

ADVANCED ECONOMETRICS I

Theory (1/3)

Instructor: Joaquim J. S. Ramalho

E.mail: jjсро@iscte-iul.pt

Personal Website: <http://home.iscte-iul.pt/~jjсро>

Office: D5.10

Course Website: <https://jjsrामalho.wixsite.com/advecoi>

Fénix: <https://fenix.iscte-iul.pt/disciplinas/03089>

Course Description

- This course provides an introduction to the modern **econometric techniques** used in the analysis of **cross-sectional** and **panel data** in the area of **microeconometrics**:
 - The interaction between theory and empirical econometric analysis is emphasized
 - Students will be trained in formulating and testing economic models using real data
- Pre-requisites (recommended):
 - Introductory Econometrics
- Grading:
 - Two problem sets (50%) + Final (open book) exam (50%)
 - Weighted mean of at least 9,5/20
 - Minimum grade at the exam of 7,5/20
 - No re-sit examinations

Contents - Theory

i. Introduction

1. Linear Regression Analysis
2. Nonlinear Regression Analysis
3. Discrete Choice Models
4. Models for Continuous Limited Dependent Variable Models

Textbooks

- Recommended:

- Cameron, A. and P.K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press

- Others:

- Baltagi, B. (2013), *Econometric Analysis of Panel Data*, John Wiley and Sons (5th Edition)
- Davidson, R. and J.G. MacKinnon (2003), *Econometric Theory and Methods*, Oxford University Press
- Greene, W. (2011), *Econometric Analysis*, Pearson (7th Edition)
- Verbeek, M. (2017), *A Guide to Modern Econometrics*, Wiley (5th Edition)
- Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, MIT Press (2nd Edition)
- Wooldridge, J.M. (2015), *Introductory Econometrics: A Modern Approach*, South Western (6th Edition).

Contents - Illustrations

1. Determinants of Firm Debt
2. Estimating the Returns to Schooling
3. Explaining Individual Wages
4. Explaining Capital Structure
5. Modelling the Choice Between Two Brands
6. Health Care Expenses and Consultations
7. Explaining Firm's Credit Ratings
8. Travel Mode Choice
9. Health Care Expenses and Consultations (revisited)
10. Determinants of Firm Debt (revisited)

Software

- Recommended:

- Stata: <http://www.stata.com>
- R: <https://cran.r-project.org>

- Others:

- Gauss: <http://www.aptech.com/products/gauss-mathematical-and-statistical-system>
- Matlab: <https://www.mathworks.com/products/matlab>

i. Introduction

i.1. Econometric Methodology

i.2. The Structure of Economic and Financial Data

i.3. Dependent Variables and Econometric Models

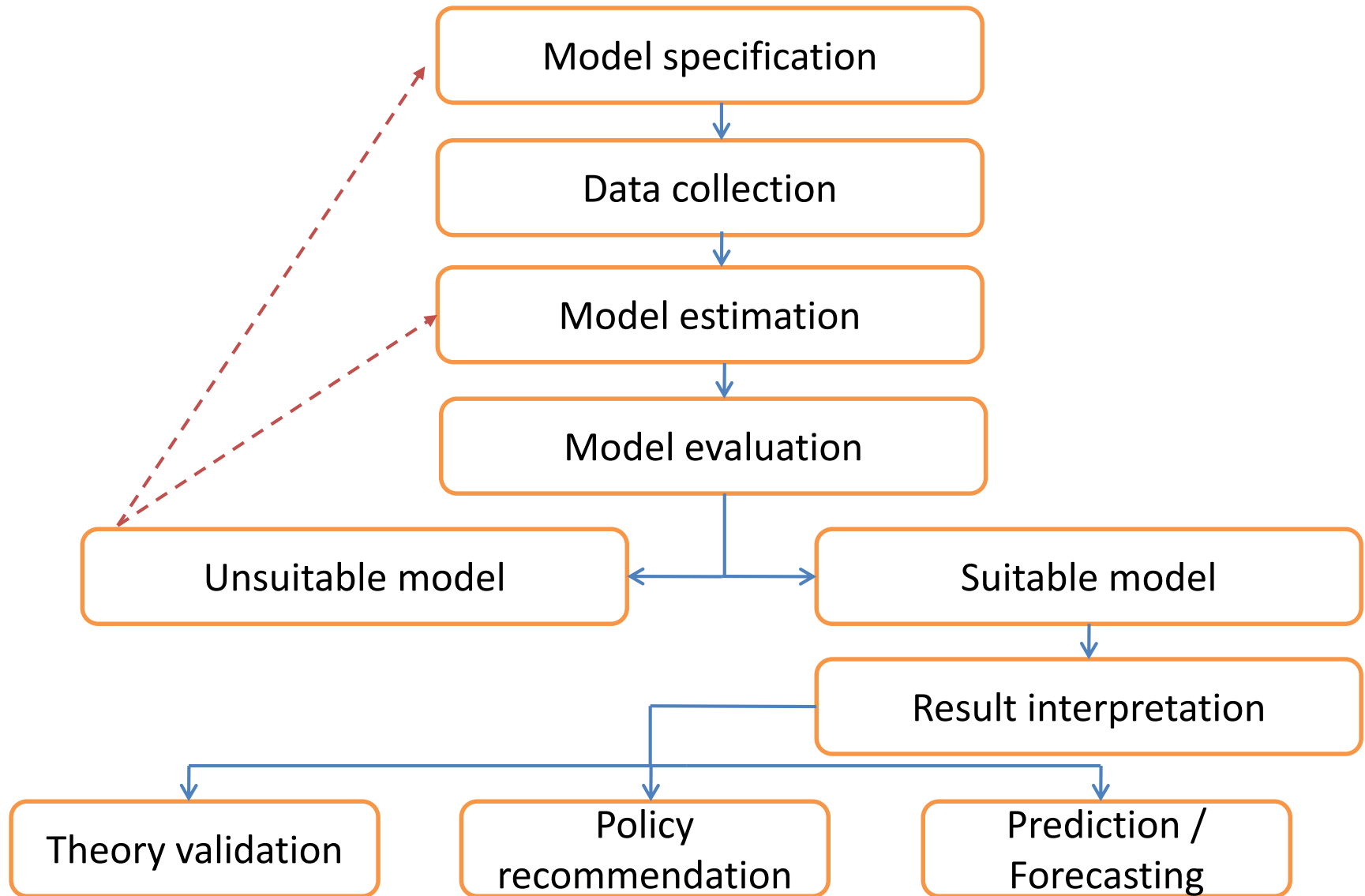
i.4. Types of Explanatory Variables

Econometrics:

- Definition:
 - Application of statistical techniques to the analysis of economic, financial, social... data aiming at estimating relationships between a **dependent variable** and a set of **explanatory variables**
- Ultimate purpose:
 - Theory validation
 - Prediction / Forecasting
 - Policy recommendation

i. Introduction

i.1. Econometric Methodology



Cross-sectional data:

- N cross-sectional units (individuals / firms / cities / ...)
- 1 time observation per unit

Time series:

- 1 unit
- T time observations per unit

Panel data

- N units
- T time observations per unit

Main types of econometric models:

- Regression Model:
 - Aim: explaining $E(Y|X)$
- Probabilistic model:
 - Aim: explaining $Pr(Y|X)$
 - Usually incorporates also a regression model for $E(Y|X)$

Y : dependent variable

X : explanatory variables

$E(Y|X)$: expected value for Y given X

$Pr(Y|X)$: probability of Y being equal to a specific value given X

Each type of econometric model has many variants

The numeric characteristics of the dependent variable restricts the variants that may be applied in each case:

Y	Type of outcome	Main model
$] - \infty, +\infty[$	Unbounded data	Linear
$[0, +\infty[$	Nonnegative data	Exponential
$[0,1]$	Fractional data	Fractional Logit,...
$\{0,1\}$	Binary choices	Logit,...
$\{0,1,2, \dots, J - 1\}$	Multinomial choices	Multinomial logit,...
$\{0,1,2, \dots, J - 1\}$	Ordered choices	Ordered logit,...
$\{0,1,2, \dots\}$	Count data	Poisson,...

