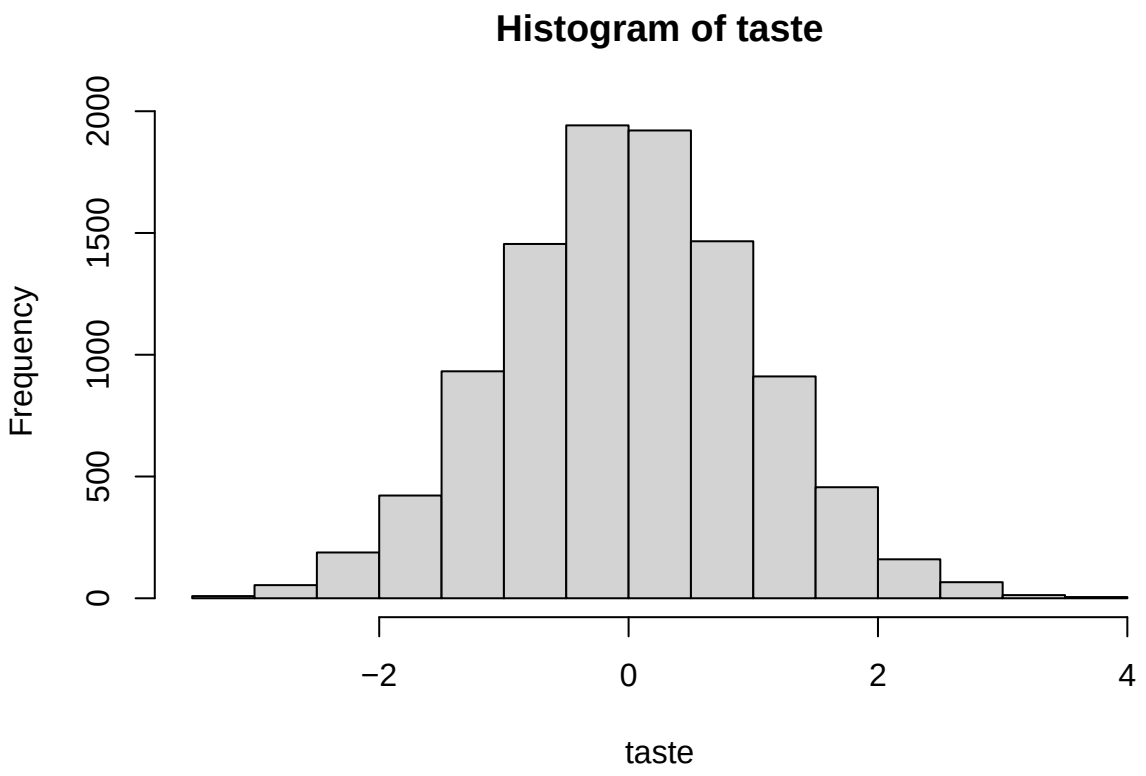
```
setwd("/Users/wangpei/OneDrive - National University of Singapore/Curriculum/Sem_04/BT3102/Assignments/
getwd()
```

```
## [1] "/Users/wangpei/OneDrive - National University of Singapore/Curriculum/Sem_04/BT3102/Assignments/
```

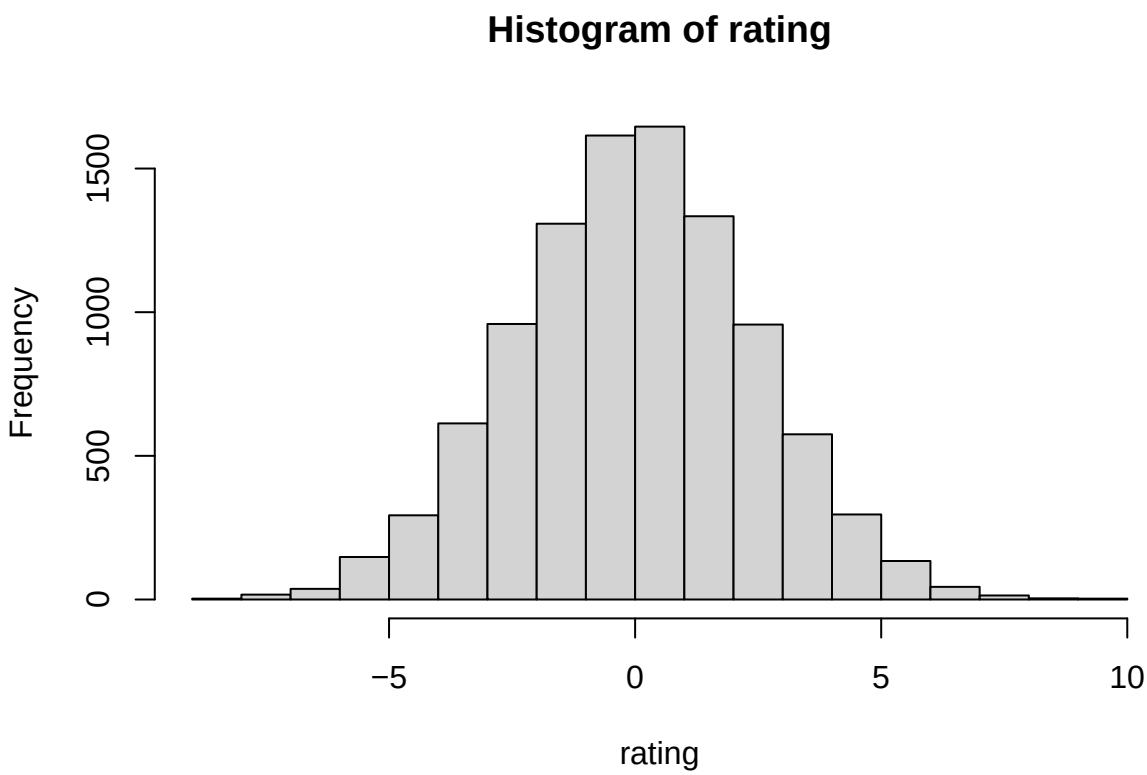
Q1. You study how sales depend on prices for wine. You believe that rating (i.e., expert ratings) can be an imperfect measure of taste (i.e., true quality). Taste is unobserved because there is no ideal measure for it.

