Labour Econometrics Seminar

Prof. François Rycx
Francois.Rycx@ulb.be
http://homepages.ulb.ac.be/~frycx/

Prof. Guillaume Vermeylen
Guillaume.Vermeylen@umons.ac.be

Prof. Mélanie Volral
Melanie.Volral@umons.ac.be

✓ **Aim of the seminar ?**

Investigate a research question in labour economics with econometric techniques.

Groups of two students.

Work of maximum 15 pages (without references and appendices).

✓ How to proceed ?

a) Define a **research question** **in labour economics**(potentially related to your PhD), after a careful analysis of the existing scientific literature.

b) Obtain the **data.**

c) Specify the **theoretical (mathematical) model &** corresponding **hypotheses.**

d) Specify the **statistical model & econometric techniques** to be used.

e) Estimate the **model & test** the relevant **hypotheses.**

f) Compare your results **with** the existing **literature.**

✓ Research question ?

a) Is there a gap in literature that could to be filled?

- ✓ **SBS-EM electronic library**: http://www.bibeco.ulb.ac.be/

  It contains notably :

  - Databases : **ECONLIT**, OCDE, EUROSTAT, IMF, etc.

    ECONLIT is a bibliographical database produced by the American Economic Association. It provides references (with abstracts) of books and articles coming from approximately 400 economic and management journals.

  - More than 5,000 electronic journals in economics and management.

- ✓ **Google Scholar**: http://scholar.google.com

- ✓ Time to publication in journals is quite long. To have access to more recent research have a look at **Working Papers** (some available through the SBS-EM library):

  - CEPR, ECB, IMF, IZA, NBB, NBER, etc.

b) What is the value added of your paper?

❖ **First time a** given research **question is investigated empirically**?

❖ **New theoretical model or hypotheses** being tested?

❖ **Better or different data** allowing to address potential econometric issues in current literature? Data for a different statistical unit (e.g. country/region/industry/company) or period?

❖ **Better or different statistical model or econometric technique**.

❖ **Other comparative advantage?**

✓ **How to obtain data?**

    ✓ **SBS-EM electronic library**: http://www.bibeco.ulb.ac.be/

    Some examples of interesting data:
- **AMADEUS / ORBIS** (pan-European financial information on more than 13 million private and public firms in 41 countries).
- **BELFIRST** (longitudinal financial information on more than 500,000 firms in Belgium).
- **Essential science indicators** (enable to examine international science research, e.g. to study research performance and trends).
- **IMF** (Trade and finance statistics).
- **International Bureau of Fiscal Documentation**.
- **OECD** (various data sets on: banks and insurances, trade, education, the labour market, public expenditures, taxes, national accounts, telecommunications and internet, **STAN** sector-level data for 32 countries in several years (includes e.g. output, value-added, profits, investments, employment, exports, imports, R&D, etc.).

    ✓ **Growing number of journals ask their authors to make their data publicly available** (e.g. American Economic Review, American Economic Journal: Applied Economics / Microeconomics / Macroeconomics, etc.)

    ✓ **Other sources: European Social Survey** (Micro EU data on socio-economic, political and health-related issues): www.europeansocialsurvey.org; **NBER data** (http://www.nber.org/data/), **DEVECONDATA** (for development studies), links to various free data sets: www.economicsnetwork.ac.uk/links/data_free.

✓ **Which econometric software ?**

I recommend to use **STATA** (but there is no obligation).

What is freely available?

Mons: STATA is freely available for students in the computer room.
ULB: Econometric software freely available (at least for ULB students).
      https://sisc.ulb.ac.be
Liège: ?.

How can you obtain STATA?

You can buy a Student STATA/IC license "monoposte semestrielle" (i.e. for six month) at the price of approximately 50 EUR (+ VAT). No transportation costs: the license will be sent to you by e-mail within one or two weeks. A manual on how to use STATA is available on demand.

Contact person: Delphine Grassot, RITME Informatique,
tel.: 02 2013210, e-mail : belgique@ritme.com.

✓ **Which econometric technique ?**

(Source: Gujarati et al., 2017).

a) It depends notably on the type of data used:

Cross-sectional data are data on one or more variables collected **at a given point in time for several statistical units (e.g. workers, households, sectors, regions, countries)**. Example: *Structure of Earnings Survey* (provides information on a large number of workers at a given point in time regarding their wages, individual and job characteristics).

Time series data are data on one or more variables collected **at different points in time** [e.g. on a daily (weather reports), weekly (money supply figures), monthly (the unemployment rate) or annual basis (GDP)] **for a single statistical unit** (e.g. a firm, an industry, a region, a country).

Pooled data are data on one or more variables collected **at different points in time (e.g. years) for several statistical units (e.g. firms).** They thus combine elements of both time series and cross-section data.

Multiple cross-section data are a special type of pooled data in which the *cross-sectional units* (e.g. households, firms) are *completely different at each point in time* (e.g. year). You cannot track information for a given cross-sectional unit over time.

Panel data are a special type of pooled data in which *the same cross-sectional units* (e.g. households, firms) *are surveyed over time*. You can track information for a given cross-sectional unit over time (if for all periods: **balanced panel** data, if for a restricted number of periods: **unbalanced panel** data).

b) A general reminder (in the context of the CLNRM):

❖ **Multicollinearity**

It refers to a situation where there is either an exact or approximately exact linear relationship among the explanatory variables (X's).

If there is perfect collinearity among the X's, OLS regression coefficients are undetermined and their standard errors are infinite.

If collinearity is high but not perfect (which is often the case), OLS estimators are still BLUE (best linear unbiased) but the standard errors of the regression coefficients tend to be large, i.e. **the population values of the coefficients cannot be estimated precisely**.

