MICROECONOMETRICS
CLASS 1

Wiktor Budziński
Marek Giergiczny

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 in room J at the Faculty

- 09:30 am - 11:30 am
- 12:00 pm - 01.30 pm
- 01.45 pm - 14:45 pm

Learning by doing workshop

- Presentation with basic theory and necessary R function and packages
- Individual work on different dataset
  - Codes with comments and some summary of the results should be send to a teacher till Wednesday next week after given class (midnight)

CLASS DATES

Teacher: Wiktor Budziński

- 2021-04-14
  2021-04-21
  2021-04-28
  2021-05-05
  2021-05-12 (Juwenalia)

Teacher: Marek Giergiczny

- 2021-05-19
  2021-05-26
  2021-06-02
  2021-06-16
  2021-06-23 (End of the semester)

COURSE ORGANIZATION

- Class 1: Brief introduction to R, OLS, building econometric models, OLS extensions
- Class 2: Generalized Linear Models, Quantile regression, continous variables with 0 outcomes
- Class 3: Models for count, ordinal and binary data
- Class 4: Discrete choice models I
- Class 5: Discrete choice models II
- Class 6: Simulation methods
- Class 7: Endogeneity
- Class 8: Sample selection, treatment effects
- Class 9: Panel data methods
- Class 10: Programming econometric models in R

R LANGUAGE

R is a programming language primarily used for statistical computation

- Open source from 1995
- Currently 9th most popular programming language
- 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 that "nice" in basic R

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

MODELS FOR CONTINUOUS VARIABLES

Introduction to R and linear regression

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

Next class:

- Generalized linear models
- Quantile regression
- What to do when we have continuous variable with zeros
- Censoring of the continuous variable

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 $\mathbf{E}\left(y_i \mid \mathbf{X}_i\right) = \mathbf{X}_i\boldsymbol{\beta}$
- Model coefficients can be estimated using Ordinary Least Squares method:

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

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

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, $\hat{\boldsymbol{\beta}}$, and calculate fitted values $\mu_i = \mathbf{X}_i \hat{\boldsymbol{\beta}}$
  - Estimate second regression of the form: $y_i = \mathbf{X}_i \boldsymbol{\beta} + \alpha_1 \mu_i^2 + \alpha_2 \mu_i^3 + \varepsilon_i$
- 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: $\left(\widehat{\varepsilon}_i\right)^2 = \mathbf{X}_i \boldsymbol{\gamma} + u_i$
- Significance of gamma's inform us whether the form is correct

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: BASIC REGRESSION

1. Read data from *wine.xlsx* into R
2. Plot the price variable and calculate basic statistics for it
3. Estimate basic regression model
4. Test basic assumptions of the OLS
    1. Use formal tests as well as graphical analysis

EXERCISE 1: BASIC REGRESSION

1. Read data from *wine.xlsx* into R
2. Plot the price variable and calculate basic statistics for it
3. Estimate basic regression model
4. Test basic assumptions of the OLS
    1. Use formal tests as well as graphical analysis

Let's go to R

EXERCISE 2: BASIC REGRESSION

1. Use nonlinear transformations to improve model's functional form
2. Add interactions between covariates

EXERCISE 2: BASIC REGRESSION

1. Use nonlinear transformations to improve model’s functional form
2. Add interactions between covariates

Let’s go to R

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: $DFFITS_i = \frac{\mu_i - \mu_{i(i)}}{s_{(i)}}$
- DFBETA measures how much coefficient estimates would change without a given observation: $DFBETA_{ij} = \frac{\beta_j - \beta_{j(i)}}{se\left(\beta_{j(i)}\right)}$

EXERCISE 3: BASIC REGRESSION

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

EXERCISE 3: BASIC REGRESSION

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

Let's go to

WORKBOOK 1

Now try to conduct a similar analysis for the exercises in Workbook1.R

Let's go to R

BOX-COX TRANSFORMATION

Instead of trying to fit various different nonlinear specifications it is possible to just estimate it using some more flexible function

- One often used transformation is Box-Cox transformation in the form of:

$$x^{(\lambda)} = \frac{x^{\lambda} - 1}{\lambda}$$

- Could be used for both, dependent and independent variables
- For $\lambda \to 0$ it becomes a logarithmic transformation
- For example, we fit the model: $\frac{y_i^{\lambda} - 1}{\lambda} = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i$
- Which is equivalent to: $y_i = \left(\lambda \mathbf{X}_i \boldsymbol{\beta} + \lambda \varepsilon_i + 1\right)^{\frac{1}{\lambda}}$
- Has to be done with Maximum Likelihood estimator

BOX-COX TRANSFORMATION

The same transformation could be applied to the independent variables

For example we could fit the following model: $y_i = \mathbf{X}_i\boldsymbol{\beta} + Z_i^{(\lambda)} + \varepsilon_i$

- There is one additional parameter to estimate

Usually also done with Maximum Likelihood method

MAXIMUM LIKELIHOOD ESTIMATOR

We assume some probabilistic model, with density given by $f(y_i \mid \mathbf{X}_i, \boldsymbol{\beta})$

