UD5: INFERENCE

Part 1: Distributions in sampling

Part 2: Inference about one population

Comparison of 2 populations

Part 3: ANOVA (Analysis of Variance)

Design of experiments

Part 4: Regression

UD 5 part 4

REGRESSION

TWO-DIMENSIONAL RANDOM VARIABLES

WHEN TWO RANDOM NUMERIC CHARACTERISTICS ARE OBSERVED FROM EACH INDIVIDUAL, WE HAVE A TWO-DIMENSIONAL RANDOM VARIABLE.

- **E.g. In the population of students, we observe the height (cm) and weight (kgs) of each student.**
- **For the control of energy consumption in a factory, we record every day the CONSUMPTION and the daily temperature (ºC).**

Study of 2 QUALITATIVE VARIABLES:

BY MEANS OF A CONTINGENCY TABLE

REPEAT GENDER	YES 1	NO 2	Row Total
MALE 1	5 83.3 10.9	41 63.1 89.1	46 64.8
FEMALE 2	1 16.7 4.0	24 36.9 96.0	25 35.2
COLUMN TOTAL	6 8.5	65 91.5	71

Marginal frequency of gender

Marginal frequency of repeat

Relative frequency of gender conditined to repeat

Relative frequency of repeat conditioned to gender

Marginal frequencies:

Frequency of each value of one variable without taking into account the other

Relative conditional frequencies:

Relative frequency of the value of one variable in relation to each value of the other

Study of 2 QUANTITATIVE VARIABLES:

1) BY MEANS OF A CONTINGENCY TABLE AFTER GROUPING THE DATA IN INTERVALS.

2) Scatter plot: graphical representation of the relationship.

3) Covariance and linear correlation: quantifies the “degree” of linear relationship between x , y

4) Simple regression: models the relationship for predictive purposes.

HEIGHT

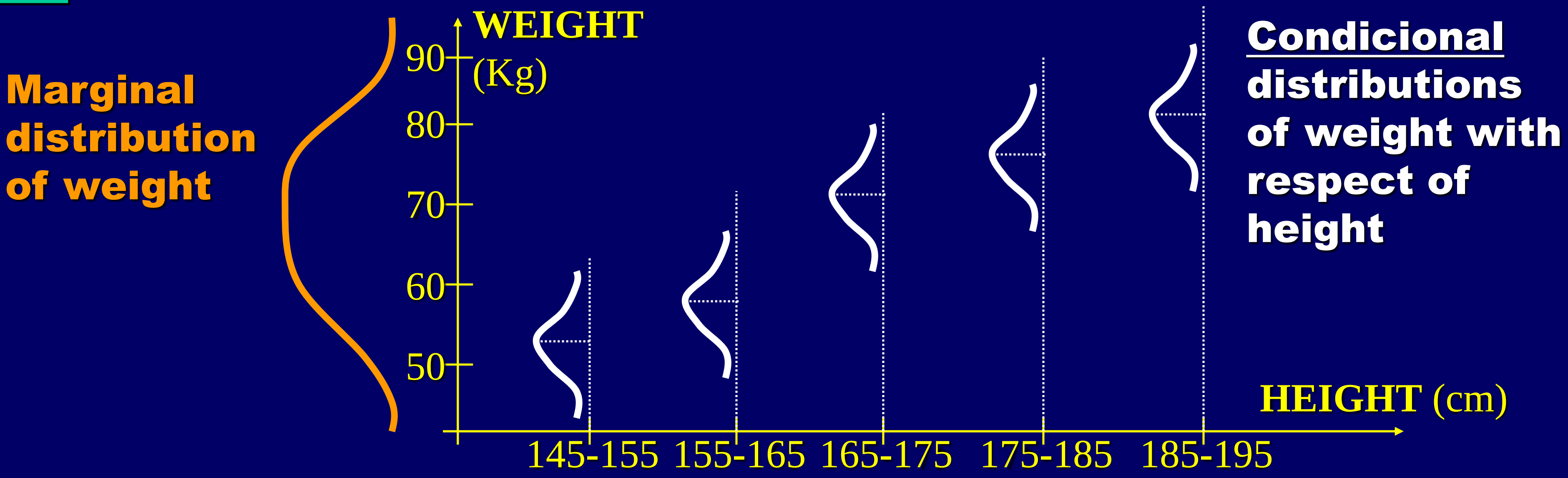
WEIGHT	145–155	155–165	165–175	175–185	185–195	Row Total
40–55	9 75.0	17 44.7	0 .0	0 .0	0 .0	26 20.0
55–70	3 25.0	18 47.4	31 53.4	5 29.4	0 .0	57 43.8
70–85	0 .0	3 7.9	24 41.4	12 70.6	3 60.0	42 32.3
85–99	0 .0	0 .0	3 5.2	0 .0	2 40.0	5 3.8
Column Total	12 9.2	38 29.2	58 44.6	17 13.1	5 3.8	130 100

