Barcelona Graduate School of Economics

Econometrics sample exam Solution

Pan Yiming

November 27 2020

1. Explain in plain words (no formulas) and with as much intuition as possible, the following concepts.

a)The Hausman test in instrumental variables estimation:

We evaluate the consistency of an estimator when compared with an alternative less efficient estimator which is already known to be consistent. For the instrumental variables estimation, we use this method to test whether a explanatory variable is exdogenous. We run one-stage OLS and 2SLS by adding instrumental variables in 2SLS. If x is exdogenous, the estimators we get from OLS and 2SLS would be similar. Or else there exist endogneity, the 2OLS with instrumental variable can elimate the effect of endogneity. Therefore, the estimators should be significantly different. With endogneity, 2OLS estimator will be consistent but OLS will not.

b) The random effects estimator in panel data:

In the panel data, the individual specific effect is a random variable. We assume that it is uncorrelated with explanatory variable of all past, current, future of same individual. In addition, we assume that the convariance of individual specific effect is constant. Under these assumptions, we can estimate the convariance of the error term, so we can use the GLS estimator and the estimator is consistent and the most efficient estimator.

c) Generalized Least Squares estimation:

Estimating the unknown paremeters when there exists heteroscedasticity in the regression model. In this case, OLS can be inefficient and give misleading inference. Observations with larger variance should weight less because they contain too much noisy and less information. GLS overcome this issue by using the information of the convriance structure of model to obtain more efficiency. A special case for this is WLS.

d) The two conditions necessary for validity of an instrument in IV estimation

---

First, the IVs should be highly correlated with explanatory variables, which means IV should capture the convariance of the explanatory variables as much as possible. In this way we can get a more efficient estimator.

Second, The IVs should be exdogenous, which means it is uncorrelated with error term. If this is not true, 2OLS fails to provide a consistent estimator.

e) Homoskedasticity in a regression model:

The variance of the error term is the same accross obsevations and in particular does not depend on the values of the explanatory variable x.

g) The difference between a consistent and an unbiased estimator

---

Consistency means that as the sample size goes to infinite, the estimator will converge in probability to thr true paremeter.

Unbaised estimator means that on average the estimator hits the true parameter value. That is the mean of the smapling distribution of the estimators is euqal to the true paremeter.

h) A Chow test

Test whether the coefficients in two linear regression on different subset are equal. If equal, no structure change. Or else, we need take the structure change into consideration. We usually do Chow test by compararing the parameters of the restricted and unrestricted model. If there exists structure change, the parameter will be significantly differnt.

4. You have to estimate the effects of education on wages, and you are interested in estimating an equation that looks like:.

a) What is the interpretation of your parameter of interest, β_1?

β_1 means the education year increases one yeay, the wage will increase β_1.

b) You now have collected data for a sample of people and run the above model by using simple OLS regression (with traditional standard errors). Then your boss tells you that your data have six possible "characteristics" which may affect the validity of your analysis. For each of the characteristics 1-6 below (taken independently):
- Give a technical name to the problem they may cause.
- Describe how the problem will affect the results of the estimation (i.e. whether it affects the properties of the parameter estimates, the standard errors, etc.). - Identify and briefly describe what test you could use (if any exists!) to detect whether indeed that problem is present.
- Describe how you can solve the problem.

---

1) "It is well known that people with more education are more intelligent to begin with, and intelligence affects wages, but you do not have a measure of 'intelligence' in your model".

Omitted variable bias. You lost some relevant variables. Here the intelligence is not taken into consideration.

Effect:

It will make the estimator biased and inconsistent. In this problem, we do not include the intelligence, so will the estimate of coefficients is not unbiased. The direction of the bias depends on the estimators as well as the covariance between the regressors and the omitted variables.

How to test:

If you include different combinations of independent variables in the model, and you see the coefficients changing, you're watching omitted variable bias in action.

Solution:

Find an instrument for the education variable. The instrument should be highly correlated with education variable. But instrument should exogenous to to the wages, which means it is not correlated with error term in this model.

2) "The dispersion of wages is much bigger for people in executive positions than for people in non- executive positions".

Heteroscedasticity.

The variance of the error term is not the same accross obsevations and in particular does depend on the values of the explanatory variable x, which is the position is executive positions or not.

Effect:

It will cause the estimates of the variance of the coefficient to be biased. So the bias will lead to biased inference, which will further lead to the hypothesis test maybe wrong.

Test:

White-test: We first run the normal OLS. And then regress the residual errors on all explanatory variables, the square of the explanatory variables and product between explanatory variables. Then we test whether all the coefficients in the model all jointly equal to zeros. If yes, we accept the null hypothesis that there does not exit heteroscedasticity. Or else, we will say there exists heteroscedasticity. But in this way, we can not know exactly which one of the explanatory concerned with heteroscedasticity.

Breusch-pagan test:

We first run the normal OLS. And then then regress the residual errors on explanatory variable you choose or the estimate of y. And then test whether the coefficient is equal to zeros. If yes, we accept the null hypothesis that there does not exit heteroscedasticity. Or else, we will say there exists heteroscedasticity. And in this way, it is possible to know exactly which one of the explanatory concerned with heteroscedasticity.

Solution:

We use robust standard error to estimate the variance of error. So we can get more correct hypothesis test and inference. In addition we can take the heteroscedasticity into consideration by using GLS model, which will give us more efficient estimator with less variance. Observations with larger variance should weight less because they contain too much noisy and less information. GLS overcome this issue by using the information of the convriance structure of model to obtain more efficiency. A special case for this is WLS.

3) "The return of education (impact of education on wages) is likely to be different for males and for females".

Specification.

The model you choose is not right, in which we overlook the effect of interaction between gender and the education.

Effect:

It will cause the estimates will be biased and inconsistent.