Model Transformations and Adaptations:

- Bounded continuous outcomes may often be transformed in such a way that they give rise to unbounded outcomes which may be modelled using a linear model
- Any econometric model may require adaptations:
 - Data structure: cross-section, time series, panel
 - Non-random samples: stratified, censored, truncated
 - Measurement error
 - Endogenous explanatory variables
 - Corner solutions

Explanatory variables:

- Their characteristics are not relevant for the choice of econometric model, but affect the interpretation of the results
- Quantitative variables (examples):
 - Levels (Euro, kilograms, meters,...)
 - Levels and squares
 - Logs
 - Growth rates
 - Per capita values
- Qualitative variables
 - Binary (dummy) variables:
$$X = \{0,1\}$$
 - Interaction variables:
$$X = \text{Dummy var.} * (\text{Quantitative or dummy var.})$$

1. Linear Regression Analysis

1.1. The Linear Regression Model with Cross-Sectional Data

1.1. The Linear Regression Model with Cross-Sectional Data

1.1.1. Exogenous Explanatory Variables

Specification

Estimation

Interpretation

Inference

Model Evaluation

RESET Test

Tests for Heteroskedascity

Chow Test

1.1.1. Exogenous Explanatory Variables Specification

Model Specification:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i \quad (i = 1, \dots, N)$$

or

$$y = X\beta + u$$

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & X_{N2} & \cdots & X_{Nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

u : error term

β : parameters

k : n. explanatory variables

p : n. parameters ($= k + 1$)

N : n. observations

1.1.1. Exogenous Explanatory Variables Estimation

Model estimation:

- β is unknown and needs to be estimated
- Most popular estimation method - Ordinary Least Squares (OLS)

$$\min \sum_{i=1}^N \hat{u}_i^2, \quad \hat{u}_i = Y_i - \hat{Y}_i, \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik}$$

or

$$\min \hat{u}'\hat{u}, \quad \hat{u} = y - \hat{y}, \quad \hat{y} = X\hat{\beta}$$

\hat{u} : residuals

\hat{Y} : fitted values of Y

$\hat{\beta}$: estimator for β

1.1.1. Exogenous Explanatory Variables Estimation

- OLS estimators:

$$\frac{\partial \hat{u}'\hat{u}}{\partial \hat{\beta}} = -2X'(y - X'\hat{\beta}) = 0 \quad (\text{Note: implies } X'\hat{u} = 0)$$

$$X'y = (X'X)\hat{\beta}$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

Stata
regress $Y X_1 \cdots X_k$

1.1.1. Exogenous Explanatory Variables Estimation

Model assumptions:

1. Linearity in parameters
2. Random sampling
- 3. $E(\mathbf{u}|X) = \mathbf{0}$**
4. No perfect collinearity
5. Homoskedasticity: $Var(u|X) = \sigma^2 I$
6. Normality: $u \sim \mathcal{N}(0, \sigma^2 I)$

σ^2 : error variance

1.1.1. Exogenous Explanatory Variables Estimation

Estimator properties:

- Finite samples:
 - Assumptions 1-4: Unbiasedness
 - Assumptions 1-5: Unbiasedness and efficiency
 - Assumptions 1-6: Unbiasedness, efficiency and normality
- Asymptotically:
 - Assumptions 1-4: Consistency
 - Assumptions 1-5: Consistency, efficiency and normality

Unbiasedness: $E(\hat{\beta}) = \beta$

Efficiency: in the group of linear unbiased estimators, OLS displays the smallest variance [$\sigma_{\hat{\beta}}^2$ or $Var(\hat{\beta})$]

Normality: $\hat{\beta} \sim \mathcal{N}(\beta, \sigma_{\hat{\beta}}^2)$

Consistency: $\lim_{N \rightarrow \infty} E(\hat{\beta}) = \beta$

1.1.1. Exogenous Explanatory Variables Estimation

Goodness-of-fit:

- Sums of squares – measures of the sample variations in y , \hat{y} and \hat{u} :
 - Total Sum of Squares: $SST = \sum_{i=1}^N (Y_i - \bar{Y})^2$
 - Explained Sum of Squares: $SSE = \sum_{i=1}^N (\hat{Y}_i - \bar{\hat{Y}})^2$
 - Residual Sum of Squares: $SSR = \sum_{i=1}^N \hat{u}_i^2$

- The total variation in y can be expressed as the sum of the explained and unexplained variation:

$$SST = SSE + SSR$$

- Coefficient of determination (R^2):
 - Proportion of the variation of the dependent variable that is explained by the explanatory variables:

$$R^2 = \frac{SSE}{SST} = r_{y,\hat{y}}^2$$

r : coefficient of correlation

1.1.1. Exogenous Explanatory Variables Interpretation

Effects from unitary changes in a given explanatory variable:

$$\Delta X_j = 1 \Rightarrow \Delta Y = \beta_j$$

- Needs adaptation for:
 - Transformed dependent variables
 - Transformed quantitative explanatory variables
 - Qualitative explanatory variables
- Aims:
 - Most of the time: testing whether the effect is null or significantly different from zero → it is equivalent to test whether a parameter or a set of parameters or a linear combination of parameters are significantly different from zero
 - Most of the time: analyze the sign of the effect (null, positive, negative)
 - Sometimes: calculate and analyze the magnitude of the effect

1.1.1. Exogenous Explanatory Variables Interpretation

Partial effect for quantitative variables:

- (*Ceteris paribus*) effect over the dependent variable originated by a unitary change in a given explanatory variable:

$$\Delta X_{ij} = 1 \text{ unit} \Rightarrow \Delta Y_i = ?$$

- It is approximated by differentiating the dependent variable in order to the explanatory variable:

- Model in levels - $Y_i = \dots + \beta_j X_{ij} + \dots + u_i$:

$$\frac{\partial Y_i}{\partial X_{ij}} = \beta_j, \text{ which implies that } \Delta X_{ij} = 1 \text{ unit} \Rightarrow \Delta Y_i = \beta_j \text{ units}$$

- Quadratic model - $Y_i = \dots + \beta_j X_{ij} + \beta_m X_{ij}^2 + \dots + u_i$:

$$\frac{\partial Y_i}{\partial X_{ij}} = \beta_j + 2\beta_m X_{ij}, \text{ which implies that}$$

$$\Delta X_{ij} = 1 \text{ unit} \Rightarrow \Delta Y_i = (\beta_j + 2\beta_m X_{ij}) \text{ units}$$

