2R03 Week 2

Tasks:

→ Read Chapters 3 and 4

→ Go to the Tutorials

→ start doing Assignment # 1
(due Thursday, Jan. 23rd)

www.probandstats3e.nelson.com
(for data sets)

→ play around with Excel and/or Minitab

Chapter 3 Bivariate data

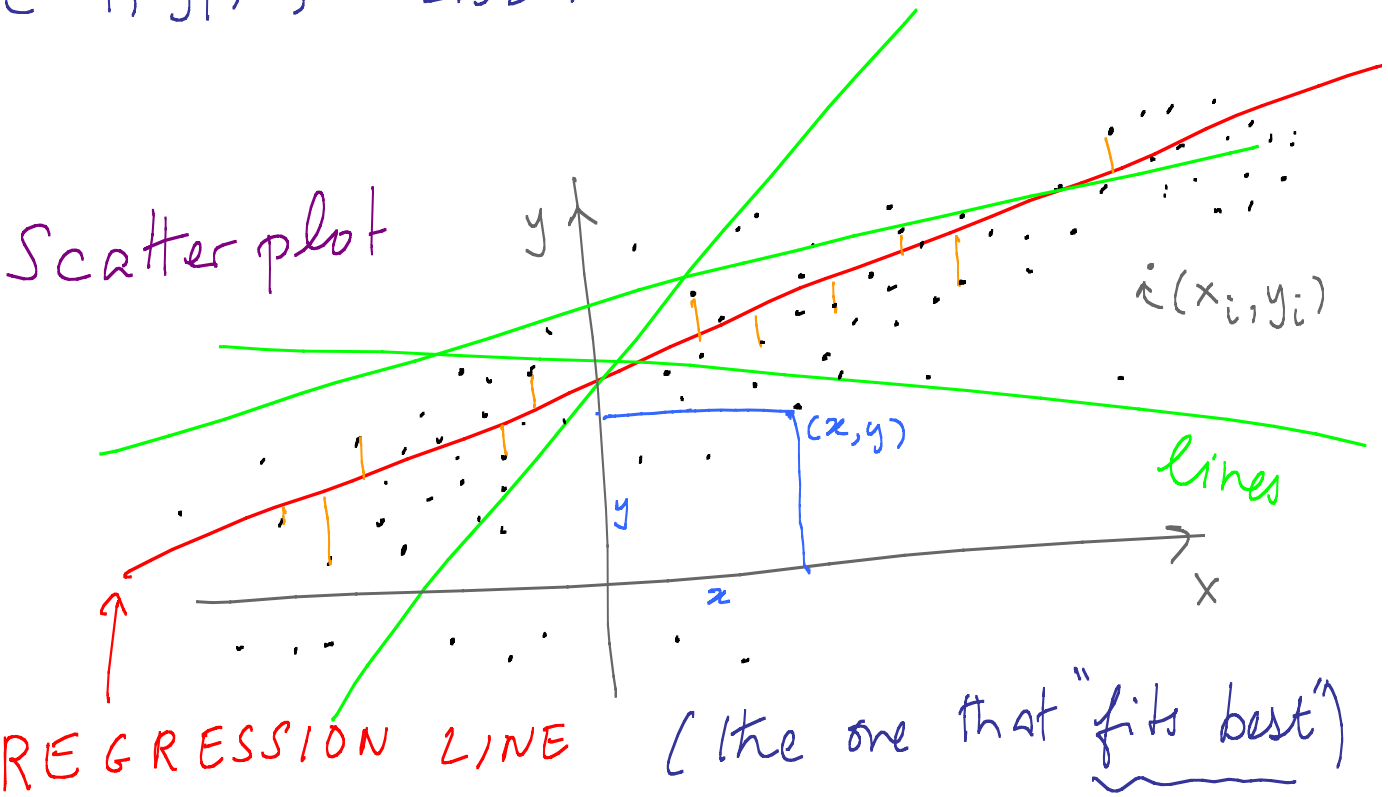
Scatter plots, REGRESSION LINE (visual)

Covariance, Correlation (numerical)

Mean, Variance

A bivariate data set is a set of n points

$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ in $\mathbb{R}^2$

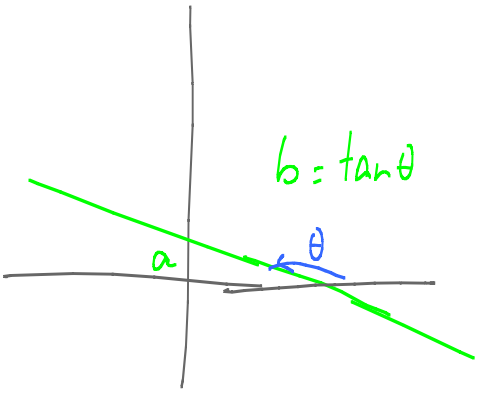


General equation of a line in the plane

$$y = a + bx$$

intercept

slope



The "best fitting" line for a given bivariate data set is given by:

$$b = \frac{S_{xy}}{S_x^2}, \quad a = \bar{y} - b\bar{x}$$

Sample variance of x

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample mean

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum y_i$$

$$\bar{y} = a + b\bar{x}$$

So what's S_{xy}? That's

(sample) COVARIANCE

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

general definition!

