INTRODUCTION TO BASIC LINEAR REGRESSION MODEL

13 September 2011

Yogyakarta, Indonesia

Cosimo Beverelli

(World Trade Organization)

LINEAR REGRESSION MODEL

- In general, regression models estimate the effect of changing one variable over another one.
- In particular, a linear regression model estimates how much the dependent (endogenous) variable *y* changes when the independent (exogenous) variable *x* changes of one unit.
- Starting from an economic model and/or an economic intuition, the purpose of such a model is to test a theory and/or to estimate a relationship.

SINGLE VARIABLE MODEL

- Having *n* observations on *x* and *y*, a simple linear regression model has the following functional form:

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad i=1,2,\ldots.n$$

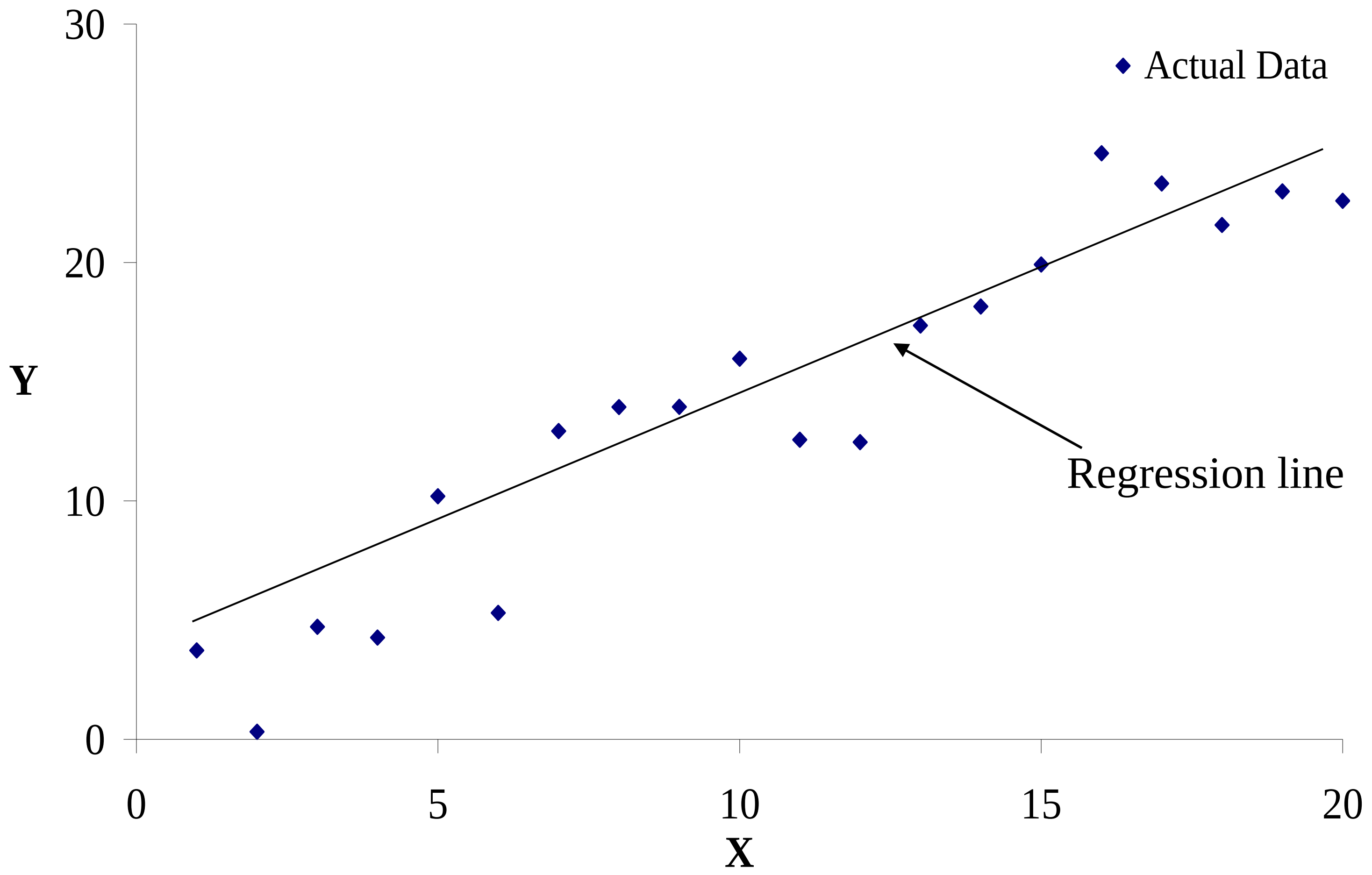
Where:

- β_0 is the **constant**.
- β_1 is the **coefficient** and describes the direction and strength of the relationship between variable *x* and *y.*
- u_i is the **error term**, which contains unobserved factors and measuring errors.

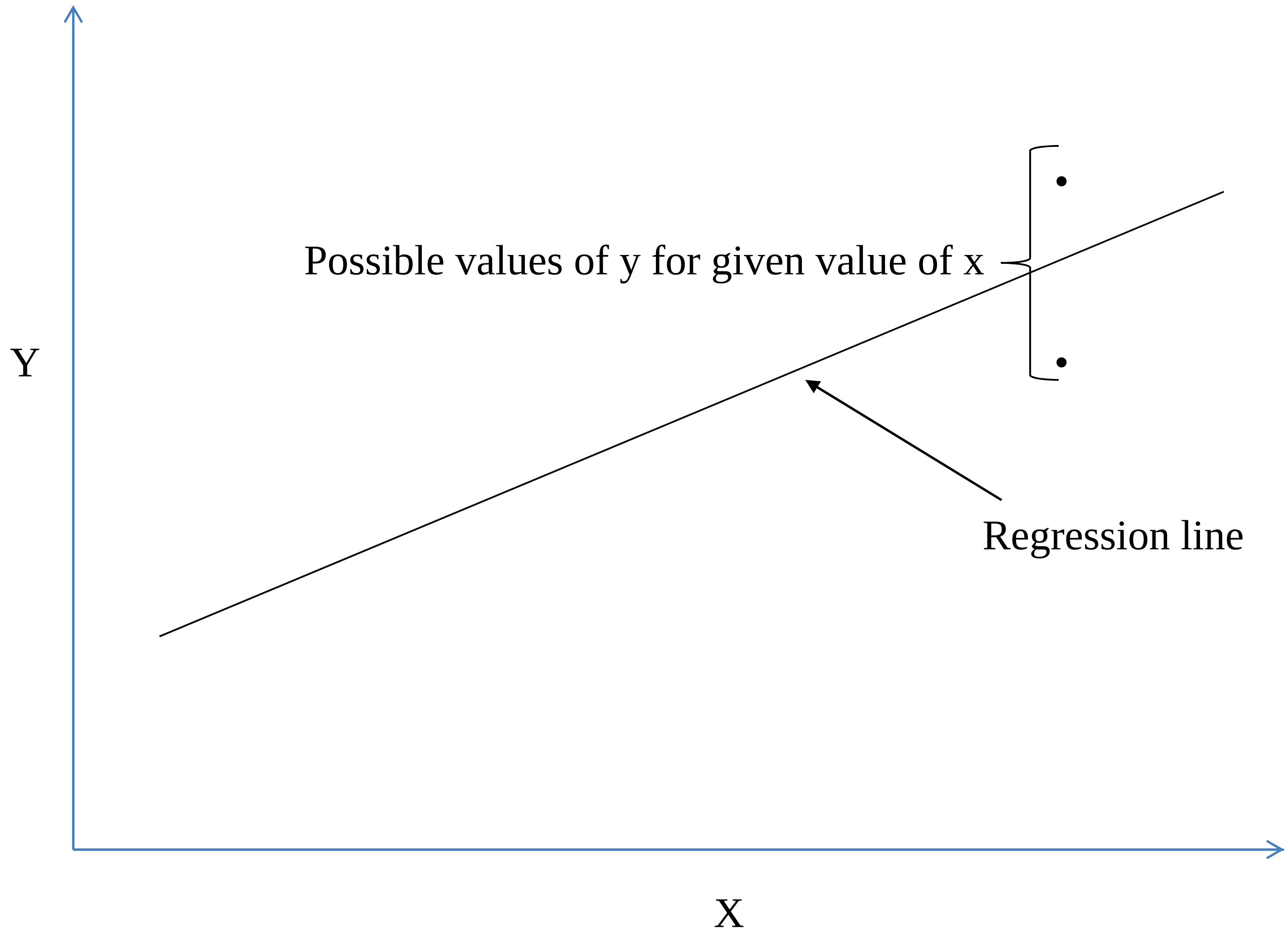
SPECIFICATION OF THE RELATIONSHIP

- DETERMINISTIC RELATIONSHIP: $y = f(x)$
  - It is possible to DETERMINE EXACTLY the values of the dependent (endogenous) variable for different values of the independent (exogenous) variable.
- STOCHASTIC RELATIONSHIP: $y = f(x) + u$
  - the values of y for different values of x cannot be determined exactly but they can be DESCRIBED PROBABILISTICALLY
  - Example: y=f(x)+u, where u= $\begin{cases} +6 \text{ with probability } 1/2 \\ -6 \text{ with probability } 1/2 \end{cases}$

DETERMINISTIC RELATIONSHIP (u=0)



STOCHASTIC RELATIONSHIP (u≠0)



WHY $u \neq 0$?

- Unpredictable element of randomness in human responses: humans are not machines!

- Effect of a large range of omitted variables: is it possible to identify/quantify everything?

- Measurement error in y: even if it is possible, it could be hard to quantify!

ASSUMPTIONS ON u

- In order to get estimates of the coefficients β_n , we need some assumptions on the error term.

1. E(u_i) = 0 for all *i.*
2. Var (u_i) = σ^2 for all *i.*
3. u_i and u_j are independent for all $i \neq j$.
4. u_i and x_j are independent for all i and j; that is, the distribution of error term does not depend on the value of the independent (exogenous) variable.
5. u_i are normally distributed for all *i*.

ASSUMPTIONS ON *u* - CONSEQUENCES

- Assumptions 1, 2, 3 and 5 together imply errors to be normally distributed:

$$u_i \sim i.i.d.(0, \sigma^2)$$

