

# MATH1905 Statistics

## Lecture 3

Lecturer: Marc Raimondo  
Carlaw 817

Lectures: Mon.11am (Chem LT1), Tue.8am (Chem LT1)

Student consultations: Tuesday 2-3pm

General information, tutes, solutions etc...

<http://www.usyd.edu.au/u/UG/JM/MATH1905/>

or

First Year Office (FYO), Carlaw 520.

## Lecture 3 (P&Q pp. 20–24)

### Sample mean and variance

For a sample  $x_1, x_2, \dots, x_n$  we define the: **Mean** (a measure of location)

$$\bar{x} = \sum_{i=1}^n x_i / n = \frac{x_1 + \dots + x_n}{n}$$

**Variance** (a measure of spread)

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

For calculations, use

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

**Standard deviation**  $s = \sqrt{s^2}$  (a measure of scale)

## Proof of the computing formula

$$\begin{aligned}
 S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 + (\bar{x})^2 - 2x_i\bar{x}) \\
 &= \sum_{i=1}^n x_i^2 + n \times (\bar{x})^2 - 2 \sum_{i=1}^n (x_i\bar{x}) = \dots + \dots - 2\bar{x} \sum_{i=1}^n x_i \\
 &\quad \text{using } \bar{x} = \sum_{i=1}^n x_i/n \rightarrow \sum_{i=1}^n x_i = n \times \bar{x} \\
 S_{xx} &= \sum_{i=1}^n x_i^2 + n \times (\bar{x})^2 - 2\bar{x}n \times \bar{x} = \sum_{i=1}^n x_i^2 - n(\bar{x})^2
 \end{aligned}$$

**Mean and variance from a frequency table** Suppose  $n$  observations take  $k$  different values,  $x_1, \dots, x_k$  say, with frequencies  $f_1, \dots, f_k$ , where  $\sum_{j=1}^k f_j = n$ . Then

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k f_j x_j$$

$$s^2 = \frac{\sum_{j=1}^k f_j x_j^2 - (\sum_{j=1}^k f_j x_j)^2 / n}{n - 1}$$

## Example: data set 2

$x_i = i$	0	1	2	3	4	5	6	7	8	9	10
$f_i$	57	203	383	525	532	408	273	139	45	27	16

$\bar{x} =$

$$\frac{57 \times 0 + 203 \times 1 + 383 \times 2 + 525 \times 3 + 532 \times 4 + 408 \times 5 + 273 \times 6 + 139 \times 7 + 45 \times 8 + 27 \times 9 + 16 \times 10}{57 + 203 + \dots + 16}$$

$$\bar{x} = \frac{10086}{2608} = 3.8673$$

$$\sum_{j=1}^k f_j x_j^2 = 203 \times 1^2 + 383 \times 2^2 + 525 \times 3^2 + 532 \times 4^2 + 408 \times 5^2 + 273 \times 6^2 + 139 \times 7^2 + 45 \times 8^2 + 27 \times 9^2 + 16 \times 10^2 = 48478$$

$$s^2 = \frac{48478 - 10086^2/2608}{2607} = 3.6333,$$

Conclusion: the sample mean is  $\bar{x} = 3.8673$  and the sample s.d. is  $s = \sqrt{3.6333} = 1.9601$

**Approximations for grouped data** If we assume that the original observations are evenly spread over each interval, with interval centres  $u_1, \dots, u_k$  say, and corresponding frequencies:  $f_1, \dots, f_k$ , we can approximate  $\bar{x}$  and  $s^2$  by  $\bar{u}$  and  $s_u^2$  where

$$\bar{u} = \frac{1}{n} \sum_{j=1}^k f_j u_j \text{ and } s_u^2 = \frac{\sum_{j=1}^k f_j u_j^2 - (\sum_{j=1}^k f_j u_j)^2 / n}{n - 1}$$

Example: rivets (Dataset 3; See P&Q p. 21):

$$\bar{u} = \frac{2 \times 13.12 + 1 \times 13.17 + \dots + 2 \times 13.67}{200} = \frac{2683.1}{200} = 13.4155$$

$$s_u^2 = \frac{2 \times 13.12^2 + \dots + 2 \times 13.67^2 - 2683.1^2 / 200}{199} = 0.01201482$$