Marginal frequency of weight

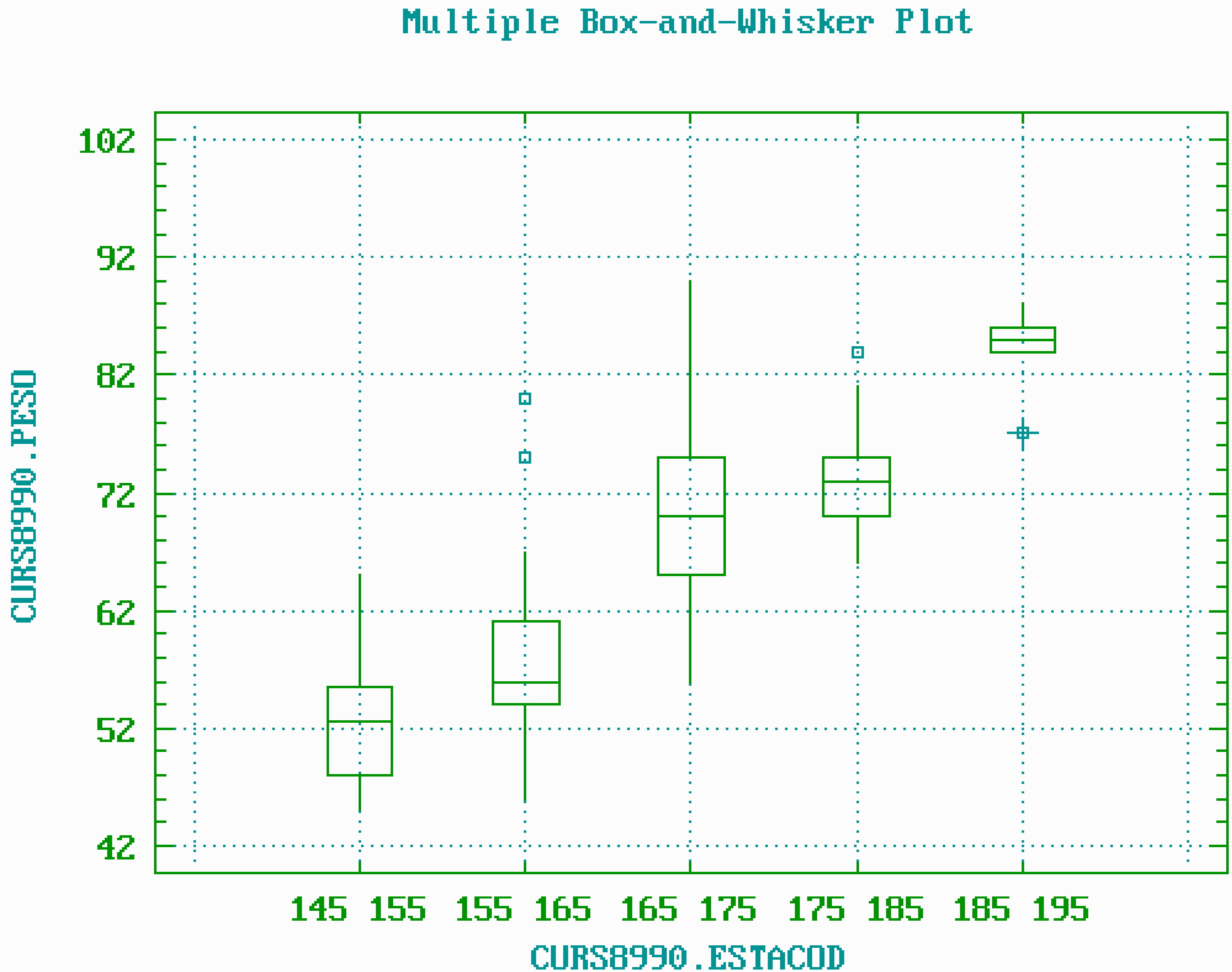
Marginal frequency of height

Relative frequency of weight conditioned to height

PROBLEM: SOME INFORMATION IS LOST IN THE TABULATION

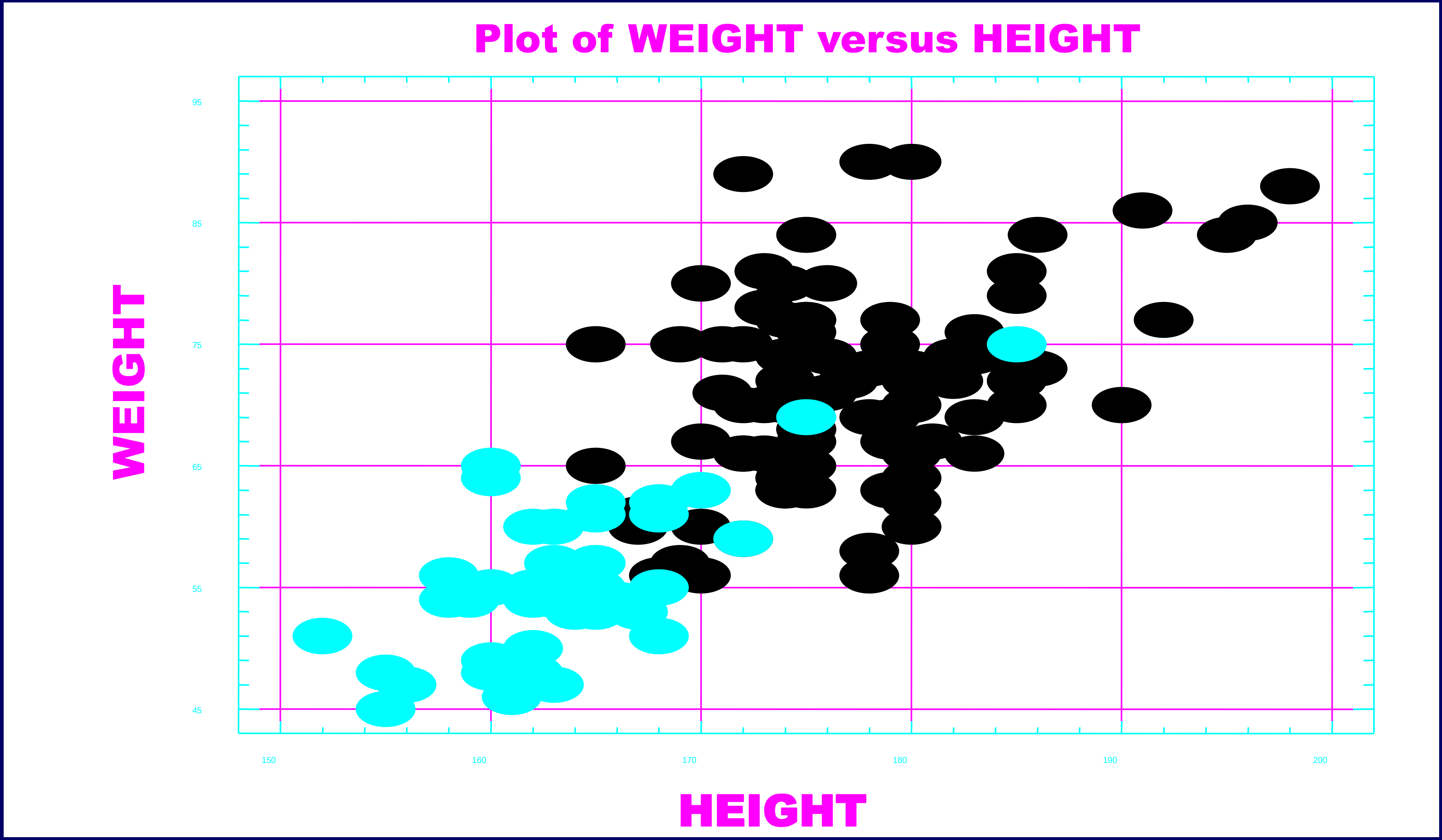


HEIGHT	Nº of cases	MEAN	STANDARD DEVIATION	MIN.	MAXIMUM
145-155	12	53.0000	6.39602	45	65
155-165	38	57.7895	7.45856	46	80
165-175	53	70.8793	7.61134	56	90
175-185	17	73.4118	4.71777	66	84
185-195	5	84.0000	4.18330	77	88



PROBLEM: SOME INFORMATION IS LOST IN THE TABULATION

SCATTERPLOT



What colors corresponds to men and women ?

TWO-DIMENSIONAL NORMAL DISTRIBUTION

EXAMPLES:

- Speed of processor and time required to execute a calculation
- Room temperature and power consumption in heating
- Weight and height of a person

... are two components of a bivariate Normal distribution

- **Vector of averages** $\vec{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$ being $m_1 = E(x_1)$ y $m_2 = E(x_2)$
- **Matrix of variances - covariances**

$$\overline{\overline{V}} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2}^2 \\ \sigma_{2,1}^2 & \sigma_2^2 \end{bmatrix} \qquad \text{being :} \quad \sigma_1^2 = \mathrm{var}(x_1), \quad \sigma_2^2 = \mathrm{var}(x_2),$$

$$\sigma_{1,2}^2 = \sigma_{2,1}^2 = \mathrm{Cov}(x_1, x_2)$$

Two-dimensional density function: f (x, y)

$$\forall S \in \Re^2 \quad P((x,y) \in S) = \iint_S f(x,y)\,dx\,dy$$

The two components X_1 and X_2 of a bivariate random variable have a bivariate normal distribution if the joint density function is:

$$f(x_1,x_2) = \frac{1}{2\pi \cdot \sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-m_1)^2}{\sigma_1^2}+\frac{(x_2-m_2)^2}{\sigma_2^2}-2\rho\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2}\right]}$$

Being:

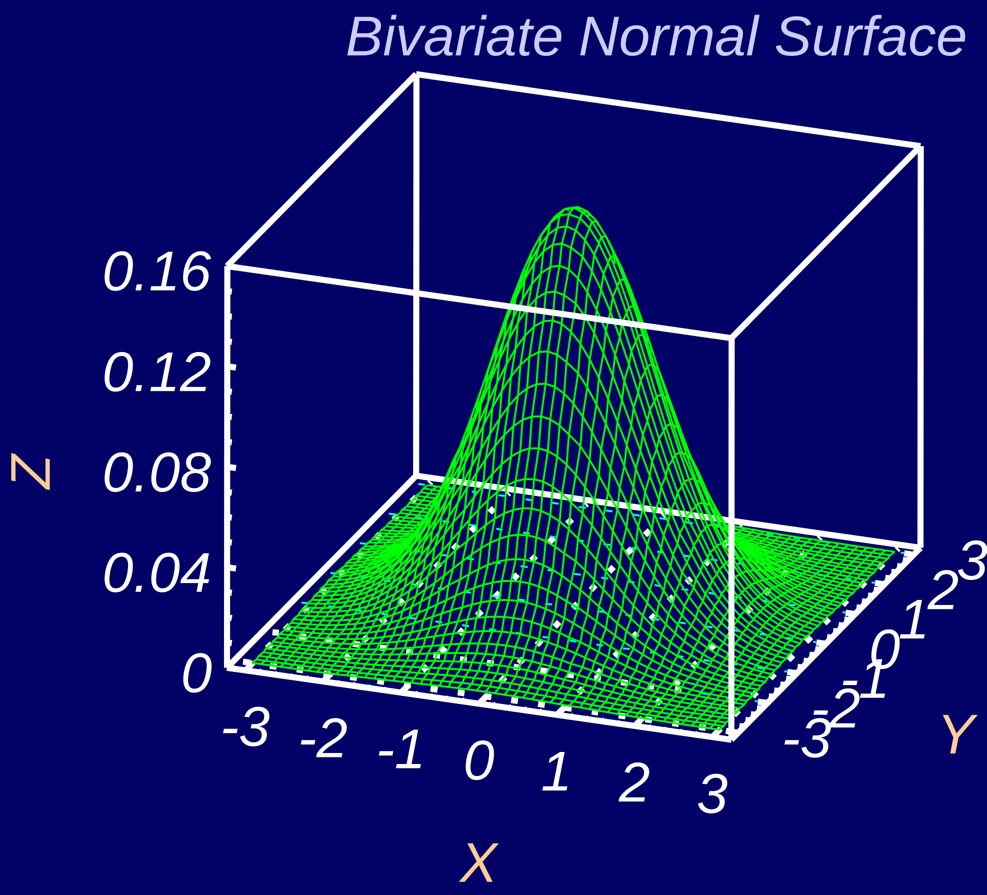
m_1 , σ_1^2 : mean and variance of the distribution of X_1

m_2 , σ_2^2 : mean and variance of the distribution of X_2

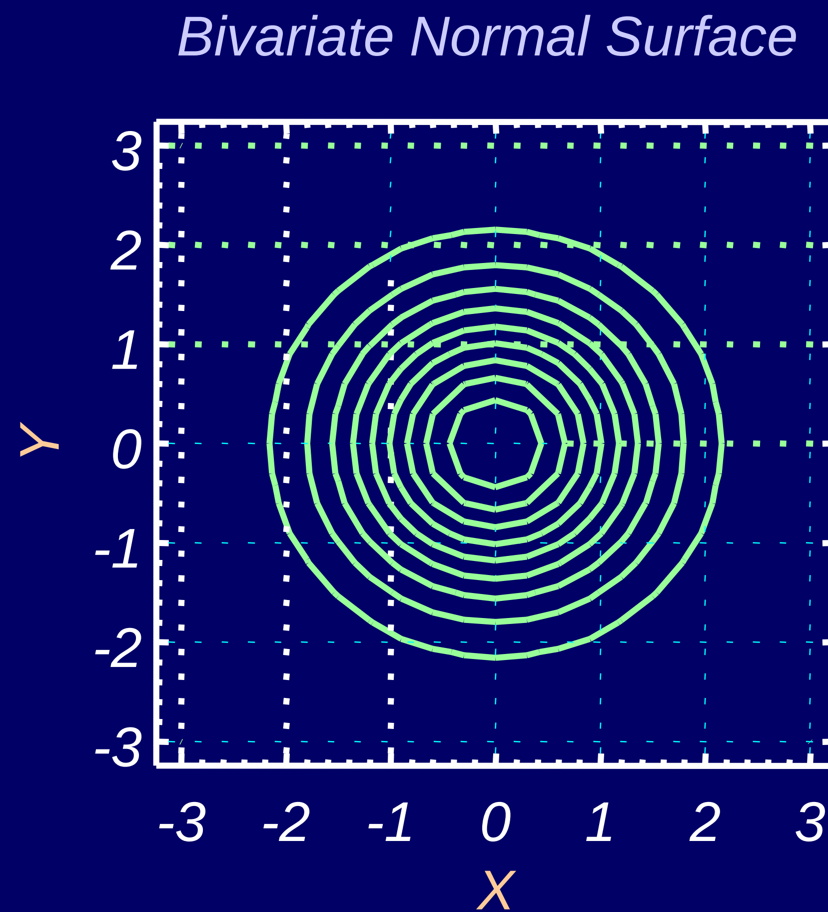
ρ : correlation coefficient between X_1 and X_2