1.1.1. Exogenous Explanatory Variables Interpretation

- Model in logs - $\ln(Y_i) = \dots + \beta_j \ln(X_{ij}) + \dots + u_i$:
 - Re-transformed model: $Y_i = e^{\dots + \beta_j \ln(X_{ij}) + \dots + u_i}$

- Derivative:

$$\frac{\partial Y_i}{\partial X_{ij}} = \frac{\beta_j}{X_{ij}} e^{\dots + \beta_j \ln(X_{ij}) + \dots + u_i}$$

$$\approx \frac{\Delta Y_i}{\Delta X_{ij}} = \frac{\beta_j}{X_{ij}} Y_i$$

$$\Leftrightarrow \frac{\Delta Y_i}{Y_i} \times 100 = \beta_j \frac{\Delta X_{ij}}{X_{ij}} \times 100$$

$$\Leftrightarrow \% \Delta Y_i = \beta_j \% \Delta X_{ij}$$

- Effect:

$$\% \Delta X_{ij} = 1\% \Rightarrow \% \Delta Y_i = \beta_j \%$$

» β_j represents an elasticity

1.1.1. Exogenous Explanatory Variables Interpretation

- Log-linear model - $\ln(Y_i) = \dots + \beta_j X_{ij} + \dots + u_i$:

- Re-transformed model: $Y_i = e^{\dots + \beta_j X_{ij} + \dots + u_i}$

- Derivative:

$$\frac{\partial Y_i}{\partial X_{ij}} = \beta_j e^{\dots + \beta_j X_{ij} + \dots + u_i}$$

$$\approx \frac{\Delta Y_i}{\Delta X_{ij}} = \beta_j Y_i$$

$$\Leftrightarrow \frac{\Delta Y_i}{Y_i} \times 100 = 100 \times \beta_j \Delta X_{ij}$$

$$\Leftrightarrow \% \Delta Y_i = 100 \beta_j \Delta X_{ij}$$

- Effect:

$$\Delta X_{ij} = 1 \text{ unit} \Rightarrow \% \Delta Y_i = 100 \beta_j \%$$

1.1.1. Exogenous Explanatory Variables Interpretation

- Semi-logarithmic model - $Y_i = \dots + \beta_j \ln(X_{ij}) + \dots + u_i$:

- Derivative:

$$\frac{\partial Y_i}{\partial X_{ij}} = \frac{\beta_j}{X_{ij}}$$
$$\approx \Delta Y_i \times 100 = \beta_j \frac{\Delta X_{ij}}{X_{ij}} \times 100$$

$$\Leftrightarrow \Delta Y_i = \frac{\beta_j}{100} \% \Delta X_{ij}$$

- Effect:

$$\% \Delta X_{ij} = 1\% \Rightarrow \Delta Y_i = \frac{\beta_j}{100} \text{ units}$$

1.1.1. Exogenous Explanatory Variables Interpretation

Partial effect for dummy variables:

- (*Ceteris paribus*) difference on the value of the dependent variable between two groups
- The effect is given by the parameter associated to the dummy variable:

- Definition of a dummy variable:

$$D_i = \begin{cases} 1 & \text{if the individual belongs to group } G_A \\ 0 & \text{if the individual belongs to group } G_B \end{cases}$$

- Example:

- Base model: $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \beta_d D_i + u_i$
- If $D_i = 1$, then $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \beta_d + u_i$
- If $D_i = 0$, then $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + u_i$
- Difference (effect of belonging to group G_A): β_d

1.1.1. Exogenous Explanatory Variables Interpretation

Partial effect for interaction variables:

- (*Ceteris paribus*) difference between two groups on the effect over the value of the dependent variable of a unitary change in a given explanatory variable
- The effect is given by the parameter associated to the interaction variable:
 - Example:
 - Base model: $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \beta_m X_{im} + \beta_{dm} (D_i * X_{im}) + u_i$
 - If $D_i = 1$, then $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + (\beta_m + \beta_{dm}) X_{im} + u_i$
 - If $D_i = 0$, then $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \beta_m X_{im} + u_i$
 - Partial effect of X_{im} for those in group G_A : $\beta_m + \beta_{dm}$
 - Partial effect of X_{im} for those in group G_B : β_m
 - Difference: β_{dm}

1.1.1. Exogenous Explanatory Variables

Inference

Estimators for the variance of the parameter estimators:

- Standard – assumes homoskedasticity: $Var(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$
- Robust – allows for heteroskedasticity:

$$Var(\hat{\beta}) = (X'X)^{-1}X'\hat{\Phi}X(X'X)^{-1}, \quad \hat{\Phi} = \begin{bmatrix} \hat{u}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{u}_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{u}_N^2 \end{bmatrix}$$

- Cluster-robust – specific for panel data
- Bootstrap – simulated variance

Stata

```
regress YX1 ... Xk  
regress YX1 ... Xk, vce(robust)  
regress YX1 ... Xk, vce(cluster clustvar)  
regress YX1 ... Xk, vce(bootstrap)
```

1.1.1. Exogenous Explanatory Variables

Inference

Hypothesis tests:

- Require specification of the:
 - Null and alternative hypothesis; typically:
 - $H_0: \text{Partial effect} = 0$
 - $H_1: \text{Partial effect} \neq 0$ (> 0 or < 0 also possible)
 - Significance level (α):
 - Probability of rejecting the null hypothesis when it is true
 - Typically, $\alpha = 0.01, 0.05$ or 0.10
- p -value:
 - The p -value of the result of a test is the probability of obtaining a value at least as extreme when the null hypothesis is true; therefore:
 - $p < \alpha \Rightarrow \text{Reject } H_0$
 - $p > \alpha \Rightarrow \text{Do not reject } H_0$

1.1.1. Exogenous Explanatory Variables Inference

- Main tests:
 - Test for the individual significance of a parameter: t test
 - Test for the joint significance of a set of parameters: F test

t test:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{N-p}^{\alpha/2}$$

$$|t| < t_{N-p}^{\alpha/2} \Rightarrow \text{Do not reject } H_0$$

$$|t| > t_{N-p}^{\alpha/2} \Rightarrow \text{Reject } H_0$$

Stata

```
regress YX1 ... Xk  
regress YX1 ... Xk, vce(robust)  
regress YX1 ... Xk, vce(cluster clustvar)  
regress YX1 ... Xk, vce(bootstrap)
```

1.1.1. Exogenous Explanatory Variables

Inference

F test:

- Model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_g X_g + \beta_{g+1} X_{g+1} + \cdots + \beta_k X_k + v$$

- Hypotheses:

$$H_0: \beta_{g+1} = \cdots = \beta_k = 0$$

$$H_1: \text{No } H_0$$

R^2 : coefficient of determination of the full model

R_*^2 : coefficient of determination of the restricted model

q : n. of parameters to be tested

- Test:

$$F = \frac{R^2 - R_*^2}{1 - R^2} \frac{N - p}{q} \sim F_{N-p}^q \rightarrow \text{valid only under homoskedasticity}$$

$$F = N \hat{\beta}'_* [\text{Var}(\hat{\beta}_*)]^{-1} \hat{\beta}_* \sim \chi_q^2 \rightarrow \text{general formula}$$

- Decision:

$$F < F_{N-p}^q \Rightarrow \text{Do not reject } H_0$$

$$F > F_{N-p}^q \Rightarrow \text{Reject } H_0$$

Stata

```
regress Y X_1 ... X_g X_{g+1} ... X_k, ...  
test X_{g+1} ... X_k
```


Specification tests:

- Model functional form: RESET test
- Heteroskedasticity: Breusch-Pagan (BP) test
- Structural breaks: Chow tests

1.1.1. Exogenous Explanatory Variables

Model Selection - RESET Test

RESET test:

- Intuition:

- Any model of the type $E(Y|X) = S(X\beta)$ may be approximated by

$$E(Y|X) = L \left[X\beta + \sum_{j=1}^{\infty} \gamma_j (X\hat{\beta})^{j+1} \right]$$

- Assume a linear form for $L(\cdot)$ and check if $\gamma_j = 0$

- Implementation:

- Estimate the original model and get $\hat{\beta}$
- Generate the variables $(X\hat{\beta})^2, (X\hat{\beta})^3, (X\hat{\beta})^4, \dots$
- Add the generated variables to the original model and estimate the following auxiliary model:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \gamma_1 (X\hat{\beta})^2 + \gamma_2 (X\hat{\beta})^3 + \gamma_3 (X\hat{\beta})^4 + \dots + v$$

- Apply an F test for the significance of the added variables:

$H_0: \gamma_1 = \gamma_2 = \gamma_3 = \dots = 0$ (suitable model functional form)

H_1 : No H_0 (unsuitable model functional form)

Stata

(only test version based on three fitted powers)

```
regress Y X_1 ... X_k  
estat ovtest
```

1.1.1. Exogenous Explanatory Variables

Model Selection - Tests for Heteroskedascity

BP Test:

- Intuition:

- Because the mean of the error term is zero, the variance of the error term is given by the sum of the squared error terms
- The squared residuals are an estimate of the squared error terms
- Check if the squared residuals and the explanatory variables are correlated

- Implementation:

- Estimate the original model and generate the variable \hat{u}^2
- Replace Y by \hat{u}^2 in the original model and estimate the following auxiliary model:

$$\hat{u}^2 = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_k X_k + v$$

- Apply an F test for the joint significance of the right-hand side variables of the previous auxiliary model:

$$H_0: \gamma_1 = \cdots = \gamma_k = 0 \text{ (homoskedasticity)}$$

$$H_1: \text{Não } H_0 \text{ (heteroskedasticity)}$$

Stata
regress Y $X_1 \dots X_k$
estat hettest, rhs fstat

1.1.1. Exogenous Explanatory Variables

Model Selection - Chow Test

Chow Test for Structural Breaks:

- Context:

- Two groups of individuals / firms / ...: G_A, G_B
- It is suspected that the behaviour of the two groups in which regards the dependent variable may have different determinants

- Implementation:

- Generate the dummy variable $D = \begin{cases} 1 & \text{if the individual belongs to } G_A \\ 0 & \text{if the individual belongs to } G_B \end{cases}$

- Estimate the original model 'duplicated':

$$Y = \theta_0 + \theta_1 X_1 + \dots + \theta_k X_k + \gamma_0 D + \gamma_1 DX_1 + \dots + \gamma_k DX_k + v$$

- Apply an F test for the significance of the terms where D is present:

$$H_0: \gamma_0 = \dots = \gamma_k = 0 \text{ (no structural break)}$$

$$H_1: \text{Não } H_0 \text{ (with a structural break)}$$

Stata
regress $Y X_1 \dots X_k D DX_1 \dots DX_k$
test $D DX_1 \dots DX_k$

1.1.2. Endogenous Explanatory Variables

Endogeneity

- Definition and Consequences

- Motivation: Omitted variables, Measurement Errors, Simultaneous Equations

- Solutions: Instrumental Variables, Panel Data

Methods based on Instrumental Variables

- Two-Stage Least Squares

- Generalized Method of Moments (GMM)

Specification Tests

- Test for the Exogeneity of an Explanatory Variable

- Test for the Exogeneity of the Instrumental Variables

- Tests for Correlation between Instrumental Variables and Explanatory Variables

1.1.2. Endogenous Explanatory Variables

Endogeneity: Definition and Consequences

Definitions:

- Exogenous explanatory variables: $E(u|X) = 0 \rightarrow$ essential assumption in any regression model
- Endogenous explanatory variables: $E(u|X) \neq 0$

Consequences:

- OLS estimators become unbiased and inconsistent

Motivation:

- Omitted variables
- Covariate measurement error
- Simultaneity

Omitted variables - example:

- True model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$, $E(v|X_1, X_2) = 0$
- Estimated model: $Y = \beta_0 + \beta_1 X_1 + u$
- As $u = \beta_2 X_2 + v$:
 - If $cov(X_1, X_2) = 0$, then $E(u|X_1) = 0 \rightarrow X_1$ is exogenous
 - If $cov(X_1, X_2) \neq 0$, then $E(u|X_1) \neq 0 \rightarrow X_1$ is endogenous

Covariate measurement error - example:

- True model: $Y = \beta_0 + \beta_1 X_1^* + v$, $E(v|X_1^*) = 0$
- e : measurement error
- Instead of X_1^* , it is observed $X_1 = X_1^* + e$
- Estimated model: $Y = \beta_0 + \beta_1 X_1 + u$
- As $u = v - \beta_1 e$:
 - If $cov(X_1, e) = 0$, then $E(u|X_1) = 0 \rightarrow X_1$ is exogenous
 - If $cov(X_1, e) \neq 0$, then $E(u|X_1) \neq 0 \rightarrow X_1$ is endogenous \rightarrow most common case because the measurement error is on X_1

Measurement error on Y has less serious consequences:

- True model: $Y^* = \beta_0 + \beta_1 X_1 + v$, $E(v|X_1) = 0$
- Instead of Y^* , it is observed $Y = Y^* + e$
- Estimated model: $Y = \beta_0 + \beta_1 X_1 + u$
- As $u = v + e$:
 - In general, $cov(X_1, e) = 0$ and $E(u|X_1) = 0$, since the measurement error is on Y and not in X_1
 - Hence, usually there are no endogeneity problems
 - However, estimation is less precise, since the error term has now two components

Simultaneity - example:

- True model:
$$\begin{cases} \text{Supply: } Q = \beta_0 + \beta_1 P + u \\ \text{Demand: } Q = \alpha_0 + \alpha_1 P + v \end{cases}$$
- Estimated model:
$$\begin{cases} \text{Supply: } Q = \beta_0 + \beta_1 P + u \\ \text{Demand: } Q = \alpha_0 + \alpha_1 P + v \end{cases}$$
- As:
$$\begin{cases} P = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{v - u}{\beta_1 - \alpha_1} \\ Q = \dots \end{cases}$$

then P is function of v and u ; hence:

- $E(u|P) \neq 0$ in the supply equation $\rightarrow P$ is endogenous
- $E(v|P) \neq 0$ in the demand equation $\rightarrow P$ is endogenous

1.1.2. Endogenous Explanatory Variables

Endogeneity: Instrumental Variables

What to do in case of endogeneity:

- Universal solution – methods based on ‘instrumental variables’:
 - Two-Stage Least Squares
 - Generalized Method of Moments (GMM)
- When data is in panel form and the endogeneity problem is caused by omitted time-constant variables:
 - Methods based on the removal of the ‘fixed effects’

Instrumental variables:

