Applied Econometrics

1st Lecture

Roman Horvath

roman.horvath@fsv.cuni.cz

http://ies.fsv.cuni.cz/en/staff/horvath

Why Is Econometrics Important?

"I have no data yet. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

Sherlock Holmes in *A Scandal in Bohemia* by Sir Arthur Conan Doyle (1859-1930)

Keynes in 1937

This brand of statistical alchemy [econometrics] is[not] ripe to become a branch of science.

Organization

- Summer semester
- Lecture: Monday, 9:30 AM (room 314)
- Seminars: Thursday, 11 AM and 6:30 PM (room 016)

- Lecturers:

  Jozef Baruník (barunik@fsv.cuni.cz)

  Jaromir Baxa (jaromir.baxa@fsv.cuni.cz)

  Roman Horvath (roman.horvath@fsv.cuni.cz)

  Jiri Kukacka (jiri.kukacka@fsv.cuni.cz)

- TAs:

  Lukáš Janásek (lukas.janasek@gmail.com)

  Lenka Nechvatalova (lenka.nechvatalova@fsv.cuni.cz)
- MA course, Core/Elective course, 6 credits

Syllabus

19.2. - OLS and Instrumental variables

26.2. - Introduction to time series models

4.3. - ARIMA

11.3. - GARCH

18.3. - GARCH - Extensions

25.3. – EY Guest lecture (No seminars, Dean's holiday on Thursday)

1.4. – Easter break

8.4. – Non-linear models

15.4. – Limited dependent variables in finance

22.4. - Vector autogregressions

29.4. - Cointegration

6.5. – Filters

13.5. - Networks

Assessment

- - Exam (80%)
  - Term paper (20%)

Term Paper

- Term paper – in general, choose any topic after consulting the lecturer or may accept a topic and data that the lecturer will propose
- You work in the team of two
- You write a short research paper
- Some examples of topics of individual assignment from the previous years:
    - The effect of spot position of exchange rate within the fluctuation band on its volatility: Evidence from Hungary
    - PX-50 stock market returns and volatility modeling
    - Daily effects of volatility of stock markets in central Europe

The course is about…

- Practical use of econometric methods
- The lectures are supplemented by computer classes, where students gain hands-on experience in applied econometric analysis
- Focus on time series techniques
- Applications to topics such as forecasting asset volatility, modeling inflation or exchange rate volatility

Readings

- Mandatory are lecture notes (presentations that you will see during lectures)
- Several applied econometrics textbooks are recommended
- BROOKS, C. INTRODUCTORY ECONOMETRICS FOR FINANCE, CAMBRIDGE UNIVERSITY PRESS.
    - Enders, W.: “Applied Econometric Time Series“, 2nd edition, 2003
      Harris, R. and R. Sollis: "Applied Time Series Modelling and Forecasting", 2003
      Stewart, K. G.: "Introduction to Applied Econometrics", 2005
      Verbeek, M.: “A Guide to Modern Econometrics”, 2nd edition, 2004
      Kratzig, M. and H. Lutkepohl ,“Applied Time Series Econometrics”, 2004
      Kocenda, E. and A. Cerny, "Elements of Time Series Econometrics", 2007, Karolinum
- Other suggested readings include journal articles (see the course website for the full list)

Software for Econometrician

- *E-views* – essentially all basic stuff, easy to use
- *PcGIVE* – essentially all basic stuff, easy to use
- *STATA* - essentially all basic stuff, easy to use, more estimation commands, many programming codes can be downloaded from internet (http://ideas.repec.org/s/boc/bocode.html)

Software for Econometrician

- *GAUSS* – difficult to use, 'must program'
- *MATLAB* – like GAUSS basically, often used by central banks for macro modeling, (excellent alternative *R*)
- *SAS* – typically used by statisticians
- *Mathematica* – typically used by mathematicians
- *Xplore*, *Splus*, *Statistica* and others

Free software

- Many partial canned packages
- In my opinion, the most comprehensive is *JMulTi* and *Gretl*,
- JMulTi downloadable from www.jmulti.com
- Gretl downloadable from http://gretl.sourceforge.net/
- Online help + help in pdf documents
- **R – will be used during seminars**

Plan of the first half of the lecture

1. OLS
    - What is it?
    - How to estimate it?
2. How to tackle some basic problems such as
    1. Autocorrelation
    2. Heteroscedasticity
    3. Multicollinearity
    4. Omitted variable bias
    5. Correct functional form
3. What makes a good model?

