Correlation & Regression

Mark Muldoon

Departments of Mathematics and

Optometry & Neuroscience

UMIST

http://www.ma.umist.ac.uk/mrm/Teaching/2P1/

Overview

Today we'll conclude our study of statistics by learning how to fit lines to data.

- **Motivating problem:** do two instruments agree?
- **On the nature of lines:** slopes, intercepts and the sense of "best".
- **Instruments again:** fitting lines and testing for significance.

Do two instruments agree?

This is a very common question in science and engineering as one often needs to know whether a new instrument or methodology agrees with existing practice. Here we examine data from a comparison of two methods for measuring peak expiratory flow rate (PEFR)

- One measurement from the Wright peak flow meter
- Other from a newer instrument, the “minimeter”
- All measurements in (litres / minute)
- Full study reported in H.G. Oldham, M.M. Bevan and M. McDermott, “Comparison of the new miniature Wright peak flow meter with the standard Wright peak flow meter”, *Thorax*, **34**, pp. 807-808.

The data

PEFR (ltrs / min)				PEFR (ltrs / min)			
Orig.	Mini	mean	difference	Orig.	Mini	mean	difference
x	y	$(x+y)/2$	$(x-y)$	x	y	$(x+y)/2$	$(x-y)$
494	512	503	-18	433	445	439	12
395	430	412.5	-35	417	432	424.5	-15
516	520	518	-4	656	626	641	30
434	428	431	6	267	260	263.5	7
476	500	488	-24	478	477	477.5	1
557	600	578.5	-43	178	259	218.5	-81
413	364	388.5	49	423	350	386.5	73
442	380	411	62	427	451	439	-24
650	658	654	-8				

The means of the two measurements in each pair, along with their differences, will prove useful later in the lecture.

Statistical questions: bias

The question “Do the two instruments agree?” has two aspects:

(1) Is there a *bias*? That is, is there an overall tendency for one of the two measurements to give larger results?

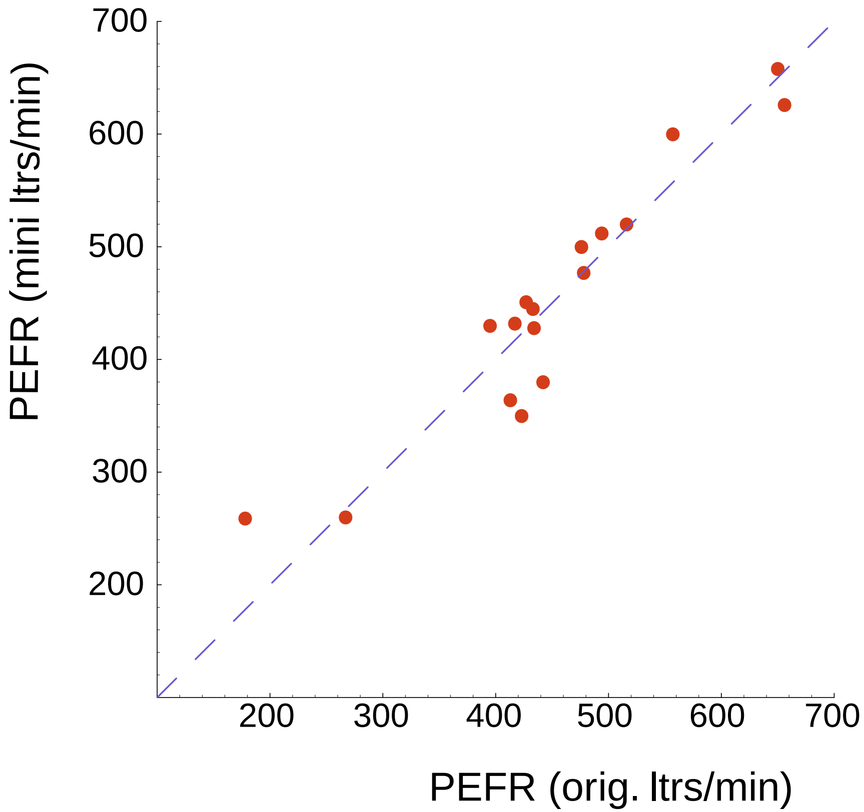
- A job for the paired-sample t-test;
- For these data $t = 2.12$ and $\nu = 16$ so we cannot reject the null: the data are consistent with the hypothesis that the two measurements are drawn from the same distribution.
- Should construct confidence interval for the mean difference and see if it is acceptable in typical applications. Original study used a much larger sample and established that any bias is very small.

Statistical questions: continued

(2) Is there any relation between differences and measurements?

