MICROECONOMETRICS
CLASS 6

Wiktor Budziński

GENERAL INFORMATION

Course will be organized on the subpage of my website https://www.acep.uw.edu.pl/budzinski/microeconometrics/

- Presentations, datasets, R codes

Meetings will take place on Tuesdays and Fridays

- 19.03.2024, room A308, 13:15 – 18:15
- 22.03.2024, room B109, 9:45 – 14:45
- 26.03.2024, room A308, 13:15 – 18:15
- 05.04.2024, room B109, 9:45 – 14:45
- 09.04.2024, room A308, 13:15 – 18:15

During class you will work on the workbook exercises, that you should finish at home (if you need) and then send to me in a week time

ENDOGENEITY IN NONLINEAR MODELS

We have been discussing methods for dealing with endogeneity for the case of continuous dependent variable in the OLS context

What to do when our dependent variable is not continuous?

- For example: binary, ordered, multinomial, etc.
- Most of the methods that we discussed in the first part of the class actually still apply
  - Control function approach is an analogous solution to 2SLS (instrumental variables methods)
  - Multiple indicator solutions is still valid approach
  - Structural equation modelling (SEM) can also be utilized

CONTROL FUNCTION APPROACH

Let say we want to estimate the model with some probability function $f\left(Y_i \mid \mathbf{X}_i\boldsymbol{\beta} + \alpha X_i^e, \theta\right)$

- For example, Poisson or multinomial logit
- X_i^e is correlated with a stochastic part of the model, which leads to endogeneity

Control function approach is estimated in two steps

- First, we use regression in the form: $X_i^e = Z_i\gamma_1 + \mathbf{X}_i\gamma_1 + u_i$
- Z works as instrumental variables
- It is used to predict residuals: $\hat{u}_i$
- Usually, OLS is used in the first, but another model could be estimated as long as we can predict residuals
- In the second step we estimate model $f\left(Y_i \mid \mathbf{X}_i\boldsymbol{\beta} + \alpha X_i^e + \lambda\hat{u}_i, \theta\right)$
- Residuals control for correlation between stochastic part of the distribution and endogenous variable

Works similarly as Wu-Hausman test for endogeneity in linear models

EXERCISE 1: CONTROL FUNCTION

1. Read *EndoChoice.Rdata* into R
    1. Simulated data, trying to recreate a wine choice
    2. Choice of wine is affected by price and other covariates
    3. Price is likely endogenous leading to suspicious results
2. Run MNL model and calculate WTP for the wine being produced by the small, family-sized vineyard
3. Use control function approach to correct for the endogeneity of price
    1. Use information on the promotion of wine as an instrumental variable

MULTIPLE INDICATORS SOLUTION

MIS utilizes indicator variables to "*impute*" the missing variable causing endogeneity

- Indicators could be collected as attitudinal questions in the survey
- Usually considered to be a linear function of the "*original*" variable

$$z_1 = \alpha_1 q + \eta_1$$

$$z_2 = \alpha_2 q + \eta_2$$

Using only one of such variables would led to biased estimates

- Due to measurement error

MULTIPLE INDICATORS SOLUTION

If we have at least two of them, then we could use one of them as an instrument for the other

- The crucial assumption is that etas (measurement errors) have to be uncorrelated

Basically, we use control function approach, but treat one of the indicator variable as endogenous

In other words, we first run: $z_{i1} = \gamma z_{i2} + u_{i1}$

And then estimate: $f\left(Y_i \mid \mathbf{X}_i \boldsymbol{\beta} + \alpha z_{1i} + \lambda \hat{u}_i, \theta\right)$

- Results may differ depending which indicator is an instrument and which is in the model

SEM

The same objective as with MIS could be obtained by using Structural Equation Models

- Instead of using indicator as an independent variable, we could treat them as dependent variables and estimate a system of equations
- The missing variable could be imputed as a latent factor

Using the previous example, we could estimate:

$$\begin{cases} f\left(Y_i \mid \mathbf{X}_i\boldsymbol{\beta} + \alpha q^*, \theta\right) \\ z_1 = \alpha_1 q^* + \eta_1 \\ z_2 = \alpha_2 q^* + \eta_2 \end{cases}$$

