CEU
Universidad
San Pablo

Regression methods

Chapter 1

Carlos Rivero Rodríguez, UCM
Madrid, July 9-13, 2007

A brief historical framework

- We can find antecedents of the regression methods on the Gauss (1777-1855) and Laplace`s (1749-1827) astronomical and physical models.
- The term “regression ” firstly appears in the Galton´s (1822-1911) biological works.
  - Galton analyzed the relationship between the size of different types of seeds and the size of the corresponding plant after one year growing.
  - "Regression Towards Mediocrity in Hereditary Stature," Journal of the Anthropological Institute, 15:246-263 (1886). Galton analyzed the heigths of 928 individuals and the mean heigths of their respective parentages, 205 in number.
- At present, regression methods are used in many scientific areas
  - Economy, biology, medicine, engineering, metheorology, sociology, psicology, etc

Regression and reverssion

- Galton works:

  “children´s heights tended to ***reverse*** to the average heigth of the population, rather than diverting from it”

- Difference between *data analysis* and *statistical inference*

- The relationship between some variables will be expressed in probabilistic terms. Consequences derived from this relationship will be also expressed in probabilistic terms.

- “The user of regression analysis attempts to discern the relationship between a dependent variable and one or more independent variables. The relation will not be a functional relation, nor can a ***cause-and-effect relationship*** necessarely be inferred”

Some applications and examples

- In Economy:
  - Relationship between salary (incomes) and education (Mincer, “Schooling, Experience and Earnings”, 1974).
  - Relationship between yield on an stock and the yield of the stock market index (Sharpe).
- In Sociology
  - Relationship between delinquency rate, public security expendidure and number of polices.
- In Pshicology
  - Characteristics of individuals that are inclinated to the violency.
  - Characteristics of the indivuduals that tend to buy a specific product.
- In Medicine
  - Relationship between time of treatment and the ammount and type of medicines.
- In Industry
  - Optimal conditions to improve yield of an industry process, in relation with temperature, preasure and reaction time of the process.

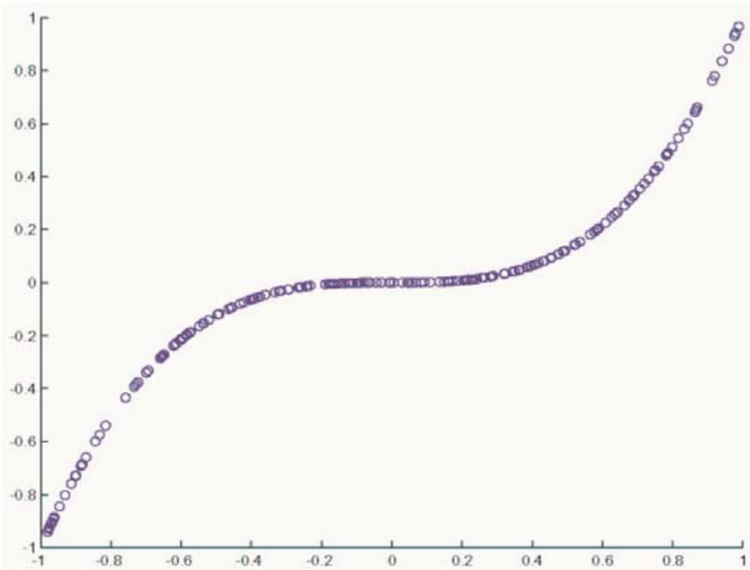
Specification of a regression model

- Objetive: "*Analyze the relationship between a response variable Y and k predictors X_1, X_2, …, X_k*"

- **Uniequational static model**
  - Simple model (with one regresor X)
    - Linear model (chapters 2-3)
    - Non linear model (chapter 11)
    - Non parametric model (chapter 13)
    - Discrete response model
      - Y is a binary variable (chapter 9)
      - Y has more than two levels
  - Multiple model (with k regressors X_1, X_2, …, X_k)
    - Linear model (chapters 4-8, 10, 12)
    - Non linear model (chapter 11)
    - Non parametric model
    - Discrete response model
      - Y is a binary variable (chapter 9)
      - Y has more than two levels
- Multiequational dymamic model

