MAP553 Regression

Chapter 1: Introduction

Contents

This introduction will allow us to motivate this course. We are interested in the following dataset

https://www.economicswebinstitute.org/data/wagesmicrodata.xls

from the `Economics Web Institute`. In this dataset, we found datas such as economic characteristics about 534 persons:

- **WAGE :** `Wage (dollars per hour)`.
- **OCCUPATION :** `Occupational category (1=Management, 2=Sales, 3=Clerical, 4=Service, 5=Professional, 6=Other)`.
- **SECTOR :** `Sector (0=Other,1=Manufacturing, 2=Construction)`.
- **UNION :** `Indicator variable for union membership (1=Union member, 0=Not union member)`.
- **EDUCATION :** `Number of years of education`.
- **EXPERIENCE :** `Number of years of work experience`.
- **AGE :** `Age (years)`.
- **SEX :** `Indicator variable for sex (1=Female, 0=Male)` .
- **MARR :** `Marital Status (0=Unmarried, 1=Married)`.

- **RACE :** Race (1=Other,2=Hispanic, 3=White).
- **SOUTH :**Indicator variable for Southern Region (1=Person lives in South, 0=Person lives elsewhere).

An extract of the data is given in the Table~1

ID	WAGE	OC.	SECT.	EDUC.	EXPER.	AGE	SEX	MARR	RACE	SOUTH
1	510.00	6	1	8	21	35	1	1	2	0
2	495.00	6	1	9	42	57	1	1	3	0
3	667.00	6	1	12	1	19	0	0	3	0
4	400.00	6	0	12	4	22	0	0	3	0
5	750.00	6	0	12	17	35	0	1	3	0
6	1307.00	6	0	13	9	28	0	0	3	0

Table 1: Extract from the dataset "Wages"

The purpose of this study (and of the course) is to evaluate the possible effect of socio-demographic characteristics on the salary of employees.

1.1. Description of the data

Before any more elaborate statistical study, a descriptive study of the data must first be carried out. The study depends on the type of the considered variables (qualitative or quantitative).

1.1.1. Quantitative variables

For the quantitative variables, we will calculate for example the mean, standard deviation, median, extremal values...

For our dataset, these statistics are regrouped in the Table~2 or graphically as in Figure~1.

	MOYENNE	ECART TYPE	MINIMUN	MEDIAN	MAXIMUM
WAGE	902.41	513.91	100.00	778.00	4450.00
EDUCATION	13.02	2.62	2.00	12.00	18.00
AGE	36.83	11.73	18.00	35.00	64.00
EXPERIENCE	17.82	12.38	0.00	15.00	55.00