OLS – quick overview

- OLS – ordinary least squares
- $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + u$
- The most popular technique, why?
- If classical assumptions hold (see the next slide), then
  - OLS is unbiased – expected value of estimated parameters is equal to their 'true' counterparts
  - Minimum variance among the class of linear estimators (BLUE)

Classical assumptions

(sometimes too restrictive)

1. Expected value of error term is zero
2. No autocorrelation between the individual realizations of error term (u_i)
3. Constant variance for all u_i
4. Independence, no correlation between u_i's and x_i, where x is explanatory variable
5. Correct specification = correct functional form

How to estimate the parameters by OLS

- Minimize the residual sum of squares
- The resulting estimators are termed least square estimators
- Theoretical characteristics of the distributions of estimated parameters depend on the assumptions made about the distribution of the error term (u_i)
- If the distributions of (u_i) assumed normal, then the distribution of LS estimators is also normal
- However, for the purposes of estimation, the assumption of normality in OLS is not required

OLS continued

- R-squared: coefficient of determination, a measure of the fit, value between zero and one
- The value of R-squared is positively associated with the number of explanatory variables, you may use adjusted R-squared
    - R-squared adjusted for the number of explanatory variables – imposes punishment for increasing the number of expl. variables
- If the error term is normally distributed, then OLS is BUE – it has minimum variance of all unbiased estimators, linear or not

OLS continued

- The estimated parameters have their standard errors
    - t-statistic = parameter/standard error
    - p-value ('exact significance'), if p-value is smaller than 0.05, then we say the variable is significant at 5% significance level (t-stat greater than 1.96 in other words)
- You may also test the joint significance of more than one explanatory variables by F-test

Multicollinearity

- Perfect multicollinearity is where there is exact linear relationship between explanatory variables included in the regression
- Perfect multicollinearity never happens in practice, unless you fall in a dummy variable trap
    - For example, you include the dummy both for large and small firms and the constant in your regression
- But you may encounter some degree of multicollinearity

What to do about multicollinearity

- Again, rules of thumb
  - no best test or remedy
  - use in combination
- Sample problem ⇒ get more data!
  - multicollinearity = problem of lack of data
    - pooling a possibility
- Time series: 1st difference the data
  - 1st differences usually less correlated
- Principal components
  - Form linear combination of correlated variables such that it maximizes its variability
      - Drawback: may lack economic reasoning

Do not just drop variables! – Omitted variables problem

Omitted variables

- If you omit important explanatory variable, the estimates are biased, and also inconsistent (i.e., the bias does not disappear, if you increase the sample size)
- Standard errors biased – no inference valid
- You are then trying to force the regression line where it does not want to go!
- If you include irrelevant variable, the estimates are unbiased, but standard errors greater than necessary

Worrying signs of omitted variables – correct functional form

- Low R-squared
- Insignificant variables
- ‘Wrong’ signs
- Bad Durbin-Watson values

