

Block 4: Hypothesis testing

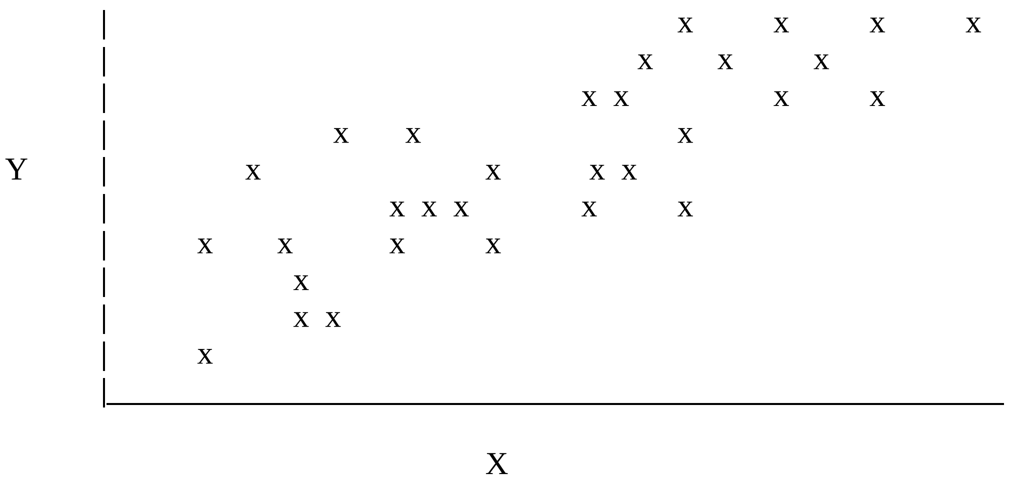
4.5.1 Visual aid for regression and correlation

[Draft only: 7 March 2015]

As a non-statistician, I developed an idiosyncratic teaching tool to introduce the idea of regression and correlation using imaginary elastic bands. Statisticians among you may be horrified, but my (non-numerate) students immediately got the idea and even anticipated the likely end result. It would be even better if there were applets available to create an animated version. I used to have transparencies and an overhead projector for this, but I also did a work-through on the chalk-board using coloured chalk (yes, chalk!)

In class I started by drawing a rough scattergram with a few data points illustrating an apparent positive linear relationship:

Fig 1: Original handmade scatterplot



I then drew in a horizontal line representing the mean of Y and a vertical line representing the mean of X and circled the point where the intersected.