- Context:
 - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$ (structural model)
 - $E(u|X_1) \neq 0 \rightarrow X_1$ is endogenous
- Definitions of instrumental variable (IV_A, \dots, IV_M):
 - $E(u|IV_A) = \dots = E(u|IV_M) = 0$
 - $cov(IV_A, X_1) \neq 0, \dots, cov(IV_M, X_1) \neq 0$
- Exogenous variables: $Z = [1 \ X_2 \ \dots \ X_k \ IV_A \ \dots \ IV_M]$
- The number of instrumental variables must be equal or larger than the number of endogenous explanatory variables

1.1.2. Endogenous Explanatory Variables

Methods based on Instrumental Variables: Two-Stage Least Squares (2SLS)

Implementation:

1. Estimate the reduced form of the model by OLS:

$$\underbrace{X_1}_{\text{End. Expl. Var.}} = \pi_0 + \underbrace{\pi_2 X_2 + \dots + \pi_k X_k}_{\text{Ex. Expl. Var.}} + \underbrace{\pi_A IV_A + \dots + \pi_M IV_M}_{\text{Instrumental Variables}} + w$$

and get $\hat{X}_1 = \hat{\pi}_0 + \hat{\pi}_2 X_2 + \dots + \hat{\pi}_k X_k + \hat{\pi}_A IV_A + \dots + \hat{\pi}_M IV_M$

2. Estimate the structural model, with X_1 replaced by \hat{X}_1 , by OLS:

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$$

Stata

(by default, variances are estimated in a standard way; to use another estimator, use the option `vce(robust)` or similar)

```
ivregress 2sls Y (X1= IV_A... IV_M) X2 ... X_k
```

1.1.2. Endogenous Explanatory Variables

Methods based on Instrumental Variables: GMM

- Very general estimation method:
 - Applies to a large variety of cases
 - Includes as particular cases OLS, maximum likelihood, etc.
- Formulation:
 - Moment conditions: $E[g(Y, X, IV; \beta)] = 0$
 - m moment conditions
 - p parameters: $\beta_0, \beta_1, \dots, \beta_k$
 - Optimization:
 - $m = p \rightarrow$ just-identified model: $g(Y, X, IV; \hat{\beta}) = 0$
 - $m > p \rightarrow$ overidentified model:
$$\min J = g(Y, X, IV; \beta)' W g(Y, X, IV; \beta)$$
where W is a weighting matrix; the first-order conditions are given by:
$$\frac{\partial g(Y, X, IV; \hat{\beta})'}{\partial \beta} W g(Y, X, IV; \hat{\beta}) = 0$$

1.1.2. Endogenous Explanatory Variables

Methods based on Instrumental Variables: GMM

Choosing W in overidentified models:

- To get efficient estimators, W has to be defined as the inverse of the covariance matrix of the moment conditions:

$$W = \Omega^{-1},$$

where:

$$\Omega = \text{Var}[g(Y, X, IV; \beta)] = E[g(Y, X, IV; \beta)g(Y, X, IV; \beta)']$$

Stata

(by default, variances are estimated in a standard way; to use another estimator, use the option `vce(robust)` or similar)

```
ivregress gmm Y (X1 = IVA ... IVM) X2 ... Xk
```

1.1.2. Endogenous Explanatory Variables

Methods based on Instrumental Variables: GMM

Particular case – OLS:

- Assumption: $E(u|X) = 0 \implies E(X'u) = 0$
- $g(Y, X, IV; \beta) = X'u = X'(y - X\beta)$
- $m = p \implies X'(y - X\hat{\beta}) = 0$

$$X'(y - X\hat{\beta}) = 0$$

$$X'y - X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'y$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

1.1.2. Endogenous Explanatory Variables

Methods based on Instrumental Variables: GMM

Instrumental Variables ($m = p$):

- Assumption: $E(u|Z) = 0 \implies E(Z'u) = 0$
- Moment conditions: $g(Y, X, IV; \beta) = Z'u = Z'(y - X\beta)$
- If $m = p \implies Z'(y - X\hat{\beta}) = 0$

$$\hat{\beta} = (Z'X)^{-1}Z'y$$

- In this case, GMM is identical to 2SLS estimation

1.1.2. Endogenous Explanatory Variables

Methods based on Instrumental Variables: GMM

Instrumental Variables ($m > p$):

- Optimization problem: $\min J = (y - X\beta)'Z\Omega^{-1}Z'(y - X\beta)$
- First-order conditions:

$$\begin{aligned} -X'Z\tilde{\Omega}^{-1}Z'(y - X\hat{\beta}) &= 0 \\ X'Z\tilde{\Omega}^{-1}Z'y &= X'Z\hat{\Omega}^{-1}Z'X\hat{\beta} \\ \hat{\beta} &= (X'Z\tilde{\Omega}^{-1}Z'X)^{-1}X'Z\hat{\Omega}^{-1}Z'y \end{aligned}$$

where $\tilde{\Omega}$ is a preliminar estimate of $\Omega = E(Z'uu'Z)$, which requires a preliminar estimate for β ; typically, $\tilde{\beta}$ is obtained from $\min J = (y - X\beta)'ZZ'(y - X\beta)$ (assumes $W = I$)

1.1.2. Endogenous Explanatory Variables

Specification Tests

Tests relevant for 2SLS / GMM:

- Tests for the exogeneity of an explanatory variable
 - If the explanatory variable is exogenous, it is better to use OLS in order to get efficient estimators
 - Methods based on IV's should be used only if really necessary, since there may be a substantial loss in precision
- Tests for the exogeneity of the instrumental variables
 - To act as an IV, a variable has to be exogenous → when based on “IV's” that actually are endogenous, 2SLS and GMM are inconsistent
- Tests for correlation between instrumental variables and explanatory variables
 - To act as an IV, a variable has to be correlated with the endogenous explanatory variable → when based on “IV's” uncorrelated with the endogenous regressors, 2SLS and GMM are inconsistent
 - If the IV's are only weakly correlated with the endogenous regressors, then the model will be poorly identified → in such a case, 2SLS and GMM may display a huge variability

1.1.2. Endogenous Explanatory Variables

Tests for the Exogeneity of an Explanatory Variable

2SLS - Wu-Hausman test:

1. Estimate the reduced model by OLS:

$$X_1 = \pi_0 + \pi_2 X_2 + \dots + \pi_k X_k + \pi_A IV_A + \dots + \pi_M IV_M + w$$

2. Calculate the residuals \hat{w}

3. Add \hat{w} to the structural model and re-estimate it, by OLS:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \delta \hat{w} + u$$

4. t test:

$$H_0: \delta = 0 \text{ (} X_1 \text{ is exogenous)}$$

$$H_1: \delta \neq 0 \text{ (} X_1 \text{ is endogenous)}$$

Stata

```
ivregress 2sls Y (X1= IVA... IVM) X2 ... Xk  
estat endogenous
```

GMM - Eichenbaum, Hansen e Singleton's (1988) C test

- Based on the difference of two J statistics (see the next page)

Stata

```
ivregress gmm Y (X1= IVA... IVM) X2 ... Xk  
estat endogenous
```

1.1.2. Endogenous Explanatory Variables

Tests for the Exogeneity of the Instrumental Variables

- These tests can be applied only when the model is overidentified
- If the model is just-identified, then it is only possible to justify the exogeneity of the IV's using theoretical arguments
- Hansen's J test (also called test for overidentifying restrictions; this is an extension of the Sargan test, very common in the 2SLS framework under homoskedasticity):
 1. Estimate the model by GMM
 2. Test

$$H_0: E(u|Z) = 0 \text{ (IV's are exogenous)}$$

$$H_1: E(u|Z) \neq 0 \text{ (IV's are not exogenous)}$$

using the J statistic:

$$\hat{J} = g(Y, X, IV; \hat{\beta})' \hat{W} g(Y, X, IV; \hat{\beta}) \sim \chi_q^2$$

q : number of overidentifying restrictions

Stata

(applies also to the 2SLS estimator, with the obvious adaptation)

```
ivregress gmm Y (X1 = IVA ... IVM) X2 ... Xk  
estat overid
```