This latter question involves the subjects of today's lecture: correlation and line-fitting.

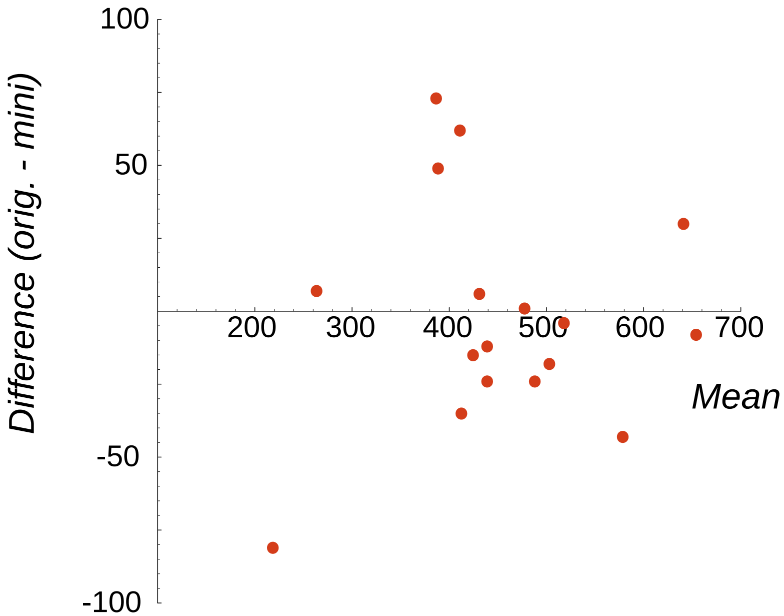
A first, graphical approach



Plot pairs of measurements (orig, mini)

- If agreement were perfect all data would fall on line $y = x$.
- Interesting details have to do with fluctuations about this ideal

A better graphical approach



Plot pairs of the form

(mean, difference)

or, using the coordinates from the previous plot

$$\left(\frac{(x+y)}{2}, (x-y)\right)$$

- Now ideal is a *horizontal* line with constant difference of zero.
- Graph highlights any systematic variation in differences.

About lines

Lines give a relationship between two variables, conventionally called x (on the horizontal axis) and y (on the vertical) of the form:

$$y = mx + b$$

- The parameter b, called the y-intercept, gives the value of y at which the line crossed the vertical axis.
- The parameter m, also called the *slope* describes how y changes when x increases.