Fig 2: Scatterplot showing mean X and mean Y

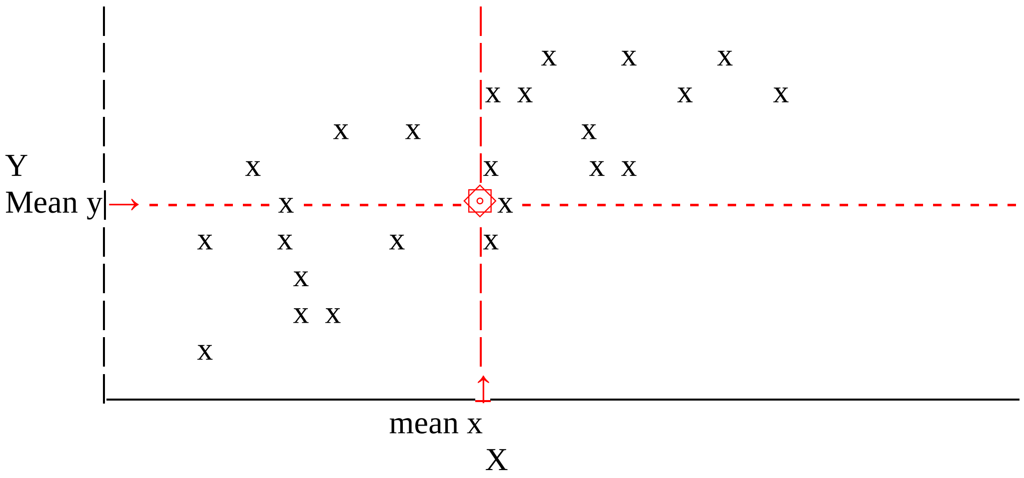
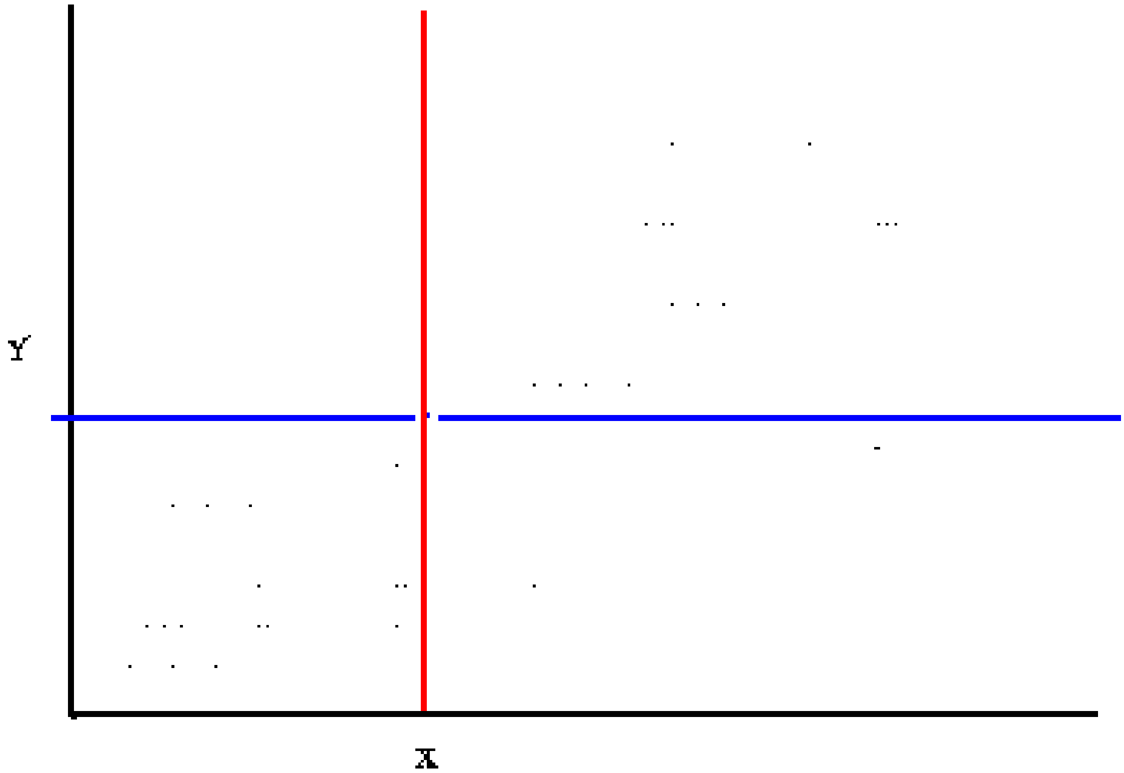
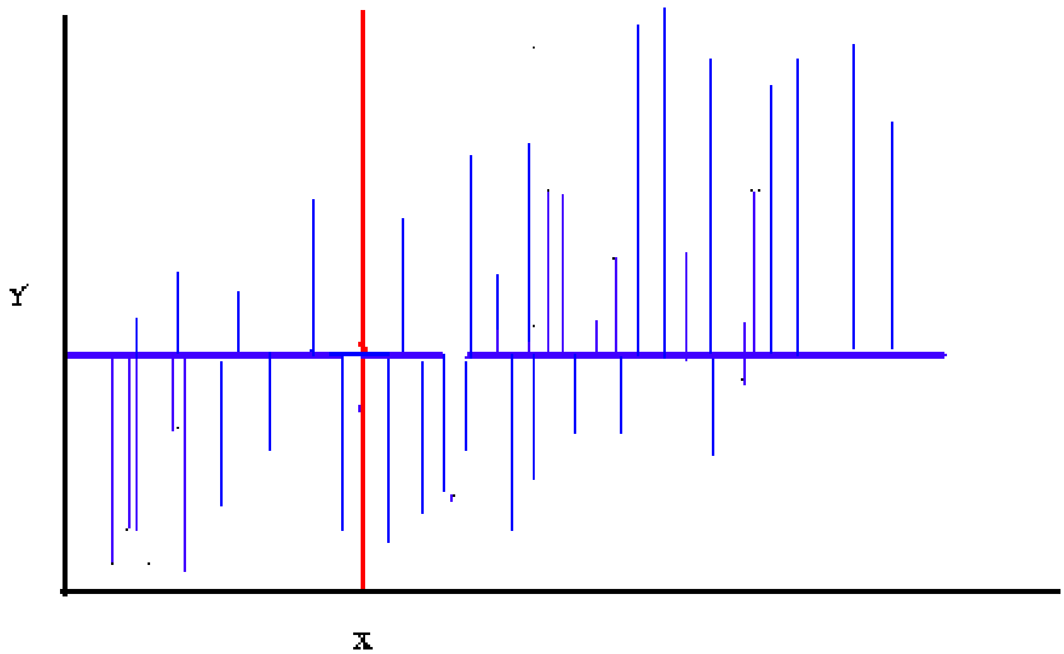