1.1.2. Endogenous Explanatory Variables

Tests for Correlation between Instrumental Variables and Explanatory Variables

Alternatives:

- F tests for the significance of the IV's in the reduced form model
- Criteria / tests for 'weak instruments'

F tests:

1. Estimate the reduced form model:

$$X_1 = \pi_0 + \pi_2 X_2 + \cdots + \pi_k X_k + \pi_A IV_A + \cdots + \pi_M IV_M + w$$

2. Test the hypothesis:

$$H_0: \pi_A = \cdots = \pi_M = 0 \text{ (IV's and } X_1 \text{ are not correlated)}$$

Stata

(applies also to the GMM estimator, with the obvious adaptation)

```
ivregress 2sls Y (X1= IVA... IVM) X2 ... Xk  
estat firststage
```

Tests for 'weak instruments':

- Even when the previous tests reveal that IV 's and X_1 are correlated, that correlation may be so weak that 2SLS / GMM estimators are very little precise
- Criterium: $F < 10 \rightarrow$ Suggest a high probability of having 'weak instruments'
- Tests: Cragg and Donald (2005), Kleibergen and Paap (2006)

1.2. The Linear Regression Model with Panel Data

1.2.1. Static Models

1.2.2. Dynamic Models

1.2.1. Static Models

Definitions

Panel data:

- N cross-sectional units: $i = 1, \dots, N$
- T time observations per unit: $t = 1, \dots, T$

Econometric analysis more complex:

- Cross-sectional data: different units \Rightarrow independent observations
- Panel data: same units \Rightarrow dependent observations over time

Advantages:

- Make possible the analysis of the dynamics of individual behaviours
- Generate more efficient estimators, since samples are larger
- Allow for endogenous explanatory variables in some special cases
- It is simple to get instrumental variables

Limitations:

- Prediction and calculation of partial effects not possible in some models

Temporal dimension:

- Short panels:
 - Sample comprises many individuals ($N \rightarrow \infty$), but there is a reduced number of time observations (small T)
 - The temporal correlation of the observations for each individual is an issue, but across individuals it is assumed independence
- Long panels:
 - Large number of time observations for each individual ($T \rightarrow \infty$)
 - Need to take into account time series issues (stationarity, cointegration, etc.)

Cross-sectional composition:

- Balanced panels:
 - Every individual is observed in all time periods ($T_i = T, \forall i$)
- Unbalanced panels:
 - Some individuals are not observed in some time periods ($T_i \neq T$)
 - Motivation: some individuals refuse to continue providing information after some time periods \rightarrow 'attrition' problem
 - Most estimators work with unbalanced panels, provided that one may assume that there is no endogenous selection (the data are missing at random)

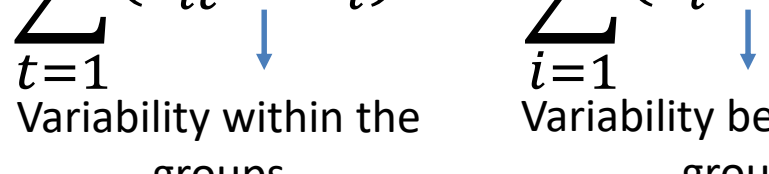
1.2.1. Static Models

Definitions

Variability over time and across individuals:

- With panel data, the variance of Y_{it} can be decomposed as follows:

$$\begin{aligned}\sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \bar{Y})^2 &= \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \bar{Y}_i + \bar{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \bar{Y}_i)^2 + \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2\end{aligned}$$



Variability within the groups Variability between the groups

- To obtain the the total, 'within' and 'between' variance, divide by, respectively, $NT - 1$, $N(T - 1)$ and $N - 1$

Base Model – Model with individual effects:

$$Y_{it} = \alpha_i + x'_{it}\beta + u_{it} \quad (i = 1, \dots, N; t = 1, \dots, T)$$

- α_i : individual effects, time-constant
- x_{it} - explanatory variables, including:
 - x_{it} : changes across individuals and over time
 - x_i : time-constant
 - d_t : temporal dummy variable
 - t : trend (one may also have t^2, t^3, \dots)
 - $d_t \cdot x_{it}$: interaction term
- u_{it} : idiosyncratic error term – changes randomly across individuals and over time

Other models:

- ‘Pooled’ or ‘Population-averaged’ model:

$$Y_{it} = \alpha + x'_{it}\beta + u_{it}$$

- Direct extension of cross-sectional models
- Popular in the area of Statistics but not in Economics, where individual heterogeneity is always an issue

- Random Coefficients Model:

$$Y_{it} = \alpha_i + x'_{it}\beta_i + u_{it}$$

- More complex than the base model, allowing for a different coefficient also for the explanatory variables
- Computationally-intensive model, hard to estimate

The individual effects model may be re-written as:

$$Y_{it} = x'_{it}\beta + (\alpha_i + u_{it})$$

- The error term has now two components, α_i and u_{it}
- The individual effects α_i may be correlated, or not, with the explanatory variables

Fixed effects:

- α_i and x_{it} are correlated $\rightarrow x_{it}$ is endogenous
- Direct estimation of the model is not possible

Random effects:

- α_i and x_{it} are not correlated $\rightarrow x_{it}$ is exogenous
- Direct estimation of the model is possible

Most panel data estimators:

- Are based on transformed versions of the base model
- Differ of the type of exogeneity required for the explanatory variables:
 - Contemporaneous: $E(x_{it}u_{it}) = 0$
 - Weak (pre-determined variables): $E(x_{it}u_{i,t+j}) = 0, j \geq 0$
 - Strict: $E(x_{it}u_{is}) = 0, \forall s, t$

List of estimators:

- Estimators for the random effects model:
 - Pooled OLS
 - Between estimator
 - Random effects estimator
- Estimators for the fixed effects model:
 - Fixed effects or Within estimator
 - Least squares dummy variables (LSDV) estimator
 - First-differences estimator
- Estimators based on instrumental variables:
 - General estimators
 - Hausman-Taylor estimator

1.2.1. Static Models

Estimators for the Random Effects Model

Pooled OLS estimator:

- Model:

$$Y_{it} = \alpha + x'_{it}\beta + \underbrace{(\alpha_i - \alpha + u_{it})}_{v_{it}}$$

- Assumption: $E[x_{it}(\alpha_i + u_{it})] = 0$
 - Requires random effects: α_i e x_{it} must be uncorrelated
 - Requires contemporaneous exogeneity between u_{it} and x_{it}
- Estimation:
 - OLS with a cluster-type estimator for the variance

Stata
`regress Y X_1 ... X_k , vce(cluster clustvar)`

1.2.1. Static Models

Estimators for the Random Effects Model

Between estimator:

- Model:

$$\bar{Y}_i = \alpha + \bar{x}_i' \beta + \underbrace{(\alpha_i - \alpha + \bar{u}_i)}_{\bar{v}_i}$$