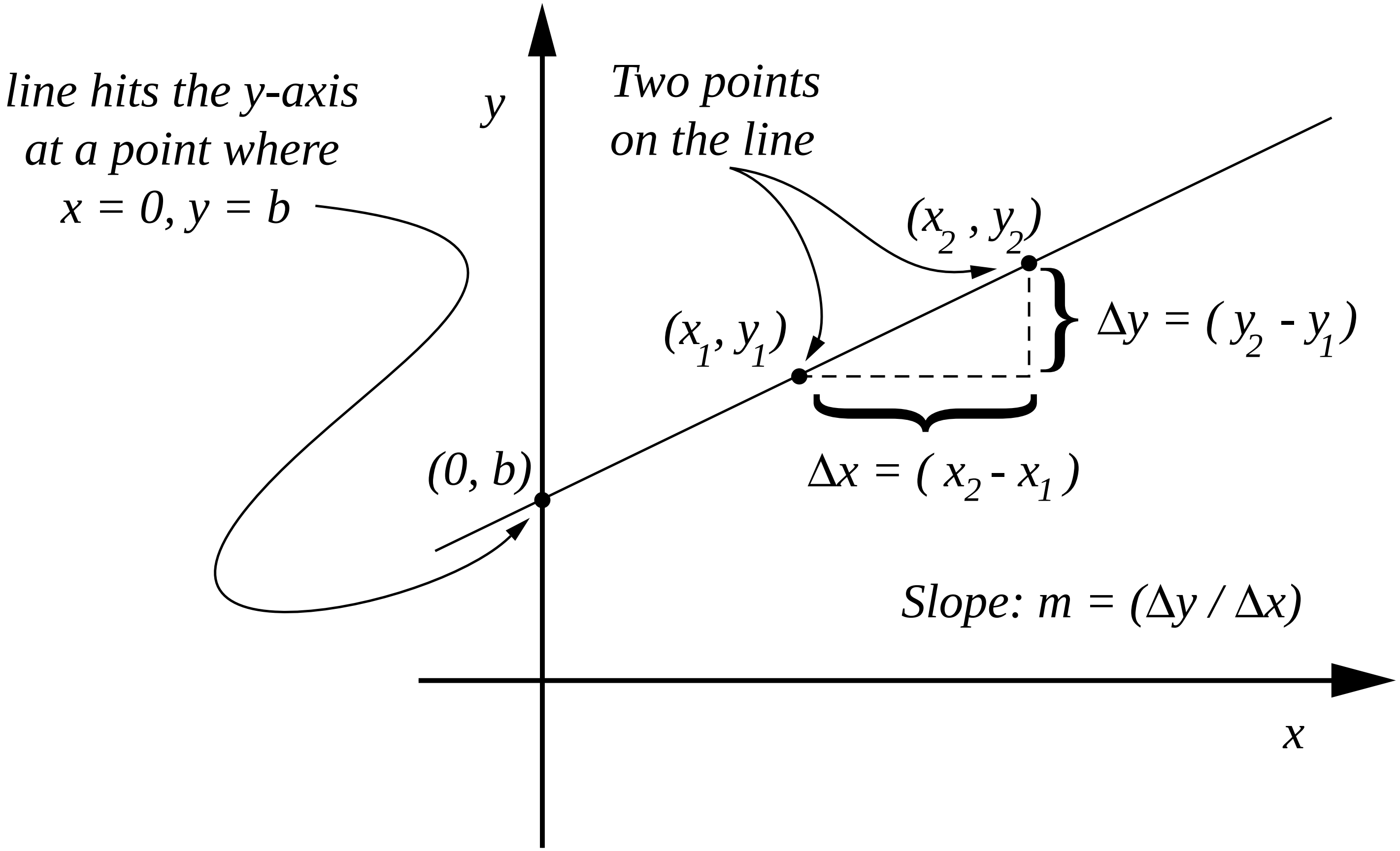
About the slope

- Slope gives rate at which line rises as one moves left-to-right across the plot: larger slope means steeper rise.
- Zero slope means a horizontal line; negative slope means line descends from left to right.
- Given two points (x_1, y_1) and (x_2, y_2) on the line one can compute the slope according to the following formulae

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} = \frac{\text{(vertical) rise}}{\text{(horizontal) run}}$$

- Slope has units given by the ratio of the units on the vertical and horizontal axes.

Lines and slopes in pictures



We will be interested in fitting lines to a set of data points and, given a set of data, will find a formula describing the line that is the “best”-fitting of all possible lines.

Lines and data

Given a collection of N data points $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, and some a particular choice of slope m and intercept b we get, for each observed x_j