- For discrete variables it should be a probability function

If we have *N* independent observations in the sample, that under this model the probability of drawing the given sample is:

$$L = \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} f(y_i \mid \mathbf{X}_i, \boldsymbol{\beta})$$

MLE looks for betas which maximize this likelihood function

- Which model is most likely to produce such data

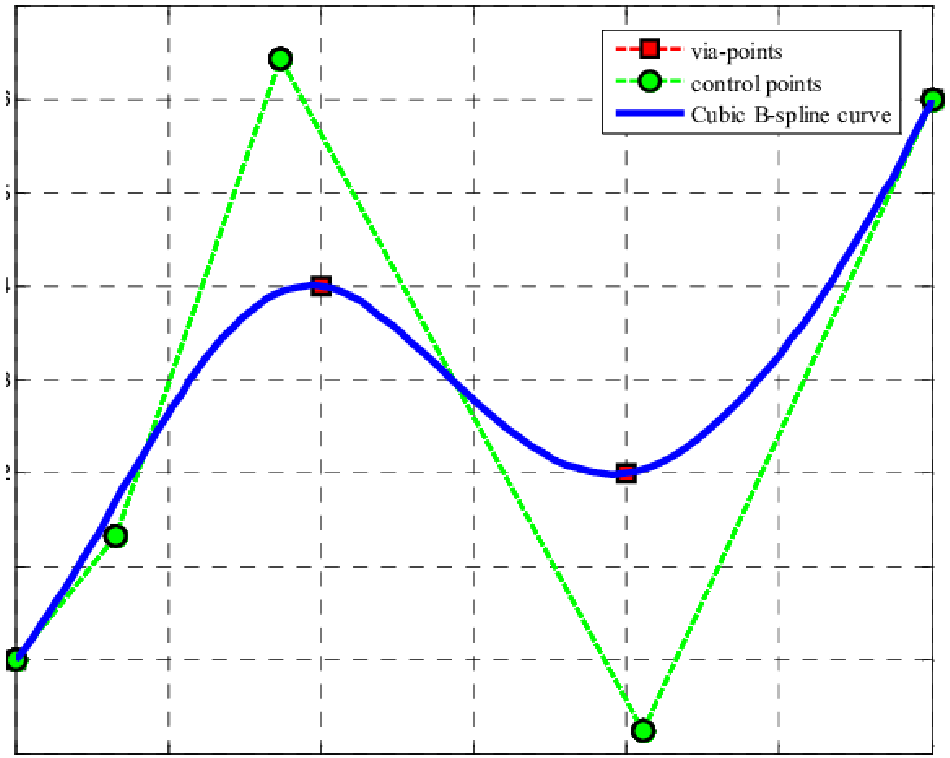
This optimization problem is equivalent to maximizing it's logarithm: $LL = \log(L) = \sum_{i=1}^{N} \log\left(f(y_i \mid \mathbf{X}_i, \boldsymbol{\beta})\right)$

- This is usually easier numerically

SPLINES

Splines are piecewise polynomials functions

- For a given number of knots splines fit a polynomial function between every two subsequent knots
- Different splines differ in terms of their smoothness and variability
- Cubic splines mean that cubic polynomial is fit between every two subsequent knots
  - Natural cubic spline adds additional constraints, namely that the function is linear beyond the boundary knots
- Optimization procedure could be used to choose the optimal number of knots and polynomials order



EXERCISE 4: NONLINEAR TRANSFORMATIONS

1. Use the Box-Cox transformation to find which transformation of price would fit data best
   1. Check whether it helps with a functional form of the model, and compare the results with a logarithmic regression
2. Use different splines to fit more complex relationships between price and no. of produced cases, as well as price and the taste score.

EXERCISE 4: NONLINEAR TRANSFORMATIONS

1. Use the Box-Cox transformation to find which transformation of price would fit data best
    1. Check whether it helps with a functional form of the model, and compare the results with a logarithmic regression
2. Use different splines to fit more complex relationships between price and no. of produced cases, as well as price and the taste score.

Let's go to

EXERCISE 4B: SPLINES IN SPATIAL ANALYSIS

Sometimes we want to model how some variable vary in space

- This relationship is likely to be nonlinear with some clusters (hot-spots and cold spots)
- Splines can be useful to identify and visualize such relationships

We used it in this paper to investigate spatial clusters of willingness to pay

- Czajkowski, M., Budziński, W., Campbell, D., Giergiczny, M., & Hanley, N. (2017). Spatial heterogeneity of willingness to pay for forest management. *Environmental and Resource Economics*, *68*(3), 705-727.

1. Look at *WTP_example.r* for visualization of this procedure

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: $\left(\mathbf{X}'\mathbf{X}\right)^{-1} \times \sum_{i=1}^{n} e_i^2 X_i X_i' \times \left(\mathbf{X}'\mathbf{X}\right)^{-1}$
- 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
- Important if the dependent variable is in logarithms

EXERCISE 5: 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

EXERCISE 5: 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

Let's go to R

WORKBOOK 1

Now try to conduct a similar analysis for the rest of the exercises in Workbook1.R