- Because of Assumption 1 (E(u_i) = 0 for all *i*) we can rewrite:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad as \quad E(y_i) = \beta_0 + \beta_1 x_i$$

- Through the population regression function (on the right) we estimate the coefficients, and substituting in the linear regression function (on the left), we obtain the sample regression function.

$$\boldsymbol{y_i = \widehat{\beta_0} + \widehat{\beta_1}\, x_i + u_i}$$

THE OLS METHOD

- The least squares method minimizes the sum of squared deviations of regression values from the observed values , that is, the residual sum of squares:

$$\text{Min} \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

- With the OLS method we find coefficients that equals:

$$argmin_{\beta_0\beta_1} \sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)^2,$$

- OLS gives:

$$\widehat{\beta_1} = \frac{\sum x_i y_i}{\sum {x_i}^2} \quad \text{and} \quad \widehat{\beta_0} = y_i \text{-} \widehat{\beta_1} x_i$$

OLS: PROPERTIES

- An estimator is BLUE (best linear unbiased estimator) if:

1. It is a linear function of the random variables.
2. It is unbiased.
3. Has the minimum variance within the class of linear and unbiased estimators.

- Proof of UNBIASEDNESS (if $y_n = \beta_n x_n + \varepsilon_n$):

$$\left.\begin{aligned} \hat{\beta} &= \frac{\sum y_n x_n}{\sum x_n^2} = \frac{\sum (\beta x_n + \varepsilon_n) x_n}{\sum x_n^2} \\ &= \frac{\beta \sum x_n^2 + \sum \varepsilon_n x_n}{\sum x_n^2} \\ &= \beta + \frac{\sum \varepsilon_n x_n}{\sum x_n^2} \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} E(\hat{\beta}) &= \beta + E\left[\frac{\sum \varepsilon_n x_n}{\sum x_n^2}\right] \\ &= \beta + \frac{\mathrm{Cov}(\varepsilon_n x_n)}{\mathrm{Var}(x_n)} \\ &= \beta \end{aligned}\right.$$

equals zero!