(so $S_{xx} = S_x^2$!)

other ways of writing the equation of the regression line:

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x})$$

$$\frac{y - \bar{y}}{S_y} = r_{xy} \frac{x - \bar{x}}{S_x}$$

← my favourite way of writing it! (using z-scores)

so what is r_{xy}? That's called

CORRELATION

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

$$= \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}$$

$$-1 \leq r_{xy} \leq +1$$

(it's actually the cosine of an angle!)

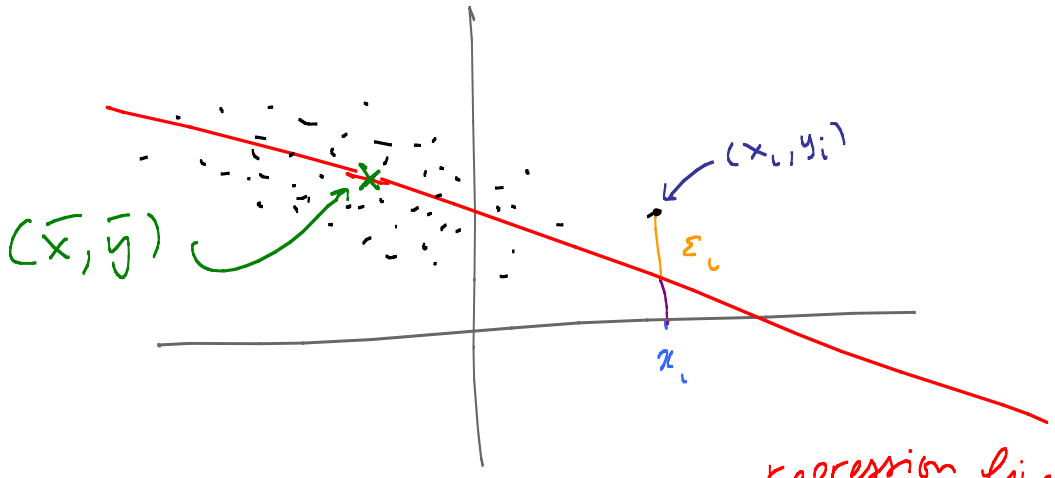
$$\cos \measuredangle(v,w) = \frac{\langle v, w \rangle}{|v|\,|w|}$$

$\langle v, w \rangle$ ← scalar product (inner)

$|v|\,|w|$ ← lengths

Important geometrical fact:

The regression line always passes through the point $(\bar{x}, \bar{y})$ which is the mean or the centre of the data points! $\bar{y} = a + b\bar{x}$



The "errors"

regression line

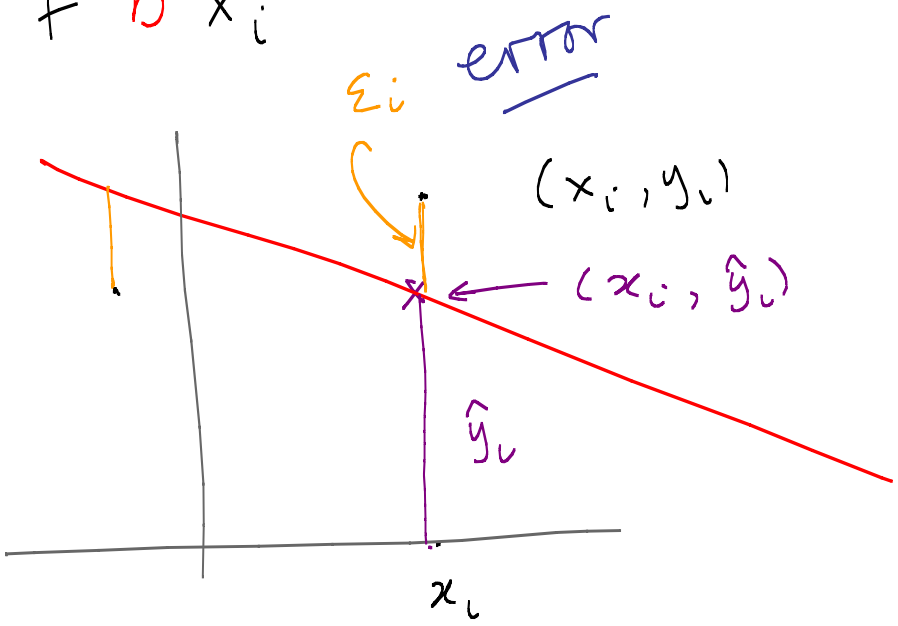
$$\varepsilon_i = y_i - (a + b x_i)$$

$\hat{y}_i$ estimate

$$\varepsilon_i = y_i - \hat{y}_i$$

error = reality − estimate

$$\hat{y}_i = a + b\,x_i$$



The ε_i's could be positive or negative

but they add up to zero.

