

MICRO DATA ANALYSIS

CLASS 1

Marek Giergiczny
Wiktor Budziński

GENERAL INFORMATION

Materials for the class will be send by email

- Presentations, datasets, R codes
- Links to meetings
- Recordings of the classes

Meetings will take place on Zoom

- Learning by doing workshop
 - ~ 60 minutes of presentation with basic theory and necessary R function and packages
 - ~ 30 minutes of individual work on different dataset
 - 60 minutes break for lunch around noon
-

CLASS DATES

- 2024-06-04 11:00 : 15:00
- 2024-06-06 11:00 : 15:00
- 2024-06-09 11:00 : 15:00
- 2024-06-11 11:00 : 15:00
- 2024-06-13 11:00 : 15:00

COURSE ORGANIZATION

Class 1: Brief introduction to R, OLS, building econometric models, OLS extensions

Class 2: Endogeneity, censoring

Class 3: Binary and ordered choice

Class 4: Count data

Class 5: Multinomial choice

R LANGUAGE

R is a programming language primarily used for statistical computation

- Open source from 1995
- Can be download at <https://www.r-project.org>

Although a lot of functionalities are installed with R, it relies on user-written packages for specific techniques

- We will install them as we go along the course

User interface is not very convenient in basic R

- We will use RStudio IDE: <https://rstudio.com/products/rstudio/download/>
-

TODAY'S CLASS

Introduction to R

- RStudio
- Reading data
- Data manipulation
- Some data-based calculations

Linear regression

- Model specification and general assumptions
 - Model's interpretation and testing of the assumptions
 - Building of the econometric model
 - Extensions: modelling heteroscedasticity
-

CASE STUDY — DETERMINANTS OF WINE PRICES

Costanigro, M., Mittelhammer, R. C., and McCluskey, J. J. (2009). ***Estimating class-specific parametric models under class uncertainty: local polynomial regression clustering in an hedonic analysis of wine markets***. Journal of Applied Econometrics, 24(7), 1117-1135.

- Data regarding prices and others characteristics of wine produced in California and Washington
- Hedonic analysis – decomposition of the value of the good (usually approximated by price) on the value of each characteristic
 - How much is each characteristic contributing to the price?
 - Can be used in marketing, but also often used in environmental economics

EXERCISE 1: INTRODUCTION TO R

1. Read data from *wine.xlsx* into R
2. Perform basic transformations and calculate some summaries of the data

LINEAR REGRESSION

Model form is as follows: $y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$

- We model linear relationship between dependent variable y and independent variables \mathbf{X}
- Model coefficients, $\boldsymbol{\beta}$, are unknown but we can estimate them from the data
- Linear regression models how mean of y depends on \mathbf{X} , namely $\mathbf{E}(y_i) = \mathbf{X}_i\boldsymbol{\beta}$
- Model coefficients can be estimated using Ordinary Least Squares method:

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \left\{ \sum_i (y_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right\}$$

- Analytical solution can be easily found: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'y$

LINEAR REGRESSION

If certain assumptions are met then OLS estimator is:

- Unbiased (on average parameters are equal to true values)
- Consistent (precision increases with the sample size)
- Efficient (characterized by the lowest variation of the estimates)

LINEAR REGRESSION

Most important assumptions:

- Linear functional form
- Spherical error terms
 - No correlation between error terms
 - Homoskedasticity – constant variance of the error term
- Exogenous covariates
 - Dependent variables are not correlated with an error term
- Normally distributed error terms

LINEAR REGRESSION

Assumption about functional form can be tested with RESET test:

- Conducted in two steps:
 - Estimate model parameters, $\hat{\beta}$, and calculate fitted values $\mu_i = \mathbf{X}_i \hat{\beta}$
 - Estimate second regression of the form: $y_i = \mathbf{X}_i \hat{\beta} + \alpha_1 \mu_i^2 + \alpha_2 \mu_i^3 + \varepsilon_i$
- Significance of alpha's inform us whether form is correct

Assumption about homoskedasticity can be tested with Breush-Pagan test:

- Second regression is estimated in which squared residuals are explained by covariates: $(\hat{\varepsilon}_i)^2 = \mathbf{X}_i \gamma + u_i$
- Significance of gamma's inform us whether the form is correct

EXERCISE 2: BASIC REGRESSION

1. Estimate basic regression model
2. Test basic assumptions of the OLS
 1. Use formal tests as well as graphical analysis
3. Use nonlinear transformations to improve model's functional form
4. Add interactions between covariates

INFLUENTIAL OBSERVATIONS

Some observation could affect model estimates more than others

- If these observations are outliers and not correct data points this could lead to model misspecification

Some useful measures to detect such observations:

- Leverage (h_i) is a weight that a given response (y_i) has on it's own fitted value (μ_i)
- Cook's distance is a combination of residuals and leverage: $D_i = \frac{r_i^2}{p} \frac{h_i}{1-h_i}$
- DFFITS measures how much fitted value changes without a given observation: $DFITS_i = \frac{\mu_i - \mu_{i(i)}}{s_{(i)}}$
- DFBETA measures how much coefficient estimates would change without a given observation: $DFBETA_{ij} = \frac{\beta_j - \beta_{j(i)}}{se(\beta_{j(i)})}$

EXERCISE 3: BASIC REGRESSION

1. Check whether there any influential observations in the model
2. Is there any dependence between price and influence measures?

HETEROSKEDASTICITY

OLS is still consistent and unbiased estimator even if error terms are heteroskedastic

- It is no longer efficient, and error terms are calculated with a wrong formula

Easy fix is to use robust covariance matrices

- The most popular one is White's matrix: $(\mathbf{X}'\mathbf{X})^{-1} \times \sum_{i=1}^n e_i^2 X_i X_i' \times (\mathbf{X}'\mathbf{X})^{-1}$
- There are other alternatives in R

HETEROSKEDASTICITY

Heteroskedasticity can also be directly accounted for by parametrizing it, and estimating how variance depends on other covariates

Instead of modelling $y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$ with $Std(\varepsilon_i) = \sigma$, we can assume some nonlinear relationship, for example: $Std(\varepsilon_i) = \sigma_i(\mathbf{X}_i) = \exp(\mathbf{X}_i\boldsymbol{\gamma})$

- Such model can be estimated with maximum likelihood method
- Could be useful if we care about predictions

EXERCISE 4: HETEROSKEDASTICITY

1. Estimate model with a robust covariance matrix
 1. Compare results with a standard estimates
2. Estimate model in which heteroskedasticity is directly controlled for.
 1. Conduct Breush-Pagan to test whether the model accounts for the whole heteroskedasticity