$$\vec{Y} \sim \mathbf{N}_2(\vec{0}, \overset{=}{\mathbf{I}})$$

TWO-DIMENS. NORMAL DISTRIBUTION



ISODENSITY CURVES



Volume under the surface = 1

$$\iint_{R^2} f(x,y)\,dx\,dy = 1$$

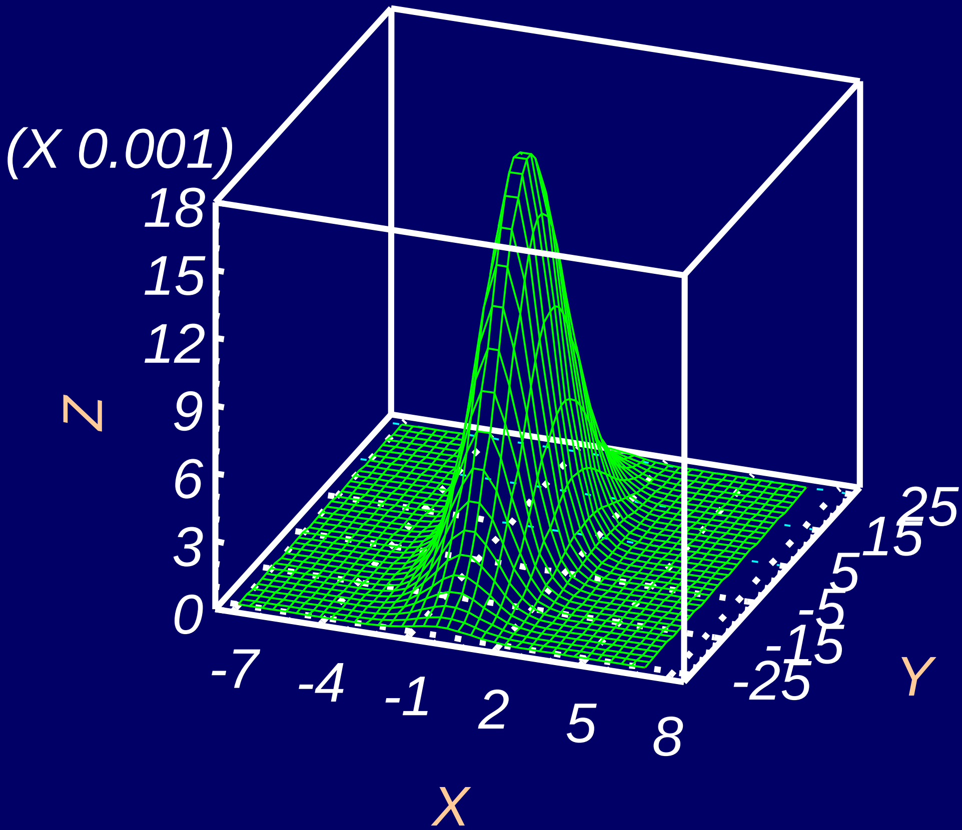
Probabilities are obtained integrating

$$\vec{Y} \sim \mathbf{N}_2(\vec{0}, \overset{=}{\mathbf{V}}_{\vec{Y}})$$

$$\overset{=}{\mathbf{V}}_{\vec{Y}} = \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix}$$

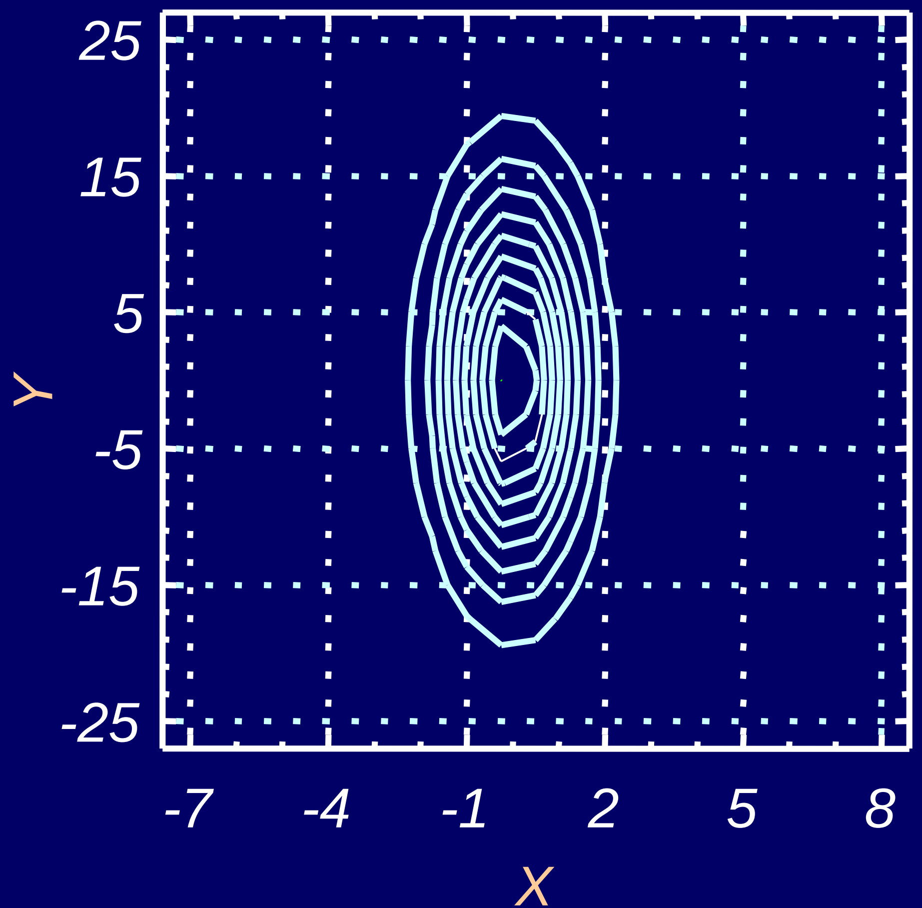
TWO-DIMENS. NORMAL DISTRIBUTION

Bivariate Normal Surface



ISODENSITY CURVES

Bivariate Normal Surface

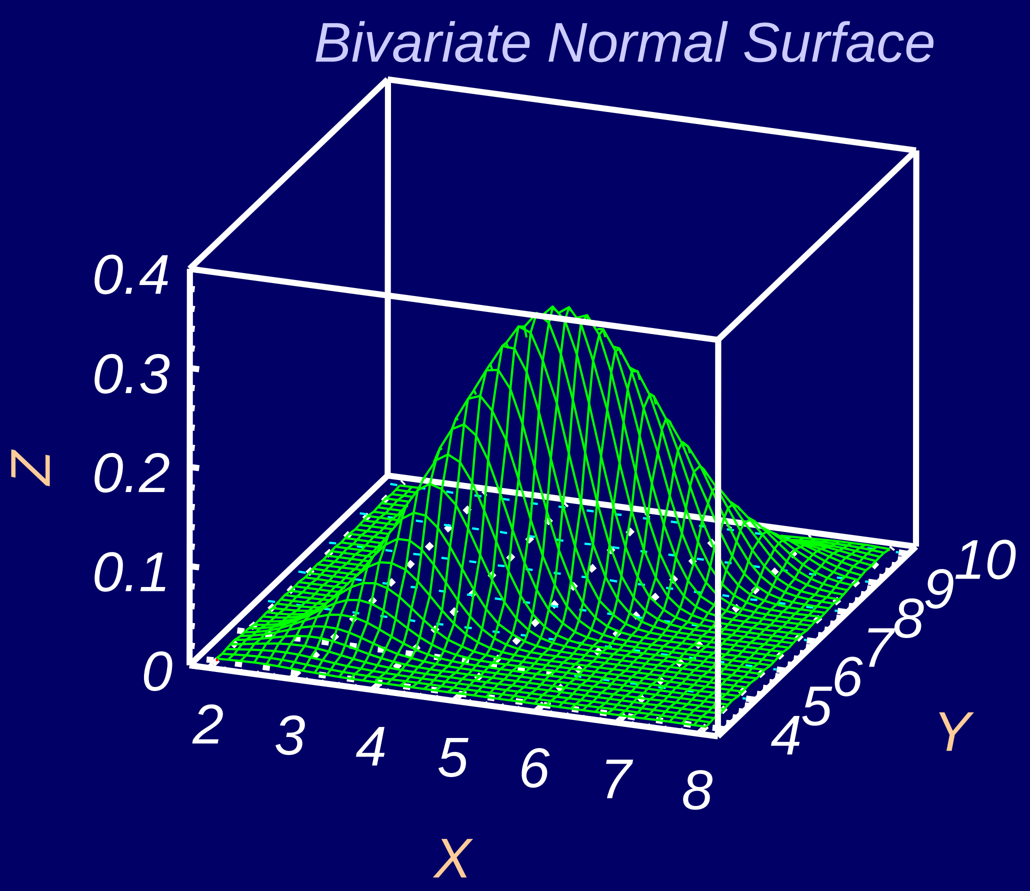


$$\vec{Y} \sim \mathbf{N_2}(\vec{\mathbf{m}}, \bar{\bar{\mathbf{V}}}_{\vec{Y}})$$