An estimator is said to be BLUE if:

1) It is **linear**, i.e. it is a linear function of a random variable.
2) It is **unbiased**, i.e. its average or expected value is equal to its true population value.
3) It has minimum variance in the class of all linear unbiased estimators. An unbiased estimator with the least variance is known as an **efficient** estimator.

❖ **Heteroscedasticity**

It refers to a situation where the conditional variance of the error term (and of the dependent variable) is not constant.

For instance, if you regress individual hourly wages on firm size, you find that bigger firms pay higher wages and that the variability of wages increases with firm size. The conditional variance of the error term (and of the dependent variable) thus increases with firm size.

Graph. 1: Heteroscedastic versus homoscedastic residuals

Heteroskedastic Residuals

Homoskedastic Residuals
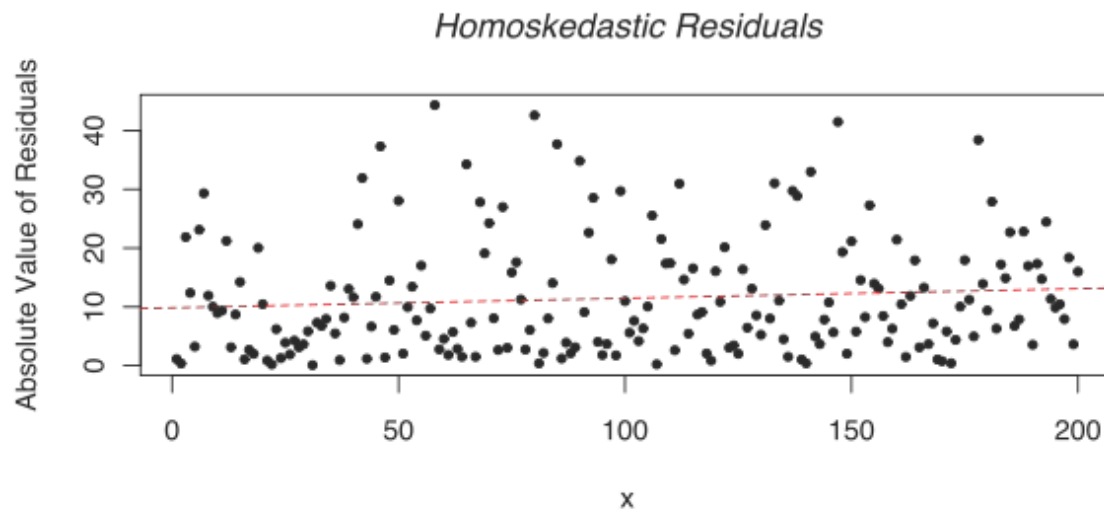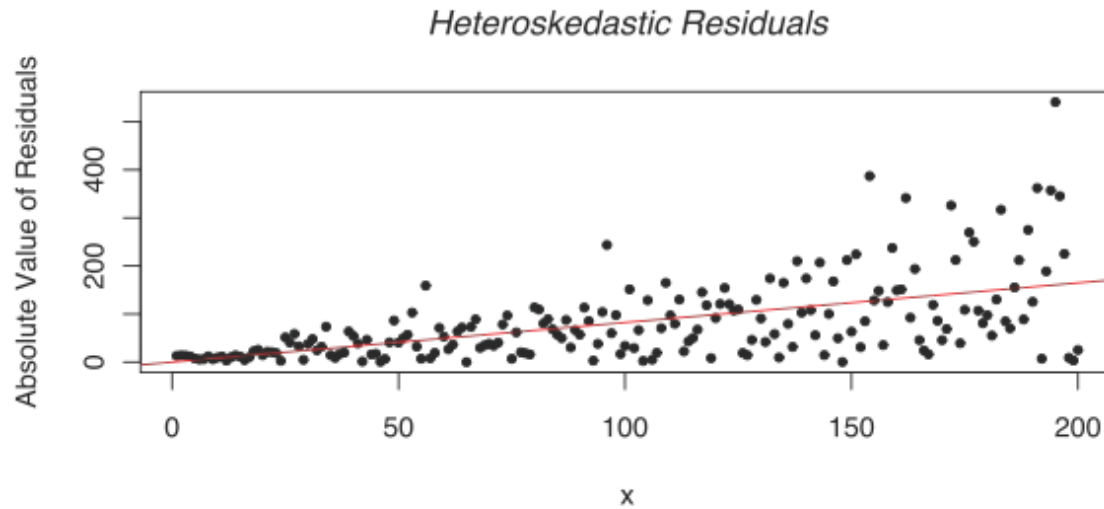
Heteroscedasticity is a standard issue, especially with cross-sectional and panel data.

If there is heteroscedasticity, OLS estimators are still unbiased and consistent (i.e. they converge to the true population value as sample size tends to infinity) but they have no longer minimum variance. They **are not efficient** anymore (i.e. they are not BLUE). Moreover, given that in practice it is not possible to determine whether the variance of the estimator is over- or under-estimated, standard **statistical inference becomes impossible**. Hence, the usual t, F and χ^2 statistics become invalid.

How to detect heteroscedasticity?

Various tests, e.g. Park test, White test, etc.

How to solve this issue?

If the functional form of heteroscedasticity is known (e.g. that the conditional variance of the error term is proportional to firm size), you can use Weighted Least Squares (i.e. apply OLS to the transformed data).