I. Assume that causal Diagram 1 is correct. Choose sensible parameter values and simulate a data set of N = 10000 observations for 3 variables: ratings, prices, and sales (taste data is removed after the simulation because it is unobserved to the analyst).

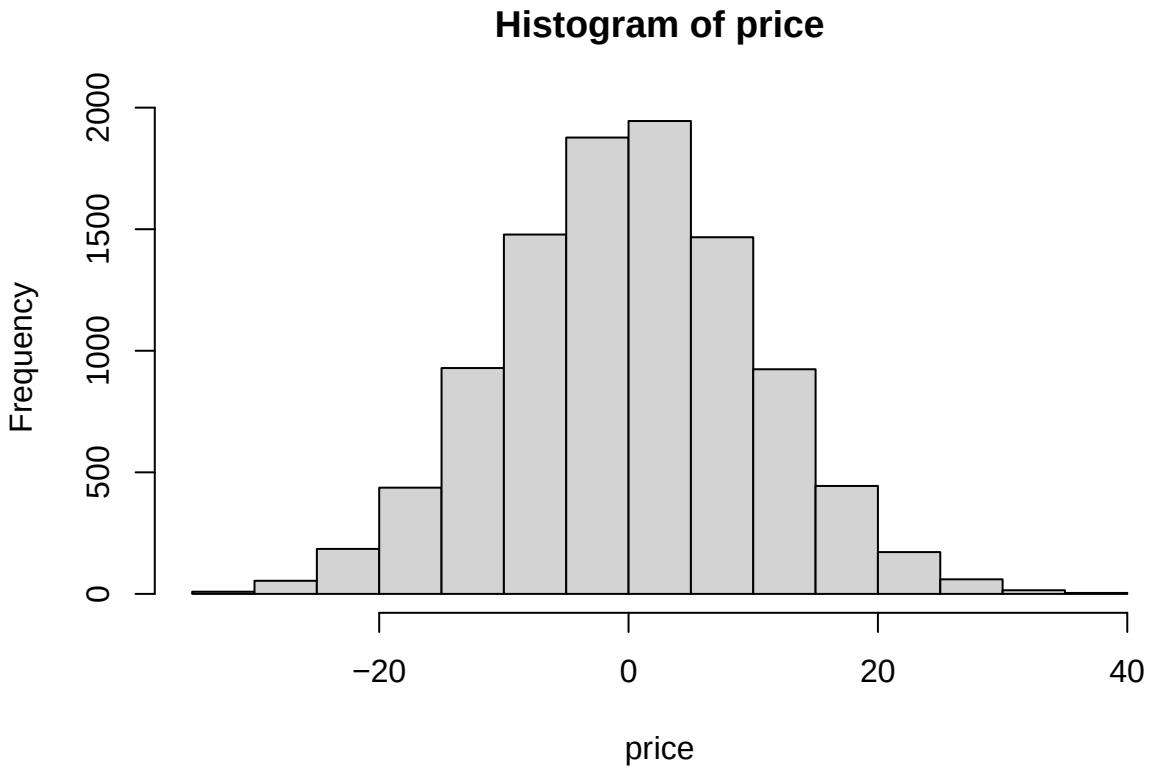
```
set.seed(37)
N = 10000
# taste in (0, 1), 2 decimal places
taste = rnorm(N)

hist(taste)
```