OLS: ASYMPTOTIC PROPERTIES

- CONSISTENCY: the probability that OLS estimate is different from the true value of the parameter is null when the sample size tends to infinite.

$$lim_{n \to \infty} prob\{|\widehat{\beta}_i - \beta_i| > \delta\} = 0 \ \ \forall \delta > 0$$

- NORMALITY (HINT FOR IN-DEPTH EXAMINATIONS): if the Gauss-Markov assumptions holds, the OLS estimators are:

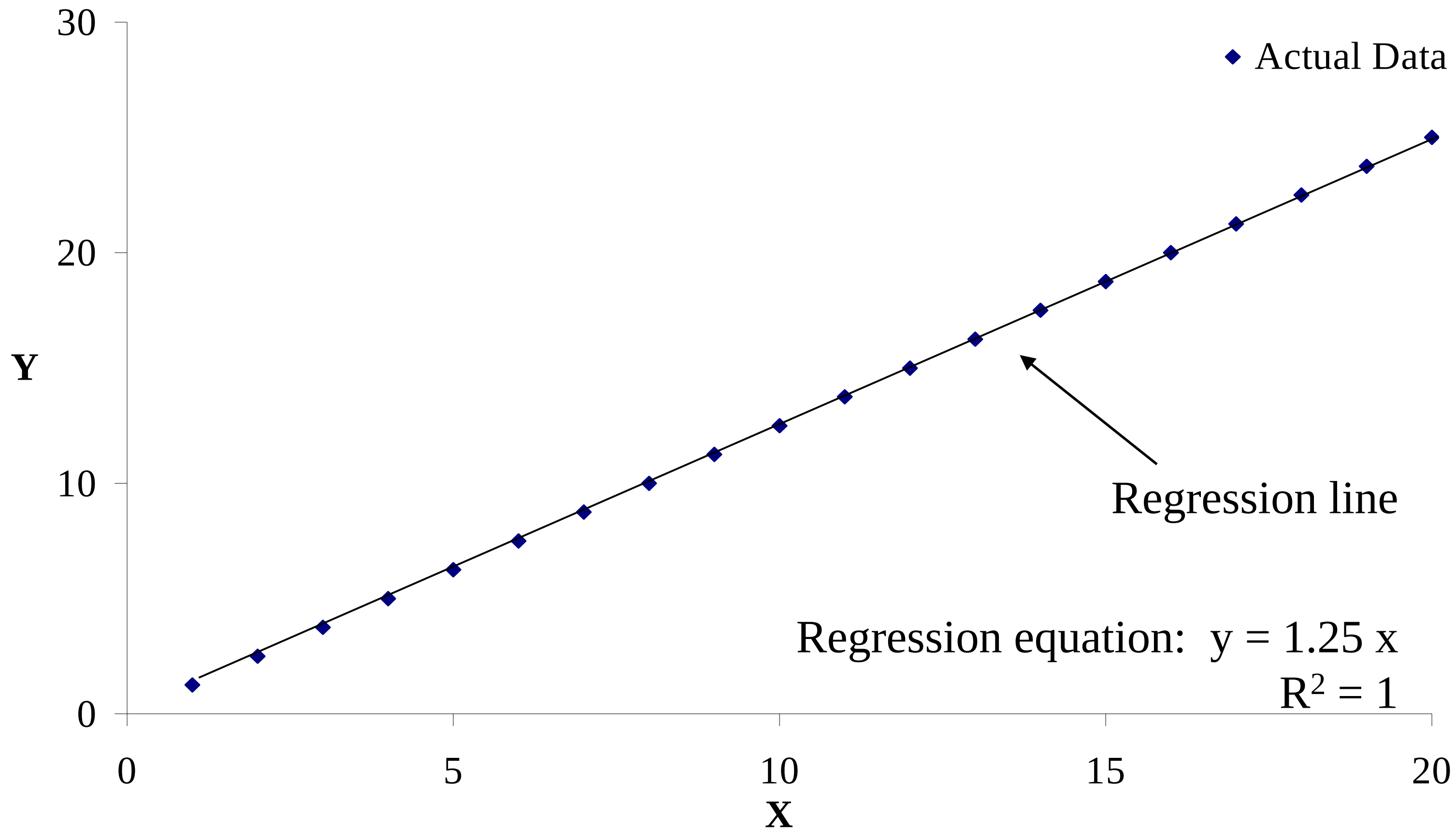
$$\sqrt{N}\left(\widehat{\beta}_i - \beta_i\right) \to N\left(0, \frac{\sigma^2}{\sum x^2}\right)$$

VISUAL PRESENTATION (DETERMINISTIC RELATIONSHIP)

- When we estimate the model, R^2 measures how much of the variation of y (in %) is due to a one unit variation of x.