If not, but sample size is sufficiently large, you could rely on White heteroscedasticity consistent standard errors.
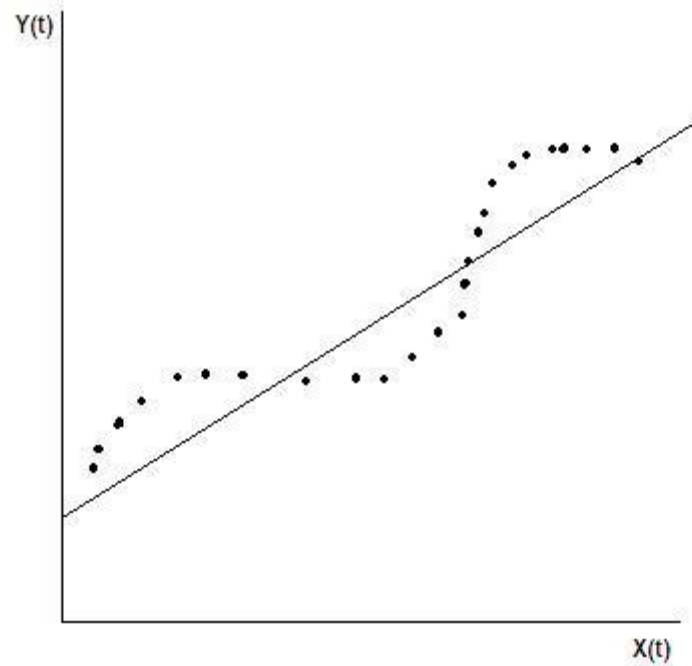
❖ Serial correlation

It refers to a situation where the error term is autocorrelated, i.e. where the error term of an observation (at time t) is influenced by the error term of any observation (at time $t-j$).

This is a standard issue, especially with time-series and panel data.

The reasons for the occurrence of serial correlation are numerous: inertia of most economic time series (e.g. GDP, employment, unemployment inflation are characterized by cycles), omission of an important regressor, use of an incorrect functional form (e.g. linear instead of quadratic), etc.

Fig. 2: Serial correlation due to inertia of GDP
(positive correlation between successive residuals)

Note: Y = annual GDP, X = annual consumption, t = time.

When using cross-sectional data, autocorrelated error terms (i.e. spacial autocorrelation) are much less likely. For instance, using cross-sectional household data (to examine consumption behavior in relation to wages), there is no prior reason to believe that the error term pertaining to one household will be correlated to the error term of another household. If this would be the case, given that there is generally no logic order in cross-sectional observations, the problem can be solved simply by randomly changing this order.

If there is serial correlation, **OLS estimators** are still unbiased and consistent (i.e. they converge to the true population value as sample size tends to infinity) but they have no longer minimum variance. They **are not efficient** anymore (i.e. they are not BLUE). Moreover, given that in practice it is not possible to determine whether the variance of the estimator is over- or under-estimated, standard **statistical inference becomes impossible**. Hence, the usual t, F and χ^2 statistics become invalid.

How to detect serial correlation?

Various tests, e.g. Durbin Watson test, Breusch-Godfrey LM-test, etc.

How to solve this issue?

Does the autocorrelation issue result from a misspecification of the model (e.g. omitted variable bias, incorrect functional form)? If yes, try to fix it.

If no, there are two possibilities:

i) Transform original model (i.e. variables) to suppress autocorrelation (as in the presence of heteroscedasticity this boils down to apply Generalized Least Squares).

ii) If sample size is large enough, still use OLS but correct standard errors for autocorrelation (and heteroscedasticity) by the Newey-West HAC procedure. HAC stands for heteroscedasticity and autocorrelation consistent standard errors.

❖ **Normality of the error term**

It is under the normality assumption that we are able to establish that the OLS estimators of the intercept and slope coefficients follow a normal distribution and that the OLS estimator of the variance of the error term $(\hat{\sigma}^2)$ is related to the chi-square distribution.

So, it is under the normality assumption that statistical inference can be done, i.e. that hypotheses can be tested using the usual *t* and *F* statistics.

Fig. 3: Normally distributed error term
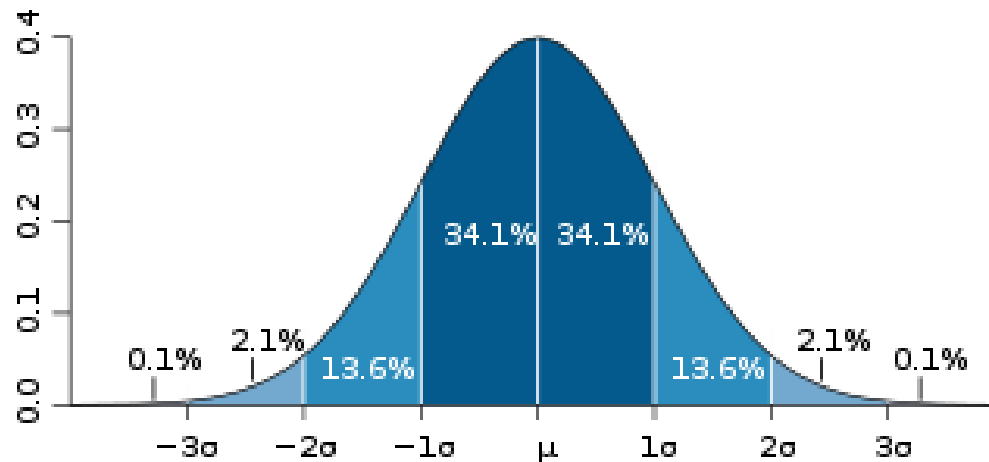
Notes:

μ (the mean) of the error term is equal to zero.

68% of the distribution is in the interval $[\mu - \sigma, \mu + \sigma]$,

95% in the interval $[\mu - 2\sigma, \mu + 2\sigma]$, and

99,7% in the interval $[\mu - 3\sigma, \mu + 3\sigma]$.

