Endogeneity and Instrumental Variables (24 Questions)

Question 1

True, false or uncertain: Unlike simple regression-based methods, difference-in-differences and regression discontinuity designs *always* provide us with reliable causal estimates.

- ◯ True
- ◉ False
- ◯ Uncertain

✔

Explanation
The interpretation of your estimates depends on the validity of your assumptions in whatever design you use. If you cannot credibly claim that the parallel trends assumption is satisfied, then estimates obtained from a differences-in-differences design cannot be interpreted causally. Similarly, for regression discontinuity design, if there are reasons to believe that people can manipulate the threshold to decide what side of the cutoff they are on, then that will undermine your estimates.

Ultimately, the causal interpretation of your coefficients hinges on the credibility of the set of assumptions you are making irrespective of your design.

Question 2

Select all that apply: If it is plausible that the outcome variable affects the regressor of interest, then

- [x] There is an endogeneity problem
- [ ] One should add more control variables
- [ ] One should use difference-in-differences
- [ ] One should use regression discontinuity

✔

Explanation
There is an endogeneity problem, because the coefficient will measure the effect of the regressor on the outcome variable AND the effect of the outcome variable on the regressor.

Endogeneity and Instrumental Variables (24 Questions)

Question 1

Suppose you are interested in estimating the following relationships using a simple OLS model. In each case, you regress the first variable on the second variable. For which of these are you likely to have an omitted variable problem?

- ◯ Cancer and smoking
- ◯ Grades and class attendance
- ◯ Incarceration rates and police presence
- ◉ All of the above

✔

Explanation
In all the cases above, it is highly likely that there are variables that are correlated with both the outcome and the regressor that are missing from this simple model. So your estimates are likely to be biased because there is an omitted variable problem.

For example, consider the following possible stories in each of these scenarios:

For incarceration rates and police presence, places with higher incarceration rates might have more crime, leading to more arrests, and law enforcement might be more present in places where there is more crime.

For grade and attendance, better (smarter) students might be more likely to attend class because they are smart, and it is easier and more enjoyable for them. And smarter students are probably more likely to get better grades irrespective of whether they attend class.

For cancer and smoking, smokers in the US are more like to be low-educated and live below the poverty line. Belonging to this demographic may in turn expose them to other carcinogens (cancer-causing substances), either through their occupation or their environments. Alternatively, other behaviors, such as increased alcohol consumption along with smoking further increase the risk of cancer.

Question 2

Suppose you are interested in estimating the following relationships using a simple OLS model, for which of the following cases are you likely to have an endogeneity problem? Again, your regress the first variable on the second variable. Select all that apply.

- ☑ Income and democracy ✔
- ☐ Mothers' education and children's immunization rates
- ☑ Sleep and coffee consumption ✔
- ☐ Cancer and smoking

Explanation
Mothers' education and children's immunization rates, cancer and smoking, and price of fish and weather are unlikely to have an endogeneity problem because the relationship is unlikely to go both ways. A mother's education might affect childrens' immunization rates, but whether or not a child is immunized probably doesn't influence their mothers' education (although there may still be an omitted variable bias).

On the other hand, estimating income and democracy as well as sleep and coffee consumption using a simple OLS model might conflate the effect of the regressor on the outcome with the effect of the outcome on the regressor. Democracy might increase GDP, but richer countries might be more likely to democratize. Drinking coffee might lead to sleeplessness and a poor night's sleep. In turn, a poor night's sleep makes people more tired, which might lead to higher coffee consumption.

Question 1

In the example, what is used as the instrument for education?

- ◯ Ability
- ◉ Randomly assigned scholarships
- ◯ Cognitive score
- ◯ Yearly earnings

✔

Explanation
Yearly earnings and cognitive score are the outcome variables. Ability was an example of a source of omitted variable bias in a naïve regression. Randomly assigned scholarships to seniors in high school are used as an instrument for education.