Ramsey RESET test

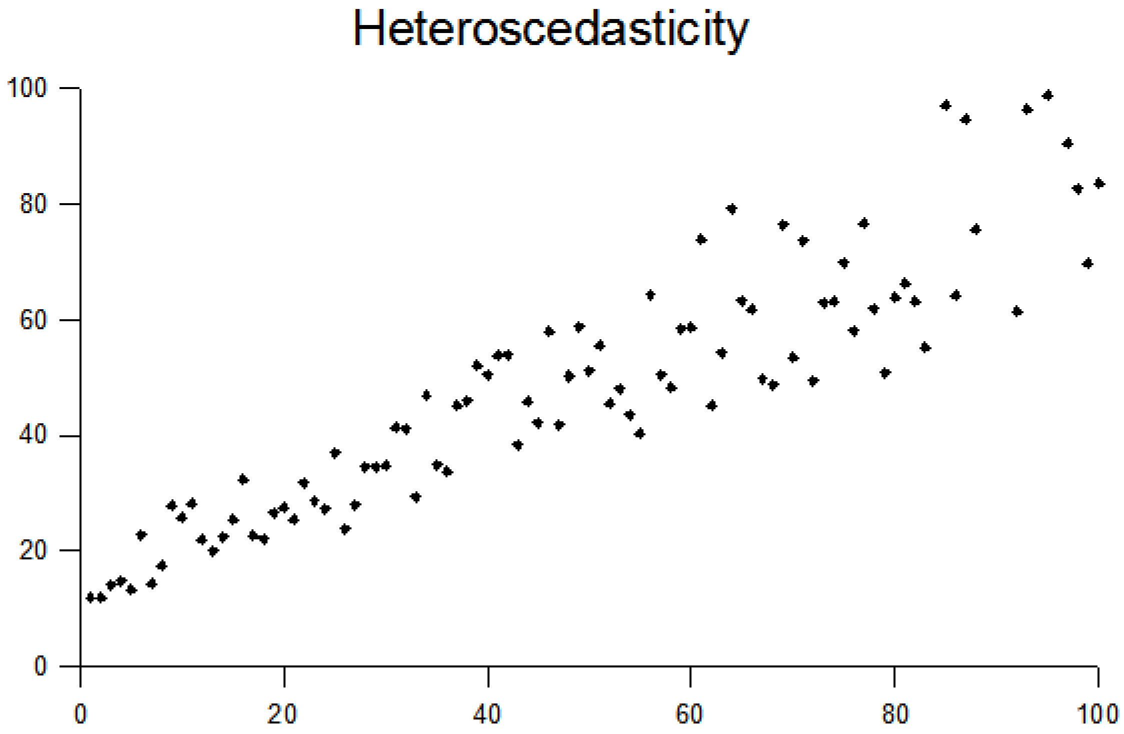
- General specification test

1. Estimate your model
2. Get the fitted value of dependent variable
3. Run auxiliary regression additionally including the square of the fitted value and the cube of the fitted value
4. Use subset F-test to examine the joint significance of these two newly included explanatory variables

Heteroscedasticity

- Violation of one of the classical assumptions
- Violates the statistical inference
- Homoscedasticity – constant variance of the error term
- More likely to occur in the cross-section
  - (say the higher income, the more discretionary spending decisions, but if you estimate the time series of consumption and income, no reason for heteroscedasticity to arise)
- Heteroscedasticity may arise in misspecified model – bad functional form. Say the correct relationship is quadratic, but you estimate linear model
- Heteroscedasticity may arise, if some relevant variables are omitted

Example of heteroskedasticity (consumption and income)



Consequences of heteroscedasticity

- Still unbiased coefficients, but the standard errors not correctly estimated – so significance tests invalid
- Predictions inefficient
- Coefficient of determination not valid
- Important to test for any heteroscedasticity and remedy it if present

Detecting Heteroscedasticity (I)

- Graphical method – good starting point
- Plot the squared value of estimated error term with dependent or explanatory variables
- Any systematic pattern suggest heteroscedasticity

Detecting Heteroscedasticity (II)

- Number of formal tests:
- Koenker test – linear form of heteroscedasticity assumption
- Regress the square of the estimated error term on the explanatory variables, if these variables are jointly significant, then you reject the null of homoscedastic errors

Detecting Heteroscedasticity (III)

- White test – most popular
- Same like Koenker test, but you include also the squares of explanatory variables and their cross products (do not include cross-products, if you have many explanatory variables – multicollinearity is likely)

Corrective Measures for Heteroscedasticity

- Transform the model:
- Take logs or shares
- Or 'White-wash' them
  - Get White robust standard errors and covariance, they are robust to an unknown form of heteroscedasticity
  - Works quite well, especially if you have large sample size
- You may use Weighted LS, if you know the form of heteroscedasticity, unlikely in practice

Autocorrelation

- More likely in time-series
- Autocorrelation means that there is some kind of relationship between the *t*-th and *t*-*i*-th error term
- Say if error term is large in time *t*, it is very likely that it will remain high next period, but classical assumptions ‘want’ no relationship!