- $\bar{Y}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} Y_{it}$, etc.
- Assumption: $E[\bar{x}_i(\alpha_i + \bar{u}_i)] = 0$
 - Requires random effects: α_i e x_{it} must be uncorrelated
 - Requires strict exogeneity between u_{it} and x_{it}
- Estimation:
 - OLS

Stata
xtreg YX₁ ... X_k, be

Random effects estimator:

- Model:

$$Y_{it} = \alpha + x'_{it}\beta + \underbrace{(\alpha_i - \alpha + u_{it})}_{v_{it}}$$

- New assumptions:

- $Var(\alpha_i) = \sigma_\alpha^2$
- $Var(u_{it}) = \sigma_u^2$

- Under the new assumptions:

$$cor(v_{it}, v_{is}) = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_u^2)$$

- Taking into account this result, more efficient estimators may be obtained
- All the previous estimators did not fully exploit the panel nature of the data in the estimation process

1.2.1. Static Models

Estimators for the Random Effects Model

- Estimation:

- Generalized least squares (GLS)
- It is equivalent to apply OLS to the following equation:

$$Y_{it} - \hat{\theta}_i \bar{Y}_i = (1 - \hat{\theta}_i) \alpha + (x_{it} - \hat{\theta}_i \bar{x}_i)' \beta + v_{it}$$

- $\hat{\theta}_i = 1 - \sqrt{\hat{\sigma}_u^2 / (T_i \hat{\sigma}_\alpha^2 + \hat{\sigma}_u^2)}$
- $v_{it} = (1 - \hat{\theta}_i) \alpha_i + (u_{it} - \hat{\theta}_i \bar{u}_i)$

- Assumption: $E(\bar{x}_i v_{it}) = 0$

- Requires random effects: α_i e x_{it} must be uncorrelated
- Requires strict exogeneity between u_{it} and x_{it}

Stata
`xtreg YX1 ... Xk, re vce(cluster clustvar)`

1.2.1. Static Models

Estimators for the Fixed Effects Model

Fixed effects / within estimator:

- Model:

$$Y_{it} - \bar{Y}_i = (x_{it} - \bar{x}_i)' \beta + (u_{it} - \bar{u}_i)$$

- Corresponds to the subtraction of the model defining the between estimator from the base individual effects model

- Assumption: $E[(x_{it} - \bar{x}_i)(u_{it} - \bar{u}_i)] = 0$

- Requires strict exogeneity between u_{it} and x_{it}

- Estimation:

- OLS with a cluster-type estimator for the variance

Stata
`xtreg YX1 ... Xk, fe vce(cluster clustvar)`

1.2.1. Static Models

Estimators for the Fixed Effects Model

- Limitations:
 - It is not possible to include in the model:
 - Time-constant explanatory variables
 - (If the model includes time dummies) Explanatory variables with identical changes over time for all individuals (e.g. age)
 - Prediction not possible
 - Partial effects conditional not only on x_{it} but also on α_i ; how to calculate them?
- Main advantage:
 - Allow for (time-constant) unobserved individual heterogeneity that may be correlated with the explanatory variables, not requiring the use of instrumental variables

LSDV estimator:

- Model:

$$Y_{it} = \sum_{j=1}^N \alpha_j d_{ij} + x'_{it} \beta + u_{it}$$

where $d_{ij} = 1$ if $i = j$ and 0 if $i \neq j$

- Assumptions and estimates of β identical to those of the fixed effects estimator
- Estimates of α_j :
 - Given by $\hat{\alpha}_i = \bar{Y}_i - \bar{x}_i' \hat{\beta}$
 - Consistent only in case of a long panel, in which case it is also possible to make prediction and calculate partial effects conditional on both x_{it} and α_i

Stata

```
areg  $YX_1 \dots X_k$ , absorb(clustvar) vce(cluster clustvar)  
or (time-constant variables need to be manually dropped in the second alternative)  
regress  $YX_1 \dots X_k$  i.clustvar, vce(cluster clustvar)
```

First-differences estimator:

- Model:

$$Y_{it} - Y_{i,t-1} = (x_{it} - x_{i,t-1})' \beta + (u_{it} - u_{i,t-1}) \Leftrightarrow \\ \Delta Y_{it} = \Delta x_{it}' \beta + \Delta u_{it}$$

- Corresponds to the subtraction of the first-differences equation from the base individual effects model

- Assumption: $E(\Delta x_{it} \Delta u_{it}) = 0$

- Requires $E(x_{it} u_{it}) = E(x_{it} u_{i,t-1}) = E(x_{it} u_{i,t+1}) = 0$

- Estimation:

- OLS

Stata
`regress D.Y D.X1 ... D.Xk, vce(cluster clustvar) nocons`

- Comparison with the fixed-effects estimator:

- Identical when $T = 2$
- Does not require strict exogeneity

Endogenous explanatory variables - $E(x_{it}u_{it}) \neq 0$:

- Possible IV's for x_{it} :
 - External instruments, as in the cross-sectional case
 - Internal instruments (same explanatory variable but relative to other time periods)
- Example of internal instruments:
 - If x_{it} is weakly exogenous (apart from the current period), then:
 - All past values (lags) of x_{it} may be used as IV's
 - Possible IV's: $x_{i,t-1}$ or $(x_{i,t-1}, x_{i,t-2})$ or $(x_{i,t-1}, \dots, x_{i,t-5})$, etc.
 - If x_{it} is strictly exogenous (apart from the current period), then:
 - All past (lags) and future (leads) values of x_{it} may be used as IV's
 - Possible IV's : $x_{i,t-1}$ and/or $x_{i,t+1}$, etc.

Stata

(external instruments – 2SLS)

```
xtivreg Y (X1=IVA... IVM) X2 ... Xk, options  
(options: re, fe, be, fd)
```

Stata

(internal instruments – 2SLS)

```
xtivreg Y (X1=L.X1 L2.X1...) X2 ... Xk, options  
(options: re, fe, be, fd)
```

Hausman-Taylor estimator:

- Useful when interest lies on time-invariant explanatory variables which are correlated with α_i , since:
 - Fixed effects estimators: drop those variables from the model
 - Random effects estimators: inconsistent
- Intermediate case between the fixed and the random effects cases:
 - Explanatory variables correlated with α_i are dealt with under the assumption of fixed effects
 - Explanatory variables uncorrelated with α_i are dealt with under the assumption of random effects

1.2.1. Static Models

Instrumental Variables Estimators

- Model:

$$Y_{it} = x'_{1it}\beta_1 + x'_{2it}\beta_2 + w'_{1i}\gamma_1 + w'_{2i}\gamma_2 + \alpha_i + u_{it}$$

- x_{1it} : time-varying explanatory variables, uncorrelated with α_i
- x_{2it} : time-varying explanatory variables, correlated with α_i
- w_{1i} : time-invariant explanatory variables, uncorrelated with α_i
- w_{2i} : time-invariant explanatory variables, correlated with α_i

- Assumptions:

- All explanatory variables are strictly exogenous relative to u_{it}
- x_1 has a larger dimension than w_2
- x_1 and w_2 are correlated

1.2.1. Static Models

Instrumental Variables Estimators

- Estimation:

- 2SLS / GMM based on the following instruments for the explanatory variables correlated with α_i :
 - x_{2it} : $(x_{2it} - \bar{x}_{2i})$
 - w_{2i} : \bar{x}_{1i}

Stata
`xthtaylor Y X11 X12 ... X21 X22 ... W11 W12 ... W21 W22 ..., endog(X21 X22 ... W21 W22 ...)`

Static Panel Data Models - Summary

Estimator	Effects		Efficiency	Prediction	Exogeneity	Internal instruments for x_{it}
	Random	Fixed				
Pooled	x			x	Contemporaneous	
Between	x			x	Strict	Lags / Leads
Random Effects	x		x	x	Strict	Lags / Leads
Fixed Effects		x		Only if long panel	Strict	Lags / Leads
First-Differences		x			$E(x_{it}u_{it}) = E(x_{it}u_{i,t-1}) = E(x_{it}u_{i,t+1}) = 0$	Lags / Leads (except $t - 1$ and $t + 1$)

1.2.1. Static Models

Inference and Model Evaluation

Variance estimators – best alternatives:

- Cluster-robust
- Bootstrap

Hausman test:

- Random or fixed effects?