$$\sum_{i=1}^{n} \varepsilon_i = 0$$

("centered")

"Best fit" means we minimize

$$\sum_{i=1}^{n} \varepsilon_i^2$$

(Least squares)

second fact about the errors (residuals)

$$\sum_{i=1}^{n} x_i \varepsilon_i = 0$$

(scalar product) $= 0$

geometrically this means that the vector $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$ is

Perpendicular to $x = (x_1, \dots, x_n)$

(The vectors are not in $\mathbb{R}^2$!)

Let's centre our variables

$$\tilde{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$$

$$\tilde{y} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$$

This is equivalent to shifting the origin of the coordinate system to $(\bar{x}, \bar{y})$.

Then we have

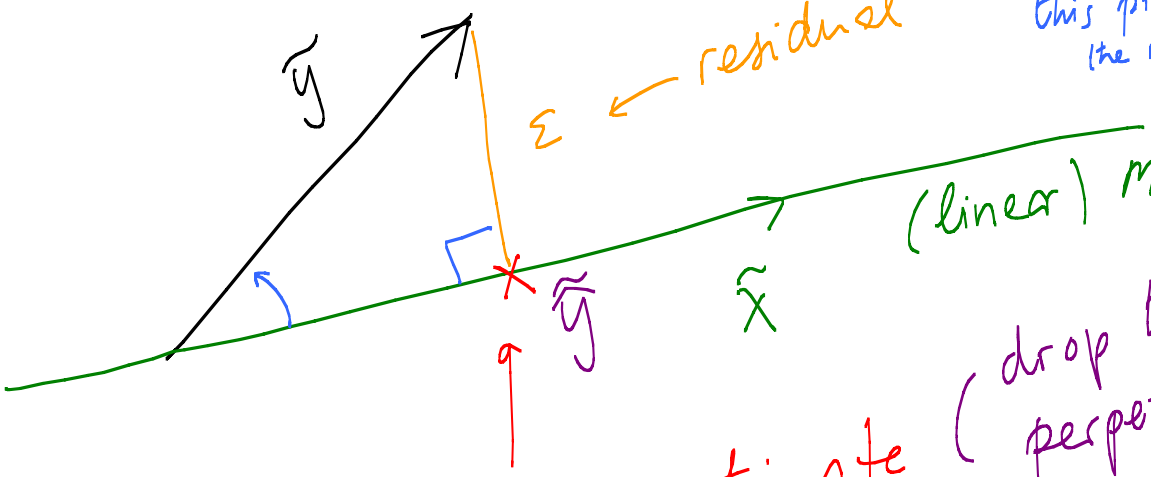
$$\sum \tilde{x}_i = 0 \quad , \quad \sum \tilde{y}_i = 0 \quad , \quad \sum \varepsilon_i = 0$$

$$\sum \tilde{x}_i \varepsilon_i = 0 \qquad \langle \tilde{x}, \varepsilon \rangle = 0$$

and they all lie in an $(n-1)$ dimensional "hyperplane" That's why the sample variances have that factor, $S_x = \frac{1}{n-1} \sum_{i=1}^{n} \tilde{x}_i^2$ etc.

The picture in $\mathbb{R}^n$ (symbolic!)

(don't confuse this picture with the regression line in $\mathbb{R}^2$)



best estimate (drop the perpendicular)

All the vectors are in $\mathbb{R}^n$: $\tilde{y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$ (centered) etc.

THEOREM OF PYTHAGORAS

(over 2,000 years old!)

$$|\tilde{y}|^2 = |\tilde{\hat{y}}|^2 + |\varepsilon|^2$$

$$\cos \sphericalangle(\tilde{x}, \tilde{y}) = r_{xy}$$ (Correlation)

This is actually the main idea behind the material in later chapters (11, 12 and 13!) For example the Pythagorian Theorem will be written

$$SST = SSR + SSE$$

SST: squared sum total; SSR: explained by model; SSE: error

and

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

is called the "coefficient of determination"

(measures the ratio $\frac{\text{Variance explained by model}}{\text{total variance}}$)

understanding linear algebra and geometry "clarifies" a lot of these formulas!