What happens if the error terms are not normally distributed?

We can rely on the following extension of the central limit theorem (Theil, 1978):

"If the disturbances (u_i) are independently and identically distributed with zero mean and (constant) variance σ^2 and if the explanatory variables are constant in repeated samples, the OLS coefficient estimators are asymptotically normally distributed with means equal to the corresponding β's.

Therefore, **usual test procedures** (t and F tests) are **still valid** asymptotically, i.e. **in large samples**.

Reminder:
Central limit theorem (which provides a theoretical justification for the normality assumption): "If there are a large number of independent and identically distributed random variables, then, with a few exceptions, the distribution of their sum tends to a normal distribution as the number of such variables increases to infinity".

In small samples, it is **very important to test for the normality assumption** to make sure that standard statistical inference (i.e. hypotheses testing and prediction) is permitted.

How to examine the normality assumption?

Various possibilities, e.g. histogram of residuals, Anderson-Darling or Jarque-Bera tests.

What if sample size is small and the normality assumption not satisfied?

You may rely on non-parametric, or distribution free, estimation methods. Also called robust estimation techniques.

c) Functional forms of regression models: some examples

How to interpret regression coefficients?

✓ **The linear model** (i.e. linear in the parameters and in the variables)

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$\beta_2 = \frac{dY}{dX} = \frac{absolute\ change\ in\ Y}{absolute\ change\ in\ X}$$

Example:

$$Ip_t = -1026,5 + 0,3016\ GDP_t$$
$$s.e. = \quad (257,6) \quad (0.0399)$$

with *Ip* (private investments) and *GDP* in billions of EUR.

If GDP increases by 1 unit (1 billion EUR), on average private investments increase by 301,6 millions EUR.

✓ **The log-linear model** (also called log-log or double-log)

$$\ln Y_t = \beta_1 + \beta_2 \ln X_t + u_t$$

$$\beta_2 = \frac{d(\ln Y)}{d(\ln X)} = \frac{dY/Y}{dX/X} = \frac{dY}{dX}\frac{X}{Y}$$

$$\Rightarrow \beta_2 = \frac{relative\ change\ in\ Y}{relative\ change\ in\ X} = elasticity\ of\ Y\ w.r.t.\ X$$

Example:

$$\ln EDUR_t = -9{,}69 + 1{,}91\ \ln CE_t$$
$$s.e. = (0.43)\quad(0.05)$$

with *EDUR* expenditures on durable goods and *CE* consumption expenditures.

If CE increases by 1 percent, on average EDUR increase by 1,91 percent.

✓ The semi-logarithmic model type I: log-lin

$$\ln Y_t = \beta_1 + \beta_2 X + u_t$$

$$\beta_2 = \frac{d \ln Y}{dX} = \frac{dY/Y}{dX} = \frac{relative\ change\ in\ Y}{absolute\ change\ in\ X}$$

$$\beta_2 * 100 = semi-elasticity\ of\ Y\ w.r.t.\ X$$

(i.e. mean % change in Y following a unit change in X)

$$\beta_2 * \overline{X} = elasticity\ of\ Y\ w.r.t.\ X\ (at\ mean\ value\ of\ X)$$

Example:

$$\ln Y_t = 7,7890 + 0,00743\ t$$
$$s.e. = (0.0023)\ \ (0.00017)$$

with Y quarterly data (for 1993:1, 1998:4) on consumption expenditures and t a linear trend (t = 1, 2, …).

Mean annual growth of cons exp = 2,97% (4 * 0,743%). Exp(7,7890) = 2413,9 = value of cons exp at t = 0, i.e. during last quarter of 1992.

✓ **The semi-logarithmic model type II: lin-log**

$$Y_t = \beta_1 + \beta_2 \ln X + u_t$$

$$\beta_2 = \frac{dY}{d \ln X} = \frac{dY}{dX/X} = \frac{absolute\ change\ in\ Y}{relative\ change\ in\ X}$$

$$\beta_2/100 = semi-elasticity\ of\ Y\ w.r.t.\ X$$

(i.e. mean absolute change in Y following a 1% change in X)

$$\beta_2/\overline{Y} = elasticity\ of\ Y\ w.r.t.\ X\ (at\ mean\ value\ of\ Y)$$

Example:

$$Y_t = -1283{,}9 + 257{,}3 \ln X_t$$
$$s.e. = (-4{,}4)\ \ (5{,}7)$$

with Y and X respectively annual household consumption and revenue in EUR. If the household revenue increases by 1%, on average consumption increases by 2,57 EUR.

d) More specific issues:

I. Time series:

Regression analysis with time series is based on the assumption that the latter are stationary. Put differently, time series have to be stationary to be able to perform standard statistical inference (based on usual t and F tests).

A stochastic process (i.e. a collection of random variables ordered in time) is said to be weakly stationary if its mean and variance are constant over time and the value of the auto-covariance (or auto-correlation) depends only on the distance or lag between the two time periods and not the actual time at which the auto-covariance (or auto-correlation) is calculated.

A stationary time series thus tends to return to its mean (called mean reversion) and fluctuations around this mean (measured by its variance) will have a broadly constant amplitude.

In contrast, a time series will be considered as non stationary if its mean and/or variance change systematically over time.