- an observed value of y-value, y_j, and
- a predicted y-value given by

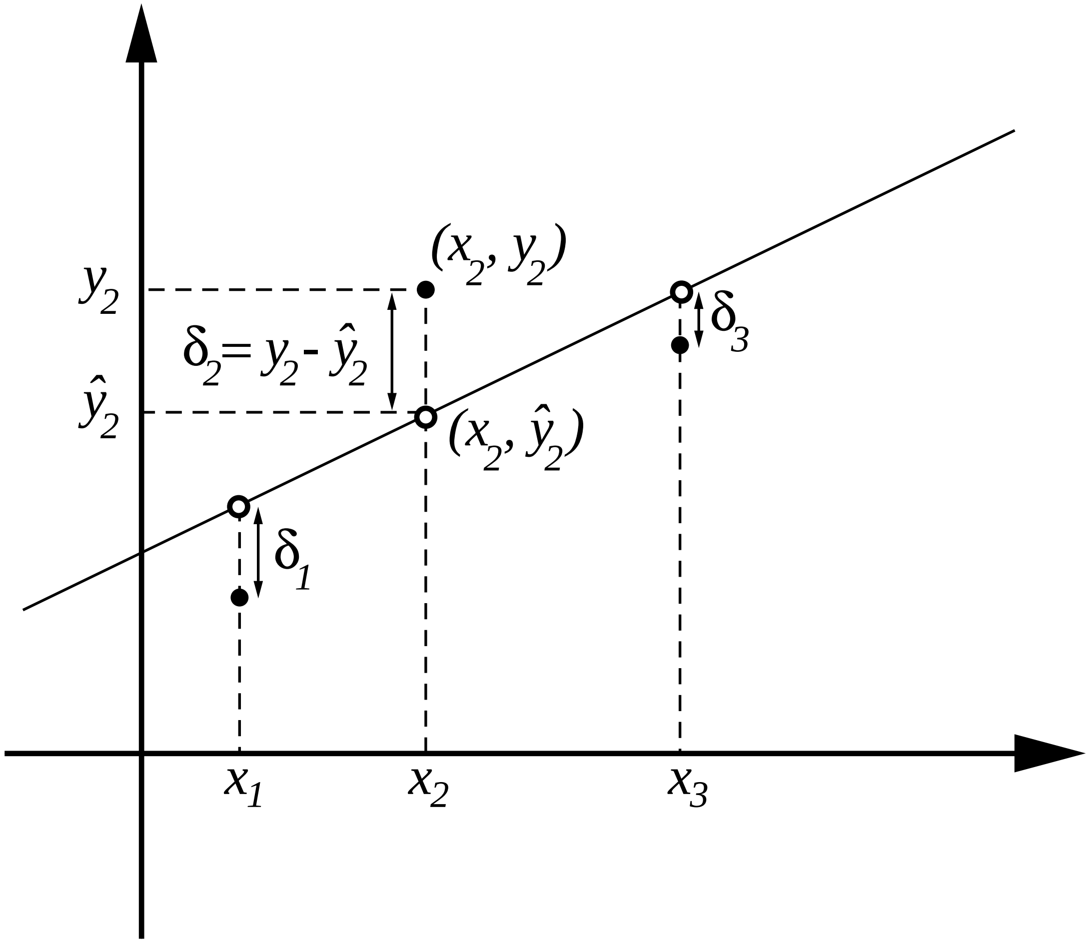
$$\hat{y}_j = mx_j + b$$

Call the difference between these two:

$$\delta_j = y_j - \hat{y}_j$$

The sense of “best”

We will learn to choose the slope m and intercept b so as to minimize the sum-of-squares of these *vertical* deviations of the data from the line.



The sense of “best”, continued

The best-fit line will be the one where m and b are chosen to minimize

$$
\begin{aligned}
\chi^2 &= \sum_{j=1}^{N} (\text{(observed } y) - \text{(predicted } y))^2 \\
&= \sum_{j=1}^{N} (y_j - \hat{y}_j)^2 \\
&= \sum_{j=1}^{N} (y_j - (mx_j + b))^2
\end{aligned}
$$

Remarks

Notice that we only pay attention to the differences between observed and predicted y-values. There are implicit assumptions that the x-values are known more accurately or are more under our control than the y-values and that the y's depend on the x's, thus

- y is sometimes called the *dependent variable*;
- x is called the *independent variable*.

Detailed calculations

The single ingredient is a list of, say, N, pairs of numbers $\{(x_1, y_1), \ldots, (x_N, y_N)\}$.

To begin with, one computes the following five sums:

$$\Sigma_x \equiv \sum_{j=1}^{N} x_j \qquad \Sigma_y \equiv \sum_{j=1}^{N} y_j$$

$$\Sigma_{xx} \equiv \sum_{j=1}^{N} x_j^2 \qquad \Sigma_{xy} \equiv \sum_{j=1}^{N} x_j y_j \qquad \Sigma_{yy} \equiv \sum_{j=1}^{N} y_j^2$$

Everything else can be derived from these.

Detailed calculations, continued

The best fit line has slope m and y-intercept b given by

$$m = \frac{N\Sigma_{xy} - \Sigma_x\Sigma_y}{N\Sigma_{xx} - (\Sigma_x)^2};$$

$$b = \frac{\Sigma_{xx}\Sigma_y - \Sigma_x\Sigma_{xy}}{N\Sigma_{xx} - (\Sigma_x)^2}.$$

The correlation coefficient

Additionally one can compute a number called the *correlation coefficient*, r, which is given by

$$r = \frac{N\Sigma_{xy} - \Sigma_x \Sigma_y}{\sqrt{N\Sigma_{xx} - (\Sigma_x)^2}\sqrt{N\Sigma_{yy} - (\Sigma_y)^2}}.$$

This number

- varies between 1 and -1;
- if $|r|$ = 1 the best-fit line runs straight through all the data without any errors;
- r = -1 indicates that y decreases as x increases while r = 1 indicates that y increases as x increases.

Testing r

One can test whether r is significantly different from zero using a t-test based on

$$t = \frac{r\sqrt{N-2}}{\sqrt{(1-r^2)}},$$

Here there are $\nu = (N-2)$ degrees of freedom and one usually does a two-sided test of the null hypothesis $r = 0$ against the alternative $r \neq 0$.