Question 2

In which of the following examples would you face the issue of being unable to impose compliance with random treatment and have to use an instrument that affects the likelihood that people make the choice of being treated instead?

- [x] The effect of education on earnings ✔
- [x] The effect of smoking on cancer ✔
- [ ] The effect of supplying electricity to villages on economic outcomes
- [ ] The effect of providing clean drinking water in existing connections on infant health

Explanation
People cannot be forced to keep their kids home from school. Instead, using scholarships (for example) as an instrument for education can help determine its effect on schooling. Similarly, people cannot be forced to take up or quit smoking, but an anti-smoking ad campaign can be used as an instrument for lower smoking levels. Randomly supplying electricity to villages, e.g. by randomizing the timing of the rollout, can be perfectly enforced and does not need an instrument per se. The random provision of clean drinking water can also be perfectly enforced by a utility.

Endogeneity and Instrumental Variables (24 Questions)

Question 1

Why would your estimate of β in the following simple OLS model be a biased estimate of the effect of education on earnings? Select all that apply.

$$Y_i = \alpha + \beta A_i + \epsilon_i$$

where Y_i denotes earnings and A_i is whether individual i goes to secondary school

- [ ] We believe that education is a bad instrument.
- [ ] We do not believe that Y and A_i are correlated.
- [x] We do not believe that A_i and the error term are uncorrelated
- [x] We believe that there are omitted variables in these regression, which are correlated both with ϵ and A_i.

Explanation
We do not believe A_i and the error term are uncorrelated because by definition, since ability is an omitted variable that means it is correlated with both A_i and the error term. So A_i and the error term not being correlated implies your $\hat{\beta}$ would be biased. Saying that there are omitted variables in these regression which are correlated both with ϵ and A_i is a more explicit version of the previous answer, so it is also correct.

Y and A_i not being correlated means your true beta is 0, it says nothing about the interpretability of your $\hat{\beta}$. Education as a bad instrument is irrelevant because we are not instrumenting anything in the model above.

Question 2

Professor Duflo mentioned that the secondary school study had a 97% tracking rate. Which of the following might be the problem if they had a low tracking rate? Select all that apply.

- [x] The final sample for which they have data might not be representative of their initial sample.
- [ ] Low tracking rates imply that your instrument was not randomly assigned.
- [ ] Low tracking rates imply that your instrument is not correlated with your regressor.
- [x] The set of people who dropped out might be different in the treatment and control groups, leading to a final sample that is not comparable in Treatment and Control groups.

✔

Explanation
There are two concerns with low tracking: first, the sample you end up with might be different from the sample you started with. Therefore, you might be worried that you are not estimating what you thought you were or on the relevant population of interest. The second reason one might worry about low tracking rates is that the set of people who couldn't be tracked might be different in both groups. This would imply that the remaining samples in each group are no longer comparable, and your estimates would be biased since they fail to account for that.

Low tracking rates do not imply that your instrument was not randomly assigned because this can happen in an RCT. I talso doesn't imply anything about instrument's correlation with your regressor.

Endogeneity and Instrumental Variables (24 Questions)

Question 3

True or False: The fact that the scholarships are randomly assigned enabled Prof. Duflo and her co-authors to identify the causal effect of assigned scholarships on school completion rates.

- (selected) True
- False

✔

Explanation
In this segment, Prof. Duflo uses the random assignment of scholarships to estimate the causal effect of assigned scholarships on secondary school completion rates

Question 1

From the example, the estimated effect of going to secondary school on cognitive scores is equal to:

- (selected) The effect of the scholarship on cognitive scores divided by the effect of the scholarship on going to secondary school.
- The effect of the scholarship on going to secondary school multiplied by the effect of the scholarship on cognitive scores
- The effect of the scholarship on going to secondary school divided by the effect of the scholarship on cognitive scores

✔

Explanation
Your estimated effect is given by $\hat{\beta} = \frac{E[Y_i \mid Z_i=1]-E[Y_i \mid Z_i=0]}{E[A_i \mid z_i=1]-E[A_i \mid z_i=0]}$ where Y_i denotes cognitive scores, Z_i is a dummy variable equal to 1 if assigned to the treatment group and 0 otherwise, and A_i denotes whether individual i goes to secondary school.