Fig. 4: Stationarity of times series

Stationary Time Series

ADF = −6.128

Non-stationary Time Series

ADF = −2.0251
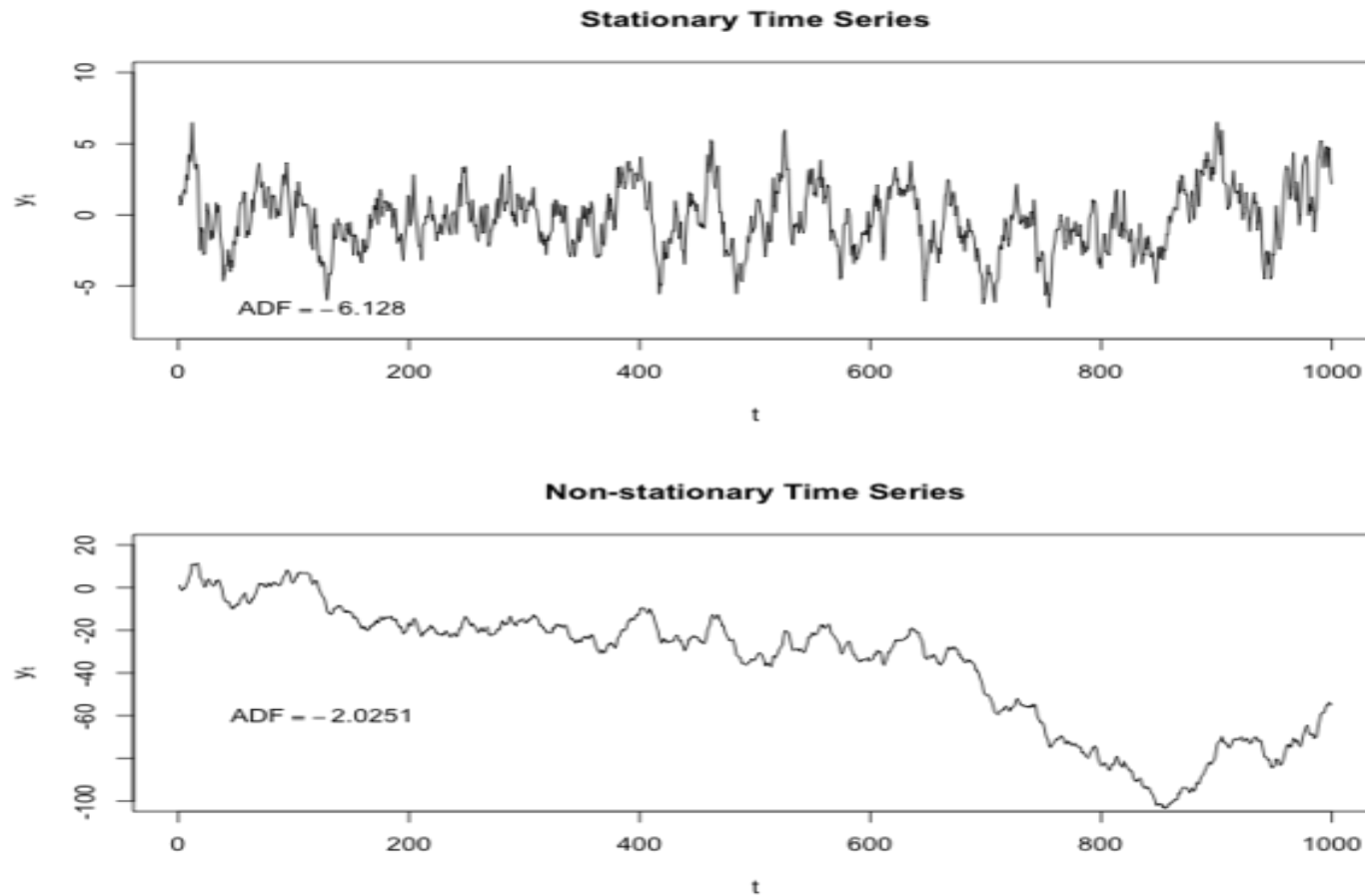
In practice, most economic time series are non stationary, e.g. GDP, unemployment, investments, consumption, exports, exchange rates, etc.

What if you run a regression with non stationary time series?

If you regress <u>two completely unrelated</u> (and non stationary) time series on each other, you may find out that the regression coefficient is highly statistically significant. From this you may be tempted to conclude that there is a significant relationship between the variables, whereas in fact there should be none.

This is called the **phenomenon of spurious** or **non-sense regression** (first discovered by Yule, 1926).

The spurious relationship derives from the fact that non stationary time series have a stochastic and/or deterministic trend that may be correlated.

How to test for stationarity?

Various possibilities

✓ Graphical analysis (Are the mean and variances of a time series stable over time?).

✓ Correlogram of a time series, i.e. graph of autocorrelation at various lags. Intuition: for non stationary time series, the auto-correlation starts from a very high value (i.e. for small lags) and tends only very slowly towards zero (i.e. for bigger lags).

✓ Unit root tests, e.g. Dickey-Fuller (DF), augmented Dickey-Fuller (ADF) and Phillips-Perron tests.

How to avoid the spurious regression problem?

✓ **Transforming non-stationary time series**

To avoid the spurious regression problem that may appear when regressing non stationary time series on each other, we have to transform non stationary time series to make them stationary.

The type of transformation depends on whether the time series are difference stationary (DSP) or trend stationary (TSP) processes.

DSP: you have a time series that is integrated of order d, noted I(d). To make it stationary, you have to differentiate it d times.

TSP: you have a time series that is stationary around a trend line. To make it stationary (i.e. to obtain a detrended time series), you have to regress it on time and to save the residuals which will be stationary.

Most macro-economic time series are DSP (integrated of order 1 or 2) rather than TSP.

✓ **Cointegration and the error correction model (ECM)**

Cointegration refers to a situation where the linear combination of two or more $I(1)$ time series is stationary.