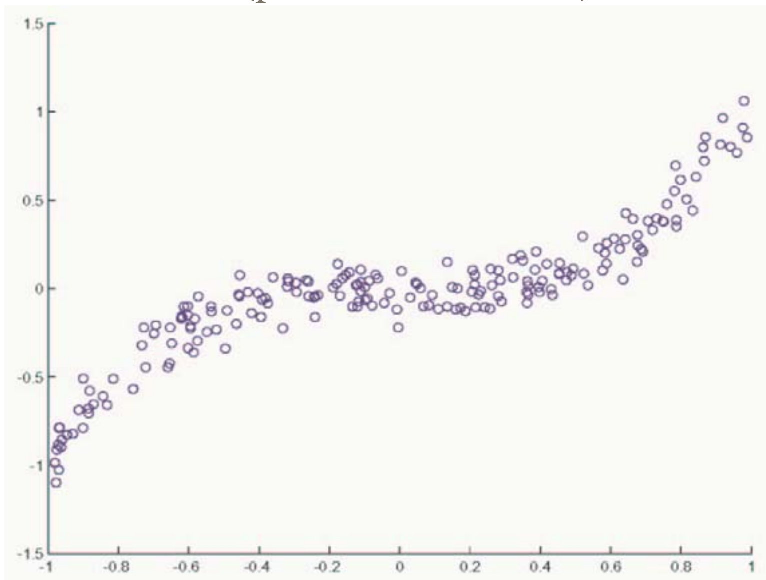
- We are interested in:
  - *“explain the variable Y as a function of X”* or equivalently,
  - *“analyze the relationship between Y and X”*

Exact dependency



Non-exact dependency

(perturbations effect)



Objetive: detect the underling function (relationship) that generates the data

Data will allow us to find this underling relationship between Y (dependent variable) and X (independent variable), after deleting the perturbations effect

CEU Universidad San Pablo

Specification of a regression model

- Simple model
  - Linear model $$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1,\ldots,N$$
  - Non-linear model $$Y_i = g(\beta_1,\ldots,\beta_r, X_i) + u_i, \quad i = 1,\ldots,N$$ $$Y_i = e^{\beta_0 + \beta_1 X_i^{\beta_3}} + u_i$$
- Multiple model
  - Linear model $$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,N$$
  - Non linear model $$Y_i = g(\beta_1,\ldots,\beta_r, X_{1i},\ldots,X_{ki}) + u_i, \quad i = 1,\ldots,N$$ $$Y_i = \log(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}) + u_i$$

Y = response scalar variable (***observable***)

$X_1,\ldots,X_k$ = regressors or independent scalar variables (***observable***)

$\beta_1, \ldots,\beta_k$ = scalar parameters (***unknown***)

u_i = random scalar perturbations (***non observables***): ADDITIVE perturbations

CEU Universidad San Pablo

Organization of the regression analysis

- **Detect the problem**.
  - Define the variables that we will relate between
  - Find data. Data sources
- **Specify an adequate model**
  - Prior knowledgement. Specify a functional relationship: linear or non linear
  - No prior knowledgement. Use a non-parametric model to suggest a parametric functional relationship
- **Estimate** the parameters $\beta_1, \ldots, \beta_k$ from the data
- **Validate** the specificated model from current data or from new or historic data
- **Make predictions and interpret** the real situation that we had modeled