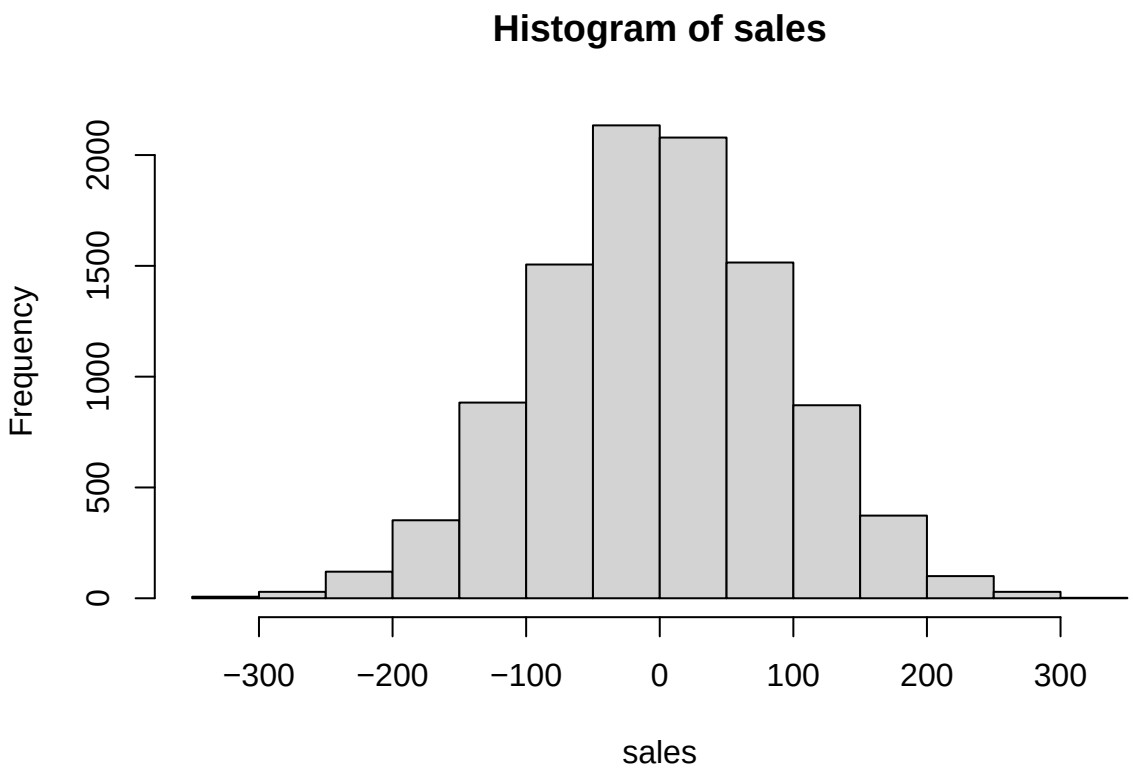

```
rating = rnorm(N,taste,2) + rnorm(N)
hist(rating)
```



```
price = 10 * taste + rnorm(N)
hist(price)
```




```
sales = 10*taste - 10*price + rnorm(N)
hist(sales)
```



II. Use the data set you just generated and regress sales on price. How does your estimate for the price coefficient differ from its true value? Does including ratings as an independent variable solve the problem? Explain why or why not.

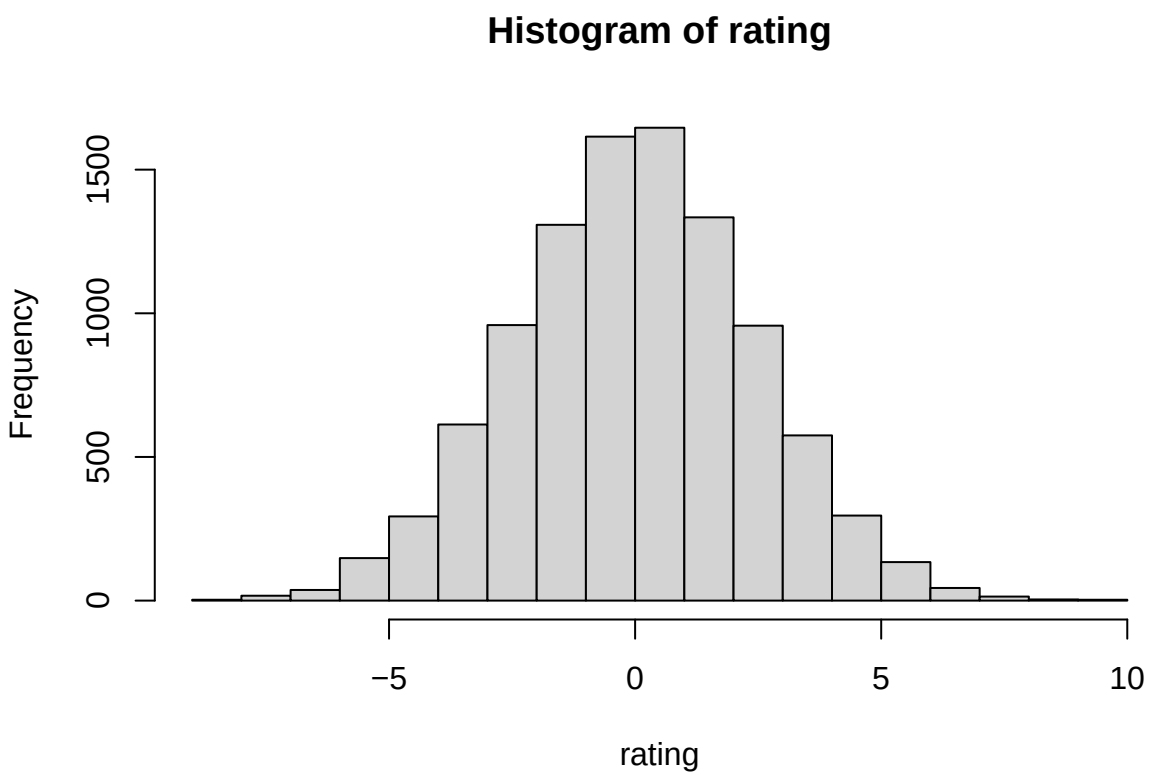
```
model1 = lm(sales ~ price)
model2 = lm(sales ~ price + rating)
stargazer(model1,model2, type="text",omit.stat=c("LL","ser","f"),
          model.numbers=TRUE, model.names = TRUE)
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                           sales
##                            OLS
##                     (1)            (2)
## ----------------------------------------
## price           -9.008***      -9.010***
##                  (0.001)        (0.002)
##
## rating                         0.020***
##                                 (0.006)
##
## Constant          0.004          0.004
##                  (0.014)        (0.014)
##
## ----------------------------------------
```