Table 2: Dataset "Wages": statistics for quantitative variables

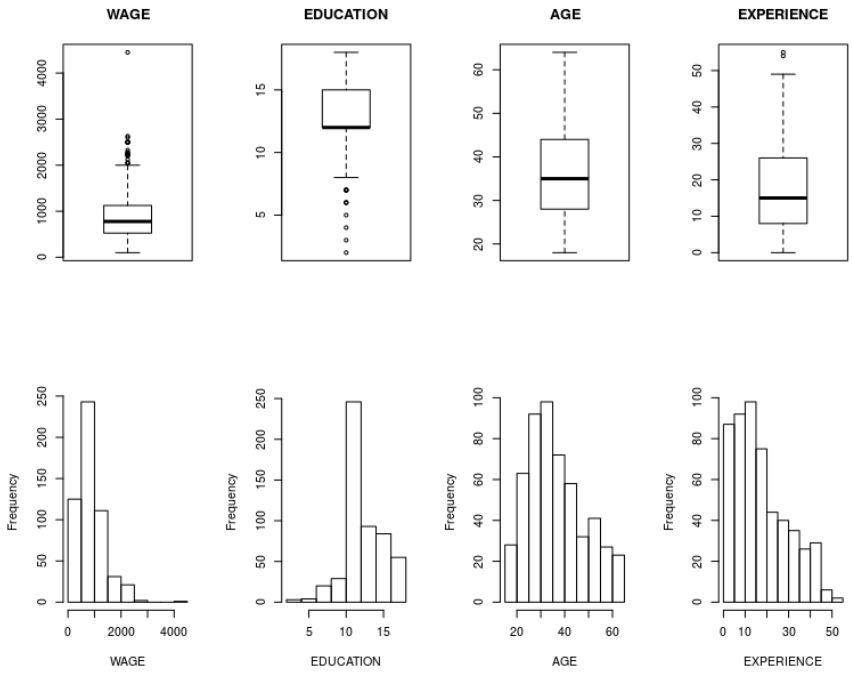

Figure 1: Dataset "Wages": Boxplots and histograms for quantitative variables

1.1.2. Qualitative variables

The qualitative variables are represented in the form of frequency tables (Table~3) or graphically as in Figure~2.

	Modalities	Effectifs	Frequencies (%)
OCCUPATION	1	55	10.30
	2	38	7.12
	3	97	18.16
	4	83	15.54
	5	105	19.66
	6	156	29.21
SECTOR	0	411	76.97
	1	99	18.54
	2	24	4.49
SEX	0	289	54.12
	1	245	45.88

Table 3: Dataset "Wages" : effectifs and frequencies for qualitative variables

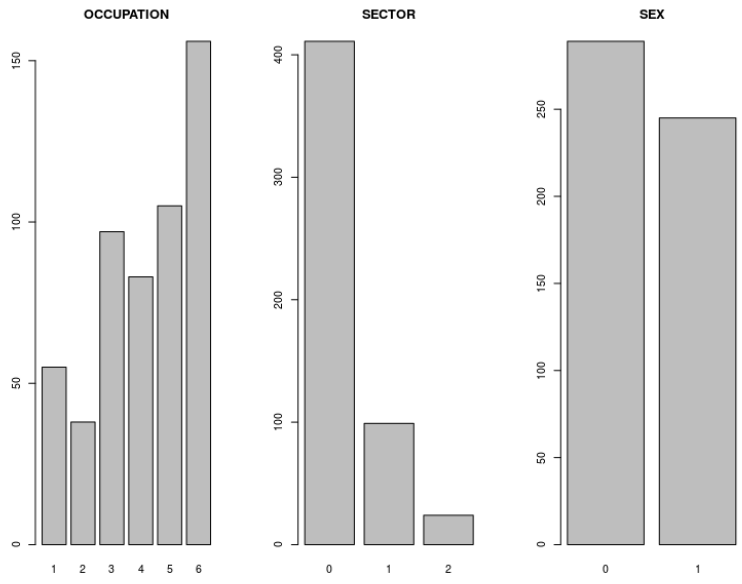

Figure 2: Dataset: "Wages": Barplot for qualitative variables

1.2. Linear regression

Now try to understand the influence of quantitative variables (`EDUCATION`, `AGE`, `EXPERIENCE`) on wages . We plot on the Figure~3 the three corresponding scatter plots.

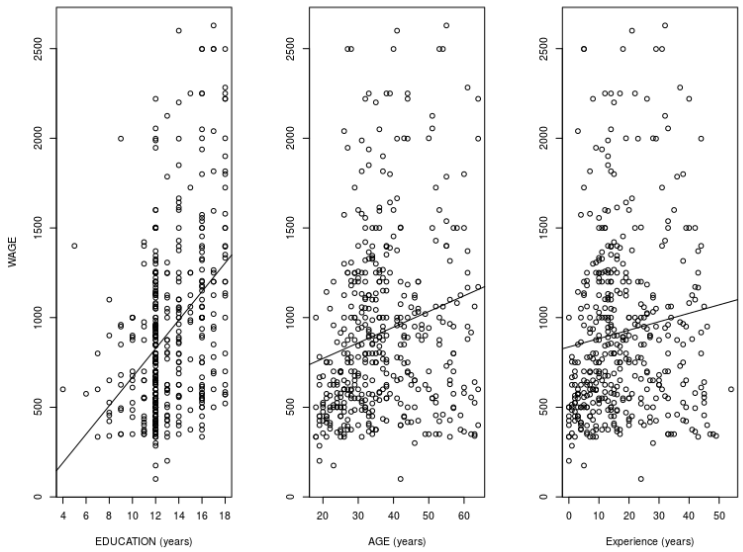

Figure 3: Dataset: "Wages": `WAGE` as functions of `EDUCATION`, `AGE` and `EXPERIENCE`.

