Assignment 2

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

3/4/2021

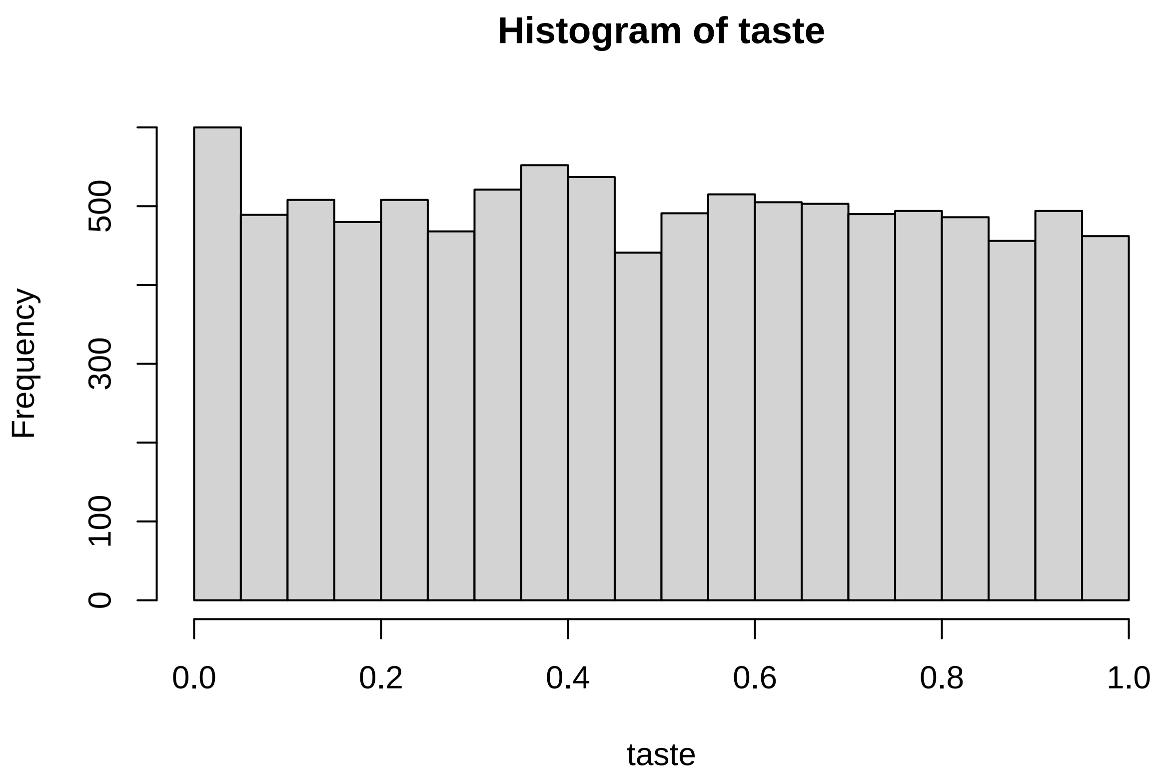
```
setwd("/Users/wangpei/OneDrive - National University of Singapore/Curriculum/Sem_04/BT3102/Assignments/
getwd()
```