Question 2

For men, the scholarship increases the chances of completing secondary school by 38.3% and increases cognitive scores by $.125$ standard deviations, so the estimated effect of secondary school on cognitive scores is:

- $.125/38.3$ standard deviations increase in cognitive scores
- $.383/.125$ standard deviations increase in cognitive scores
- (selected) $.125/.383$ standard deviations increase in cognitive scores
- $38.3/.125$ standard deviations increase in cognitive scores

✔

Explanation
This is the correct answer, because the estimated effect of secondary school on cognitive scores will be equal to the effect of the scholarship on cognitive scores divided by the effect of the scholarship on going to secondary school ($\hat{\beta} = \frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[A_i|Z_i=1]-E[A_i|Z_i=0]}$)

Question 3

Your IV estimate for the effect of education on cognitive scores is going to be _________ the effect of scholarships on cognitive scores.

- (●) Larger than or equal to
- ( ) Smaller than
- ( ) Larger than
- ( ) Smaller than or equal to

✔

Explanation
If everyone who got the scholarship went to secondary school, the denominator in the formula above would be equal to 1, and therefore, beta hat would just the same as the effect as the scholarship. But if not everybody goes to school, you are dividing the numerator by a number smaller than one, so you will get a larger number: intuitively, it "scales up" the reduced form coefficient to take into account that the difference is accounted for only a fraction of the people who actually went to school.

Question 1

A good instrument needs to: (Select all that apply)

- [x] Be correlated with your regressor of interest ✔
- [x] Randomly assigned or as good as randomly assigned ✔
- [x] Have no direct effect on the outcome variable of interest. ✔
- [ ] Positively correlated with the outcome variable

Explanation
If the instrument is not correlated with the regressor, then it will not produce variation in the regressor. If the instrument is not randomly assigned, then the instrument may not produce only exogenous variation in the regressor. If the instrument directly affected the outcome variable, then we could not distinguish between the effect from the variation in the regressor and the direct effect. Positive correlation with the outcome variable is irrelevant.

Endogeneity and Instrumental Variables (24 Questions)

Question 2

Which of the conditions for a good instrument **cannot** be tested or proven empirically? (Select all that apply)

- [ ] Your instrument is correlated with your regressor of interest.
- [ ] Your instrument is randomly or as good as randomly assigned.
- [x] Your instrument has no direct effect on Y
- [ ] They can all be tested.

Explanation

We can test the correlation very easily by regressing the regressor on the instrument. With an RCT, the fact that your instrument is random is guaranteed (in other cases, we might need to assume this, however, you can show that the instrument is uncorrelated with other observable baseline characteristics to support your claim that this assumption holds). We cannot test the instrument's direct effect on Y, because the direct effect on Y and the indirect effect (through the instrumented for regressor) cannot be disentangled. The key assumption underlying any IV strategy is that the only impact of your instrument on your outcome (test scores) is through your instruments' effect on your regressor (attendance). This is commonly known as the "exclusion restriction."

This is not an assumption you can test or prove empirically, but you can try and think of what other variables might be affected by your instrument and affect your outcome of interest and try to find evidence in support of your claim.

Question 3

Which of the following statements represent the assumption that the "exclusion restriction" is satisfied in this context?

Note: A is an indicator for receiving a secondary education, Z is an indicator for receiving a scholarship, and Y is an outcome of interest.