Possible Causes of Autocorrelation

- Inertia and cyclicality
- Functional misspecification
- Omitted variables

Consequences of autocorrelation

- Identical as for heteroscedasticity, standard errors affected – invalid inference

Detection of autocorrelation (I)

- Graphical method – good starting point
- Plot the estimated error term over time
- Any systematic pattern suggests autocorrelation

Detection of autocorrelation (II)

- Formal tests:
- Durbin Watson test for AR(1)
- D-W test statistic roughly equal to 2*(1-P)
- P is the correlation between the error term and its 1st lag
- D-W statistic symmetrically distributed around 2, if it is far from 2, there is autocorrelation. There is also a grey zone, where you cannot conclude about autocorrelation based on D-W test, negative autocorrelation if test statistic 'enough larger' than 2, opposite for positive autocorrelation
- Critical values found in most textbooks, the values depend on sample size and number of explanatory variables

D-W test requires:

- Model with intercept
- Explanatory variables exogenous
- Disturbances homoscedastic
- Possibility of AR(1) only
- Explanatory variables should not contain lagged dependent variables
  - (then you have to do Durbin's h-test)

Autocorrelation of higher order

- If you suspect AR(k) process in your residuals, use Breusch – Godfrey test
- Run auxiliary regression
  - Residual on the explanatory variables + the lagged values of residuals, test the joint significance of lagged residuals

Remedial measures

- First differencing
- Quasi differencing
- ‘Newey-West wash’, analogy of White-washing, it makes your standard errors and covariance robust to unknown form of autocorrelation

Model specification: what makes a good model?

- Parsimony
- Goodness of fit
- Theoretical consistency – has to ‘make sense’
- Predictive power

INSTRUMENTAL VARIABLES

Readings (not mandatory)

- Theory:
  - Angrist and Krueger (2001): Instrumental Variables and the Search for Identification, *Journal of Economic Perspectives*, pp.69-85.
  - Hausman (2001): Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left, *Journal of Economic Perspectives*, pp. 57-67.
- Application:
  - Horvath (2005): Exchange Rate Variability, Pressures and Optimum Currency Areas: Evidence from Developed Economies, *Applied Economics Letters*, pp. 919-922.

Instrumental Variables & 2SLS

Consider the following model:

Eq.1: $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + u$

Eq. 2: $x_1 = p_0 + p_1z + p_2x_2 + \ldots + p_kx_k + v$

Eq.1 – sometimes called structural (or second-stage) equation

Eq. 2 – sometimes called first-stage equation

Why Use Instrumental Variables?

- Instrumental Variables (IV) estimation is used when your model has endogenous *x*'s
- That is, whenever Cov(x,u) ≠ 0
- In other words, not only that *x* influences *y*, but also *y* influences *x* (have to „get rid“ of the latter influence)
- Thus, IV can be used to address the problem of omitted variable bias or simultaneity
- Additionally, IV can be used to solve the classic errors-in-variables problem

What Is an Instrumental Variable?

- In order for a variable, z, to serve as a valid instrument for x, the following must be true

1. The instrument must be exogenous

- That is, $\text{Cov}(z,u) = 0$

2. The instrument must be correlated with the endogenous variable x

- That is, $\text{Cov}(z,x) \neq 0$

- Instrument is valid, if exogenous and „well correlated“, i.e. informative

More on Valid Instruments

- We have to use common sense and economic theory to decide if it makes sense to assume Cov(z,u) = 0 (i.e. exogenous instrument)
  - Difficult to test, as we do not have residuals (u), only the estimated u)
- We can test if Cov(z,x) ≠ 0
- Just testing H_0: $p_1 = 0$ in $x = p_0 + p_1 z + v$

IV Estimation

- If instrument, z, is valid and informative

- Then, take the fitted value of x_1 from the first stage equation and include it instead of x_1 into the structural (second stage) equation

IV versus OLS estimation

- IV is consistent, while OLS is inconsistent, when Cov(x,u) $\neq$ 0

- The stronger the correlation between z and x, the smaller the IV standard errors