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

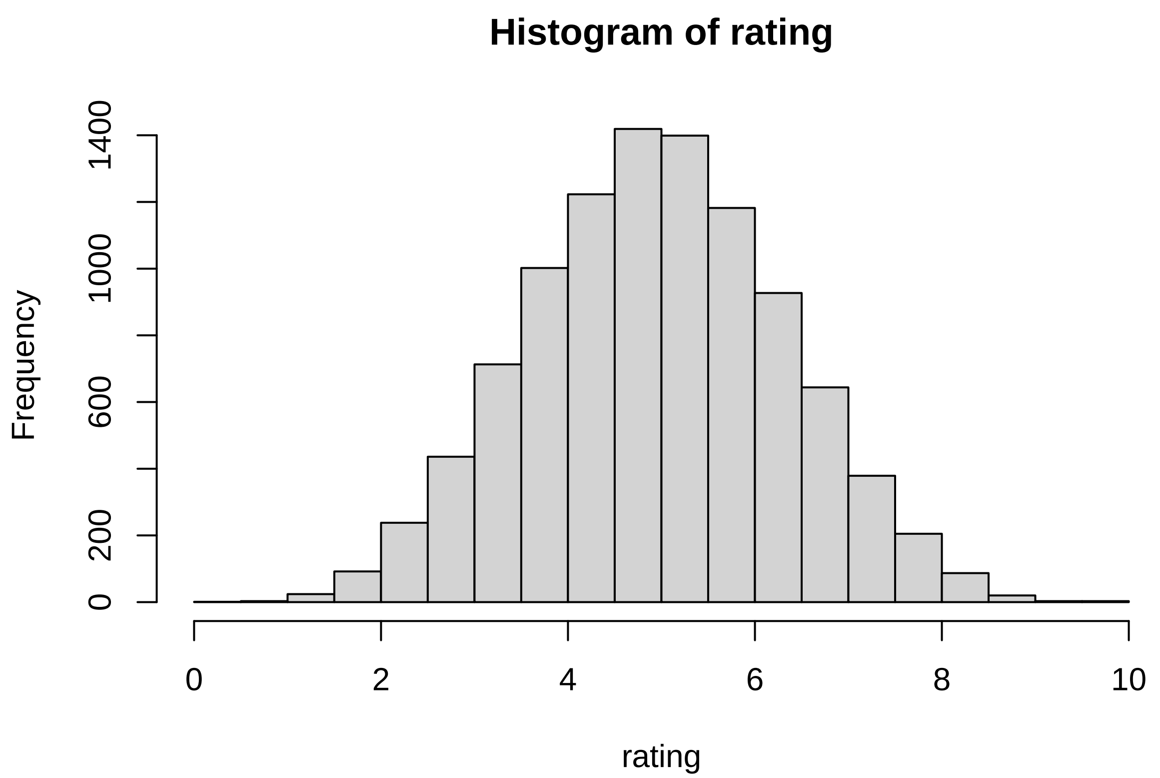
Q1. You study how sales depend on prices for wine. You believe that rating (i.e., expert ratings) can be an imperfect measure of taste (i.e., true quality). Taste is unobserved because there is no ideal measure for it.

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 = sample(seq(0,1,0.01), N, replace=T)
hist(taste)
```

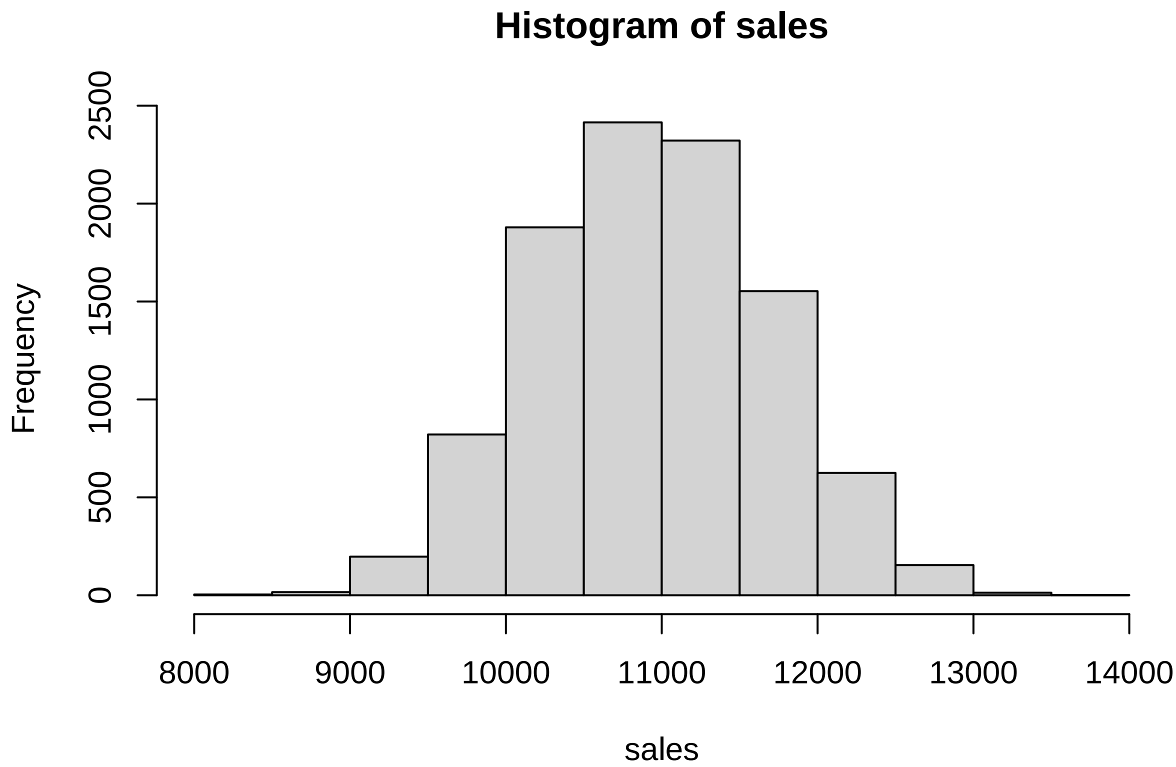