VISUAL PRESENTATION

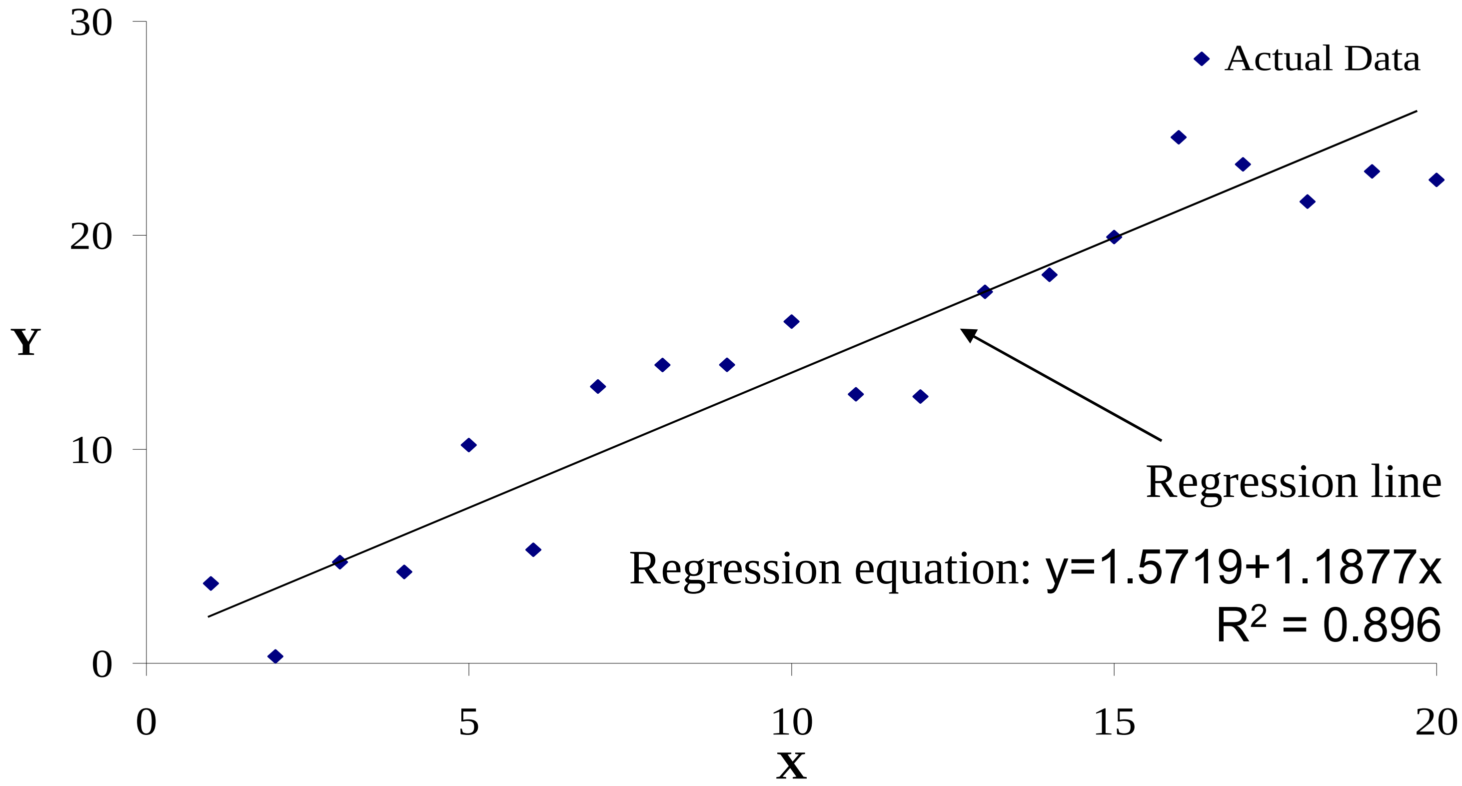
PERFECT LINEAR RELATIONSHIP BETWEEN x AND y



