Econ 21020 - Problem Set 5

Due 12/7 at 11:00 AM. Submit to Canvas. May be completed in groups of up to 6 students. Only one submission is required per group. Note that this assignment is graded for completion, only.

Problem 1

You want to estimate the proportion of UChicago students who have ever cheated. Call the true value of this proportion θ. However, you think if you just ask students this question, they might lie. Instead, you collect an iid sample of n students, where you give each student the instructions:

1. Flip a fair coin (50/50) secretly.
2. If the coin comes up heads, answer the question, "Have you every cheated?" If the coin comes up tails, give the response "Yes."

Under this procedure, you assume that everyone will answer honestly if they get a heads. Let X_i denote the response of the ith student, where

$$X_i = \begin{cases} 1 \text{ if they say "Yes"} \\ 0 \text{ if they say "No"} \end{cases}$$

However, you do not observe the outcome of the coin toss for each student.

(a) As a function of θ, what is $P\{X_i = 1\}$?

(b) Show that $\hat{\theta}_n = \frac{2}{n}\sum_{i=1}^{n} X_i - 1$ is a consistent estimator for θ.

(c) Find the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ (the distribution as $n \to \infty$).

(d) Propose an estimator, $\hat{\sigma}$, such that

$$\frac{1}{\sqrt{\hat{\sigma}}}\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0,1)$$

Show that this convergence in distribution takes place.

Problem 2

In the multivariate case of linear regression, we brushed over the idea of homoskedasticity, and went right to heteroskedasticity robust inference. However, we can define homoskedasticity analagously in the multivariate case: homoeskedasticity holds if $E[U|X] = 0$ and $Var(U|X) = Var(U)$, where X is a $(k+1) \times 1$ random vector.

(a) Show that

$$\Sigma = E[XX']^{-1} Var(XU) E[XX']^{-1}$$

is equal to

$$\Sigma^{Ho} = E[XX']^{-1} Var(U)$$

if U is homoskedastic. (Hints: Start by working with $Var(XU)$. The definitional of the conditional variance may be useful. This will in general look similar to an analagous result in the univariate case.)

(b) Show that $Var(U) = E[U^2]$, (and so $\Sigma^{Ho} = E[XX']^{-1}E[U^2]$) under any interpretation of linear regression.

Problem 3

In the case of multivariate linear regression, we can test more types of hypotheses than we did in the univariate case. Let's look at one other type: testing a hypothesis that two subcomponents of β are equal. That is, consider a regression equation:

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

and the hypotheses $H_0 : \beta_1 = \beta_2$, $H_1 : \beta_1 \neq \beta_2$. These are equivalent to the hypotheses $H_0 : r'\beta = 0$, $H_0 : r'\beta \neq 0$ for the 3×1 vector:

$$r = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

(a) Find the limiting distribution of:

$$r'\sqrt{n}(\hat{\beta} - \beta)$$

Represent the variance of the limiting distribution in terms of r and Σ, where Σ is the typical variance of the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$,

$$\Sigma = E[XX']^{-1} Var(XU) E[XX']^{-1}$$

(b) What are the dimensions of the variance of the limiting distribution of $r'\sqrt{n}(\hat{\beta} - \beta)$?

(c) Propose an estimator $\hat{\sigma}$ such that

$$\frac{1}{\sqrt{\hat{\sigma}}} r' \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0,1)$$

Show that this convergence in distribution takes place. (Hint: Take a look at the version of inference on multivariate linear regression that we did in class. You may make use of the result that $\hat{\Sigma}$, as defined in lecture, is a consistent estimator for Σ.)

(d) Think of real-world Y, X_1, and X_2 for which this type of hypothesis test may be interesting.

Problem 4

This question develops another example of a LATE, based on the paper Angrist (1990).

(a) Consider a causal model, of the effect of military service on (post-service) wages. Specifically,:

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

where

$$W = wage$$
$$M = \begin{cases} 1 \text{ if served in the military} \\ 0 \text{ if not} \end{cases}$$

Why might it be the case that $E[MU] \neq 0$, where U is defined in the causal model sense?

(b) Consider now an instrumental variables approach, using the military draft implemented by the US government during the Vietnam War as an instrument. For the sake of this example, assume the draft works very "simply" - every single US man aged 19-26 is entered into the draft and a subset are drafted.[1] Those who are drafted are called up to service by the government, under threat of legal action. Define the new variable, D:

$$D = \begin{cases} 1 \text{ if drafted} \\ 0 \text{ if not} \end{cases}$$

The LATE interpretation of IV requires three assumptions:

(a) $(W_1, W_0, M_1, M_0) \perp D$ (implies instrument exogeneity)

(b) $M_1 \neq M_0$ sometimes (analogous to instrument relevance)

[1] From now on, we will consider the "population" to be US men aged 19-26 at the time of the Vietnam War draft.