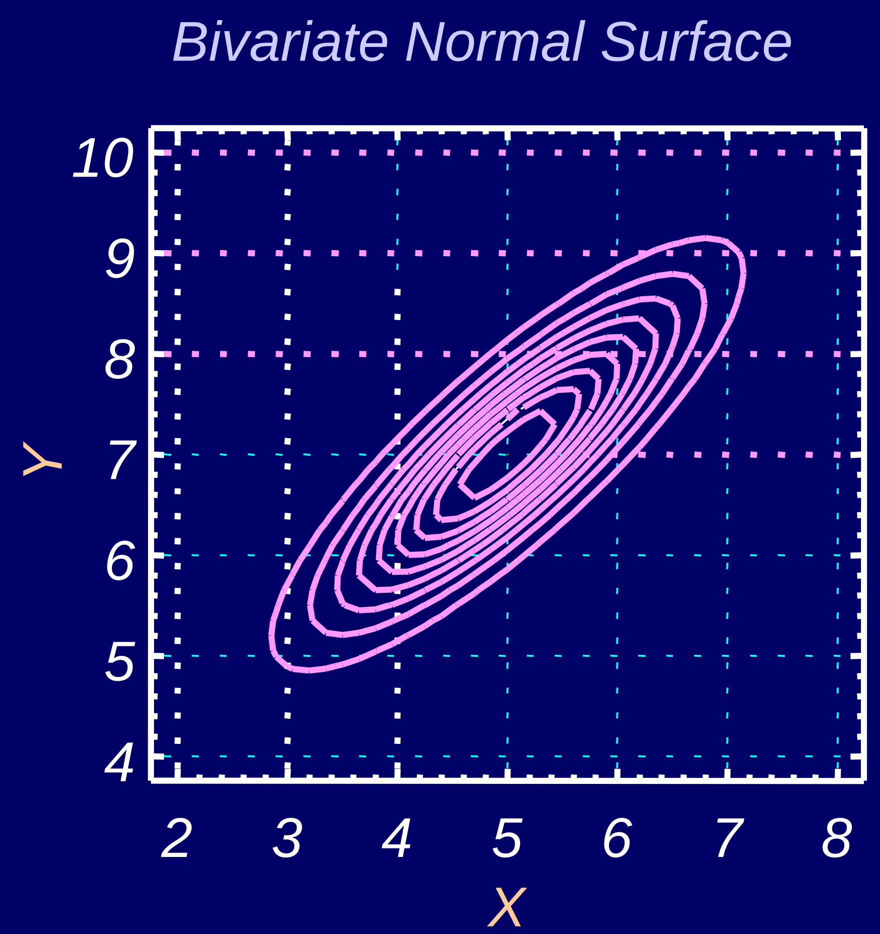
$$\vec{\mathbf{m}} = \begin{Bmatrix} \mathbf{5} \\ \mathbf{7} \end{Bmatrix}$$

$$\bar{\bar{\mathbf{V}}}_{\vec{Y}} = \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}$$

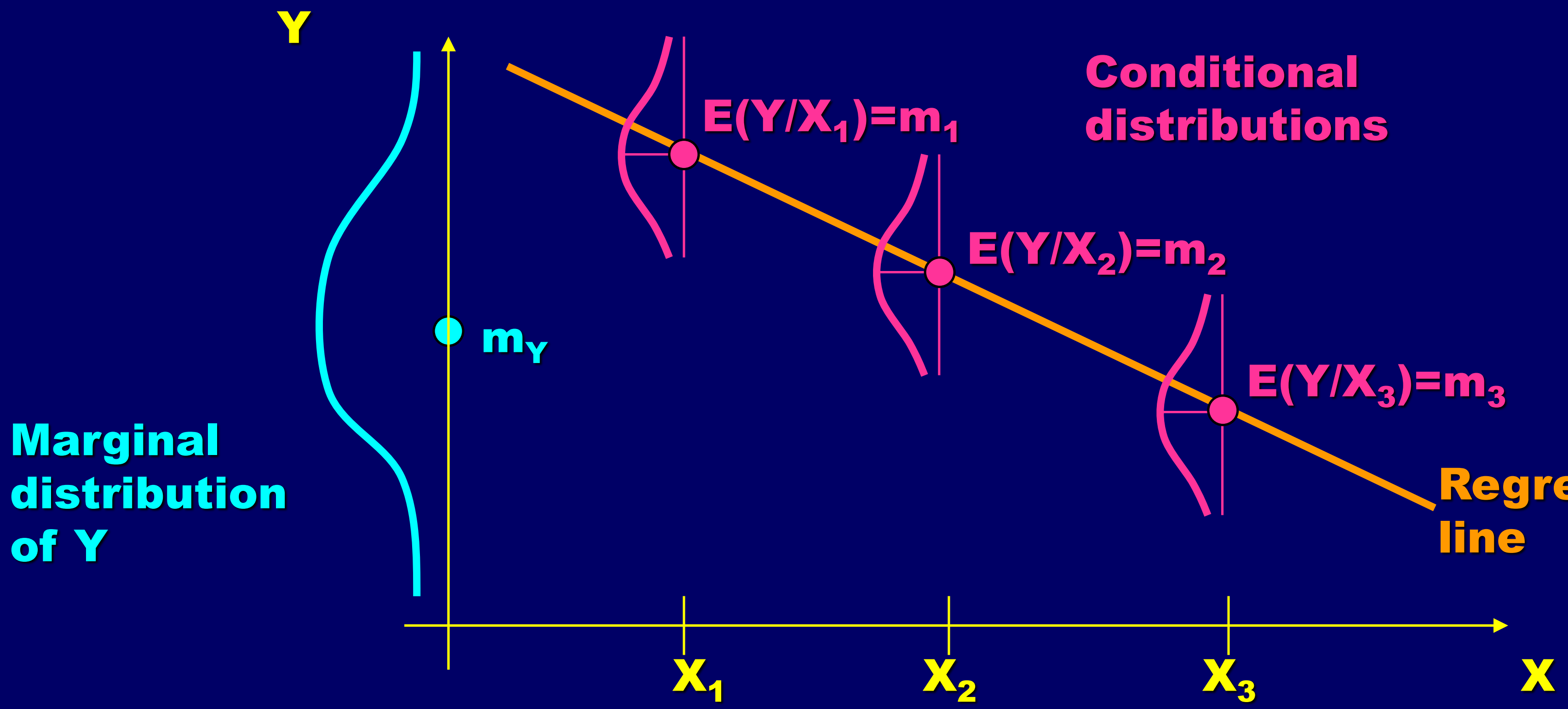
TWO-DIMENS. NORMAL DISTRIBUTION



ISODENSITY CURVES



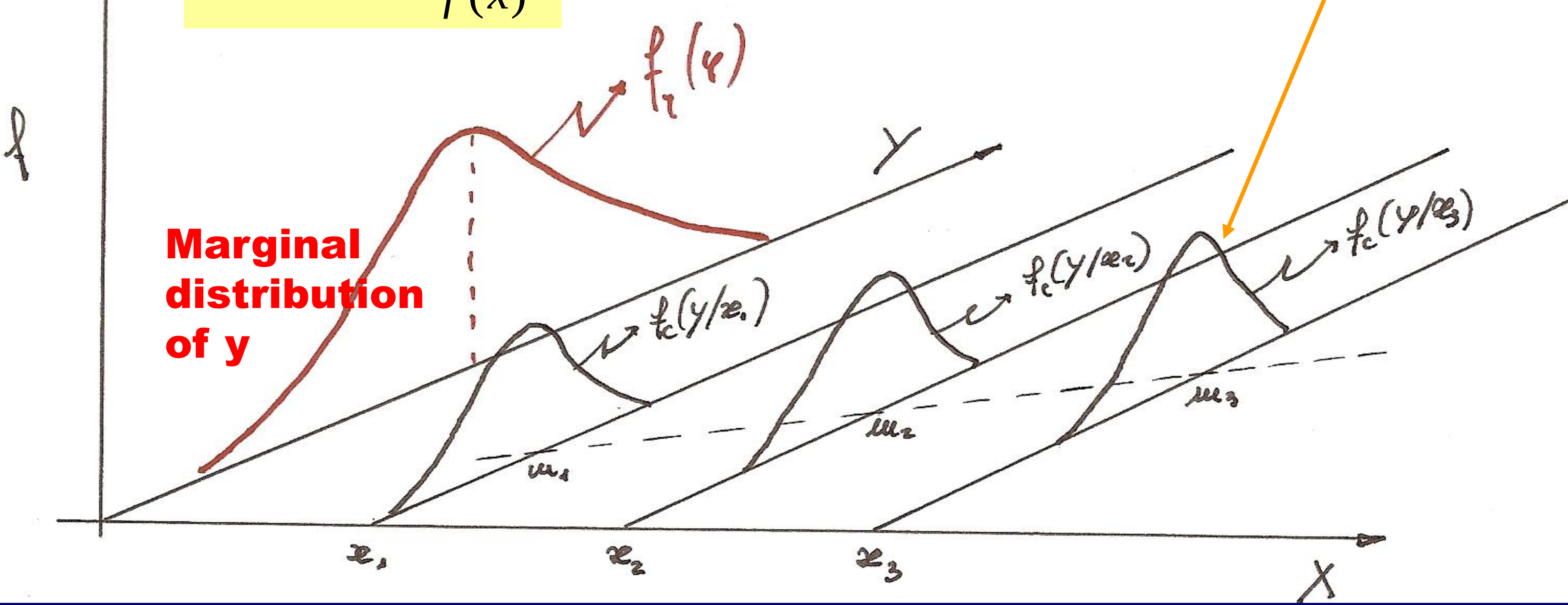
SIMPLE REGRESSION: THEORETICAL MODEL



$$\hat{Y} = E(Y/X = x_t) = \alpha + \beta \cdot x_t$$

$$f_c(y/x) = \frac{f(x,y)}{f(x)}$$

Conditional distribution y/x : distribution of y when x takes a particular value



Variance of residuals ("residual variance")

$$y_t = \alpha + \beta \cdot x_t + u_t$$

Residuals: $u_t \approx N(0, \sigma^2)$

(Variance of the conditional distribution =

= variance of residuals) ≤ variance of Y

SIMPLE LINEAR REGRESSION

Y = f (X)

THE CONDITIONAL DISTRIBUTION: Y / X=x_t is a random variable with parameters:

$$E(Y/X = x_t) = \alpha + \beta \cdot x_t$$

$$\sigma^2(Y/X = x_t) = \sigma^2_{\text{residual}} \quad \text{(constant)}$$

EXAMPLE:

Y DAILY GAS CONSUMPTION IN WINTER IN A FACTORY FOR HEATING

X TEMPERATURE OF EACH DAY (ºC)

What is α ?