## Computational shortcuts

It's easy to check that if  $x_i = a + hd_i$ ,  $i = 1, 2, \dots, n$  then the mean and variance of the  $x$ 's are

$$\bar{x} = a + h\bar{d} \quad \text{and} \quad s^2 = h^2 s_d^2$$

**Example** For data  $x = 1010, 1030, 1040, 980, 990$ , let  $a = 1000$  and  $h = 10$ .

$$\bar{d} = (1 + 3 + 4 - 2 - 1)/5 = 1,$$

$$s_d^2 = \frac{1 + 9 + 16 + 4 + 1 - 25}{5} = 6.5,$$

$$\bar{x} = 1000 + 10 \times 1 = 1010 \quad \text{and} \quad s^2 = 10^2 \times 6.5 = 650.$$

The sample s.d. is  $s = \sqrt{s^2} = 25.4951$

**Proof (change of scale and location)  $x_i = a + hd_i$**

$$\sum_{i=1}^n (a + hd_i) = n \times a + \sum_{i=1}^n (hd_i) = na + h \sum_{i=1}^n d_i$$

$$\bar{x} = \sum_{i=1}^n x_i/n = a + h \sum_{i=1}^n d_i/n = a + h\bar{d}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (a + hd_i - (a + h\bar{d}))^2 = \sum_{i=1}^n (hd_i - h\bar{d})^2$$

$$S_{xx} = h^2 \sum_{i=1}^n (d_i - \bar{d})^2 = h^2 S_{dd}$$

**Median or mean?** Both the median and the mean are measures of location, but which is preferable?

For symmetric data, the mean is usually less variable from sample to sample than the median.

For skewed data, the median is a better measure of location.

The median is not affected as much as the mean by outliers. This property of the median is known as 'robustness'.

The mean is easier to compute than the median and is much easier to handle theoretically.

Note also that the sample standard deviation  $s$  and the IQR ( $= Q_3 - Q_1$ ) are both measures of spread.

**Lecture 4 Bivariate data.** Statisticians often investigate a bivariate sample

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , either with the purpose of measuring the *strength* of the relationship or of *specifying* the relationship.

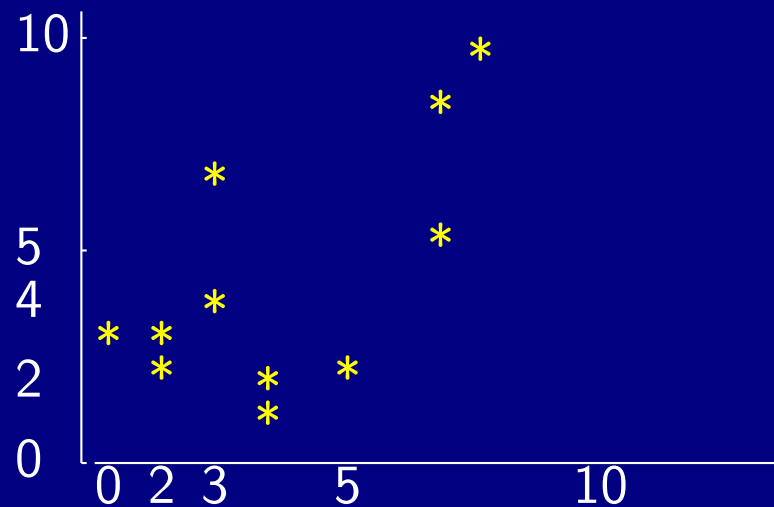
$x$	Cholesterol Level	Inflation rate	Weight
$y$	Blood Pressure	Unemployment	Height

The simplest presentation of bivariate data is a **scatterplot** of the points  $(x_i, y_i), i = 1, \dots, n$ . What do the following plots reveal?

