Econ 21020 - Problem Set 4

Due 11/18 at 11:59 PM. Submit to Canvas. May be completed in groups of up to 6 students. Only one submission is required per group.

Problem 1

In class we said that, for an n-dimensional column vector X,

$$Var(X) = E[(X - E[X])(X - E[X])']$$

is an $n \times n$ dimensional matrix where the element in the ith row and jth column is $Cov(X_i, X_j)$. Show/explain why this is the case.

Problem 2

We're interested in the relationship between two random variables $X \in \{0, 1\}$ and $Y \in \{0, 1\}$. Specifically, we're interested in something called the "odds ratio." Define following notation:

$$p(y, x) = P(Y = y, X = x)$$
$$p(y|x) = P(Y = y|X = x)$$

Suppose that $p(y, x) > 0$ for all possible combinations of (y, x). Then, we'll define the odds ratio as:

$$OR = \frac{\frac{p(1|1)}{p(1|0)}}{\frac{p(0|1)}{p(0|0)}}$$

(a) Express OR in terms of $p(0, 0)$, $p(0, 1)$, $p(1, 0)$, and $p(1, 1)$.

(b) Suppose we have a sample $(Y_1, X_1), ..., (Y_n, X_n)$ that are iid $\sim (X, Y)$. Define

$$\hat{p}_n(y, x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i = y, X_i = x\}$$
$$\hat{OR}_n = \frac{\frac{\hat{p}_n(1,1)}{\hat{p}_n(1,0)}}{\frac{\hat{p}_n(0,1)}{\hat{p}_n(0,0)}}$$

Show that $\hat{OR}_n$ is a consistent estimator for OR.

(c) What will

$$\sqrt{n}(\begin{bmatrix}\hat{p}_n(1,1)\\ \hat{p}_n(1,0)\\ \hat{p}_n(0,1)\\ \hat{p}_n(0,0)\end{bmatrix} - \begin{bmatrix}p(1,1)\\ p(1,0)\\ p(0,1)\\ p(0,0)\end{bmatrix})$$

converge to in distribution as $n \to \infty$? (The things in square brackets are 4×1 column vectors). Make sure the variance of the limiting distribution is specified (a general element-wise description is sufficient - not need to lay out the entire matrix). (Hint: look at the multivariate version of one of our familiar statistics results).

Problem 3

Suppose we're interested in studying the association of wages with other variables. Sepcifically, we consider a regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

where

$$\begin{aligned} Y &= \text{wage} \\ X_1 &= \text{age in years} \\ X_2 &= \text{years of schooling} \\ X_3 &= \text{years of experience} \end{aligned}$$

It is the case that, among our population of interest, everyone starts school at the age of 6 and everyone works every year that they are not in school (and are older than 6).

(a) We cannot estimate this regression consistently (even under a descriptive interpretation). Why not? Explain specifically why the problem you identify arises.

(b) Propose an alternative regression that avoids the problem mentioned in part (a).

(c) In our description of the context, we implicitly assume that there is no such thing as "unemployment" - everyone either works or is in school every year after the age of 6. Suppose we now say that unemployment is a possibility (it is possible for someone to neither work nor be in school). Will the issue identified in part (a) still apply?

Problem 4

One of the values of multivariate linear regression is that it allows us to specify more general types of relationships between variables than simple linear regression. In class we discussed interaction effects as an example of this. We'll now look at another type of example. We define Y as: $Y = X + X^2$ where $X \sim N(0,1)$. Then, $E[Y|X] = X + X^2$.

(a) Consider a simple linear regression:

$$Y = \beta_0 + \beta_1 X + U$$

Under the best linear predictor interpretation, β_0 and β_1 will satisfy the following first-order conditions:

$$\begin{aligned} E[Y - \beta_0 - \beta_1 X] &= 0 \\ E[X(Y - \beta_0 - \beta_1 X)] &= 0 \end{aligned}$$

Solve this system of equations for β_0 and β_1. It may be useful to note that for $X \sim N(0,1)$, $E[X^2] = 1$ and $E[X^3] = 0$.

(b) Draw pictures of the best linear approximation to $E[Y|X]$, $\beta_0 + \beta_1 X$, and the actual $E[Y|X]$ on the same graph. The graph need not be extremely precise - it just needs to capture the major features of the functions.