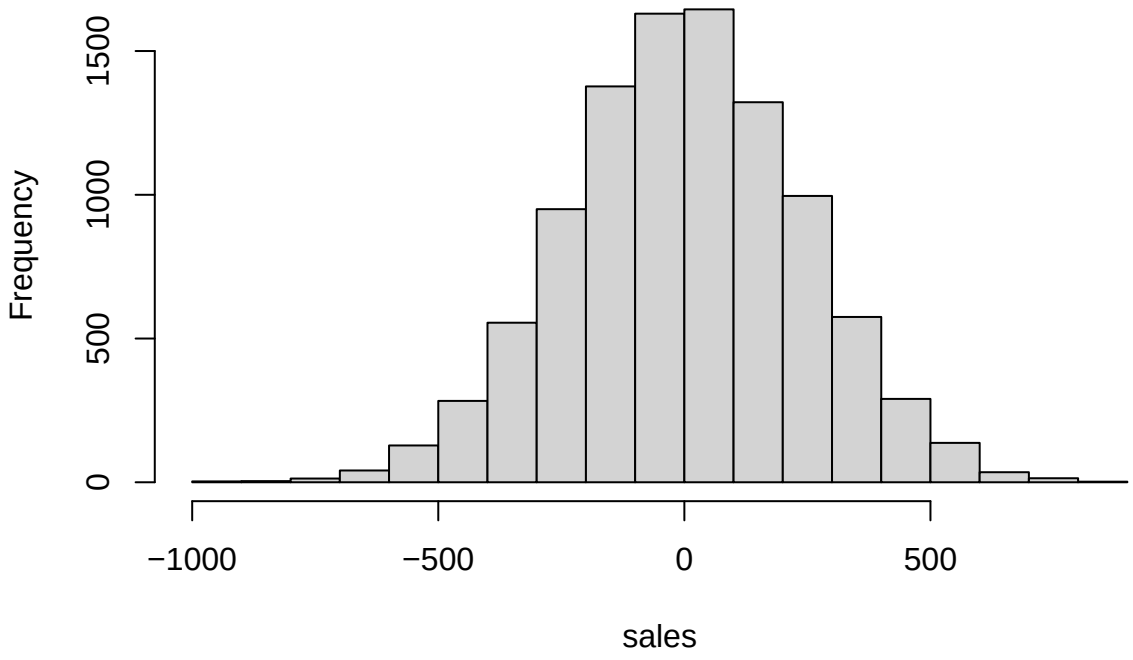
```
## Observations      10,000         10,000
## R2                1.000           1.000
## Adjusted R2       1.000           1.000
## ========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

III. Redo I-II and this time assume that causal Diagram 2 is correct.

```
set.seed(37)
N = 10000
taste = rnorm(N)
rating = rnorm(N,taste,2) + rnorm(N)
hist(rating)
```




```
price = 10*rating + rnorm(N)

sales = 10*taste - 10*price + rnorm(N)
hist(sales)
```




```
model3 = lm(sales ~ price)
model4 = lm(sales ~ price + rating)
stargazer(model3,model4,
          type="text",omit.stat=c("LL","ser","f"),
          model.numbers=TRUE, model.names = TRUE)
```

```
##
## ==========================================
##                   Dependent variable:
##               ---------------------------
##                          sales
##                           OLS
##                     (1)            (2)
## ------------------------------------------
## price           -9.831***      -10.090***
##                  (0.004)        (0.092)
##
## rating                          2.594***
##                                 (0.923)
##
## Constant          0.065          0.063
##                  (0.093)        (0.093)
##
## ------------------------------------------
## Observations     10,000         10,000
## R2                0.998          0.998
## Adjusted R2       0.998          0.998
## ==========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

Q2

2.I

Data generating process:

$$\alpha_1 = 0$$
$$\alpha_2 = \alpha_3 = \beta_1 = \beta_2 = \beta_3 = \gamma = 1$$

```
set.seed(37)
N = 10000

D = rnorm(N)
E = rnorm(N)
F = rnorm(N)
a1=0
a2=a3=b1=b2=b3=g=1

C = g*F + rnorm(N)
A = 0 + a2*C + a3*D + rnorm(N)
B = b1*A + b2*C + b3*E + rnorm(N)
C_measured = C + rnorm(N)
D_measured = D + rnorm(N)
A_measured = A + rnorm(N)
```