Figure 1.8 Typical Scatterplots

## Example 1

Produce a scatterplot for the following data  $(x_i, y_i)$ . (2,2) (3,4) (7,9) (4,1)  
(4,2) (4,4) (3,7) (7,5) (8,10) (1,3) (5,2) (5,9) (2,3)



\*

**Correlation** If a linear relationship is suggested by the scatterplot, it would be useful to have a measure of the strength of the linear relationship. **Definition** The correlation coefficient,  $r$ , is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

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

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

## Calculation formulae

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n y_i \right)^2$$

## Example 1 (cont') $n = 13$ pairs

$$(x_i, y_i) = (2,2) (3,4) (7,9) (4,1) (4,2) (4,4) (3,7) (7,5) (8,10) (1,3) (5,2) \\ (5,9) (2,3)$$

$$\sum_{i=1}^{13} x_i y_i = 2 \times 2 + 3 \times 4 + 7 \times 9 + 4 \times 1 + 4 \times 2 + 4 \times 4 + \\ 3 \times 7 + 7 \times 5 + 8 \times 10 + 1 \times 3 + 5 \times 2 + 5 \times 9 + 2 \times 3 = 307$$

$$\sum_{i=1}^{13} x_i = 2 + 3 + 7 + 4 + 4 + 4 + 3 + 7 + 8 + 1 + 5 + 5 + 2 = 55$$

$$\sum_{i=1}^{13} y_i = 2 + 4 + 9 + 1 + 2 + 4 + 7 + 5 + 10 + 3 + 2 + 9 + 3 = 61$$

$$\sum_{i=1}^{13} x_i^2 = \\ 2^2 + 3^2 + 7^2 + 4^2 + 4^2 + 4^2 + 3^2 + 7^2 + 8^2 + 1^2 + 5^2 + 5^2 + 2^2 = 287$$

$$\sum_{i=1}^{13} y_i^2 = 2^2 + 4^2 + 9^2 + 1^2 + 2^2 + 4^2 + 7^2 + 5^2 + 10^2 + 3^2 + 2^2 + 9^2 + 3^2 = 399$$

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum x_i \sum y_i = 307 - \frac{1}{13} 55 \times 61 = 48.92$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 = 287 - 55^2/13 = 54.31$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - \frac{1}{n} (\sum_{i=1}^n y_i)^2 = 399 - 61^2/13 = 112.77$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{48.92}{\sqrt{54.31 \times 112.77}} = 0.625 \quad (3dp)$$

**Interpretation of correlation** The correlation coefficient  $r$  measures the strength of the linear relationship between the points  $\{(x_i, y_i), i = 1, \dots, n\}$ . A value close to  $\pm 1$  indicates that nearly all the variability in  $y$  is explained by a linear relationship between  $x$  and  $y$ . **However, it is important to look at a scatterplot when interpreting  $r$ .**

In each of the following three plots the correlation coefficient  $r$  is approximately 0.7 ....SEE [PQ] p.27

## Proof of calculation formula

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

using  $\bar{x} = \sum_{i=1}^n x_i / n \rightarrow \sum_{i=1}^n x_i = n \times \bar{x}$

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

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

## Proof $r \geq -1$

$$A = \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sqrt{S_{xx}}} + \frac{y_i - \bar{y}}{\sqrt{S_{yy}}} \right)^2 \geq 0$$

$$A = \sum_{i=1}^n \left( \frac{(x_i - \bar{x})^2}{S_{xx}} + \frac{(y_i - \bar{y})^2}{S_{yy}} + 2 \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \right)$$

$$A = 1 + 1 + 2 \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \geq 0$$

$$A = 2 + 2r \geq 0$$

$$\rightarrow r \geq -1$$

## Proof $r \leq 1$

$$B = \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sqrt{S_{xx}}} - \frac{y_i - \bar{y}}{\sqrt{S_{yy}}} \right)^2 \geq 0$$

$$B = \sum_{i=1}^n \left( \frac{(x_i - \bar{x})^2}{S_{xx}} + \frac{(y_i - \bar{y})^2}{S_{yy}} - 2 \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \right)$$

$$B = 1 + 1 - 2 \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \geq 0$$

$$B = 2 - 2r \geq 0$$

$$r \leq 1$$

**Proof**  $r(h\mathbf{x}, \mathbf{y}) = r(\mathbf{x}, \mathbf{y})$ 

Let  $h > 0$ ,  $h\mathbf{x} = (hx_1, hx_2, \dots, hx_n)$

$$S(h\mathbf{x}, \mathbf{y}) = hS(\mathbf{x}, \mathbf{y})$$

$$S(h\mathbf{x}, h\mathbf{x}) = h^2S(\mathbf{x}, \mathbf{x})$$

$$r(h\mathbf{x}, \mathbf{y}) = \frac{hS_{xy}}{\sqrt{h^2S_{xx}S_{yy}}} = r(\mathbf{x}, \mathbf{y})$$