CEU Universidad San Pablo

Simple linear regression model

Chapter 2

Carlos Rivero Rodríguez, UCM
Madrid, July 9-13, 2007

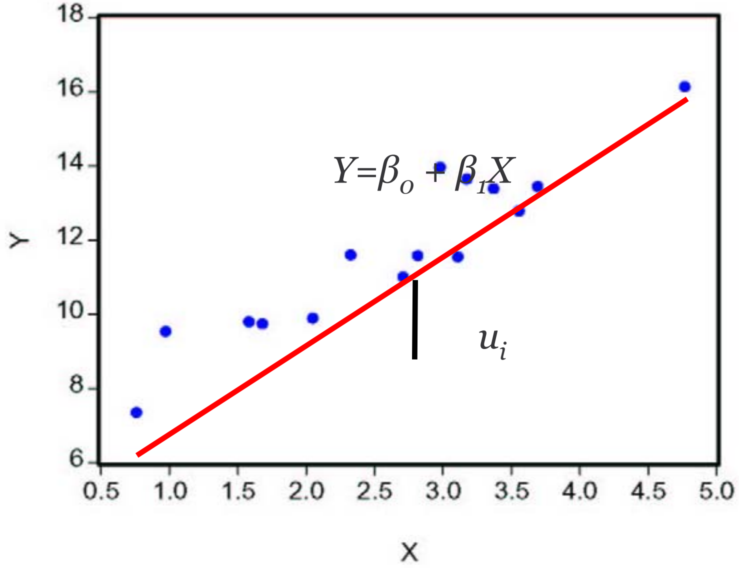
Introduction

- Relationship between a response or dependent variable Y and a single regressor or independent variable X.
- Linear regression does not mean that the relationship between Y and X can be represented as a straight line.
- Linear regression means that the model is linear in the parameters

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$Y_i = \beta_0 + \beta_1 X_i^2 + u_i$$

u_i indicates the non-exact relationship between Y and X



We estimate with past to predict future

Regression describes the relationship between Y and X

Interpretation

– Parameter β_1 represents the mean increasing of Y when X increases one unit.

– Parameter β_1 represents the mean differential increasing of Y as X increases.

$$\boxed{\beta_1 = \frac{\partial Y}{\partial X}}$$

– Parameter β_o represents the mean fixed effect which is not due to X

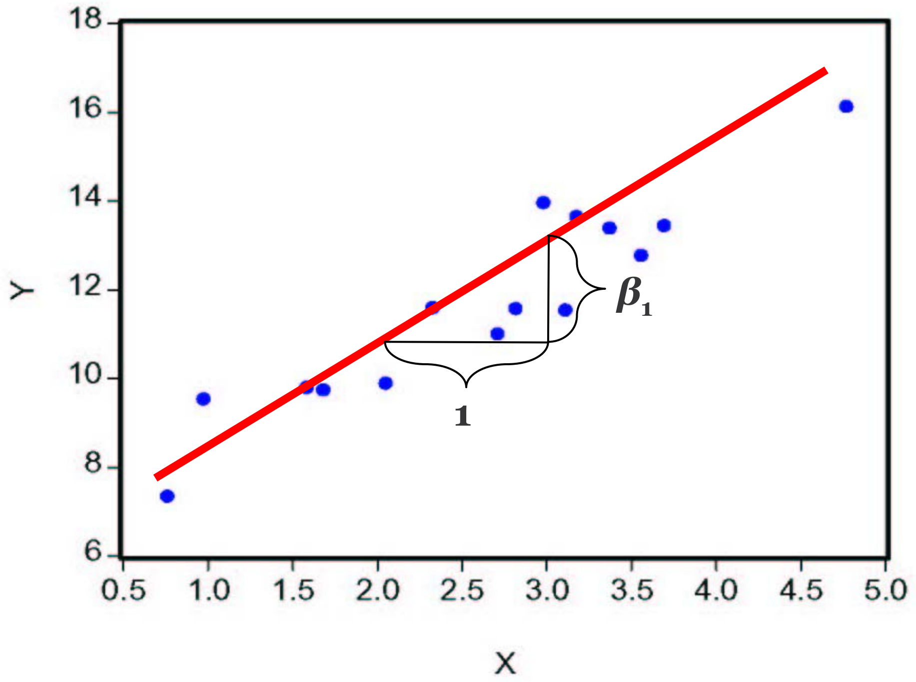
Example:
Once β_o and β_1 have been estimated:

Y=700+100X

Y=income (x1000 €);
X=education (years);

β_1=100 represents the mean increasing of the salary when the indivudual`s education increases one year.

β_o=700 represents the mean part of the income which does not depend on the education level.