(c) $M_1 \geq M_0$ always - called Monotonicity

where W_1 and W_0 are the potential outcomes corresponding to the two values of M and M_1 and M_0 are the potential treatments corresponding to the two values of D. Evaluate each of the three LATE assumptions in this context (are they reasonable to assume? Why or why not?).

(c) We can split the population into three groups:

- Always-takers: People for whom $M_1 = 1,\ M_0 = 1$
- Never-takers: People for whom $M_1 = 0,\ M_0 = 0$
- Compliers: People for whom $M_1 = 1,\ M_0 = 0$

Interpret in words who the members of each group are.

(d) Define the LATE for this regression. Interpret it in words. Is this LATE interesting (this last question is more or less wholly subjective - feel free to argue either way, demonstrating your knowledge of what a LATE is)?

Problem 6

We now conclude our discussion of the data from Angrist and Krueger (1999), following fairly close the identification strategy that they use in their paper. We will again refer to X as years of education and Y as log wage. We will continue to assume that $E[X^4], E[Y^4] < \infty$.

Consider quarter of birth as an instrument for number of years of education completed. The idea of this instrument is thus - students are legally required to attend school until a certain age. Given the cutoffs for students to be assigned to grades in US schools, students born earlier in the year tend to be older than their peers are, during each grade. Therefore, students born earlier in the year will tend to, legally, be allowed to drop out of high school in earlier grades than their peers. In this way, quarter of birth will affect years of education completed while being, potentially, independent of other causal determinants of wage. Consider the following graph, reproduced from Angrist and Kreuger (1999):

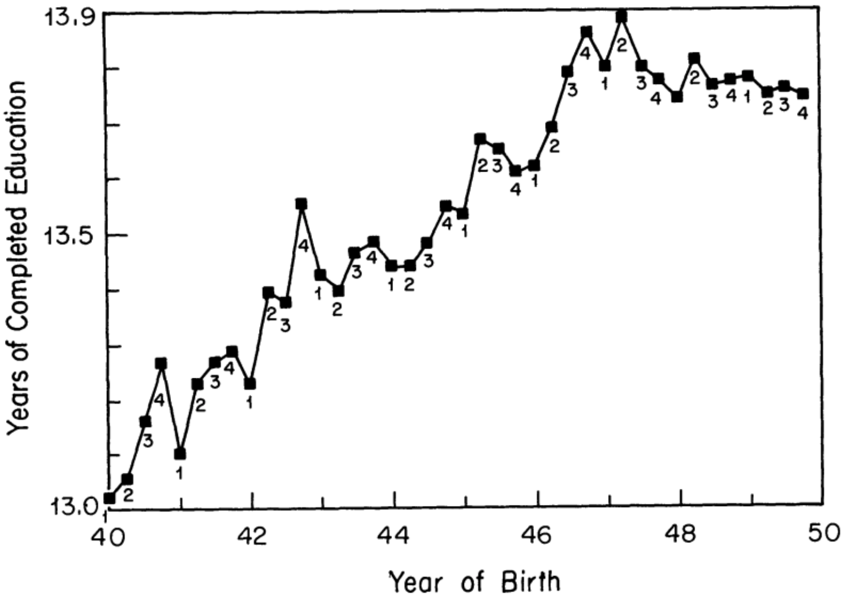

Figure II
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

For convenience, let's define our instrument, Z, as:

$$Z = \begin{cases} 1 \text{ if born in quarter 1} \\ 0 \text{ if born in quarters 2-4} \end{cases}$$

(a) For a valid IV, we need to satisfy two assumptions: instrument exogeneity and instrument relevance. Do you think these will be valid in this context? Why or why not?

(b) Calculate the IV estimand for the equation:

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

using Z as an instrument, using your preferred software. Interpret the output.

(c) Suppose we considered a binary version of our schooling variable,

$$X' = \begin{cases} 1 \text{ if graduated HS} \\ 0 \text{ if not} \end{cases}$$

Then, if we assumed our LATE assumptions are true, we could interpret the paramter β_1' from

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

using Z as an instrument, as a LATE. Write down an expression for this LATE and interpret it in words.