# Lecture 4/5 (P&Q pp. 25–30)

## Linear regression

If the scatter plot indicates a linear pattern and if the correlation coefficient is away from zero. We can fit a **Linear regression of  $y$  on  $x$ :  $y = a + bx$**  choosing  $a$  and  $b$  to minimise the sum of squared residuals:

$$M = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

to find  $a$  and  $b$  we compute  $\frac{\partial M}{\partial a}$  and  $\frac{\partial M}{\partial b}$  and set these equal to zero

## Example: Data on slide 2 of Lec.4

$$\bar{x} = (2 + 3 + \dots + 5 + 2)/13 = 55/13 = 4.2308$$

$$\bar{y} = (2 + 4 + \dots + 9 + 3)/13 = 61/13 = 4.6923$$

$$S_{xx} = 4 + 9 + \dots + 25 + 25 + 4 - 55^2/13 = 54.3077$$

$$S_{yy} = 4 + 16 + \dots + 4 + 81 + 9 - 61^2/13 = 112.7692$$

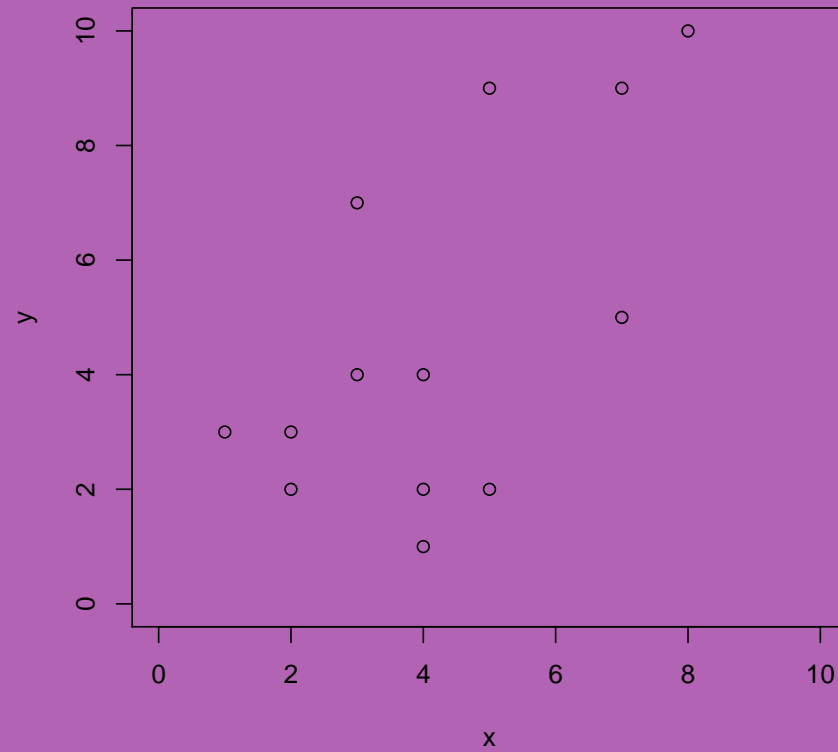
$$S_{xy} = 4 + 12 + \dots + 45 + 6 - 55 \times 61/13 = 48.9231$$