AVERAGE CONSUMPTION WHEN Tª = 0º C

What is β ? **INCREASE OF Y (AVERAGE GAS CONSUMPTION) IF TEMPERATURE INCREASES 1ºC**

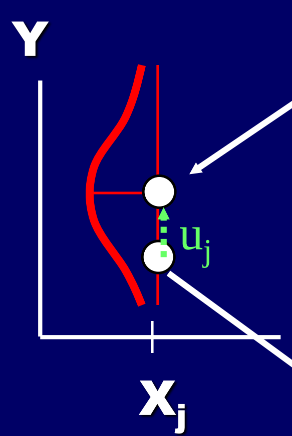
Will β be positive or negative in this case?

POPULATION

The linear relationship existing at the population level between the two quantitative components of a bivariate random variable (X,Y) is:

$$E(Y) = \alpha + \beta \cdot X$$

$$E(Y/X=x_j) = \alpha + \beta\, x_j$$



$$y_j = \alpha + \beta \cdot x_j + u_j$$

residual

SAMPLE

The linear relationship existing between the two quantitative components of a bivariate random variable, estimated from the sample, is:

$$E(Y) = a + b \cdot X$$

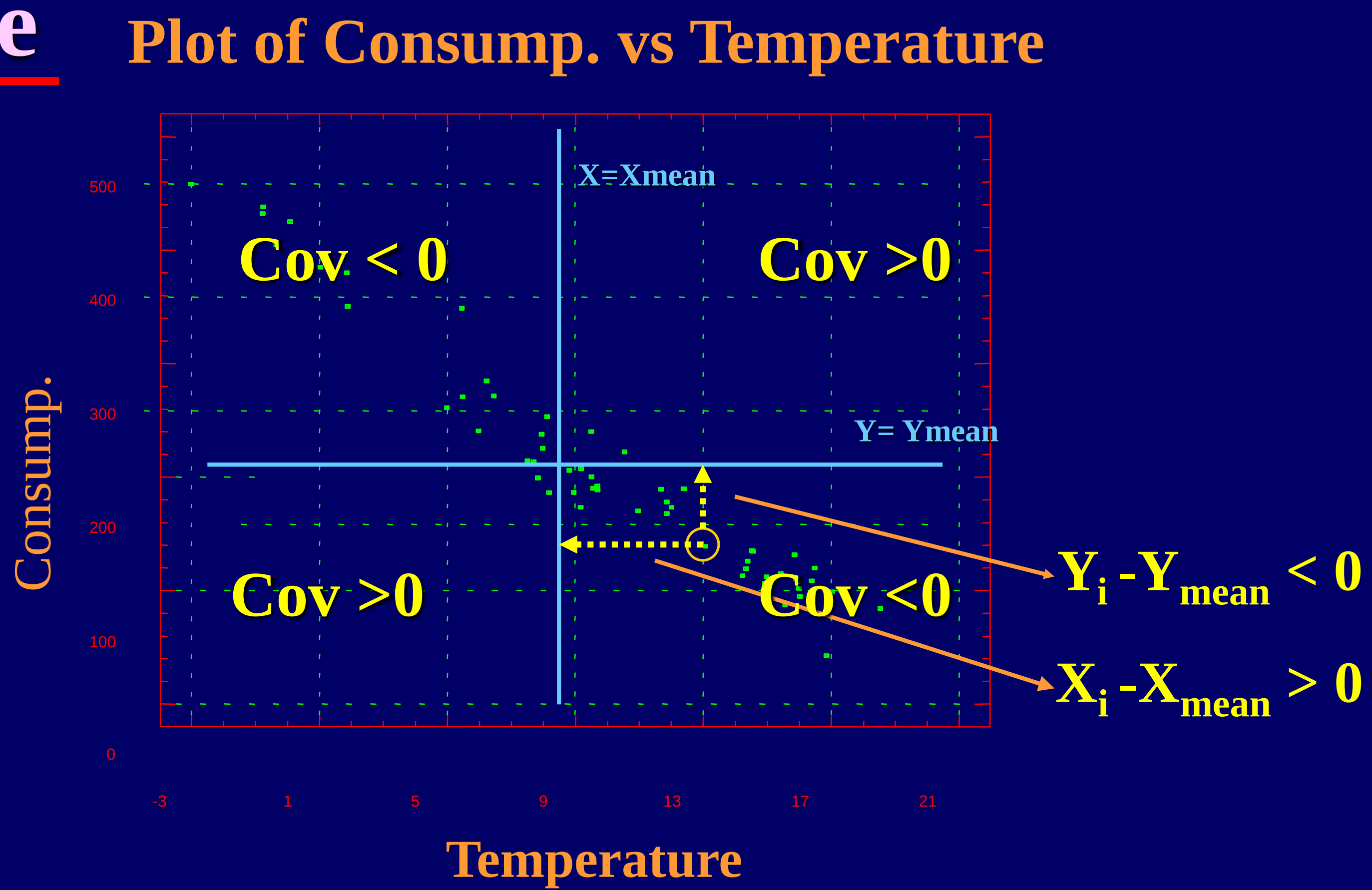
$$y_j = a + bx_j + e_j$$

residual

$$\hat{\alpha} = a$$
$$\hat{\beta} = b$$

How can we quantify the degree of correlation between 2 variables?

Covariance



$$Cov_{(X,Y)} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{N-1}$$

$$\operatorname{cov}(x,y) = E\left[(x-m_x)\cdot(y-m_y)\right] = E(x\cdot y) - m_x\cdot m_y$$

If covariance = -50, is the correlation between the 2 variables strong or weak?

Drawback: depends on the dimensions (scale) in which variables are measured.

CORRELATION CEFFICIENT

$$r_{xy} = \frac{\operatorname{cov}_{xy}}{S_x \cdot S_y}$$

$$\rho = \frac{\operatorname{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

If $\rho = 0 \Rightarrow$ no linear relationship exists between x and y

If $\rho = \pm 1 \Rightarrow \exists$ exact linear relationship between x and y

Advantage: is non-dimensional $-1 < r_{xy} < 1$

$$\mathrm{cov}(x,x) = E\left[(x - m_x)\cdot(x - m_x)\right] = E\left[(x - m_x)^2\right] = \sigma^2_x$$

Matrix of Variances - covariances

Matrix of covariances

$$\begin{pmatrix} \mathrm{cov}(x_1, x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{cov}(x_2, x_2) \end{pmatrix} \quad \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_1, x_2) & \mathrm{var}(x_2) \end{pmatrix}$$

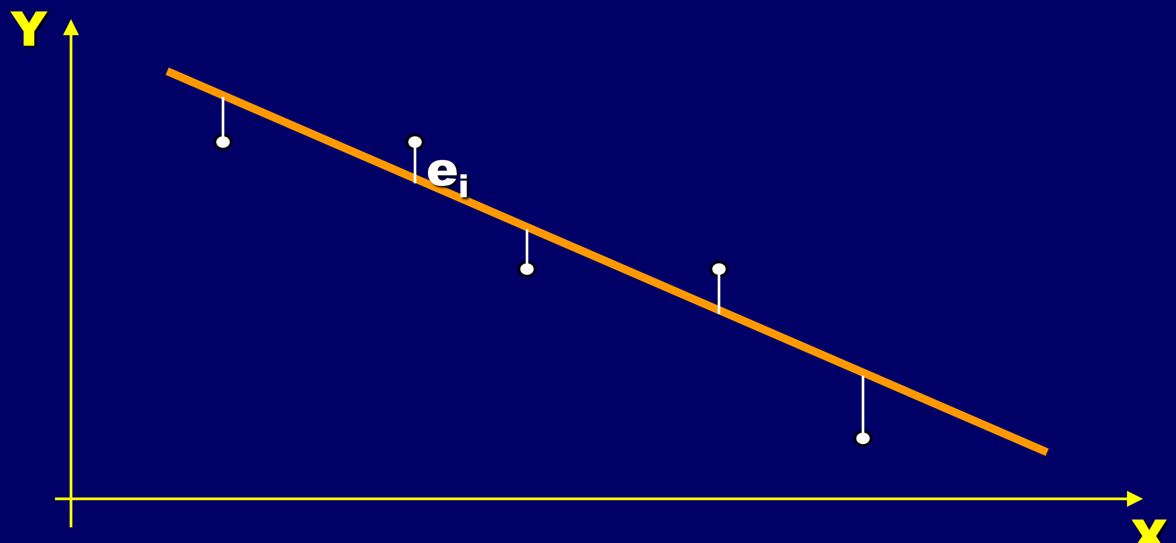
$$\begin{pmatrix} \mathrm{cov}(x_1, x_1) & \mathrm{cov}(x_1, x_2) & \mathrm{cov}(x_1, x_3) \\ \mathrm{cov}(x_2, x_1) & \mathrm{cov}(x_2, x_2) & \mathrm{cov}(x_2, x_3) \\ \mathrm{cov}(x_3, x_1) & \mathrm{cov}(x_3, x_2) & \mathrm{cov}(x_3, x_3) \end{pmatrix}$$

The matrix is symmetric !