```
trans = taste*3 + 7 + rnorm(N)
# map to (0, 10) and round to 1 decimal place
rating = round(10*(trans - min(trans))/(max(trans)-min(trans)), 1)
hist(rating)
```



```r
price = 10*taste + 1000 + rnorm(N, 0, 50)
hist(price)
```


```r
sales = 20000 + 2000*taste - 10*price + rnorm(N, 0, 50)
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)
summary(model1)
```

```
##
## Call:
## lm(formula = sales ~ price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1154.7  -493.8    -8.7   493.2  1165.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20103.8751   116.0404  173.25   <2e-16 ***
## price          -9.1178     0.1153  -79.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 580.1 on 9998 degrees of freedom
## Multiple R-squared:  0.3847, Adjusted R-squared:  0.3846
## F-statistic:  6250 on 1 and 9998 DF,  p-value: < 2.2e-16
```

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

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)
```

2.I.1 draw causal diagram

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).

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

```
set.seed(37)
C_measured = C + rnorm(N)
D_measured = D + rnorm(N)
A_measured = A + rnorm(N)

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) # ok, only coeff of C is biased
lm6 = ivreg(B ~ A|D_measured) # bad, D is weak iv
```

```r
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*** 1.007*** 1.955***                  
##                 (0.010)     (0.020)    (0.007)  (0.009)  (0.017)  (0.027)                   
##                                                                                         
## C               1.012***                                                        0.815***
##                 (0.014)                                                         (0.018) 
##                                                                                         
## E                                      1.019***                                         
##                                        (0.014)                                          
##                                                                                         
## F                                               0.673***                                
##                                                 (0.019)                                 
##                                                                                         
## C_measured                                               0.660***                       
##                                                          (0.019)                        
##                                                                                         
## A_measured                                                                 0.973*** 0.796***
##                                                                            (0.018)  (0.010) 
##                                                                                         
## Constant          0.001      -0.016     -0.001   -0.006   -0.009    0.002   -0.011   -0.002 
##                 (0.014)     (0.021)    (0.014)  (0.016)  (0.017)  (0.020)  (0.018)  (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    0.776    0.682    0.745    0.802  
## Adjusted R2       0.835      0.657      0.833    0.778    0.776    0.682    0.745    0.802  
## =========================================================================================
## 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 regressoin.

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: Both (B ~ A+F) and (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 can identify relation between A and B (but bot B and C). This is because:

$$B = b_0 + b_1 A + b_2 C + \epsilon_1,\ C^* = C + \epsilon_2 \implies B = b_0 + b_1 A + b_2 C^* + (\epsilon_1 - b_2 \epsilon_2)$$

The error term of B ~ A+C* is correlated C*, but not A. Thus, the estimation for b2 is biased but the estimation for b1 is unbiased.

f: B ~ A|D_measured cannot identify the relation between A and B because D_measured is not a good instrumental variable. As shown below, the R-squared between A and D_measured is merely 0.1284, and the regression coefficient is not significant, so D_measured is a bad instrumental variable for A.