$$r = 48.9231 / \sqrt{54.3077 \times 112.7692} = 0.6252$$

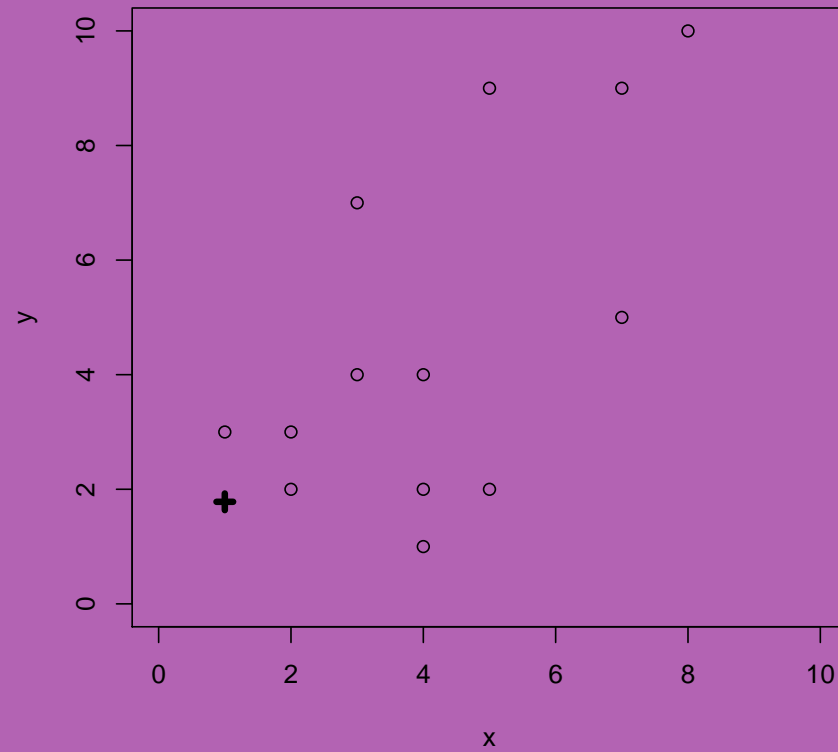
$$b = 48.9231 / 54.3077 = 0.90085$$

$$a = \bar{y} - b\bar{x} = 4.6923 - 0.90085 \times 4.2308 = 0.8810$$

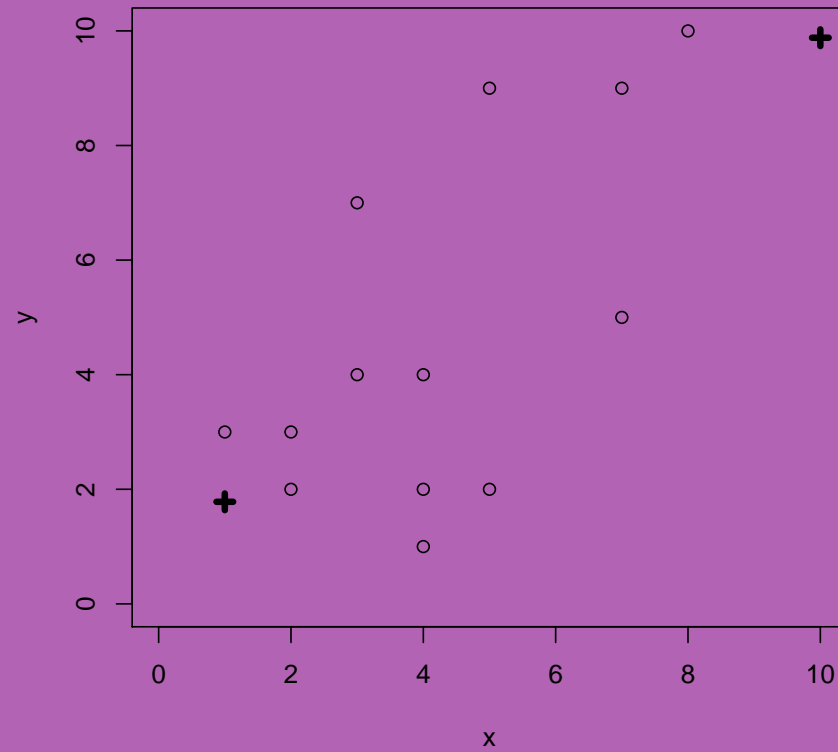
(2,2) (3,4) (7,9) (4,1) (4,2) (4,4) (3,7) (7,5) (8,10) (1,3) (5,2) (5,9)  
(2,3)