- ○ $E[Y_i|Z_i=1]-E[Y_i|Z_i=0]\neq 0$
- ○ $E[A_i|Z_i=1]-E[A_i|Z_i=0]\neq 0$
- ○ $E[X_i|Z_i=1]-E[X_i|Z_i=0]=0$, where X_i is a vector of observable baseline characteristics.
- ◉ $E[\epsilon_i|Z_i=1]-E[\epsilon_i|Z_i=0]=0$

✔

Explanation

As explained, the exclusion restriction refers to the assumption that is that the only impact of your instrument on your outcome (test scores) is through your instruments' effect on your regressor (attendance). This requires that there is no omitted variable in your model which is correlated with both assignment and test scores. This will only be true if your error term is uncorrelated with your instrument.

Endogeneity and Instrumental Variables (24 Questions)

Question 1

We might be concerned with our estimate for the impact of education on test scores from the scholarship experiment, because scholarships could __________ . Select all that apply.

- [x] Ease monetary constraints for those who would have gone to school anyways ✔
- [x] Give self-confidence ✔
- [ ] Lead to more education
- [ ] None of the above

Explanation
Those who did not receive scholarships often scraped together money in order to go to school, which may have lowered their nutrition or increased their stress. Being chosen for a scholarship may have made individuals feel special or competent, which could increase their test scores.

Question 2

Which of the following is a legitimate way to choose the set of outcomes for which you report your results if you collected data on a lot of outcomes? (select all that apply)

- [x] Pre-analysis plan
- [ ] Report everything
- [x] Write down a model ex-post justify which variables
- [ ] Whatever is significant

✔

Explanation
With a large survey with many variables, you are likely to find some significant results just by chance if you report everything or whatever is significant. By making a pre-analysis plan or an ex-post model, you can promise to focus on a few variables, which means that the chances that you'll find significant results by chance are reduced.

Endogeneity and Instrumental Variables (24 Questions)

Question 1

What is the Wald estimate? Select all that apply.

- [x] The reduced form relationship over the first stage relationship. ✔
- [ ] An estimate of the relationship between the instrument and the instrumented for regressor
- [ ] An estimate of the relationship between the instrument and the outcome variable
- [x] The simplest form of the instrumental variables estimator. ✔

Explanation
By definition, the Wald estimate is the reduced form relationship over the first stage relationship. Intuitively, you want to scale the effect of your instrument by your effect of your instrument on your regressor (so it makes sense to divide by the first stage). The estimate of the relationship between the instrument and the instrumented for regressor is the relationship you are estimating in your first stage. The estimate of the relationship between the instrument and the outcome variable is the relationship you estimate in your reduced form. It is the simplest because it uses a dummy variable as the instrument.

Question 2

In general, if some of the IV assumptions are violated, your reduced form estimates will be _______ biased relative to your Wald Estimate.

- ◯ More
- ◉ Less ✔
- ◯ Equally as

Explanation
Continuing with the scholarship example, suppose that scholarships affect test scores through both attendance and through their effect on psychological outcomes (ex. Self-esteem). This would imply that your estimate of the effect of scholarships on test scores would be a combination of the effect of attendance, and the psychological effect. This is your reduced form estimate, which is the numerator of your Wald estimate. So if self-esteem is omitted from your model, your reduced form will be biased. But your Wald estimate is just your reduced form divided by your first stage estimate, which is less than or equal to 1. This implies that the bias in your reduced form estimates will be "blown up" in your Wald estimate because of the fact that first stage is less than 1. So your Wald estimate will be more biased than your reduced form estimates.

Endogeneity and Instrumental Variables (24 Questions)

Question 1

Which of the following would imply a violation of the exclusion restriction in Guido Imbens' flu shot experiment? Recall that the experiment aims to understand the effect of flu shots on sick days.