```
summary(lm(A ~ D_measured))
```

```
##
## Call:
## lm(formula = A ~ D_measured)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8855 -1.2686 -0.0076  1.2707  6.9588
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01853    0.01880  -0.985    0.325
## D_measured   0.51311    0.01337  38.377   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.88 on 9998 degrees of freedom
## Multiple R-squared:  0.1284, Adjusted R-squared:  0.1283
## F-statistic:  1473 on 1 and 9998 DF,  p-value: < 2.2e-16
```

g: B ~ A_measured|D can identify the relation between A and B. ?????????????????? Why

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.

I.2

```
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 + e3
```

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

# increase D by 1
D2 = D+1
B2 = (e2 + b3*E + a3*b1*D2 + (a2*b1+b2)*C)/(1-a1*b1)
A2 = e1 + a1*B2 + a2*C + a3*D2
A2[1] - A[1] # Decrease by 4/3
```

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

```r
# increase E by 1
E2 = E+1
B3 = (e2 + b3*E2 + a3*b1*D + (a2*b1+b2)*C)/(1-a1*b1)
A3 = e1 + a1*B3 + a2*C + a3*D
A3[1] - A[1] # increase by 2/3
```

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

```r
# increase F by 1
F2 = F+1
C2 = g*F2 + e3
B4 = (e2 + b3*E + a3*b1*D + (a2*b1+b2)*C2)/(1-a1*b1)
A4 = e1 + a1*B4 + a2*C2 + a3*D
A4[1] - A[1] # decrease by 1/3
```

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

I.3

```r
library(stargazer)
library(AER)
data = read.csv("hw2q2.csv")
attach(data)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     A, B, C, D, E, F

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

```r
g_lm = lm(C ~ F) # ok
a3_lm = lm(A ~ D) # bad
b3_lm = lm(B ~ E) # bad
a1_a2_a3_lm = lm(A ~ B+C+D) # ok, ???????????????
a1_a2_lm = ivreg(A ~ B+C|E+F) # ok
b1_b2_b3_lm = lm(B ~ A+C+E) # bad
```

```r
b1_b2_lm = ivreg(B ~ A+C|D+F) # ok

stargazer(a3_lm,g_lm,b3_lm,a1_a2_a3_lm,a1_a2_lm,b1_b2_b3_lm, b1_b2_lm,type="text",omit.stat=c("LL","ser
          model.numbers=TRUE, model.names = TRUE)
```

```
##
## ======================================================================================
##                                        Dependent variable:
##                 ----------------------------------------------------------------------
##                     A        C         B                 A                      B
##                    OLS      OLS       OLS       OLS   instrumental   OLS   instrumental
##                                                          variable             variable
##                    (1)      (2)       (3)       (4)      (5)         (6)      (7)
## --------------------------------------------------------------------------------------
## B                                            -0.797*** -0.813***
##                                               (0.005)   (0.016)
##
## A                                                                 -0.678*** -0.504***
##                                                                    (0.005)   (0.009)
##
## C                                            -0.786*** -0.790***  -0.608*** -0.505***
##                                               (0.009)   (0.026)    (0.010)   (0.025)
##
## D              -1.353***                     -0.811***
##                 (0.019)                       (0.011)
##
## F                        0.502***
##                          (0.010)
##
## E                                  -0.832***                      -0.372***
##                                     (0.018)                        (0.011)
##
## Constant        -0.006    0.005    -0.010    -0.011     -0.016     -0.015     -0.011
##                 (0.019)  (0.010)   (0.018)   (0.010)    (0.013)    (0.011)    (0.012)
##
## --------------------------------------------------------------------------------------
## Observations    10,000    10,000   10,000    10,000     10,000     10,000     10,000
## R2               0.334    0.202    0.171     0.819      0.701      0.718      0.624
## Adjusted R2      0.334    0.202    0.171     0.819      0.701      0.718      0.624
## ======================================================================================
## Note:                                                       *p<0.1; **p<0.05; ***p<0.01
```

II

II.1

II.2

II.2.a: When D increases by 1, A will _____.

II.2.b: When E increases by 1, A will _____.

II.2.c: When F increases by 1, A will _____.

II.3

III

Q3

I

```r
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
```

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

I.1

I.2

II

```r
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)
```

```
## Warning in ivreg.fit(X, Y, Z, weights, offset, ...): more regressors than
## instruments
```

```
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
```