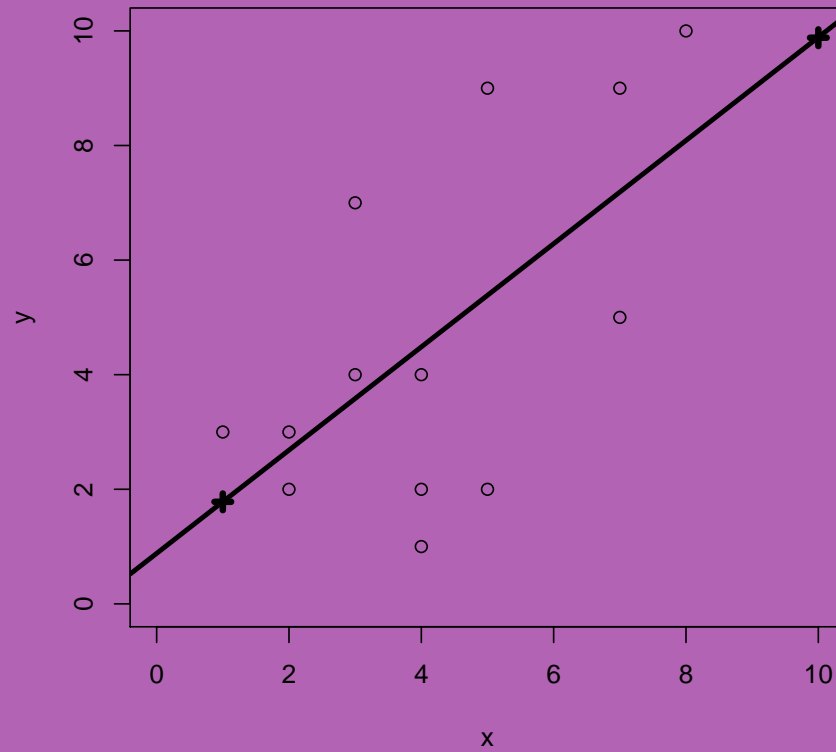
$$x = 1, y = 0.9 \times 1 + 0.881 = 1.781$$



$$x = 10, y = 0.9 \times 10 + 0.881 = 9.881$$



# Regression line



## Computing formula for **a** and **b** (proof)

$$M(a, b) = \sum_{i=1}^n (y_i - a - bx_i)^2$$
$$= \sum y_i^2 + na^2 + b^2 \sum x_i^2 + 2ab \sum x_i - 2a \sum y_i - 2 \sum x_i y_i$$

$$\frac{\partial M}{\partial b} = 2b \sum x_i^2 + 2a \sum x_i - 2 \sum x_i y_i = 0, \quad a = \bar{y} - b\bar{x}$$

$$b \sum x_i^2 + (\bar{y} - b\bar{x}) \sum x_i - \sum x_i y_i = 0$$

$$b(\sum x_i^2 - \bar{x} \sum x_i) = \sum x_i y_i - \bar{y} \sum y_i / n$$

$$bS_{xx} = S_{xy} \quad , \quad b = \frac{S_{xy}}{S_{xx}} \quad (2)$$

# In R

```
> cor(x,y)
[1] 0.6251551
> lm(y~x)
      Intercept          X
0.8810198 0.9008499
```

## Lecture 5/6 Linear regression (cont')

Recall that the *least squares regression line*  $y = a + bx$  has  $a$  and  $b$  chosen to minimise  $\sum_{i=1}^n e_i^2$ , where  $e_i = y_i - \hat{y}_i$  and  $\hat{y}_i = a + bx_i$ . The values of  $a$  and  $b$  are  $b = \frac{S_{xy}}{S_{xx}}$  and  $a = \bar{y} - b\bar{x}$ . For the previous numerical example, we can now draw the regression line on the scatterplot:

We next ask: Is the line a good fit?

The proportion of variability of  $y$ 's explained by the regression on  $x$  is  $r^2$  (proof next) Here,

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{48.9231^2}{54.3077 \times 112.7692} = 0.39$$

... so about 40% of the variability of the  $y$ 's is explained by the  $x$ 's (line). it is a reasonable fit...