Where q* is a missing variable causing endogeneity of one of the X variables

- It is modelled as a latent variable

SEM

In the case of discrete choice models the SEMs such like that are usually called hybrid choice models

- As a hybrid of choice data and attitudinal data

This class of models is usually used to estimate the effect of some psychological variables on preferences

- Not always have to involve endogeneity

EXERCISE 2: USE OF INDICATOR VARIABLES

1. Use MIS to account for the endogeneity of price variable
   - Use rating variables as indicators for the quality
2. Use a hybrid choice model specification to account for the endogeneity of price variable

MIS VS. SEM

MIS is much easier to implement than the SEM solution

The advantage of SEM is the flexibility

MIS assumes a linear model for the indicators

- Often not realistic when having Likert scale variables as indicators
- It is not clear how large the bias is when this assumption is violated

EXERCISE 3: MIS VS. SEM

1. Compare the WTP estimates from MIS and SEM using the *EndoChoiceOrd.Rdata*
    1. It is basically the same dataset, but indicator variables are measured on the Likert Scale

WORKBOOK 6

Conduct a similar analysis in Workbook6.R

1. Utilize CF, MIS and SEM to correct for endogeneity in the life satisfaction model
    1. Dependent variable is ordinal (measured on Likert scale)

CF AND MIS – STANDARD ERRORS

CF and MIS requires a two-step procedure

- We first estimate an instrumental variable equation
- Then we estimate the main equation with residuals from the first-step as an additional covariate

Although these solutions solve endogeneity problem, the maximum likelihood estimates actually do not follow the theoretical distribution

This is because in the second step we treat residuals as an exogenous covariate

- In reality, residuals are only the estimate of the error term from the first step
- The model does not take into account the variation associated with residuals being the output from the other model
- As such, standard errors in the second step are likely to be underestimated

BOOTSTRAP

If the theoretical distribution of the estimates of some statistic are unknown, one can utilize simulation techniques to try to obtain the empirical distribution

Bootstrap is the most straightforward simulation method, that is widely used in practice

- The idea is to generate N artificial samples based on the actual sample that we observe
  - We do it by drawing with replacement from the original sample
- By conducting estimation on each artificial sample separately we obtain N points from the distribution of estimate/statistic of interest

The alternative would be to estimate a joint model in which both equations are estimated simultaneously

- We can do it on later classes

BOOTSTRAP

Example:

- We have a sample of 5 observations: X_1, X_1, X_3, X_4, X_4
- Artificial samples could be

$$X_1, X_2, X_3, X_4, X_5$$

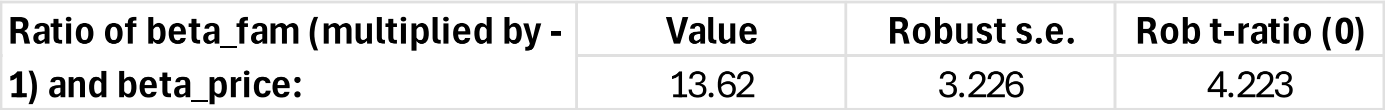
- Or

$$X_1, X_2, X_3, X_3, X_3$$

EXERCISE 4: BOOTSTRAP

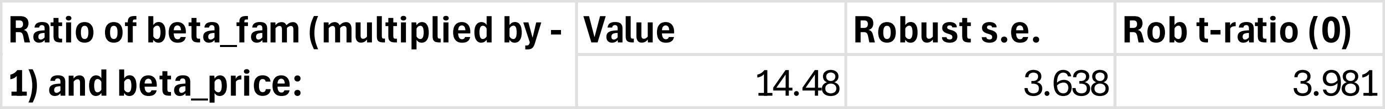
1. Compare the distribution of WTP obtained in exercise 2 from the control function and MIS, with the one obtained from bootstrap

CF:



Ratio of beta_fam (multiplied by - 1) and beta_price:	Value	Robust s.e.	Rob t-ratio (0)
	13.62	3.226	4.223

MIS:



Ratio of beta_fam (multiplied by - 1) and beta_price:	Value	Robust s.e.	Rob t-ratio (0)
	14.48	3.638	3.981