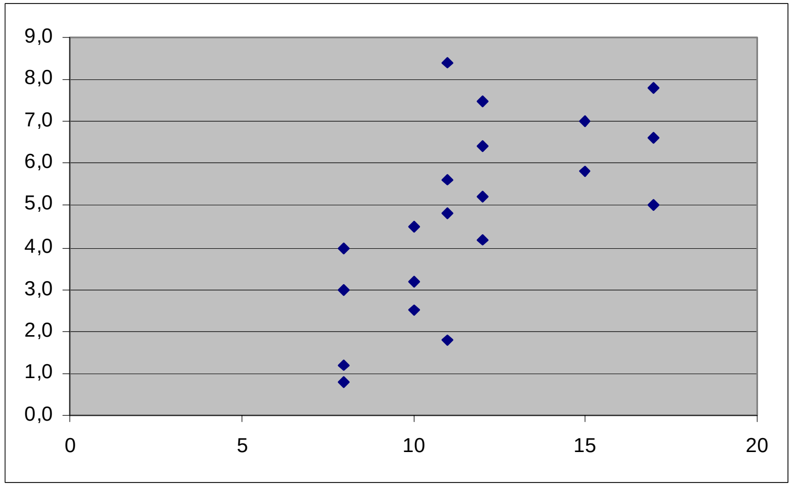


Data graphs

Salary=Y	Education=X
0,8	8
1,2	8
1,8	11
2,5	10
3,0	8
3,2	10
4,0	8
4,2	12
4,5	10
4,8	11
5,0	17
5,2	12
5,6	11
5,8	15

Salary=Y	Education=X
6,4	12
6,6	17
7,0	15
7,5	12
7,8	17
8,4	11

Grafical display of data may help us to select (specify) a useful model



Hypotesis of the model

$$Y_i = \beta_0 + X_i\beta_1 + u_i, \quad i = 1,\ldots, N$$

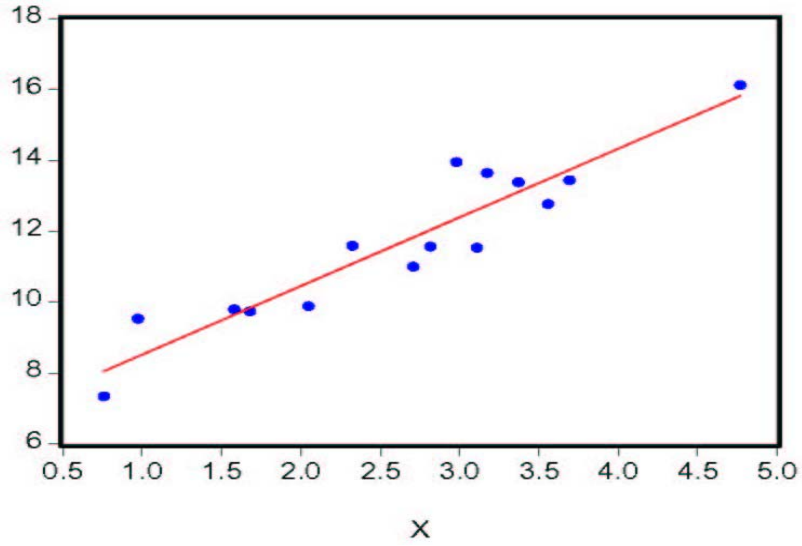
- The model is well specified
  - **Linear model**
  - *Y* depends on *X* and no other variables influences Y
  - Constant parameters
- Degrees of freedom (*N* > 2)
- Non stochastic regressors
- **Mean zero errors**: $\boxed{E(u_i) = 0}$
- **Homoskedasticity:** $\boxed{V(u_i) = \sigma^2, \quad i = 1,\ldots, N}$
- **Incorrelation**: $\boxed{E(u_i u_j) = 0, \quad i \neq j}$
- **Normal distribution:** $u_i \equiv N(0, \sigma^2) \quad i.i.d.$

En lo sucesivo, **todas las hipótesis enunciadas serán asumidas**, no siendo válidos los resultados que se expondrán si alguna de ellas no se cumple

CEU Universidad San Pablo

Parameter estimation

– We need to obtain the β_0 and β_1 values that better fit the N available observations.



– We need to define the term "fitting".

– Depending on the definition of "fitting" we have fixed, we will have the diffent parameter estimation methods.

- Ordinary Least Squares method
- Maximum Likelihood method
- Bayesian estimation method
- Moments method

OLS estimation method

– Find the β_0 and β_1 values that minimize the sum of square deviations between the observed value Y_i and the prediction $\beta_0+\beta_1 X$:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min_{\beta_0 \beta_1} SC(\beta_0, \beta_1) = \arg\min_{\beta_0 \beta_1} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$