Matrix of correlations

$$\begin{pmatrix} r_{xx} & r_{xy} \\ r_{yx} & r_{yy} \end{pmatrix} \quad \begin{pmatrix} 1 & 0{,}7404 \\ 0{,}7404 & 1 \end{pmatrix}$$

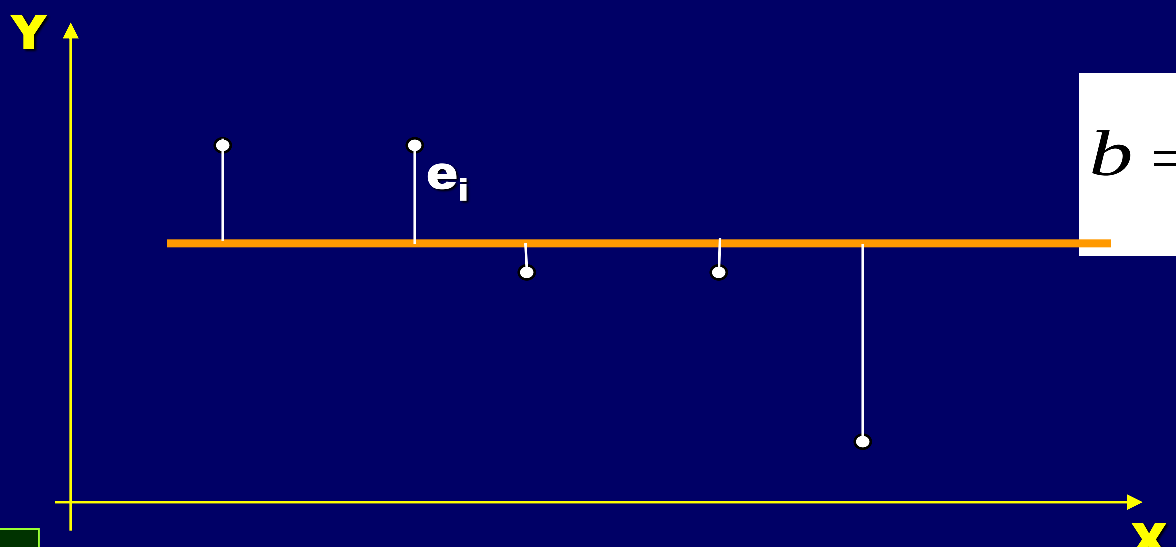
CALCULATION OF REGRESSION LINE



$$\sum e_i = 0$$

minimize $\sum e_i^2$

$$Y = a + b \cdot X$$



$$b = r \cdot \frac{S_Y}{S_X} = \frac{\mathrm{cov(x,y)}}{\mathrm{S}_\mathrm{Y}^2}$$

$$a = \overline{\mathrm{Y}} - b \cdot \overline{\mathrm{X}}$$

REGRESSION

Statistical tool used to establish a model (mathematical equation) able to predict values of a dependent variable Y as a function of one or more input variables (X_1,X_2, …,X_j).

Y = f(X) => simple regression

Y = f(X_1,X_2, …,X_j) => multiple regression (linear or non-linear)

If $$\vec{X} = \begin{Bmatrix} X \\ Y \end{Bmatrix} \approx N_2(\vec{m}, \overline{\overline{V}})$$ Then the two marginal distributions are Normal

The conditional distribution of Y when X = x is Normal with:

Mean: $$\hat{Y} = E(Y/X = x) = m_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - m_X)$$ (regression line)

Variance: $$\sigma^2(Y/X = x) = \sigma_Y^2 \cdot (1 - \rho^2)$$ (residual variance)

Does not depend on X (homoskedasticity)

INDEPENDENCE OF 2 CONTINUOUS VARIABLES

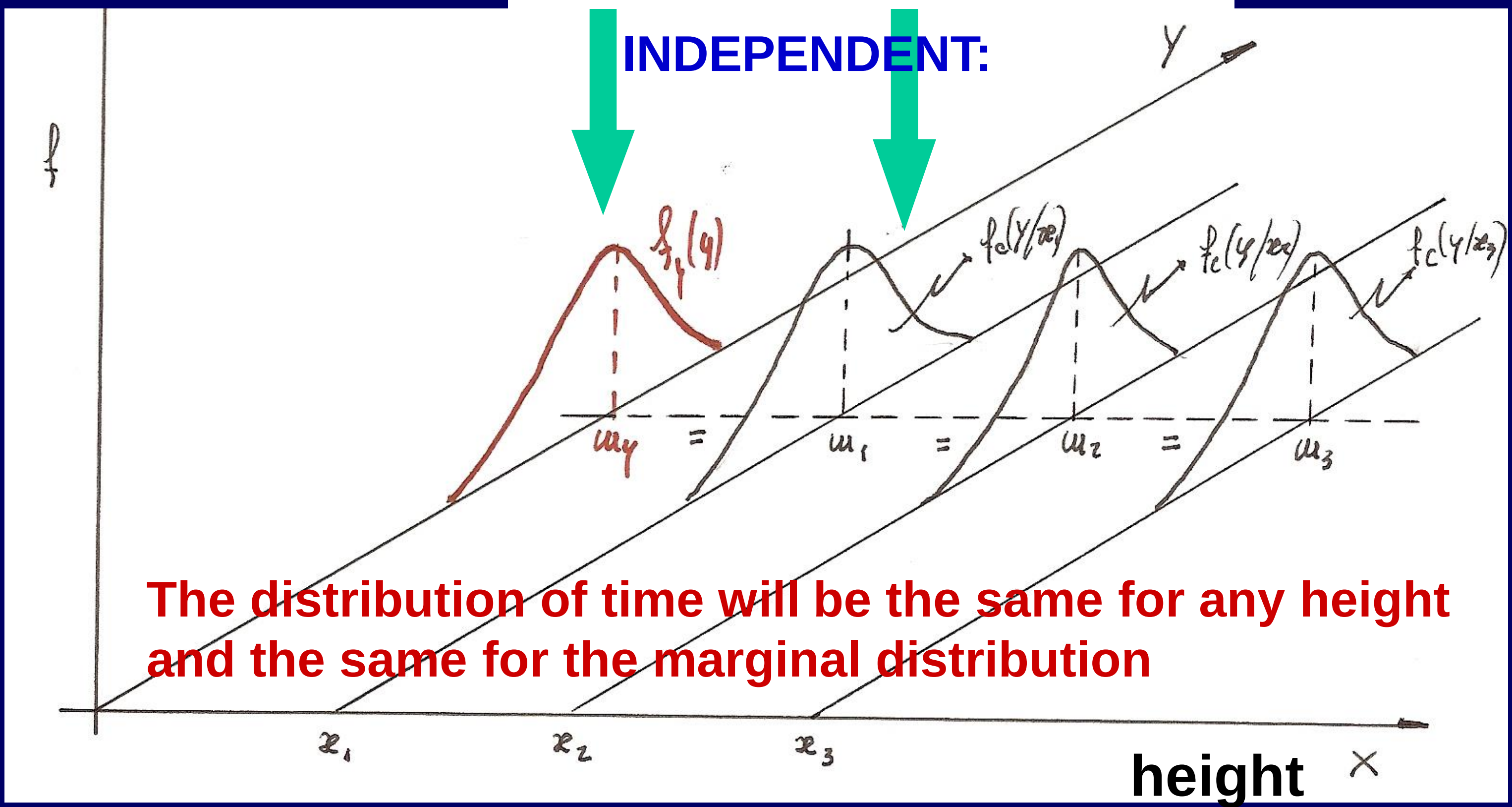
Two components X, Y of a bivariate random variable are independent if the <u>events</u> ($X \le x$) and ($Y \le y$) are independent

EXAMPLE: Two variables are studied: the height of a student and the time required to arrive at university.

Are they correlated?

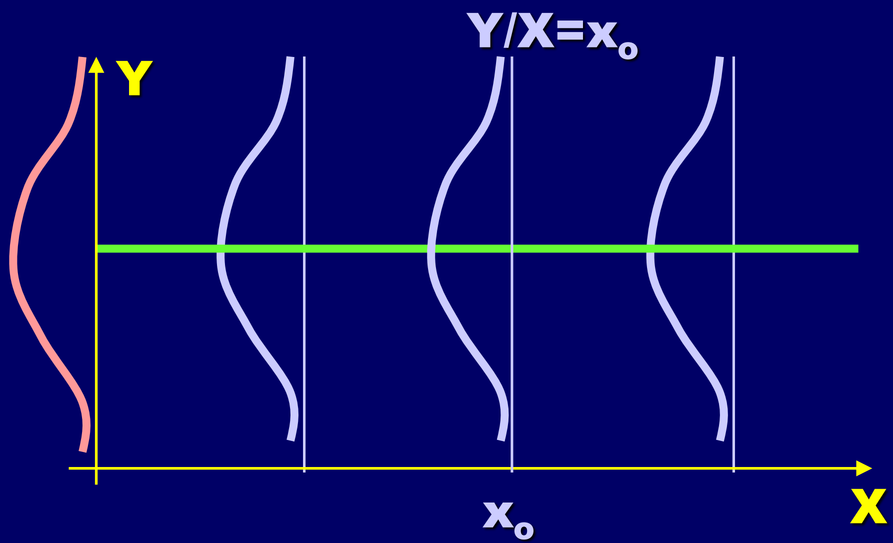
Are they independent?