2.I.1 draw causal diagram

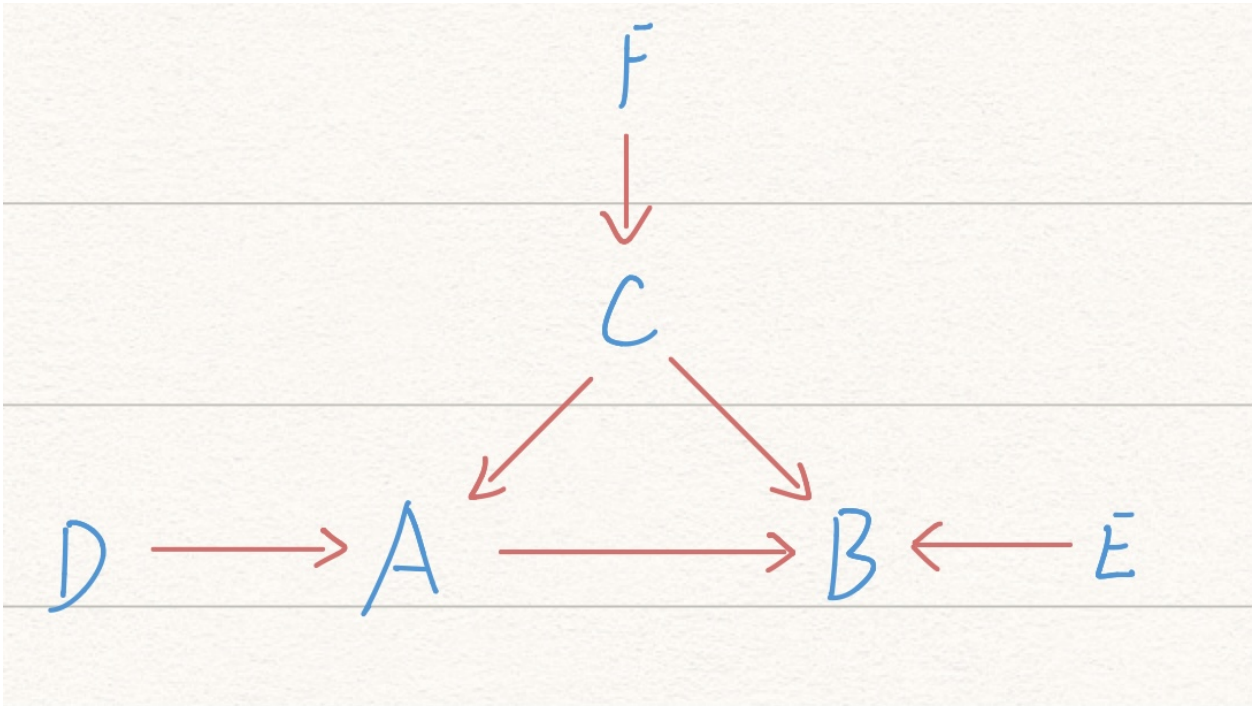


Figure 1: Q2.1 causal diagram

A is correlated with B,C,D,F; depends directly on D,C, indirectly on F.

B is correlated with A,C,D,E,F; depends directly on A,C,E,, indirectly on F,D.

C is correlated with A,B,F; depends directly on F.

D is correlated with A,B; depends on nothing (exogenous).

E is correlated with B; depends on nothing (exogenous).

F is correlated with A,B,C; depends on nothing (exogenous).

2.I.2 Show all collider variables and how they may bias estimates.

Collider variables are variables with multiple parents.

A is a collider variable, it will cause endogenous problem if added into regressions of D and C (i.e. D ~ C+A, C ~ D+A yield biased estimates.)

B is a collider variable, it will cause endogenous problem if added into regressions of A, C, E. (i.e. C ~ E+B, E ~ C+B, A ~ C+B, C ~ A+B, A ~ E+B, E ~ A+B, etc.)

2.I.3 Which variables to include to predict A? Is the model also a good causal inference model?

Regress A on D and C (A ~ D+C). It is also a good model for causal inference as it captures the true causal relationship (No endogenous problem because all parents of A are included in the regression model).