Suppose, CM_t (consumption) and DI_t (Disposable income) are $I(1)$ and that the residuals ($\hat{u}_t$) of the OLS regression of CM_t on DI_t are stationary. In that case, CM_t and DI_t are cointegrated (because $\hat{u}_t = CM_t - \hat{\beta}_1 - \hat{\beta}_2\, DI_t$).. Moreover, given the stationarity of the residuals, regression results won't be spurious and usual statistical inference will be relevant.

The Engle-Granger (EG) and Augmented Engle-Granger (AEG) tests can notably be used to find out if two or more series are cointegrated.

From an economic point of view, cointegration of two (or more) time series suggests that there is a long-run relationship between them.

In order to estimate and to reconcile both short- and long-run relationships between cointegrated variables, one may rely on the ECM developed by Engle and Granger.

II. Panel data

A panel data set consists of observations on the same cross-sectional units (e.g. individuals, households, firms, regions, countries) over several time periods.

✓ **Substantial advantages of these data** (over cross-sectional and time series data):

- Estimation techniques for panel data can take the heterogeneity of cross-sectional units explicitly into account. Put differently, these techniques can control for individual (or cross-sectional) fixed unobserved heterogeneity.

Suppose you want to estimate the impact of education on individual wages. Having panel data, you will be able to estimate this relationship *ceteris paribus*, that is, after **controlling for observed individual heterogeneity** (i.e. for individual variables included in the data set, e.g. sex, experience, occupation, industry) **and unobserved individual time-invariant heterogeneity** (i.e. for variables that are not included in the data set and that do not change over time, e.g. innate ability, family background). See below.

- They increase the sample size considerably and thus the precision of the estimates (i.e. more information, more variability, less collinearity, more degrees of freedom and more efficiency).

- They are better suited to study the "dynamics of change" (e.g. wage and labour mobility) and enable to study more complicated behavioral models (e.g. economies of scale, technological progress).

✓ **Problems with these data**:

Since such data involve both cross-section and time dimensions, problems notably associated to cross-sectional data (e.g. heteroscedasticity) and time series data (e.g. autocorrelation) have to be addressed. Another potential issue is the cross-correlation in individual units at the same point in time.

✓ **Panel data estimation techniques**:

$$y_{it} = \lambda_0 + X_{it}\beta + \alpha_i + u_{it}$$

with

y_{it} the dependent variable observed for individual i at time t (e.g. the hourly wage)

X_{it} the vector of explanatory variables (e.g. education, sex, experience, occupation, industry).

α_i the unobserved time-invariant individual effect (e.g. ability, motivation, historical or institutional factors).

u_{it} the error term.

Pooled OLS or panel data techniques?

To test for the existence of individual unobserved fixed effects, i.e. whether intercepts vary significantly among individuals, we can use a **restricted F test**:

$$F = \frac{\left(R_{NC}^2 - R_C^2\right)/m}{\left(1 - R_{NC}^2\right)/(n-k)}$$

where

- R_C^2 is the determination coefficient of the constraint regression model $\left(y_{it} = \lambda_0 + X_{it}\beta + u_{it}\right)$, in which we impose that all intercepts are identical and equal to λ_0.
- R_{NC}^2 is the determination coefficient of the non constraint regression model $\left(y_{it} = \lambda_0 + \sum_{i-0}^{N} \delta_i I^i (individual\,units)_{it} + X_{it}\beta + u_{it}\right)$, in which we allow intercepts to vary among individuals.
- m the number of linear constraints, k the number of parameters in the non constraint regression, and n the number of observations.

H_0: all intercepts are equal to λ_0.

H_1: at least one cross-sectional unit has an intercept that is significantly different from λ_0.

Decision rule:

If the *F* statistic is significant at conventional probability levels, we reject H_0, i.e. we reject the null hypothesis that there are no individual fixed effects $\Rightarrow$ you should rely on panel data estimation techniques.

Most prominent techniques used in the presence of individual fixed effects:

- The **fixed effects model** (FEM):

  In the FEM, the intercept in the regression model is allowed to differ among individual (or cross-sectional) units so as to control for fixed unobserved individual heterogeneity. In practice, to attain this goal, dummy variables can be included for the different individual units (general rule: if you have *n* units, (*n*-1) dummies should be inserted; otherwise, there is a perfect collinearity problem). In that case, the FEM is known as the least-squares dummy variables (**LSDV**) model.

The inclusion of dummies for the cross-sectional units is equivalent to the estimation of a within differentiated model (a model where the mean of each variable has been subtracted from the initial values). This model is known as the **FEM**.

$$FEM \equiv y_{it} - \overline{y_i} = \left(X_{it} - \overline{X}_i\right)\beta + \left(u_{it} - \overline{u}_i\right)$$

where

$$\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \;,\; \overline{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$$

and

$$\overline{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$$

One may also estimate a first difference (**FD**) model (a model in which all variables are expressed in first differences, i.e. by their absolute change between two consecutive periods) or a long difference (**LD**) model (a model in which all variables are expressed as long differences, i.e. by their absolute change between two points in time separated by at least 2 periods)

$$FD \equiv y_{i,t} - y_{i,t-1} = \left(X_{it} - X_{i,t-1}\right)\beta + \left(u_{i,t} - u_{i,t-1}\right)$$

$$LD \equiv y_{i,t} - y_{i,t-2} = \left(X_{i,t} - X_{i,t-2}\right)\beta + \left(u_{i,t} - u_{i,t-2}\right)$$

Drawback of these models:

- The LSDV model consumes many degrees of freedom (difference between the number of observations and parameters to be estimated) when the number of cross-sectional units *N* is large. Indeed, we have to introduce *N-1* dummy variables (in addition to the intercept).
- With the FD model, all observations of the first period are lost. The loss of observations is even more critical with the LD model.
- When applying the FD, LD or FEM to an unbalanced panel, information on all cross-sectional units observed at a single point in time are lost.