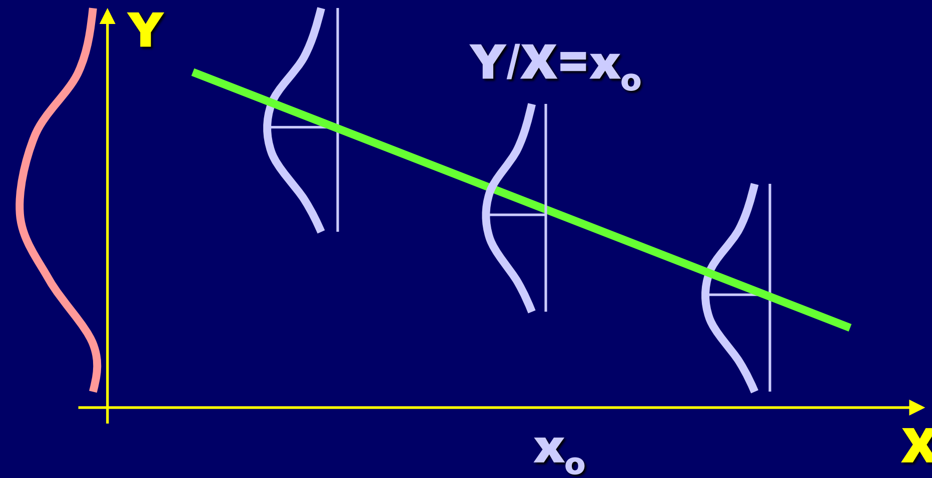
$$f_Y(y) = f\left(y/x = x_o\right)$$



The distribution of time will be the same for any height and the same for the marginal distribution

INTERPRETATION OF r^2





$r^2=0$; $b=0$; $S^2_{res} = S^2_y$

X, Y are independent

$0 < r^2 < 1$; $S^2_{res} < S^2_y$

$$S^2_{residuals} = S^2_Y \cdot (1 - r^2_{XY})$$

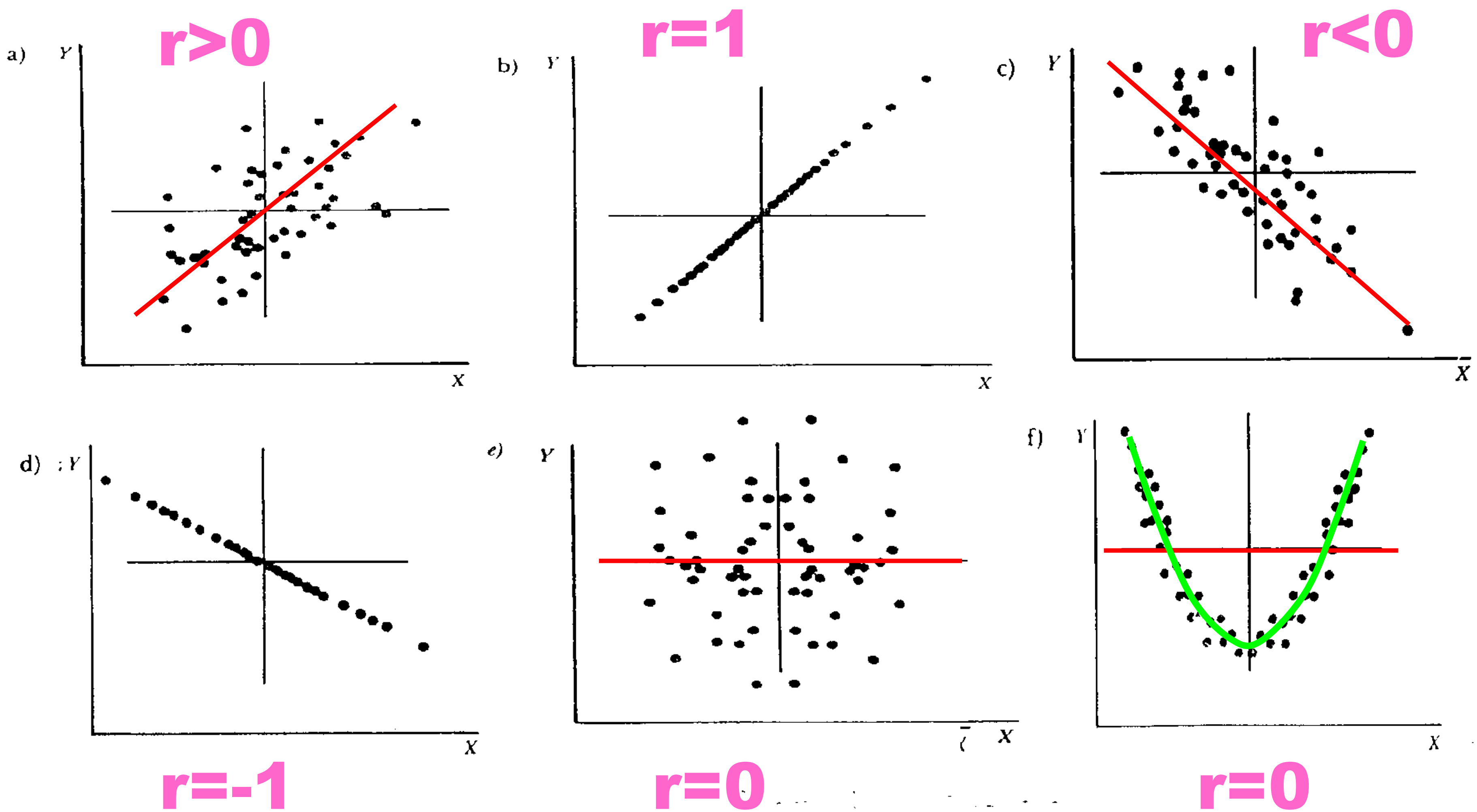
r^2 : proportion of the variability of Y explained by variable X

In simple regression:

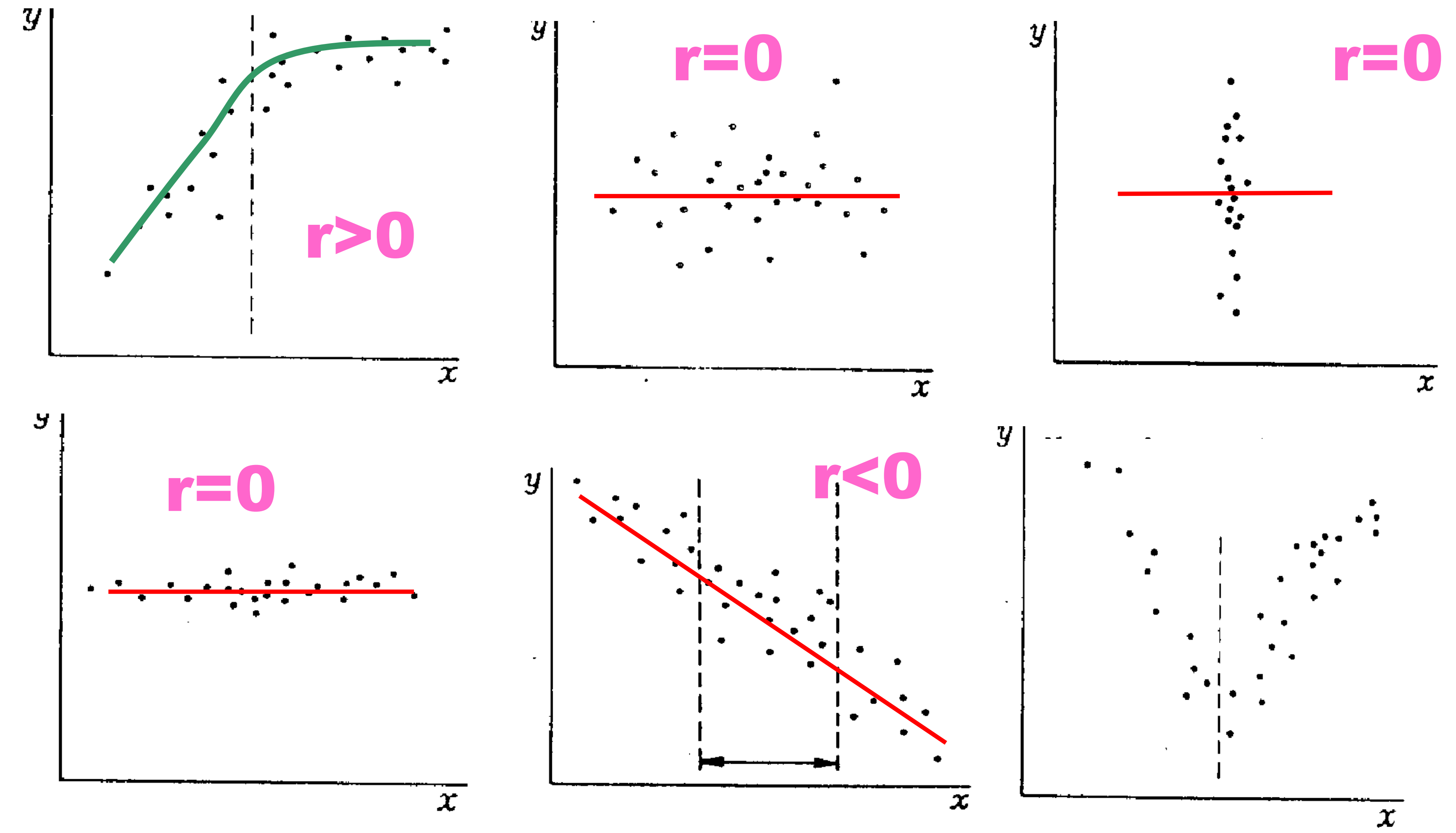
coefficient of determination (R^2) = (correlation coefficient)^2 = r^2

INTERPRETATION OF REGRESSION MODELS

TYPES OF SCATTERPLOTS



TYPES OF SCATTERPLOTS



INTERPRETATION OF RELATIONSHIPS BETWEEN 2 VARIABLES

EXAMPLE: The government wants to promote the use of computers among citizens.

- One study reveals that, in Spanish homes, the annual expenses in computers is positively correlated with the expenses in shoes.

- **CONCLUSION**: ~~the government decides to promote the expenses in shoes (by lowering the prices) in order to incentivate the use of computers.~~

↑ Nº children/family, age, level of studies, family income...