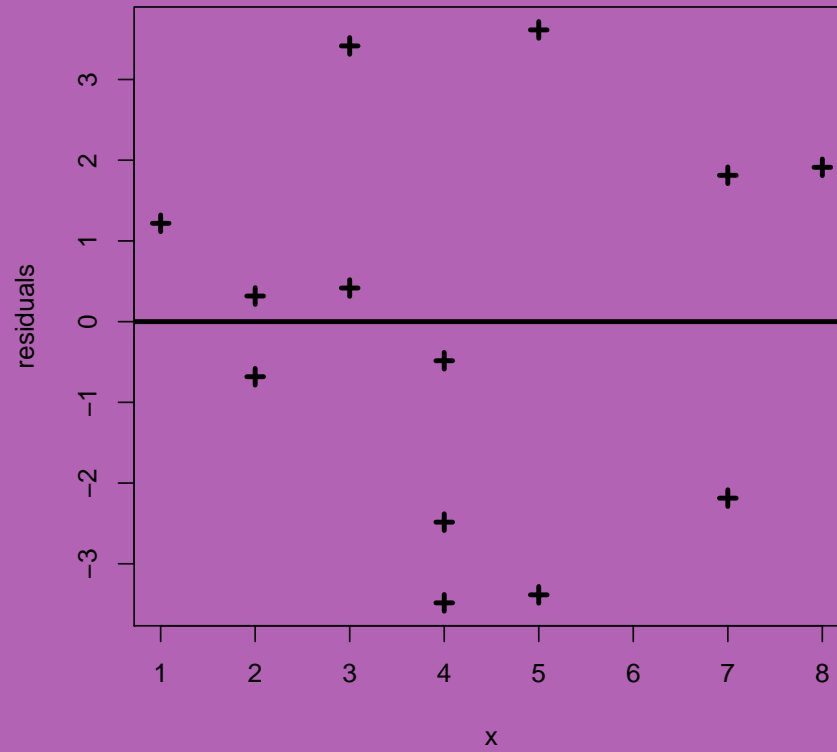
**The residuals**  $e_1, \dots, e_n$  are computed from the data; here (2,2) (3,4) (7,9) (4,1) (4,2) (4,4) (3,7) (7,5) (8,10) (1,3) (5,2) (5,9) (2,3) using the regression line.

e.g.  $y_1 - (0.881 + 0.9 \times x_1) = 2 - (0.881 + 0.9 \times 2) = -0.68$

**The residuals** should be plotted against  $x_1, \dots, x_n$  and should show no pattern, and should be evenly distributed around the zero horizontal line.

Draw this residual plot p.32

**shows no pattern.**

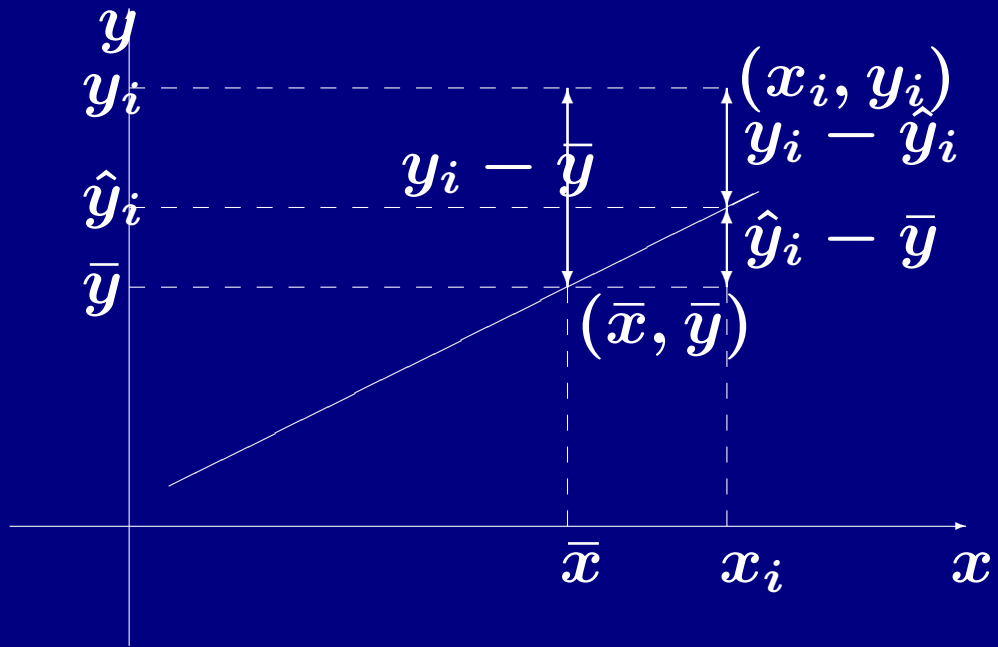


**Proportion of the variability** of  $y$ 's explained by  $x$  in least squares regression. The variance of the  $y$ 's is  $S_{yy}/(n - 1)$ . **we will show that** the *regression sum of squares* is  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = S_{xy}^2/S_{xx}$ .