$H_0: E(\alpha_i x_{it}) = 0$ (RE and FE consistent, RE also efficient)

$H_1: E(\alpha_i x_{it}) \neq 0$ (FE consistent, RE inconsistent)

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' [V(\hat{\beta}_{FE}) - V(\hat{\beta}_{RE})]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi_k^2$$

Stata

(models must be estimated using standard estimators of the variance)

```
xtreg YX1 ... Xk, fe
```

```
estimates store ModelFE
```

```
xtreg YX1 ... Xk
```

```
estimates store ModelRE
```

```
hausman ModelFE ModelRE
```


Models with lagged dependent variables:

- Include $Y_{i,t-1}, Y_{i,t-2}, \dots$ as explanatory variables:

$$Y_{it} = \gamma_1 Y_{i,t-1} + \dots + \gamma_p Y_{i,t-p} + x'_{it} \beta + \alpha_i + u_{it}, t = (p+1), \dots, T$$

- All estimators for static panel data models are inconsistent

Example – $AR(1)$ model:

$$Y_{it} = \gamma_1 Y_{i,t-1} + \alpha_i + u_{it}$$

- This equation holds for all time periods, so:

$$Y_{i,t-1} = \gamma_1 Y_{i,t-2} + \alpha_i + u_{i,t-1}$$

$$Y_{i,t-2} = \gamma_1 Y_{i,t-3} + \alpha_i + u_{i,t-2}$$

etc.

- Random effects estimator:
 - Required assumption: $E(Y_{i,t-1}\alpha_i) = 0$
 - However, α_i is one of the components of $Y_{i,t-1}$, so $E(Y_{i,t-1}\alpha_i) \neq 0$
 - All other lags of Y_{it} are also correlated with α_i
- Fixed effects model:
 - Required assumption: $E[(Y_{i,t-1} - \bar{Y}_i)(u_{it} - \bar{u}_i)] = 0$
 - However, $u_{i,t-1}$ (included in \bar{u}_i) is one of the components of $Y_{i,t-1}$, so $E[(Y_{i,t-1} - \bar{Y}_i)(u_{it} - \bar{u}_i)] \neq 0$
 - All other lags of Y_{it} , which are included in \bar{Y}_i , are also correlated with the corresponding element of \bar{u}_i

- First-differences method:
 - Required assumption: $E(\Delta Y_{i,t-1} \Delta u_{it}) = 0$
 - However, $u_{i,t-1}$ (included in Δu_{it}) is one of the components of $Y_{i,t-1}$ (included in $\Delta Y_{i,t-1}$), so $E(\Delta Y_{i,t-1} \Delta u_{it}) \neq 0$
 - Unlike the previous estimators, $Y_{i,t-2}, Y_{i,t-3}, \dots$ are not correlated with Δu_{it} , provided that there is no autocorrelation
 - Solution: using $Y_{i,t-2}, Y_{i,t-3}, \dots$ (ou functions of these variables) as instruments for $\Delta Y_{i,t-1}$

1.2.2. Dynamic Models

Instrumental Variable Estimators

Base Dynamic Panel Data Model:

$$\Delta Y_{it} = \gamma_1 \Delta Y_{i,t-1} + \Delta x'_{it} \beta + \Delta u_{it}, t = 3, \dots, T$$

Assumption:

- u_{it} has no autocorrelation $\Rightarrow \Delta u_{it}$ has first-order autocorrelation:

$$\begin{aligned} Cov(\Delta u_{it}, \Delta u_{i,t-1}) &= Cov(u_{it} - u_{i,t-1}, u_{i,t-1} - u_{i,t-2}) \\ &= -Cov(u_{i,t-1}, u_{i,t-1}) \neq 0 \end{aligned}$$

Main estimators:

- Anderson-Hsiao (1981)
- Arellano-Bond (1991) – ‘Difference GMM’
- Blundell-Bond (1998) – ‘System GMM’

1.2.2. Dynamic Models

Instrumental Variable Estimators

Anderson-Hsiao (1981):

- Two alternative instruments:
 - $Y_{i,t-2}$

Stata
`ivregress gmm D.Y(DL.Y = L2.Y) D.X1 ... D.Xk`

- $\Delta Y_{i,t-2}$ (one observation is lost but in general seems to produce more efficient estimators)

Stata
`xtivreg D.Y(DL.Y = DL2.Y) D.X1 ... D.Xk`
or
`xtivreg Y(L.Y = L2.Y) X1 ... Xk, fd`

Arellano-Bond (1991):

- Proposed using all available lags of $Y_{i,t}$ as instruments:
 - $t = 3: Y_{i,1}$
 - $t = 4: Y_{i,2}, Y_{i,1}$
 - ...
 - $t = T: Y_{i,T-2}, \dots, Y_{i,2}, Y_{i,1}$
- Total number of instruments: $(T - 1)(T - 2)/2$
- It is possible to use only a subset of the available instruments
- More efficient than Anderson-Hsiao's (1981) estimators

Stata
`xtabond $YX_1 \dots X_k$, maxldep(#) twostep vce(robust)`

1.2.2. Dynamic Models

Instrumental Variable Estimators

Blundell-Bond (1998):

- Often, lags of $Y_{i,t}$ are not good instruments for $\Delta Y_{i,t}$
- Proposed adding the following instruments: $\Delta Y_{i,2}, \dots, \Delta Y_{i,T-1}$
- Total number of instruments: $\frac{(T-1)(T-2)}{2} + (T-2)$
- It is possible to use only a subset of the available instruments
- More efficient than Arellano-Bond's (1991) estimator
- Requires heavier assumptions than Arellano-Bond's (1991) estimator

Stata
`xtdpdsys Y X1 ... Xk, maxldep(#) twostep vce(robust)`

1.2.2. Dynamic Models

Tests for Instruments Validity

Most common tests:

- Hansen's J test of overidentifying restrictions

Stata

(after `xtabond` or `xtdpdsys`, with variance estimated in a standard way)
`estat sargan`

- Test for autocorrelation

- All estimators for dynamic panel data models assume first-order autocorrelation:

- $Cov(\Delta u_{it}, \Delta u_{i,t-l}) \neq 0$
- $Cov(\Delta u_{it}, \Delta u_{i,t-l}) = 0, l > 1$

Stata

(after `xtabond` or `xtdpdsys`, with variance estimated in a standard way)
`estat abond, artests(3)`