- (x) If the letter reminded doctors to give advice regarding handwashing or other ways to prevent the flu
- ( ) If some doctors did not read the letter
- ( ) If some patients did not follow ignored the doctor's advice
- ( ) If the letter reminded doctors to give flu shots

✔

Explanation
The letter reminding the doctors to give flu shots is a necessary assumption for the validity of the instrument, otherwise the first stage would be 0, and both your estimate and its standard errors would blow up to infinity. If some doctors did not read the letter, it might weaken the first stage relationship but would not make the instrument invalid. If some patients ignore their doctors, it also might weaken the first stage relationship but would not make the instrument invalid. If the letter reminded doctors not only about flu shots, but also about handwashing, then this would violate the exclusion restriction by affecting sick days through a channel other than the flu shot.

Question 1

What would make the lottery outcome or being offered to attend a charter school an invalid instrument for the effect of charter schools on test scores for the pool of applicants?

- ( ) Some who win the lottery choose not to go to the charter school
- ( ) It is a perfect instrument
- (x) Winning the lottery gave the winners psychological benefits relative to the losers
- ( ) The pool of applicants is smarter than the general population

✔

Why? (Select all that apply)

- [ ] It is not randomly assigned
- [ ] It does not change where students go to school
- [x] It is correlated with omitted variables that can affect outcomes
- [x] It violates the exclusion restriction

✔

Explanation
The pool of applicants being smarter could affect generalizability but not internal validity. If the lottery gave psychological benefits to winners (for example, some people may think winning the lottery signals that they are destined to do well in school) this could increase their test scores independently, rather than through the effect of attending a charter school, which would bias the results. Some people not choosing to go to charter school would only affect the strength of the first stage regression, though that would not necessarily imply your instrument is invalid (unless a huge number of people do so, or your sample size is tiny).

Question 1

Why would a child being in a school that was randomly assigned to receive deworming pills (only about half the kids in a treatment school actually receive the pills) be an invalid instrument for the effect of being dewormed on cognitive scores?

Hint: worms are highly contagious

- ○ Even the children who are not dewormed in the treatment school benefit because the children around them are more likely to give them worms.
- ○ The treatment and control schools might be different
- ○ It is a valid instrument
- ◉ Even the children who are not dewormed in the treatment school benefit, because the children around them are less likely to give them worms

✔

Explanation
Since the children who are not dewormed benefit, simply being in a treatment school (the instrument) directly affects cognitive scores (the outcome) which violates the exclusion restriction. The treatment and control schools should not be different because of random assignment. It makes no sense that children who are not dewormed benefit from the children around them being more likely to give them worms.

Endogeneity and Instrumental Variables (24 Questions)

Question 2

Say you are interested in the effect of years of education on test scores. Why would the random assignment of laptops to students probably be a bad instrument for years of education?

- ◯ Some laptops may break
- ◉ Laptops may improve cognitive abilities independently of their effect on years of education
- ◯ Laptops may also improve the test scores of the siblings of the child
- ◯ Some children may not know how to use the laptops

✔

Explanation
Lack of knowledge about how to use the laptops and laptops breaking would only weaken the first stage regression. If laptops improved cognitive abilities independent of their effect on years of education, then this would violate the exclusion restriction. Laptops improving the test scores of siblings should not be a problem as long as the siblings are not in our sample.

Question 1

The Local Average Treatment Effect (LATE) represents the effect on the:

- ◯ Treatment group
- ◯ Never takers
- ◉ Compliers
- ◯ Always takers

✔

Explanation
The LATE represents the effect on the compliers, because the compliers are the individuals who could be affected by the treatment.

Endogeneity and Instrumental Variables (24 Questions)

Question 2

What assumption would need to hold for the Wald estimate to equal the LATE?

- ◯ There are only "Always Takers"
- ◯ There are no "Never takers"
- ◉ There are no "Defiers"
- ◯ There are only "Compliers"

✔

Explanation
For the scholarship example, defiers are those who would go to secondary high school even if not offered a scholarship, but would not go if offered a scholarship. If these individuals existed, then we could not separate the effect on defiers from the LATE (the effect on the compliers).