**we will show that**

$S_{yy} = \sum_{i=1}^n e_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \text{Res SoS} + \text{Reg SoS}$  Thus the variability of the  $y$ 's is reduced from  $S_{yy}$  to the *residual sum of squares*  $\sum_{i=1}^n e_i^2$  and the proportional reduction is

$$\frac{S_{xy}^2/S_{xx}}{S_{yy}} = r^2.$$



**Assessing the regression fit** We conclude that the linear-regression-fit is appropriate if

- The residual plot shows no pattern
- The residual boxplot is symmetric w.r.t 0 and the residuals histogram is bell shaped.
- The smaller the residuals IQR (or standard deviation) is the better is the fit (larger  $r^2$ )

**Prediction** If the regression seems satisfactory then the  $y$ -values can be predicted for given  $x$ -values by plugging in. Note that the new  $x$ -values should not be too far away from the interval  $(\min x_i, \max x_i)$ .

### Example (ctd.)

Recall for our example, the regression line of  $y$  on  $x$  is

$$y = 0.881 + 0.9x.$$

So for  $x = 0$  and  $10$  we predict  $y$ -values of **0.88**, **9.881**

## R commands



```
y.lm=lm(y~x)
```



```
y.lm.res=lm(y~x)$res
```



```
plot(x,y.lm.res)
```

- `boxplot(y.lm.res)`

- `summary(y.lm.res)`

- `hist(y.lm.res)`

## Proof $S_{yy} = \text{Res SoS} + \text{Reg SoS}$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

the result will follow if we prove that the last term is zero

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum e_i(\hat{y}_i - \bar{y}) = (a - \bar{y}) \sum e_i + b \sum e_i x_i$$

first we show that  $\sum e_i$  is zero

$$e_i = y_i - a - b x_i = y_i - \bar{y} + \bar{x} - b x_i$$

$$\sum e_i = \sum (y_i - \bar{y}) + n \times b \bar{x} - b \sum x_i$$

$$= \sum y_i - n \sum y_i / n + n \times b \bar{x} - n \times b \sum x_i / n = 0$$

This proof fits p.39-42

now we show that  $\sum(e_i x_i)$  is zero

$$x_i e_i = x_i y_i - x_i \bar{y} + b \bar{x} x_i - b x_i^2$$

$$\sum_{i=1}^n x_i e_i = \sum x_i y_i - \sum x_i \sum y_i / n + \frac{S_{xy}}{S_{xx}} (\sum x_i)^2 / n - \frac{S_{xy}}{S_{xx}} \sum x_i^2$$

$$\sum_{i=1}^n x_i e_i = S_{xy} - \frac{S_{xy}}{S_{xx}} (\sum x_i^2 - (\sum x_i)^2 / n) =$$

$$\sum_{i=1}^n x_i e_i = S_{xy} - \frac{S_{xy}}{S_{xx}} \times S_{xx} = 0$$

This proof fits p.39-42

$$\text{Proof } \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = S_{xy}^2 / S_{xx}$$

$$(\hat{y}_i - \bar{y})^2 = (a + bx_i - \bar{y})^2 = (\bar{y} - b\bar{x} + \frac{S_{xy}}{S_{xx}}x_i - \bar{y})^2$$

$$(\hat{y}_i - \bar{y})^2 = (\bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} + \frac{S_{xy}}{S_{xx}}x_i - \bar{y})^2 = ((x_i - \bar{x}) \frac{S_{xy}}{S_{xx}})^2$$

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum (x_i - \bar{x})^2 \left(\frac{S_{xy}}{S_{xx}}\right)^2 = \frac{S_{xy}^2}{S_{xx}}$$

from which we deduce that

$$\frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx} \times S_{yy}} = r^2$$