2.I.4 Show whether or not each of following data is enough to identify relation between A and B.

```
lm1 = lm(B ~ A+C) # ok
lm2 = ivreg(B ~ A|D) # ok
lm3 = lm(B ~ A+E) # bad, omitted variable bias by C
lm4 = lm(B ~ A+F) # bad, omitted variable bias by C
lm5 = lm(B ~ A+C_measured) # bad, measurement error
lm6 = ivreg(B ~ A|D_measured) # ok, but not so accurate
lm7 = ivreg(B ~ A_measured|D) # ok
lm8 = lm(B ~ A_measured+C) # bad, measurement error by A

stargazer(lm1,lm2,lm3,lm4,lm5,lm6,lm7,lm8,
          type="text",omit.stat=c("LL","ser","f"),
          model.numbers=TRUE, model.names = TRUE)
```

```
##
## ================================================================================================
##                                          Dependent variable:
##              ------------------------------------------------------------------------------
##                                                    B
##                 OLS    instrumental            OLS               instrumental      OLS
##                          variable                                    variable
##                 (1)        (2)        (3)      (4)      (5)      (6)      (7)      (8)
## ------------------------------------------------------------------------------------------------
## A            0.996***   0.974***   1.504*** 1.335*** 1.242*** 0.948***
##              (0.010)    (0.020)    (0.007)  (0.009)  (0.010)  (0.029)
##
```

```
## C            1.012***                                                    1.327***
##              (0.014)                                                     (0.015)
##
## E                              1.019***
##                                (0.014)
##
## F                                      0.673***
##                                        (0.019)
##
## C_measured                                     0.517***
##                                                (0.011)
##
## A_measured                                                        0.983*** 0.678***
##                                                                   (0.023)  (0.009)
##
## Constant      0.001      -0.016     -0.001   -0.006   -0.005   -0.016   -0.022   -0.005
##              (0.014)    (0.021)    (0.014)  (0.016)  (0.016)  (0.021)  (0.023)  (0.016)
##
## ------------------------------------------------------------------------------------
## Observations  10,000     10,000     10,000   10,000   10,000   10,000   10,000   10,000
## R2            0.835      0.657      0.833    0.778    0.793    0.647    0.586    0.783
## Adjusted R2   0.835      0.657      0.833    0.778    0.793    0.647    0.586    0.783
## ====================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

Comments on models:

- a: B ~ A + C can identify relation between A and B because E is exogenous and C, which is endogenous, is included in the regression.
- b: B ~ A | D can identify relation between A and B because D is a good instrumental variable as it strongly correlates with A and it correlates with B only through A.
- c: B ~ A + E cannot identify relation between A and B due to omitted variable bias. C effects both A and B and is omitted in the regression.
- d: B ~ A + F cannot identify relation between A and B because C is omitted in the regression, causing omitted variable bias. F cannot be a instrumental variable as it correlates with B not only through A.
- e: B ~ A + C_measured cannot identify relation between A and B. This is because:

$$
\begin{aligned}
& A = a_0 + a_1 C + \epsilon_1 \\
& B = b_0 + b_1 A + b_2 C + \epsilon_2 \\
& C^* = C + \epsilon_3 \\
\Longrightarrow\ & A = a_0 + a_1 C^* + (\epsilon_1 - a_1 \epsilon_3) \\
& B = b_0 + b_1 A + b_2 C^* + (\epsilon_2 - b_2 \epsilon_3)
\end{aligned}
$$

The error term of B ~ A+C* is correlated C* and A. Thus, the estimation for b1 and b2 are biased.

- f: B ~ A | D_measured can identify the relation between A and B but D_measured is a weaker instrumental variable than D. As shown below, the correlation between A and D_measured is 0.3518, so the estimation bias is larger than B ~ A|D.

```
cor(A, D_measured)
```

```
## [1] 0.3517814
```

- g: B ~ A_measured | D can identify the relation between A and B because D (the instrumental variable) can fix the attenuation effect of measurement error on A.
- h: B ~ A_measured + C cannot identify the relation between A and B because the error term of the regression is correlated with A_measured, causing biased estimation of b1.

2.II

Data generating process:

$$\alpha_1 = \alpha_2 = \alpha_3 = -0.8$$
$$\beta_1 = \beta_2 = \beta_3 = -0.5$$
$$\gamma = 0.5$$

2.II.1

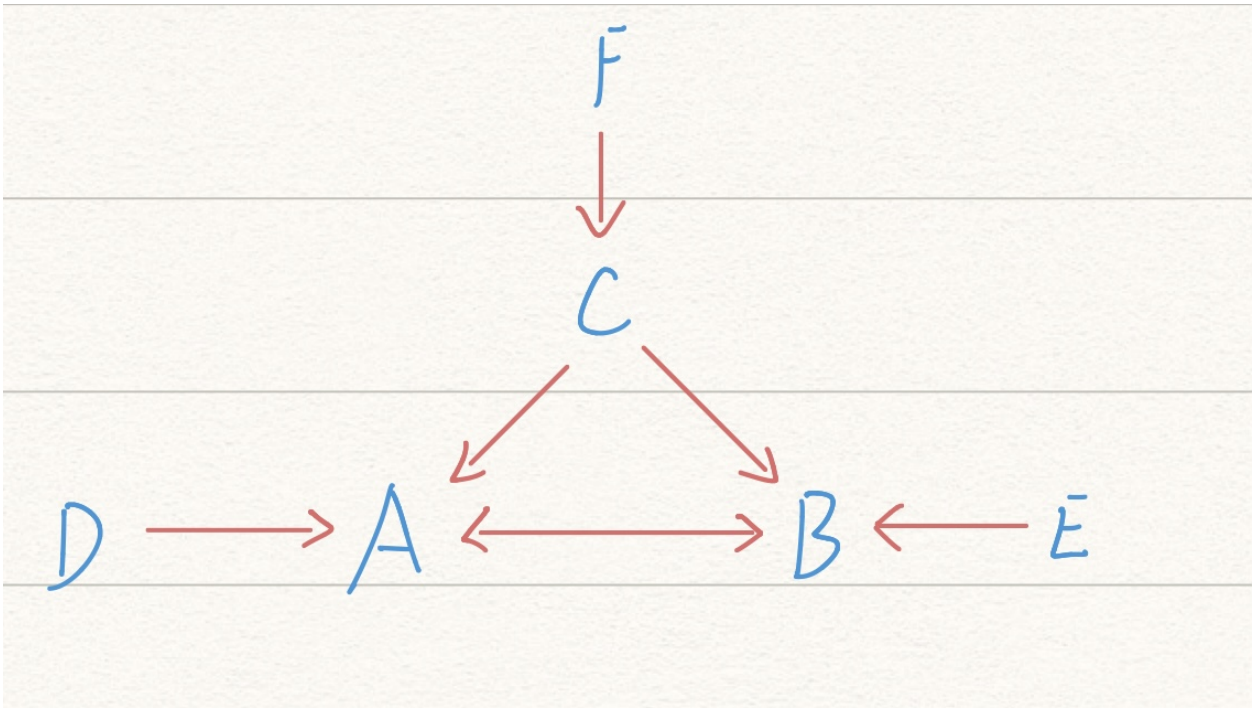


Figure 2: Q2.2 causal diagram

A is correlated with B,C,D,E,F; depends directly on B,C,D, indirectly on E,F.

B is correlated with A,C,D,E,F; depends directly on A,C,E,, indirectly on D,F.

C is correlated with A,B,F; depends directly on F.

D is correlated with A,B; depends on nothing (exogenous).

E is correlated with A,B; depends on nothing (exogenous).

F is correlated with A,B,C; depends on nothing (exogenous).

2.II.2

Solution of the simultaneous equations:

$$A = \frac{(a_1b_2 + a_2)C + a_3D + a_1b_3E + a_1\epsilon_3 + \epsilon_2}{1 - a_1b_1}$$

$$B = \frac{(a_2b_1 + b_2)C + a_3b_1D + b_3E + b_1\epsilon_2 + \epsilon_3}{1 - a_1b_1}$$