(c) Now consider the multivariate linear regression:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U$$

Under the best linear predictor interpretation, β_0, β_1, and β_2 will satisfy the following first-order conditions:

$$\begin{aligned} E[Y - \beta_0 - \beta_1 X - \beta_2 X^2] &= 0 \\ E[X(Y - \beta_0 - \beta_1 X - \beta_2 X^2)] &= 0 \\ E[X^2(Y - \beta_0 - \beta_1 X - \beta_2 X^2)] &= 0 \end{aligned}$$

Solve this system of equations for β_0, β_1, and β_2. Make use of the previously given moments of a standard normal, and that $E[X^4] = 3$ for $X \sim N(0,1)$.

(d) Draw pictures of the new best linear approximation to $E[Y|X]$, $\beta_0 + \beta_1 X + \beta_2 X^2$, and the actual $E[Y|X]$ on the same graph. The graph need not be extremely precise - it just needs to capture the major features of the functions.

(e) We see that the approximation improves by cleverly allowing for the nonlinearity in Y. Think of a real-life Y and X where allowing for nonlinearities in such a way may be useful. That is, think of a Y and X that you think might be (very approximately) described by $Y = X + X^2$ Neither math nor accuracy to the real world are required - just think of an example that you think might fit and give some economic intuition.

Problem 5

Let's consider another example of the selection on observables identification strategy (based on Fagereng et al (2021)). This paper considers the question of why wealthy parents tend to have wealthy children. Specifically, the paper is interested in the extent to which wealthiness passes from parents to children due to favorable genetic characteristics versus monetary endowments (buying stuff for the kid after they are born, like better schooling, or simply giving the kid money).

(a) Consider a causal model:

$$Y = \beta_0 + \beta_1 W + U$$

where:

$$\begin{aligned} Y &= \text{child's wealth (upon reaching adulthood)} \\ W &= \text{parents' wealth} \end{aligned}$$

(We assume that W will here stand in for the totality of parental characteristics). Fagereng et al were concerned that they could not consistently estimate the causal parameter β_1 for this model. Explain the specific concern given the description of the research question.

(b) To answer this question, the paper considers a situation in the 20th century in which many Norwegian families adopted Korean children through a centralized agency. The agency did not allow the adoptive families to request any kind of characteristics of their adoptive children. Instead, the agency would simply match families with the next child in line for adoption, in the order that families were approved for adoption (where the order depends on when the family applied to adopt and how long it took them to get approved).

Define new variable

$$T = \text{Measure of when the adoptive family's application was approved}$$

What assumption can we make about this new variable in order to enable us to identify β_1 from part (a)? Interpret this assumption in words. (Hint: Follow the class example of using a control variable to identify a causal parameter).

(c) Consider the new regression equation:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 W + \tilde{\beta}_2 T + \tilde{U}$$

Using the assumption that you made in part (b), show that $\tilde{\beta}_1$ will consistently estimate the causal β_1 from part 1. (Hint: Again, follow the example from class).

(d) What will $\tilde{\beta}_2$ estimate? Is it causal?

Problem 6

Lets return to the data from Angrist and Krueger (1999), as was used in Problem 6 of the previous problem set. We will again refer to X as years of education and Y as log wage. We will continue to assume that $E[X^4], E[Y^4] < \infty$.

(a) The regression equation:

$$Y = \beta_0 + \beta_1 X + U \tag{1}$$

is likely difficult to interpret causally. Pick one “component” of causally-defined U that you would expect to be correlated with X (thereby preventing us from consistently estimating causal β_1). For your chosen component of U, guess what direction you think that omitting that variable will “bias” the estimate of OLS $\hat{\beta}_1$ (relative to causal β_1), appealing to the formula for omitted variable bias and your economic intuition.

(b) Suppose someone proposed using variable: $A =$ year of birth (“year_of_birth” in the data set) as a control variable, and claims that including this in the regression:

$$Y = \beta_0 + \beta_1 X + \beta_2 A + U \tag{2}$$

will allow us to estimate β_1 as a causal parameter consistently. Do you think this idea makes sense? Why or why not?

(c) Perform regressions according to equations (1) and (2) using R or another language of your choice.

(d) Appealing to knowledge rather than computational outputs, what will happen to the R^2 going from (1) to (2)? Will the same thing necessarily happen to the adjusted R^2?

(e) (5 Points Bonus) Calculate the R^2 and adjusted R^2 for (1) and (2). Interpret what you find (pay close attention to the formulae for those two statistics).