↑ Expenses in shoes

↑ Expenses in computers

Not causal relationship

Partial dependence

VERY IMPORTANT RULE:

THE CORRELATION OBSERVED BETWEEN 2 VARIABLES DOES NOT IMPLY NECESSARILY A CAUSE-EFFECT RELATIONSHIP

INTERPRETATION OF RELATIONSHIPS

One-way causal dependence

CAUSE	EFFECT
Room temperature	Power consumption for heating
Amount of rain	Amount of water in reservoirs
Speed of computer processor	Time required to carry out a computational operation

INTERPRETATION OF RELATIONSHIPS

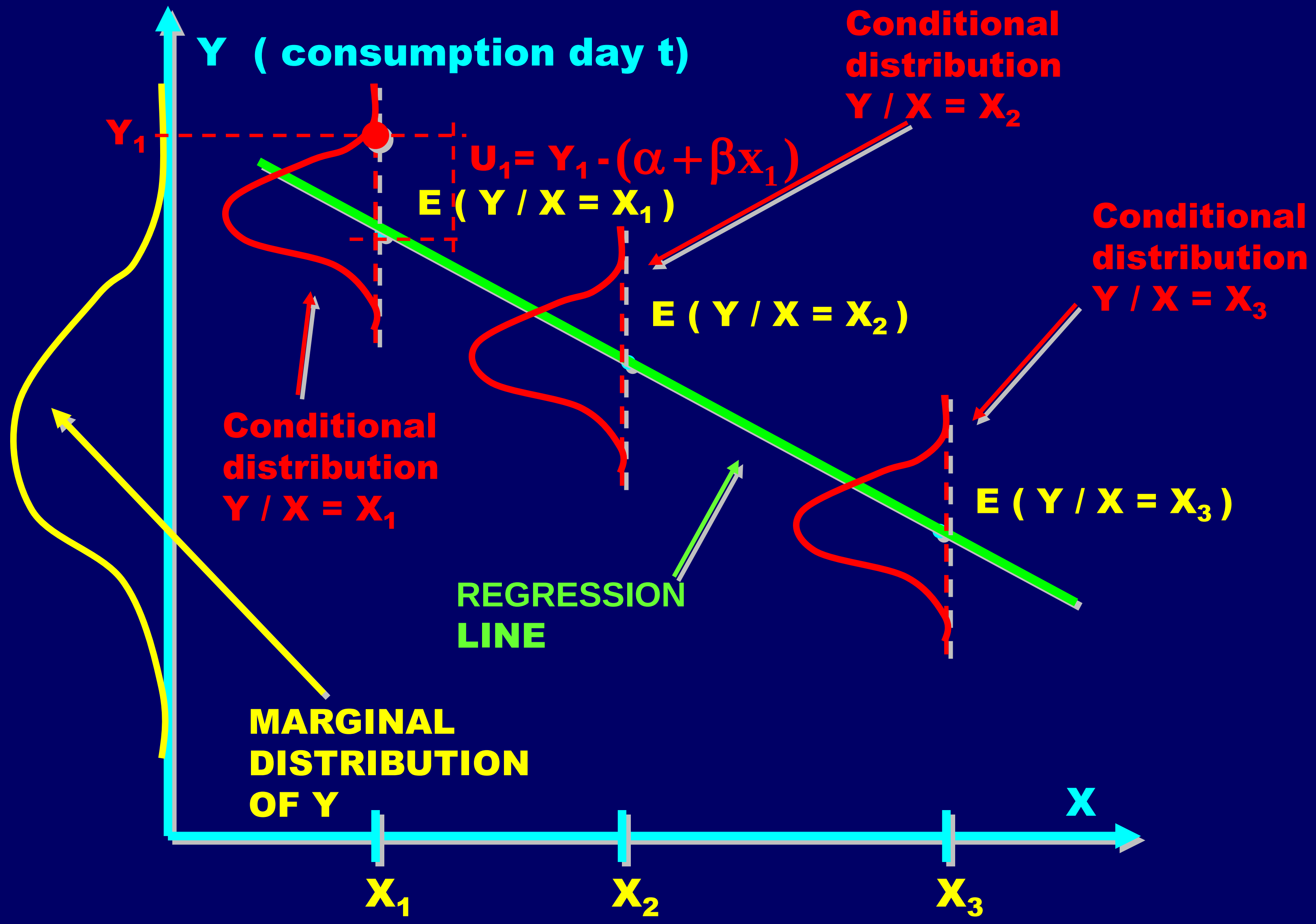
Partial dependence with one or more variables:

CAUSE	EFFECT
Genetic characteristics	Height and weight
Family income	Expenses in shoes of families
Attendance of theory classes	Final score in the exam
Nº hours of study of Statistics	Final score in the exam

Interdependence between both variables:

- Supply and demand of a product
- Levels of sales and expenses in advertising
- Age of husband and wife in a couple

RESIDUALS



Y (consumption day t)
Y_1
$U_1 = Y_1 - (\alpha + \beta x_1)$
$E (Y / X = X_1)$
Conditional distribution $Y / X = X_2$
$E (Y / X = X_2)$
Conditional distribution $Y / X = X_3$
$E (Y / X = X_3)$
Conditional distribution $Y / X = X_1$
REGRESSION LINE
MARGINAL DISTRIBUTION OF Y
X
X_1
X_2
X_3

Residual $u_t = y_{observed} - y_{predicted}$

(example of gas consumption vs Tª):

u_t = consumption observed at day t (y_t) MINUS average consumption estimated when temperature = x_t

$$E(Y/x = x_t) = \alpha + \beta x_t \Rightarrow Y_t = \alpha + \beta x_t + u_t$$

$$\boxed{u_t = y_t - (\alpha + \beta x_t)}$$

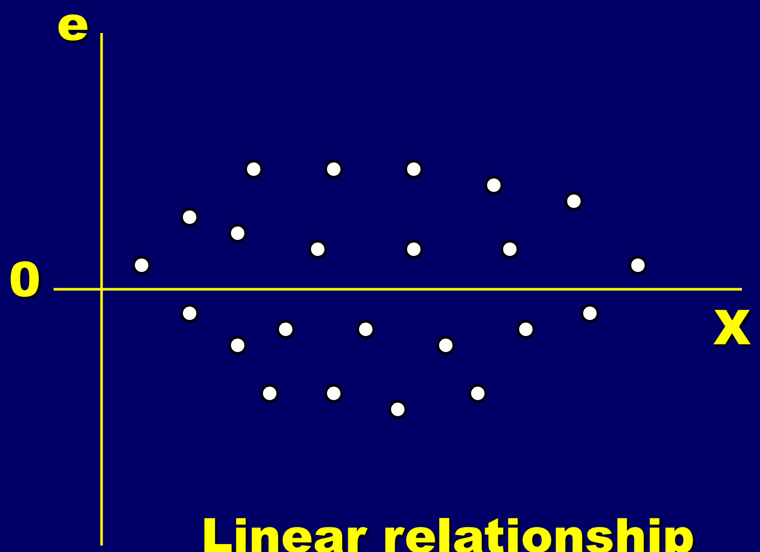
$$E(u_t) = 0 \qquad \sigma^2(u_t) = \sigma^2$$

- IT IS ASSUMED THAT u_t :
  - ARE NORMALLY DISTRIBUTED
  - ARE INDEPENDENT BETWEEN THEM

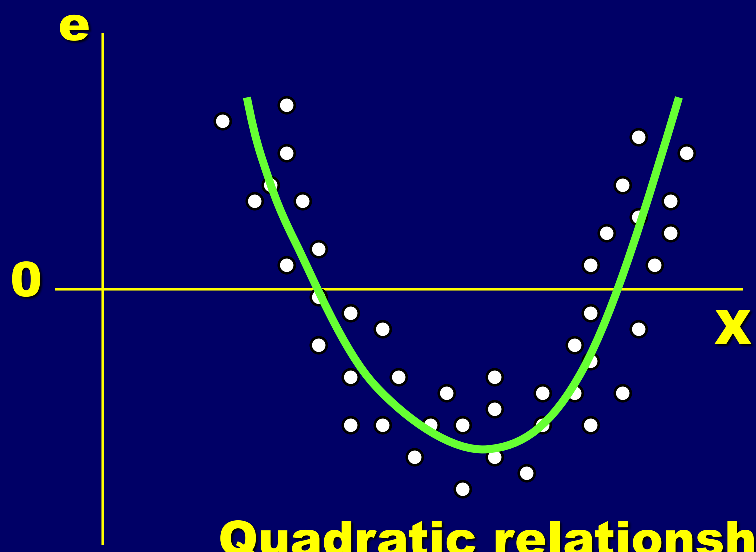
U_t COLLECTS THE EFFECT OF ALL REMAINING FACTORS (NOT INCLUDED IN THE MODEL) OVER GAS COMSUPTION IN A GIVEN DAY t.

ANALYSIS OF RESIDUALS

- **Outliers: are identified with a Normal Probability Plot**
- **Lack of normality in the data: it can be studied by plotting residuals in a Normal Probability Plot.**
- **Lack of linearity in the relationship between E(Y) and X: it can be studied by plotting e_i as a function of X_i**



Linear relationship



Quadratic relationship

ANALYSIS WITH STATGRAPHICS

Data of weight (kg) and height (cm) were collected from students registered in this university certain year.

Data were analyzed by means of a regression analysis using Statgraphics.

a) Weight = a + b · height — What is the interpretation of a?

b) Weight = a + b · (height - 150) — What is the interpretation of a?

What model is more convenient for an easier interpretation of the regression coefficients?

Regression Analysis - Linear Model: Y = a + b X

Dependent Variable: WEIGHT

Independent Variable: HEIGHT - 150

Parameter	Estimate	Standard error	T statistic	P-value
Intercept	46,343	1,7078	27,1355	0,0000
Slope	0,869	0,0695	12,5124	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	8094,44	1	8094,44	156,56	0,0000
Residual	6669,58	129	51,70		
Total	14764,00	130			

Correlation Coeficient 0,7404

R-squared: 54,83 %

Standard Error of Est. =

a) What is the standard deviation of weight for those students with a height of 175 cm?

b) Obtain for the 95% of cases, the weight of students with a height of 175 cm.