- The **random effects model** (REM).

  An alternative to the FEM is the REM. In the REM it is assumed that the intercept of an individual unit (i.e. the unobserved fixed effect α_i) is a random draw from a much larger population with a constant mean value. The individual intercept is then expressed as a deviation from this constant mean value.

Fixed effects *versus* random effects model?

It depends on the correlation between the unobserved individual fixed effects (α_i) and the regressors (X_{it}).

If α_i and X_{it} are *not* correlated, the REM should be preferred (they provide more efficient estimates).

If α_i and X_{it} are correlated, the FEM should be preferred (REM estimates are biased, while FEM estimates are not).

Why might α_i and X_{it} be correlated?

Suppose we want to estimate the impact of education on individual wages with panel data. If α_i corresponds to innate ability or family background, when estimating the wage equation including α_i, it is very likely that α_i will be correlated with education given that innate ability and family background are often major determinants of education.

How to decide between both models?

You can use the Hausman (asymptotic) test.

H_0: α_i and X_{it} are *not* correlated.
H_1: α_i and X_{it} are correlated.

If the Chi-square test statistic is significant at conventional probability levels, you reject H_0, i.e. you reject the null hypothesis that α_i and X_{it} are *not* correlated $\Rightarrow$ you should rely on the FEM.

III. Qualitative response models:

Suppose we want to:

i) **Study the labour force participation decision of adults.**
The dependent variable can take two values:
 - Y=1 if the person is in the labour force.
 - Y=0 if the person is not.
 ⇒ Binary or dichotomous dependent variable.

ii) **Examine the determinants of votes during US elections.**
If only two political parties (Democratic and Republican),
the dependent variable can take two values:
 - Y=1 if the vote is for a Democratic candidate.
 - Y=0 if the vote is for a Republican candidate.
 ⇒ Binary or dichotomous dependent variable.

These are some examples of response models in which the dependent variable is qualitative.

Difference between regressions models with quantitative and qualitative dependent variables (Y's)?

i) Objective when Y is quantitative = estimate the expected or mean value of Y, given the values taken by the regressors $\Rightarrow$ estimate: $E(Y/X_{1i}, X_{2i}, ..., X_{ki})$.

ii) Objective when Y is qualitative = find the probability of something happening, such as voting for a Democratic candidate, owning a house or belonging to a union. Hence, qualitative response model are often known as *probability models*.

How to estimate qualitative response models?

A) The linear probability model (LPM)

The simplest possible binary regression model is the linear probability model (**LPM**). In this model the binary response variable is regressed on the relevant explanatory variables by using the standard OLS methodology.

Example:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

where:

Y = 1 if the family owns a house, and 0 otherwise,

X = the household revenue.

We thus estimate by OLS the expected value of Y_i given X_i, i.e. $E(Y_i / X_i)$, which can be interpreted as the conditional probability that something is happening given X_i, i.e. $\Pr(Y_i = 1/X_i)$

The principal weakness of the LPM is that it assumes that the probability of something happening increases linearly with the level of the regressor.

This very restrictive assumption can be avoided if we use the logit and probit models.

B) The logit and probit models

In the **logit model** the dependent variable is the log of the odds ratio, which is a linear function of the regressors.

$$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = \beta_1 + \beta_2 X_i + u_i$$

where

- L_i is called the logit.
- $P_i/(1 - P_i)$ is the odds ratio in favor of owning a house, i.e. the ratio of the probability that a family will own a house to the probability that it will not own a house. Thus, if $P_i = 0.8$, it means that odds are 4 to 1 in favor of the family owning a house.

- The probability function P_i is the logistic distribution: $P_i = \dfrac{1}{1 + e^{-(\beta_1 + \beta_2 X_i)}}$).

Although L is linear in X, the probabilities themselves are not. This property is in contrast with LPM.

If the data are available in grouped form, we can first estimate the value of P_i (i.e. the probability to own a house in each group e.g. in each class of revenue), by taking the fraction of people in each group (e.g. in each class of revenue) that owns a house. Next, we compute the odds ratio for each group (provided that the number of observations per group is large enough). Finally, we estimate the parameters of the logit model using weighted least squares (WLS) in order to control for the heteroscedastic nature of the error term

If the data are available at the individual (or micro) level, we cannot estimate the value of the odds ratio. The point is that at the individual level the probability to own a house P_i is equal to 0 or 1 (either the person owns a house or s/he doesn't). Indeed, note that:

i) If the person owns a house: $L_i = \ln\left(\dfrac{1}{1-1}\right) = \ln\left(\dfrac{1}{0}\right) = \ln(\infty) = \infty$.

ii) If the person doesn't own a house: $L_i = \ln\left(\dfrac{0}{1-0}\right) = \ln\left(\dfrac{0}{1}\right) = \ln(0)$ which

is undefined.

OLS or WLS cannot be used, non-linear estimation procedures are required (i.e. the maximum likelihood estimator).

The standard R^2 is not very interesting for qualitative response models. Various alternative statistics can be used to estimate the quality of the fit (called pseudo-R^2's): R^2 of McFadden, Count R^2, etc.

To test the null hypothesis that all the slope coefficients are simultaneously equal to zero, we rely on the likelihood ratio (LR) statistic (under the null it follows a Chi-square distribution with df equal to the number of explanatory variables). It is equivalent to the F test used for the standard linear regression model.