VISUAL PRESENTATION

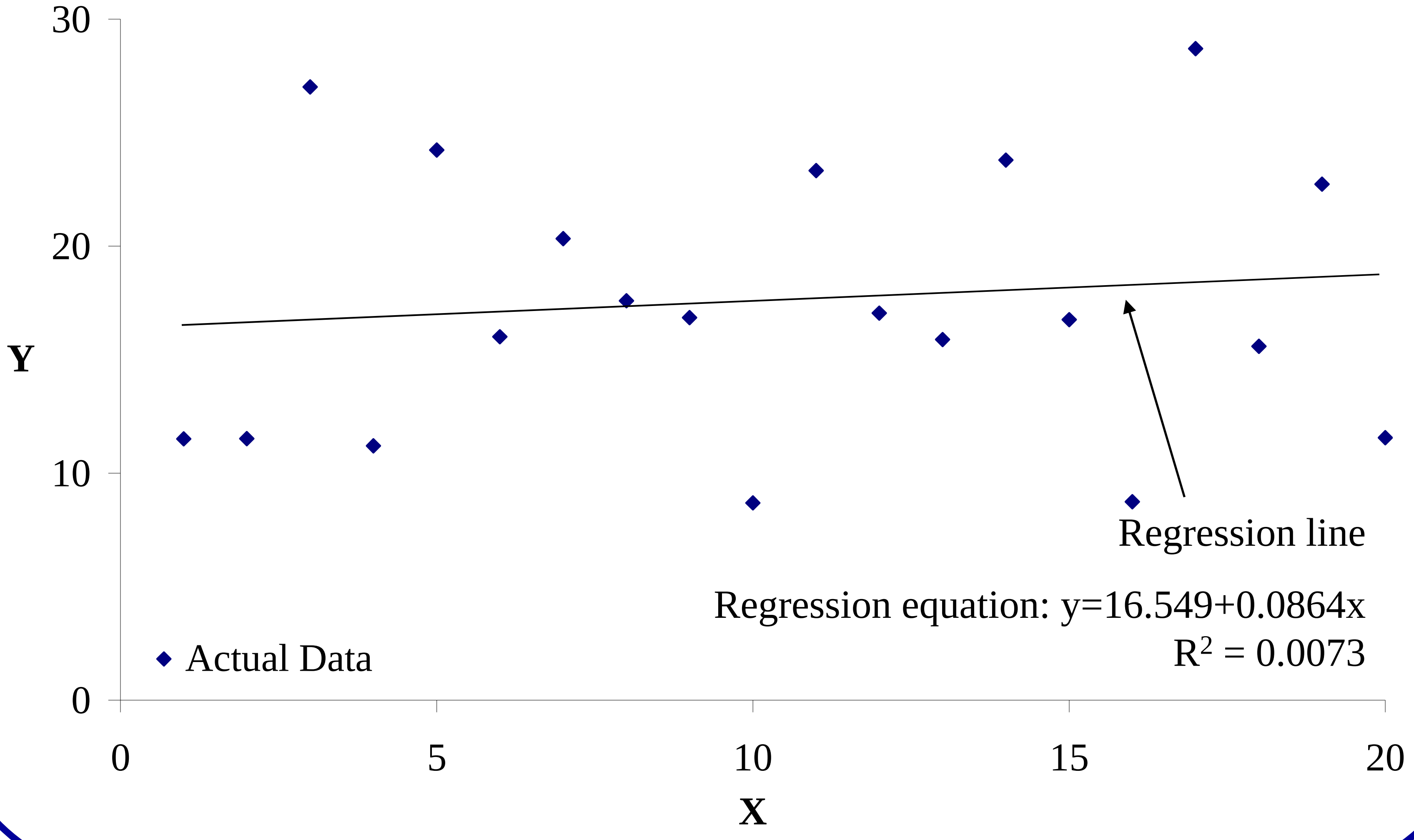
STRONG LINEAR RELATIONSHIP BETWEEN x AND Y

- The regression line minimizes the sum of squared residuals between the regression line itself and the data points (OLS)



VISUAL PRESENTATION

NO OBSERVABLE RELATIONSHIP BETWEEN x AND Y



HOW TO INTERPRET?