Linear correlation coefficient In order to quantify the linear relationship between two quantitative variables X and Y, the linear correlation coefficient can be calculated r_{XY}:

$$r_{X,Y} = \frac{s_{x,y}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i,j}(y_i - \bar{y})(x_j - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{1}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. By construction, $|r_{XY}| \leq 1$. If the points are perfectly aligned then $|r_{XY}| = 1$. On this dataset, we get

r_{XY}	EDUCATION	AGE	EXPERIENCE
WAGE	0.40	0.21	0.12

Question : Is it big? small? Significant?

Comment:

- ☛ Statistical tests can be used (non-parametric tests based on the rank method). Under **R**, these tests are implemented in the function `cor.test` in which to specify the desired test (`method = "kendall" or "spearman"`).

Consider again the scatter plot on the left of Figure~3. If we try to summarize it by a line, called *simple linear regression line*, we write:

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + e_i$$

where e_i is an error term between the line of the observation y_i.

In this very simple model, the variable of interest (quantitative) is explained by a quantitative variable called *explonatory variable* (or *covariate* or *predictor* or *regressor*...). The slope (β_1) and the intercept (β_0) of the line are *estimated* from the observations to properly "set" the line. In this course, we will see how to estimate these parameters, what are the properties of such an estimator. In addition, we want to know if the slope is significantly different from 0, *i.e.* we will try to write tests on the parameters of the models.

We can also try to explain the salary as a linear combination of the other quantitative variables:

$$\begin{aligned} \text{Wage}_i &= \beta_0 + \beta_1 \text{Education}_i + \beta_2 \text{Age}_i + \beta_3 \text{ Experience}_i + e_i \\ &= \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + e_i \end{aligned}$$

where x_{ik} is the value of the k-th predictor variable of the individual i. Then, we speak of *multiple linear regression*. The same questions as before are: Is β_k significance? How to select the most relevant predictor variables? And so on.

1.3. Analysis of the variance

Analysis of the variance with one factor It may also be interesting to study the relationship between the WAGE (variable of quantitative interest) and the qualitative variables, for example SEX, or OCCUPATION (type of occupation). Graphically, we can draw boxplots by modality of the qualitative variable as in Figure~4.

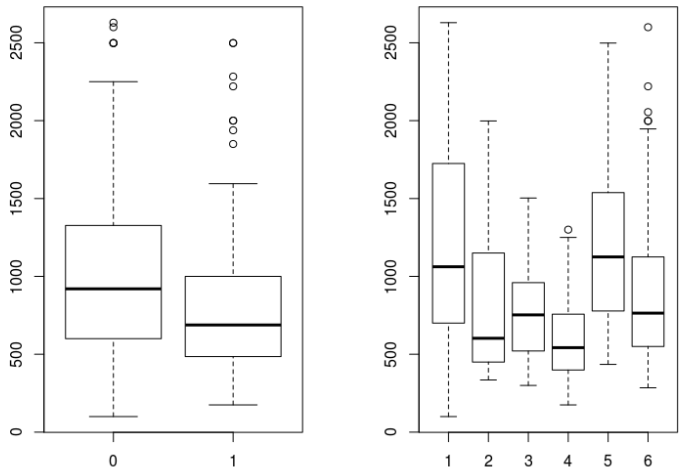

Figure 4: Dataset: "Wages": Wages as functions of SEX (0 = Male) or OCCUPATION category.