The **probit model** differs from the logit model in that the probability function P_i follows a normal distribution. From a practical point of view, both models give similar results (in terms of marginal effects, see below). Researchers have often chosen for the logit model because of its comparative mathematical simplicity. However, nowadays, this is not an issue anymore given the availability of sophisticated econometric softwares.

How to interpret logit and probit estimates?

To compute the *change in the probability of an event occurring as the result of a unit change of a regressor*, all other things being equal, we must compute **marginal effects**.

Intuition:

In both the **logit** and **probit model**:
$$\beta_2 = \frac{\Delta\left[\ln\left(\dfrac{P_i}{1-P_i}\right)\right]}{\Delta X_i}$$

The slope coefficient β_2 measures the variation in the log of the odds following a unit change in the variable X, all other things being equal.

Note that Amemiya has shown that: $\beta_{\log it} = 1,6 * \beta_{probit}$. This is due to the fact that the logistic distribution has slightly fatter tails.

Now, if we want to estimate **the marginal effect**, i.e. the rate of change in the probability of an event occurring as a result of a unit change in X, we must compute the following expression:

For the **logit model**: $\left| \dfrac{\Delta P_i}{\Delta X_i} = \beta_2 \ P_i\left(1 - P_i\right) \right|$

with $P_i = \dfrac{1}{1 + e^{-(\beta_1 + \beta_2 X_i)}}$ (the cumulative logistic distribution function)

For the **probit model**: $\left| \dfrac{\Delta P_i}{\Delta X_i} = f\left(\beta_1 + \beta_2 X_i\right) \right|$

with $f()$ the density function of a standard normal variable.

In both the logit and probit models, the computation of marginal effects involves all the regressors.

C) Some other models:

✓ **Tobit model:** developed for censored samples, i.e. samples in which information on the dependent variable is only available for some observations.

Suppose we want to estimate the impact of socio-economic variables on the amount spent to buy a house. Our sample will be censored given that information regarding the amount spent to buy a house is obviously not available for those who didn't buy a house.

✓ **Ordinal logit and probit models**: to be used when the dependent variable is qualitative and contains more than two ordinal (i.e. ranked or ordered) outcomes.

Exemple: survey reponses such as: i) "strongly agree", "somewhat agree", "somewhat disagree", or "strongly disagree". These reponses are generally coded as 1 "strongly agree", 2 "somewhat agree", 3 "somewhat disagree" and 4 "strongly disagree".

This is an ordinal scale because there is a clear ranking among the categories (i.e. logical order) but we cannot say that 4 "strongly disagree" is four times 1 "strongly agree" or 2 "somewhat disagree" is twice 1 "strongly agree".

✓ **Multinomial logit and probit models**: to be used when the dependent variable is qualitative and contains more than two nominal (i.e unordered or unranked) outcomes (i.e. outcomes without logical order).

Examples: i) choice of transportation mode to work, e.g. car, bus, bicycle, train or motorbike; ii) labour market status, e.g. employed, unemployed, inactive.

✓ **Duration models**. To consider questions such as: What determines the duration of unemployment spells, of a strike, or of a marathon race?

✓ **Additional mini course taught by Mélanie Volral**

Date TBD (at UMONS)

How to get started with Stata?

This session is devoted to students who do not yet know how to use Stata. For those who are already familiar with Stata, this session is therefore not required.

For more details: see web site of doctoral school

http://www.edtgestion.hec.ulg.ac.be/

✓ Some useful textbooks

Baltagi, B.H. (2002), *Econometrics*, 3rd ed., Springer, NY.

Baum, C.F. (2006), *An Introduction to Modern Econometrics Using Stata*, Stata Press, College Station, Texas.

Greene, W.H. (2005), *Econométrie*, 5ème éd., Pearson Education, Paris (also available in English as *Econometric Analysis*).

Gujarati, D.N., Porter, D. and Gunasekar, S. (2017), *Basic Econometrics*, McGraw Hill Higher Education, NY.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton University Press, Princeton.

Kennedy, P. (1998), *A Guide to Econométrics*, 4th ed., MIT Press, Cambridge (Mass.).

Murray, P.M. (2006), *Econometrics: A Modern Introduction*, Pearson Education, Boston.

Rabe-Hesketh, S. and Skrondal, A. (2008), *Multilevel and Longitudinal Modeling Using Stata*, Stata Press, College Station, Texas.

Sevestre, P. (2002), *Econométrie des données de panels*, Dunod, Paris.

Wooldridge, J.M. (2002), *Econometric Analysis of Cross-section and Panel Data*, MIT Press, Cambridge (Mass.).

✓ **Practical details**

a) **Groups of 2 students**.

b) Please **provide today**:

✓ the **composition of your group** (names, forenames, e-mail addresses, university affiliation, PhD supervisor)

&

✓ the **subject** you will be working on

c) Papers should be **at most 15 pages long** (without references and appendices).

d) **Contents and lay out are both very important**. Make sure: i) your research question is clearly explained, ii) the literature review is up to date, iii) the econometric model and results are clearly reported, iv) all references are correctly cited (1: 1 equivalence), etc. **Plagiarism is an act of fraud**.

e) **Office hours: by appointment only (via Teams)** (francois.rycx@ulb.be, guillaume.vermeylen@umons.ac.be **or** melanie.volral@umons.ac.be).

f) **Deadline:** seminars should be send by e-mail to both Mélanie Volral & François Rycx on the **3rd of June at 12 am**.

g) **Defense:** June (exact date will be fixed later on):
   – Presentation: maximum 10 minutes.
   – Discussion: 10 - 15 minutes (at least one question for each member of the group).