- The higher the value of $R^2 \in [0,1]$, the higher the relevance of a variation in the independent (exogenous) variable *x* in explaining a variation of the dependent (endogenous) variable *y*.
- The coefficient β_1 captures the effect of the independent variable on the dependent one, and, in the single variable case, it also represents the slope of the regression line.

- NOTICE THAT in the last slide the coefficient is statistically not different from zero; that is, it is not statistically significant.

MULTIVARIATE MODEL

- In the general case, the regression model has more than one independent (exogenous) variable:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

- In this case the coefficient β_1 measures the partial effect of x_1 on the dependent (endogenous) variable *y*, after controlling for all other independent (exogenous) variables x_2 and x_3.
- NOTICE THAT in the multivariate model the coefficient does not represent the slope of the regression line, but, exactly as in the single variable model, a coefficient statistically not different from zero it is not statistically significant.

MULTIVARIATE MODEL

- Assuming that assumptions on the error terms hold (slide 8), we obtain the following sample regression function:

$$y_i = \widehat{\beta_0} + \widehat{\beta_1}\, x_{1i} + \widehat{\beta_2}\, x_{2i} + \widehat{\beta_3} x_{3i} + u_i$$

- In the next two slides, we illustrate a numerical example in order to show how to interpret the results of a standard linear regression model (OLS ostimate).

AN EXAMPLE

- The aim of our analysis is to study the determinants of women labour market participation:

- Dependent (endogenous) variable:
  - y = women labour market participation (hours worked)
- Independent (exogenous) variables:
  - x_1 = education
  - x_2 = number of children
  - x_3 = age

```
reg part educ child age
```

AN EXAMPLE

P-value of the model. It indicates the reliability of x's to predict y.
p-value < 0.05 → statistically significant relationship.

R-square shows the amount of variance of y explained by x.

Number of obs = 24590
F (---,---) = ---
Prob > F = 0.003
$R^2 = 0.54$
Root MSE = ---

The t-values test the hypothesis that the coefficient is different from 0. To reject this, you need a t-value greater than 1.96 (at 0.05 confidence). You can get the t-values by dividing the coefficient by its standard error. The higher t value, the higher the significance of the variable

Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05. In this case, education, children significant; age not significant.

PARTICIPATION	coef.	robust. s. e.	t	t > lpl	95% conf. int.
Education	3.47	---	8.32	0.000	--- ---
Children	-1.12	---	-2.93	0.001	--- ---
Age	0.28	---	0.77	0.268	--- ---
Constant	4.18	---	10.56	0.000	--- ---

Coefficient >(<)0 ⟹ positive (negative) effect of x on y

Part = 4.18 Cons + 3.47 Educ – 1.12 Child

PANEL DATA ANALYSIS
FIXED EFFECTS – RANDOM EFFECTS

A PANEL DATASET

- Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behaviour of entities (countries, firms, individuals) are observed across time.

PROS

- Allows control for variables you cannot observe or measure like cultural factors or variables that change over time but not across entities.
- Allows to include variables at different levels of analysis (individuals, neighbourhoods, cities…).

CONS

- Difficulties in designing panel surveys (data collection and data management issues).
- Cross-country dependency in case of macro panels.

FIXED EFFECTS / RANDOM EFFECTS

- "…the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the independent (exogenous) variables in the model"
  [William H. Greene, Econometric Analysis, 6th ed., 2007]
- If you have reason to believe that individual time invariant effects are not correlated with the time varying independent (exogenous) variables then random effect model is consistent and efficient.
- If you have reason to believe that individual time invariant effects are correlated with x_{it} then fixed effect method is consistent.

BASIC FIXED EFFECTS MODEL

- Having *n* observations on *x* and *y* for T periods, a basic fixed effects model has the following functional form:

$$y_{it} = x`_{it}\beta + \alpha_i + \varepsilon_{it} \qquad i=1,2,\ldots.n \qquad t=1,2,\ldots.T$$

Where:

- $x`_{it}$ can contain observable variables that changes across i only, across t only or across i and t.
- α_i is the unknown intercept (the individual effect) for each entity (so there are *n* entity-specific intercepts).
- ε_{it} are the idiosyncratic errors and change both across entity (i) and time (t).