In a natural way, to compare the wages within the different populations, one will try to compare the averages within the groups:

OCCUPATION	1	2	3	4	5	6
Mean by OCCUPATION	1254.27	790.68	758.99	591.87	1205.73	879.75

SEX	0	1
Mean by SEX	1032.7	788.47

Tests which compare means can then be applied. However, it is actually possible to write a linear model to study salary according to SEX:

$$\text{Wage}_i = \underbrace{\beta_0}_{\text{Wage of Male}} \mathbb{1}_{\text{Sex}_i=0} + \underbrace{\beta_1}_{\text{Wage of Female}} \mathbb{1}_{\text{Sex}_i=1} + e_i.$$

Similarly, we can write a linear model to study salary according to OCCUPATION:

$$\text{Wage}_i = \sum_{l=1}^{l} \underbrace{\beta_l}_{\text{Wage of Occupation } l} \mathbb{1}_{\text{Occupation}=l} + e_i.$$

These models are called Anova models one factor[1]. We want here to explain the variation of the WAGE according to a single factor (SEX or OCCUPATION).

Analysis of the variance with two factors One may wonder if there is not a joint effect of the two factors. We will then seek to cross factors, for example by calculating mean of WAGE as follows:

	OCCUPATION	1	2	3	4	5	6	Mean by SEX
SEX	0	1442.11	972.05	795.29	612.52	1277.87	941.92	1032.71
	1	944.88	531.57	750.92	580.74	1132.02	605.17	788.47
Mean by	OCCUPATION	1254.27	790.68	758.99	591.87	1205.73	879.75	919.77

We will try to study the influence of each factor on WAGE but also their joint influences (possible interactions). We will see in this course how to write and study a linear model.

1.4. Analysis of covariance

One can think that the link between WAGE and EDUCATION is not the same depending on whether one is a MALE or a FEMALE. We will then want to write a regression model for each sex (see Figure~5.) In the same way, we will see in this course that it is possible to write a linear model also answering this need.

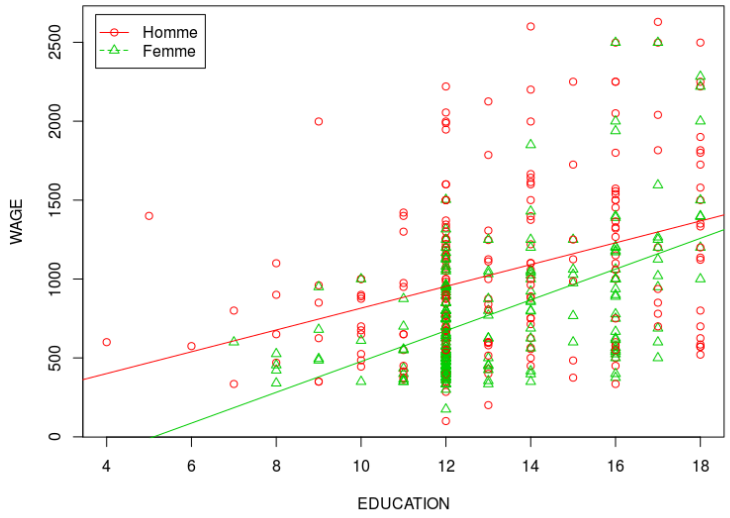

Figure 5: Dataset: "Wages": Wages as functions of SEX and EDUCATION

[1] *Analysis of variance model with with a single factor*

1.5. Matrix form

Let $y_1, \ldots, y_n$, n independent observations of a quantitative variable. For each observation/individu i, we have p real quantities $(x_{i1}, \ldots, x_{ip})$ (quantitative or indicator). We try to explain **the response** y_i as a linear function of the p **predictors** $(x_{i1}, \ldots, x_{ip})$. So, for all i, we write

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i, \qquad \forall i = 1, \ldots, n \tag{2}$$

where e_i is the error term.

Set:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}}_{p+1 \text{ columns}}, \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \tag{3}$$

Then, we can write a matrix version of the equations (2).

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e} \tag{4}$$

$\boldsymbol{y}$ and X are observed. The parameter $\boldsymbol{\beta}$ is unknown and must be estimated. Estimation and statistical tests will be the subject of next chapters.