The Effect of Poor Instruments

- What if our assumption that Cov(*z*,*u*) = 0 is false?
- The IV estimator will be inconsistent, too
- We can compare asymptotic bias in OLS and IV
- In theory, prefer IV if Corr(*z*,*u*)/Corr(*z*,*x*) < Corr(*x*,*u*)

$$\text{IV}: \text{plim}\hat{\beta}_1 = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \bullet \frac{\sigma_u}{\sigma_x}$$

$$\text{OLS}: \text{plim}\ \tilde{\beta}_1 = \beta_1 + Corr(x,u) \bullet \frac{\sigma_u}{\sigma_x}$$

IV Estimation in the Multiple Regression Case

- IV estimation can be extended to the multiple regression case
- Call the model we are interested in estimating the structural model
- Our problem is that one or more variables are endogenous and in addition to those variables there could be also one or more explanatory variables which are exogenous
- Of course, for each endogenous *x*, you need at least one *z*.

Multiple Regression IV (cont)

- Consider the structural model as
- $y_1 = b_0 + b_1x_1 + b_2x_2 + u_1$,

  where x_1 is endogenous and x_2 is exogenous
- Let z_1 be the instrument, so $\text{Cov}(z_1,u_1) = 0$ and
- $x_1 = p_0 + p_1x_2 + p_2z_1 + v_2$
- This reduced form equation regresses the endogenous variable on all exogenous ones

Two Stage Least Squares (2SLS)

- It's possible to have multiple instruments
- Consider our original structural model, and let $x_1 = p_0 + p_1 x_2 + p_2 z_1 + p_3 z_2 + v_2$
- Here we're assuming that both z_1 and z_2 are valid instruments – they do not appear in the structural model and are uncorrelated with the structural error term, u_1
- The best instrument is a linear combination of all of the exogenous variables

More on 2SLS

- Method extends to multiple endogenous variables – need to be sure that we have at least as many excluded exogenous variables (instruments) as there are endogenous variables in the structural equation

Weak instruments

- In case of one endogenous variable
  - advisable to have F-value from first stage regression greater than 10 to assure that instrument is informative
- In case of more explanatory variables
  - quite difficult, you may use so-called Shea partial R-sqr. (which lacks distribution theory = you cannot test if instrument is weak or not with this test) or simply throw away insignificant instruments in the first stage regression

Testing for Endogeneity

- If we do not have endogeneity, both OLS and IV are consistent (but IV is biased)
- Idea of Hausman test (for endogeneity) is to see if the estimates from OLS and IV are different

Testing for Endogeneity (cont)

- While it's a good idea to see if IV and OLS have different implications, it's easier to use a regression test for endogeneity
- If x_1 is endogenous, then v_2 (from the reduced form equation) and u_1 from the structural model will be correlated
- The test is based on this observation

How to proceed with endogeneity test

- Save the residuals from the first stage regression
- Include the residual in the structural equation (which of course has x_1 in it)
- If the coefficient on the residual is statistically different from zero, reject the null hypothesis of exogeneity
- If multiple endogenous variables, jointly test the residuals from each first stage

Testing Overidentifying Restrictions

- If there is just one instrument for our endogenous variable, we can’t test whether the instrument is uncorrelated with the error
- We say the model is just identified
- If we have multiple instruments, it is possible to test the overidentifying restrictions – to indicate if some of the instruments are correlated with the error

The OverID Test

- Estimate the structural model using IV and obtain the residual
- Regress the residual on all the exogenous variables and obtain the R^2 to form nR^2, where n is the no. of observations
- Under the null that all instruments are uncorrelated with the error, LM ~ c_q^2 where q is the number of extra instruments

2SLS - Procedure

1. Estimate the structural equation by OLS
2. Any explanatory variable susceptible to be endogenous?
3. If yes, run Hausman test and test which explanatory variables are endogenous
4. Regress each endogenous variable on your instruments and see which instruments are significant (to get informal idea about their relevance)
5. Estimate the equation by 2SLS (use the fitted values of endogenous variables)
6. Check the model specification
7. If the number of instruments ‘out of model’ is larger than the number of endogenous explanatory variables, use test for overidentifying restrictions to examine if the estimated residual is correlated with your instruments