BT3102 Assignment 2

Mar 4 to Mar 25, 2021

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.

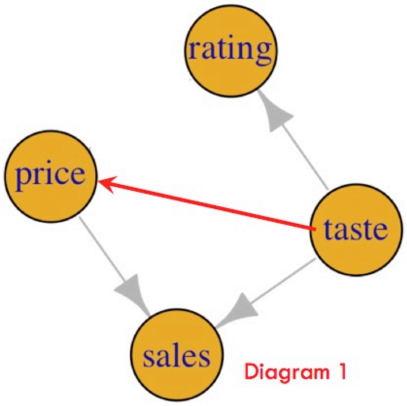

Diagram 1

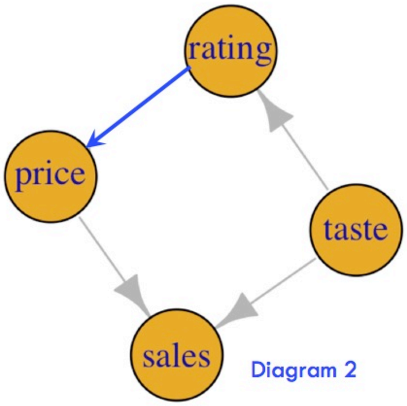

Diagram 2

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).

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.

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

Q2. Assume that $\alpha_1\beta_1 \neq 1$. Let N=10000. Consider the following DGP:

D=rnorm(N)

E=rnorm(N)

F=rnorm(N)

C= γ $*$F+rnorm(N)

A= α_1*B+α_2*C+α_3*D+rnorm(N)

B= β_1*A+β_2*C+β_3*E+rnorm(N)

I. Assume that $\alpha_1 = 0$ is known. Simulate a data set with $\alpha_2 = \alpha_3 = \beta_1 = \beta_2 = \beta_3 = \gamma = 1$.

1) Draw the causal diagram. Discuss the correlation and dependence of the variables.

2) Show all the collider variables and discuss how they may bias estimates.

3) Suppose that you want to build a regression model to <u>predict</u> the value of A. Which variables do you include in the regression model? Is the regression model you proposed for prediction also a good model for causal inference? Comment.

4) Suppose that you want to identify the causal relation between A and B. Show whether or not each of the following data is enough. Comment.

   a. A, B and C.
   b. A, B and D.
   c. A, B and E.
   d. A, B and F.
   e. A, B and C_measured, where C_measured=C+ rnorm(N)
   f. A, B and D_measured, where D_measured=D+ rnorm(N)
   g. A_measured, B and D, where A_measured=A+ rnorm(N)
   h. A_measured, B and C, where A_measured=A+ rnorm(N)

II. Simulate a data set with $\alpha_1 = \alpha_2 = \alpha_3 = -0.8, \beta_1 = \beta_2 = \beta_3 = -0.5$ and $\gamma = 0.5$.

1) Draw the causal diagram. Discuss the correlation and dependence of the variables.

2)

   a. When D increases by 1, A will ___.
   b. When E increases by 1, A will ___.
   c. When F increases by 1, A will ___.

3) With (all) the data you just simulated, show how you identify the values of $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$, and γ.

III. Download "hw2q2.csv." The data were generated by the DGP. Identify the values of $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$, and γ.

Q3. Download the data file "Attend.csv."

I. To determine the effects of attending lecture on final exam performance, estimate a regression model relating ***stndfnl*** (the standardized final exam score) to ***atndrte*** (the percentage of lectures attended). Include the dummy variables *fresh* (indicator for a freshmen student) and ***soph*** (indicator for a sophomore student) as explanatory variables.

1) Interpret the regression coefficient on ***atndrte*** and discuss its significance.

2) How confident are you that the OLS estimate is estimating the causal effect of student attendance? Explain your answer.

II. As proxy variables for a student's ability, add to your regression model in part (I) the variables ***priGPA*** (prior cumulative GPA) and ***ACT*** (achievement test score).

1) Now what is the effect of the ***atndrte*** variable? Discuss how this effect differs from that in part (I).

2) What happens to the statistical significance of the dummy variables ***fresh*** and ***soph*** now as compared with part (I)? Explain how this may have come about.

III. Use the ***hwrte*** variable as an instrumental variable (IV) for ***atndrte***. Perform an IV estimation based on your regression model in part (II). Comment on the results of this IV regression estimation. Comment on the validity of using the ***hwrte*** variable as an IV for ***atndrte*** if you suspect that the ***atndrte*** variable is endogenous.