BASIC FIXED EFFECTS MODEL

- Given:

$$\boldsymbol{y_{it}} = \boldsymbol{x`_{it}\,\beta} + \boldsymbol{\alpha_i} + \boldsymbol{\varepsilon_{it}} \qquad \boldsymbol{i=1,2,\ldots.n} \qquad \boldsymbol{t=1,2,\ldots.T}$$

- We can estimate: $\hat{\beta}_1^{\,within}$ by applying the within transformation that gets rid of individual unobserved effects $\boldsymbol{\alpha_i}$:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)`\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

- Or, alternatively, we can insert n-1 dummy variables for each entity, then estimate through the least square method obtaining: $\hat{\beta}_1^{\,LSDV}$.

BASIC RANDOM EFFECTS MODEL

- Having *n* observations on *x* and *y* for T periods, a basic random effects model has the following functional form:

$$\boldsymbol{y_{it} = x`_{it}\beta + \alpha_i + \varepsilon_{it}} \qquad \boldsymbol{i=1,2,\ldots.n} \qquad \boldsymbol{t=1,2,\ldots.T}$$

Where (difference with respect to basic FE model):

- $\boldsymbol{\alpha_i}$ is assumed to be uncorrelated with x_{it}'
- We assume $\alpha_i \sim i.i.d.(0, {\sigma_\alpha}^2)$ to be the between-entities component of the error term
- $\boldsymbol{\varepsilon_{it}}$ is the within-entity component of errors

BASIC RANDOM EFFECTS MODEL

- RE model:

$$y_{it} = x`_{it}\beta + \overbrace{\alpha_i + \varepsilon_{it}}^{u_{it}} \qquad i=1,2,\ldots.n \qquad t=1,2,\ldots.T$$

- $\varepsilon_{it} \sim i.i.d.(0, {\sigma_\varepsilon}^2) \quad AND \quad \alpha_{it} \sim i.i.d.(0, {\sigma_\alpha}^2)$

  so u_{it} is an equicorrelated error term.

- If $\boldsymbol{u_{it}}$ is not correlated with the regressors then we can consistently estimate ${\hat{\beta}_1}^{GLS-RE}$ by applying GLS to the model (GLS is more efficient than OLS because $var(\boldsymbol{u_{it}}) \neq (\sigma^2 I)$).

FIXED EFFECTS OR RANDOM EFFECTS?

- To decide whether to use Fixed Effects or Random Effects, you need to test if the errors are correlated or not with the exogenous variables.
- The standard test is the Hausman Test: null hypothesis is that the preferred model is random effects vs. the alternative the fixed effects.
- To run this test, you need to run both a Fixed Effects and a Random Effects:

```
xtreg y x1, fe
estimates store fixed
xtreg y x1, re
estimates store random
hausman fixed random
```

Command to run panel datasets

Fixed Effects option

Random Effects option

THE HAUSMAN TEST

```
hausman fixed random
```

	Coefficient (b) fixed	Coefficient (B) random	Difference (b – B)	---
x1	---	---	---	---

- We are testing the null hypothesis that difference in coefficients is not systematic.
- If Prob>chi2 > 0.05 then we use Fixed Effects
- If Prob>chi2 < 0.05 then we use Random Effects

FIRM-LEVEL ANALYSIS

- The use of firm level analysis implies several difficulties; they can be summarized as follows:

1. **Data collection**: many information are needed for an econometric analysis at firm-level. There are several dataset but some of them are (partially) incomplete. For example, on the geographic location or the sector specificity of firms. [Aitken, Hanson and Harrison, 1997]
2. **Nature of the variables**: in firm-level analysis, the dependent variable is often a dummy (for example, export Y/N) or a probability (for example, export intensity), therefore the linear regression method is not the most suitable procedure (as it can give estimates negative or greater than one for a probability, for example).

- In the next sessions, we illustrate suitable procedures for firm-level analysis with limited dependent variables