```
set.seed(37)
N = 10000

D = rnorm(N)
E = rnorm(N)
F = rnorm(N)
a1=a2=a3 = -0.8
b1=b2=b3 = -0.5
g = 0.5
e1 = rnorm(N)
e2 = rnorm(N)
e3 = rnorm(N)

C = g*F + e1
A = ((a1*b2+a2)*C + a3*D + a1*b3*E + a1*e3 + e2)/(1-a1*b1)
B = ((a2*b1+b2)*C + a3*b1*D + b3*E + b1*e2 + e3)/(1-a1*b1)
```

```
# increase D by 1
D2 = D+1
A2 = ((a1*b2+a2)*C + a3*D2 + a1*b3*E + a1*e3 + e2)/(1-a1*b1)
A2[1] - A[1] # decrease by 4/3
```

```
## [1] -1.333333
```

```
# increase E by 1
E2 = E+1
A3 = ((a1*b2+a2)*C + a3*D + a1*b3*E2 + a1*e3 + e2)/(1-a1*b1)
A3[1] - A[1] # increase by 2/3
```

```
## [1] 0.6666667
```

```
# increase F by 1
F2 = F+1
C2 = g*F2 + e1
A4 = ((a1*b2+a2)*C2 + a3*D + a1*b3*E + a1*e3 + e2)/(1-a1*b1)
A4[1] - A[1] # decrease by 1/3
```

```
## [1] -0.3333333
```

- a: When D increases by 1, A will increase by $\frac{a_3}{1-a_1b_1} = -\frac{4}{3}$
- b: When E increases by 1, A will increase by $\frac{a_1b_3}{1-a_1b_1} = \frac{2}{3}$
- c: When F increases by 1, A will increase by $\frac{\gamma(a_2+a_1b_2)}{1-a_1b_1} = -\frac{1}{3}$

2.II.3 Show how you can identify the DGP coefficients.

```
g_lm = lm(C ~ F)  # regress C on F
a1_a2_a3_lm = ivreg(A ~ B+C+D|E+C+D)  # Using E as iv for B
b1_b2_b3_lm = ivreg(B ~ A+C+E|D+C+E)  # Using D as iv for A

stargazer(g_lm, a1_a2_a3_lm, b1_b2_b3_lm,
          type="text",omit.stat=c("LL","ser","f"),
          model.numbers=TRUE, model.names = TRUE)
```

```
##
## ================================================
##                       Dependent variable:
##               ---------------------------------
##                   C          A            B
##                  OLS    instrumental instrumental
##                            variable     variable
##                  (1)        (2)          (3)
## ------------------------------------------------
## F             0.507***
##               (0.010)
##
## B                        -0.809***
##                           (0.012)
##
## A                                      -0.496***
##                                         (0.008)
##
## C                        -0.799***     -0.486***
##                           (0.009)       (0.010)
##
## D                        -0.789***
##                           (0.013)
##
## E                                      -0.487***
##                                         (0.011)
##
## Constant       -0.011     -0.008        0.006
##               (0.010)    (0.010)       (0.010)
##
## ------------------------------------------------
## Observations  10,000      10,000        10,000
## R2             0.204      0.866         0.779
## Adjusted R2    0.203      0.866         0.779
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

- We can identify γ by regress C on F (C ~ F).
- We can identify $\alpha_1, \alpha_2, \alpha_3$ by regress A on C and D and using E as instrumental variable for B.
- We can identify $\beta_1, \beta_2, \beta_3$ by regress B on C and E and using D as instrumental variable for A.