Application to the data

Taking the means of the pairs as x and their differences as y:

	Mean	Diff.			
	x	y	x^2	xy	y^2
	503	-18	253009	-9054	324
	412.5	-35	170156.25	-14437.5	1225
	⋮	⋮	⋮	⋮	⋮
	386.5	73	149382.25	28214.5	5329
	439	-24	192721	-10536	576
Sums	7674	-36	3668712	-10382	24120

Slope and intercept

Thus the sums one needs to compute best-fit lines are

$$\Sigma_x = 7674 \qquad \Sigma_y = -36$$

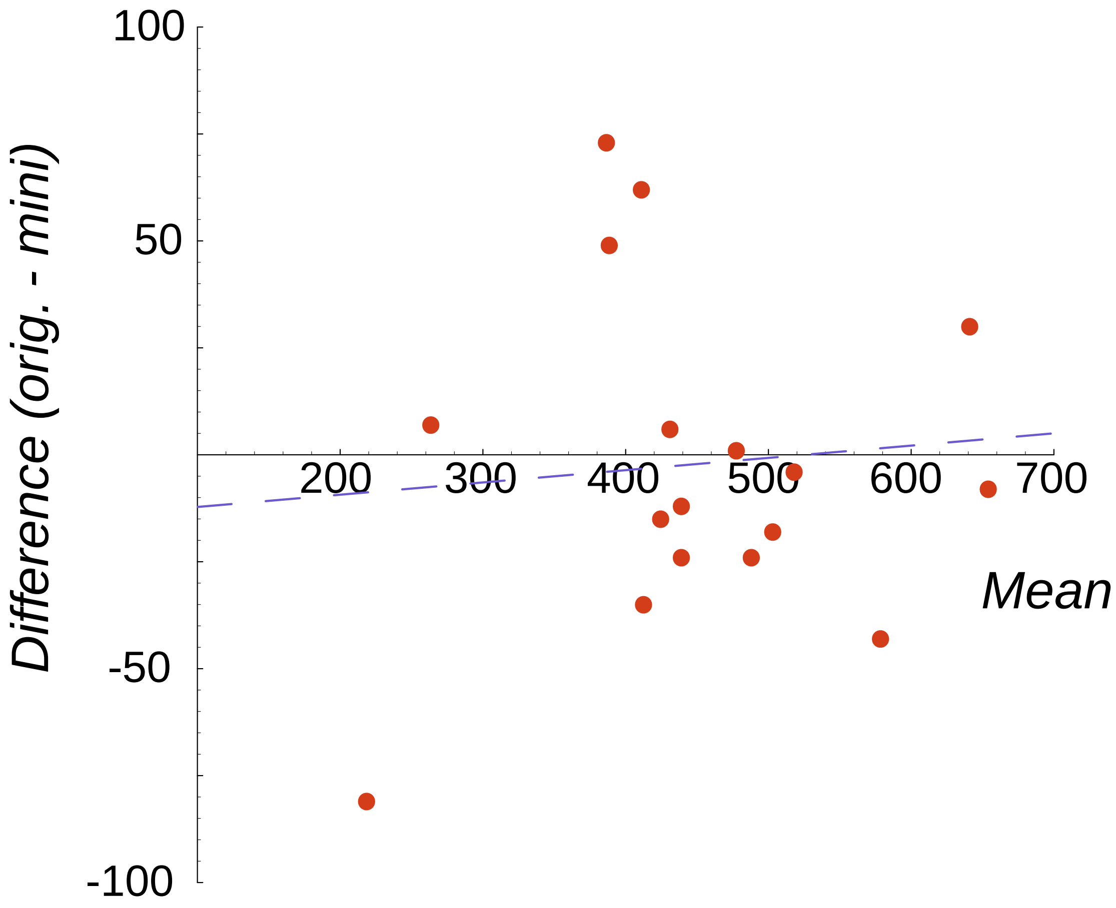
$$\Sigma_{xx} = 3668712 \qquad \Sigma_{xy} = -10382 \qquad \Sigma_{yy} = 24120$$

and so

$$m = \frac{17 \times (-10382) - 7674 \times (-36)}{17 \times 3668712 - (7674)^2} \approx 0.0287$$

$$b = \frac{3668712 \times (-36) - 7674 \times (-10382)}{17 \times 3668712 - (7674)^2} \approx -15.1$$

Plotting the line



Correlation coeffi cient:

$$r = \frac{17 \times (-10382) - 7674 \times (-36)}{\sqrt{17 \times 3668712 - (7674)^2}\sqrt{17 \times 24120 - (-36)^2}} \approx 0.0837$$

Testing r

This value of r leads to a t-value of

$$t = \frac{0.0837\sqrt{15}}{\sqrt{1-(0.0837)^2}} \approx 0.346$$

This is far less than the critical value for $\nu = 15$, $\alpha = 0.05$, and so we cannot reject the null hypothesis: the data are consistent with the view that there is no correlation between the PEFR value and the disagreement between the two instruments at that value.