Test:

By adding the product of female and education, we run OLS, when we see the obvious change in the coefficients, then there exits the effect of the interaction variable.

Solution:

Take the interaction of the female and education into consideration. And then get the new model.

4) "The people in your sample come from the customer database of a company that sells software products".

Sample selection bais:

We just sample the people from of selling the software products. So the sample does not cover the all population.

Effect:

It will make predictability of the model very poor for the out of sample data.

Solution:

Change your data by taking more general samples, which can represent the whole population.

5) "Total working experience and tenure in a company are variables that are very much related"

Collinearity:

The high correlation between the explanatory variables. In this model, the total work experience is correlated with tenure in a company.

Effect:

It will make the estimators have high variance, which mean low significance.

Test:

Regress the explanatory variable(here is total work experience) on the other explanatory variables. If VIF$>$10 we can say there exist serious collinearity. If VID is close to 1, the collinearity is not serious.

Solotion:

We can solve it by selecting the explanatory variables again. Or change the data. In addition, we can collect more data.

6) "Education and wages are two 'characteristics' of a person that are the joint result of common factors such as the education of parents, etc...".

Endogeneity:

The explanatory variable is correlated with error term. In this case, education and wage are joint result of common factors such as the education of parents. So we can see that the education cause the endogeneity.

Effect:

It will cause the biased and inconsistent estimators.

Test:

Hausman Test:

Run OLS and 2SLS by adding instrumental variables in 2SLS. If x is exdogenous, the estimators we get from OLS and 2SLS would be similar. Or else there exist endogneity, the 2OLS with instrumental variable can eliminate the effect of endogneity. Therefore, the estimators should be significantly different. With endogneity, 2OLS estimator will be consistent but OLS will not.

Solution:

the 2OLS with instrumental variable possiblely eliminates the effect of endogneity.

6.3) "The bigger that investment is, the less it seems to affect ROA".

Non-linearity:

The effect of the explanatory variable is different when it changes.

Effect:

It will cause the inconsistent estimators. Actually, if you estimate a linear model with non-linear data you will get nothing of interest. It is totally wrong.

Test:

Reset Test:

By add the polynomial of explanatory variables in the model, we run the regression. And then test there the coefficients of the polynomial of explanatory variables are jointly equal to zero. If so we can say that the non-linearity does not exits. Or else, we must take the non-linearity into account.

Solution:

We can solve this problem by using polynomial of explanatory variable or using log transformation. Specifically speaking, we should determine this by watching the graph.

7. The government is planning to subsidize language courses for immigrants since not knowing the language limits the access to good jobs. In order to understand the potential impact of this policy, you are required to estimate the effect of speaking Spanish on the wages that immigrants (from non- Spanish-speaking countries) earn in Spain. The National Survey on Immigrants contains data collected in 2007on wages and language proficiency (and of other characteristics, such as age, education, gender, etc) for a random sample of immigrants.

a) You first estimate a linear regression where the regressor (Span) is a binary indicator of language proficiency (1 if the person speaks fluent Spanish, 0 otherwise) and the dependent variable (wage) is the net monthly wage (see the results below). What is the estimated effect of speaking fluent Spanish on wages? Comment on the coefficient and interpret the test on significance of the effect.

Compared to these who can not speak Spainish fluently, the wages these who can speak fluently will be 177.757 higher. The p-value for it is 0.0, so we can say it is significantly different from zero. And we are 95% confident that the coefficient value will be in the interval [128.94, 226.58]

b) You then extend the model to include additional regressors: age (in years), woman (a dummy which takes value 1 for female immigrants), and a binary variable (univ) which is 1 if the immigrant has a university degree. Use the results (below) to predict the salary of a 20-year old female immigrant who speaks fluent Spanish but has no university degree.

Predicted wage is :

$$144.73 * 1 + 10.92 * 20 - 445.02 * 1 + 468.85 * 0 + 704.49 = 622.6$$

c) Someone tells you that Span is probably endogenous. Explain why it is reasonable to expect that Span (speaking fluent Spanish) is endogenous to wages.

The intuition is that both wages and the Spanish fluency can be influenced by the intelligence, but we do not include intelligence in the model. As we all know if one is more intelligent, they will have good ability to learn an language. At the same time, a smarter person tends to get higher wage in his or her job. So we can see that the Spanish fluency may be correlated with the error term in this model, which will cause the endogeneity.

d) You believe that the age at which the immigrant arrived in Spain could be a valid instrument for Span (speaking fluent Spanish). Explain why the age of arrival may be a valid instrument.

Intuitively, I think it is good to be a instrumental variable.

Firstly, it is highly correlated with the Spanish fluency, because if the person was less than 14 years old when he or she arrived at Spain, it is much easier for him or her to study the Spanish. If the one was older than 14, the one would miss the golden time of learning a language.

Secondly, it has no direct effect on the people's wages. So it is not correlated with the error term, which means it is exogenous.

e) The following table shows the results of estimating the above regression by 2SLS (the instrument arrive is a dummy which takes value 1 if the person was less than 14 years old when he/she arrived in Spain and a 0 otherwise). Is the instrument weak? Explain.

We can see form the first stage OLS that the F-stat is 51.57 which is much bigger than 10. So we can see that the coefficients of instrument in first stage OLS is significantly different from zeros. So instrument is highly correlated with Spin. So this instrument is very strong.

g) Compare the estimated effect of Span on wages in the OLS and 2SLS regressions. Why are they so different?

As we can see the R2 in the first stage is very low, which means that the instrument variable captures very little variation of the Span. At the same time we can find the variance of the efficient of Span in the second stage is very high, which is 7 times compared to the OLS in question b). So the estimator here is not very efficient. So we can say that the instrumental variable is not good.