2.III Identify model coefficients of given data.

```
rm(list=ls())
data = read.csv("hw2q2.csv")
attach(data)
```

```
## The following object is masked from package:base:
##
##     F
```

```
g_lm = lm(C ~ F)  # regress C on F
a1_a2_a3_lm = ivreg(A ~ B+C+D|E+C+D)  # Using E as iv for B
b1_b2_b3_lm = ivreg(B ~ A+C+E|D+C+E)  # Using D as iv for A

stargazer(g_lm, a1_a2_a3_lm, b1_b2_b3_lm,
          type="text",omit.stat=c("LL","ser","f"),
          model.numbers=TRUE, model.names = TRUE)
```

```
##
## ================================================
##                        Dependent variable:
##                -----------------------------------
##                   C           A            B
##                  OLS    instrumental instrumental
##                              variable     variable
##                  (1)         (2)          (3)
## ------------------------------------------------
## F              0.729***
##                (0.003)
##
## B                         -0.322***
##                            (0.006)
##
## A                                       0.311***
##                                         (0.006)
##
## C                          0.431***     -0.263***
##                            (0.003)      (0.004)
##
## D                         -0.581***
##                            (0.003)
##
## E                                       0.539***
##                                         (0.003)
##
## Constant       -0.0001     0.005*       0.001
##                (0.003)     (0.003)      (0.003)
##
## ------------------------------------------------
## Observations 100,000     100,000      100,000
## R2             0.346       0.355        0.165
## Adjusted R2    0.346       0.355        0.165
```

```
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The identified values are:

$$\begin{aligned}\alpha_1 &= -0.32\\ \alpha_2 &= 0.43\\ \alpha_3 &= -0.58\\ \beta_1 &= 0.31\\ \beta_2 &= -0.26\\ \beta_3 &= 0.54\\ \gamma &= 0.73\end{aligned}$$

Q3

I

```
library(stargazer)
library(AER)
data = read.csv("Attend.csv")
data$fresh = as.factor(data$fresh)
data$soph = as.factor(data$soph)
attach(data)

lm1 = lm(stndfnl ~ atndrte+fresh+soph)
summary(lm1)
```

```
##
## Call:
## lm(formula = stndfnl ~ atndrte + fresh + soph)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76165 -0.68039 -0.02466  0.65886  2.54299
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.521253   0.193459  -2.694 0.007228 **
## atndrte      0.008407   0.002171   3.872 0.000118 ***
## fresh1      -0.269192   0.114164  -2.358 0.018661 *
## soph1       -0.110904   0.097584  -1.136 0.256153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9626 on 674 degrees of freedom
## Multiple R-squared:  0.02986,    Adjusted R-squared:  0.02554
## F-statistic: 6.914 on 3 and 674 DF,  p-value: 0.0001372
```

```
# Not so confident? May have confounding vars.
```

I.1

I.2

II

```
lm2 = lm(stndfnl ~ atndrte+fresh+soph+priGPA+ACT)
summary(lm2)
```

```
##
## Call:
## lm(formula = stndfnl ~ atndrte + fresh + soph + priGPA + ACT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.40928 -0.55632 -0.02683  0.58124  2.26979
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.295971   0.303556 -10.858  < 2e-16 ***
## atndrte      0.005415   0.002347   2.307   0.0213 *
## fresh1      -0.030822   0.106121  -0.290   0.7716
## soph1       -0.151856   0.088246  -1.721   0.0857 .
## priGPA       0.427452   0.080685   5.298 1.59e-07 ***
## ACT          0.083580   0.010985   7.608 9.41e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8698 on 672 degrees of freedom
## Multiple R-squared:  0.2103, Adjusted R-squared:  0.2045
## F-statistic:  35.8 on 5 and 672 DF,  p-value: < 2.2e-16
```

II.1

II.2

III

```
lm3 = ivreg(stndfnl ~ atndrte + fresh + soph + priGPA + ACT|hwrte)

stargazer(lm1, lm2, lm3, type="text",omit.stat=c("LL","ser","f"),
          model.numbers=TRUE, model.names = TRUE)
```

```
##
## ================================================
##                       Dependent variable:
##               ---------------------------------
##                             stndfnl
##                       OLS         instrumental
##                                     variable
```

```
##                 (1)       (2)        (3)
## ---------------------------------------------
## atndrte      0.008***   0.005**    0.012***
##               (0.002)   (0.002)    (0.004)
##
## fresh1       -0.269**   -0.031
##               (0.114)   (0.106)
##
## soph1         -0.111    -0.152*
##               (0.098)   (0.088)
##
## priGPA                  0.427***
##                          (0.081)
##
## ACT                     0.084***
##                          (0.011)
##
## Constant     -0.521*** -3.296***  -0.968***
##               (0.193)   (0.304)    (0.296)
##
## ---------------------------------------------
## Observations    678       678        672
## R2             0.030     0.210      0.021
## Adjusted R2    0.026     0.204      0.020
## =============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```