Fig 3: Scatterplot at beginning



I then built up the diagram by drawing vertical lines from each data point to mean Y:

Fig 4: Scatterplot with vertical lines from points to mean Y



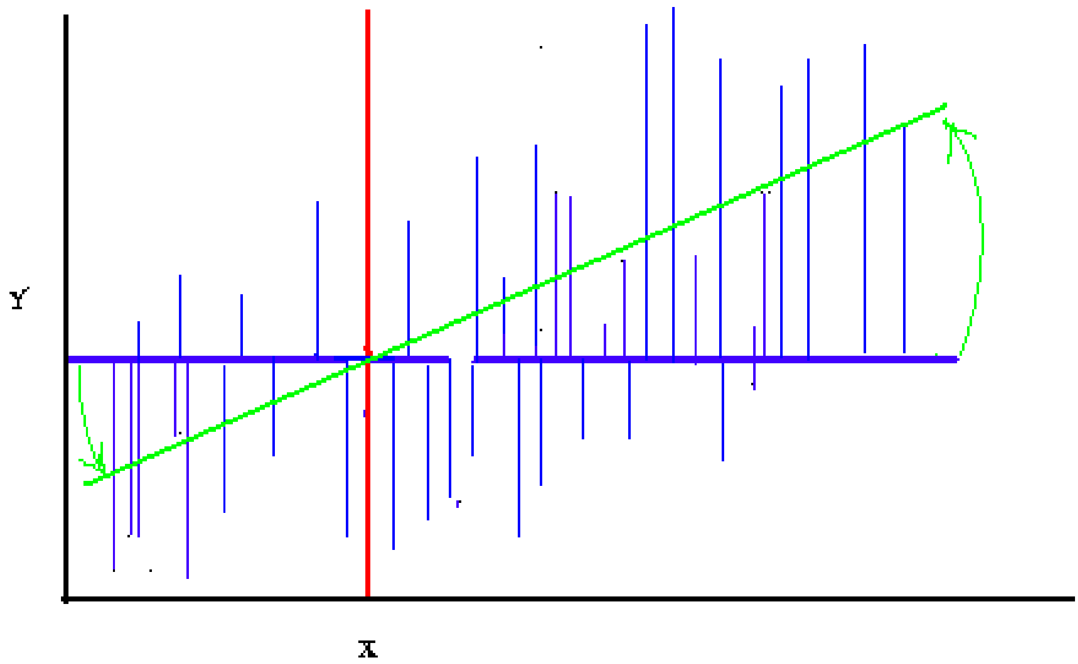
Somewhere in here I used the diagrams to explain why you have to square the distances from each data point to the mean before adding them together, otherwise the negative ones cancel out

the positive ones, before calculating variance and standard deviations. Going step by step through the process, one data point at a time, helped to build up an equation (very gently).
At the same time on the other side of the double-width board, I built up step by step elements of the formulae for variance and standard deviation of Y. This helped to explain why negative deviations need to be squared otherwise they just cancel out the positive ones. Students intuitively understood this.

At this point I asked students if it was possible to draw a straight line through the data cloud such that the vertical distance of each point from the line of mean Y was minimised. I the asked them to imagine the whole scattergram as a wooden board and the horizontal **blue line** as a rigid pole with a hole in it and a nail hammered through the hole into the board at the intersection of mean Y and mean X so that it is free to rotate. Hold the pole steady and imagine the vertical blue lines as elastic bands attached vertically from each data point to the pole. What will happen if I let go of the pole?

Yup, even social work students get this one!

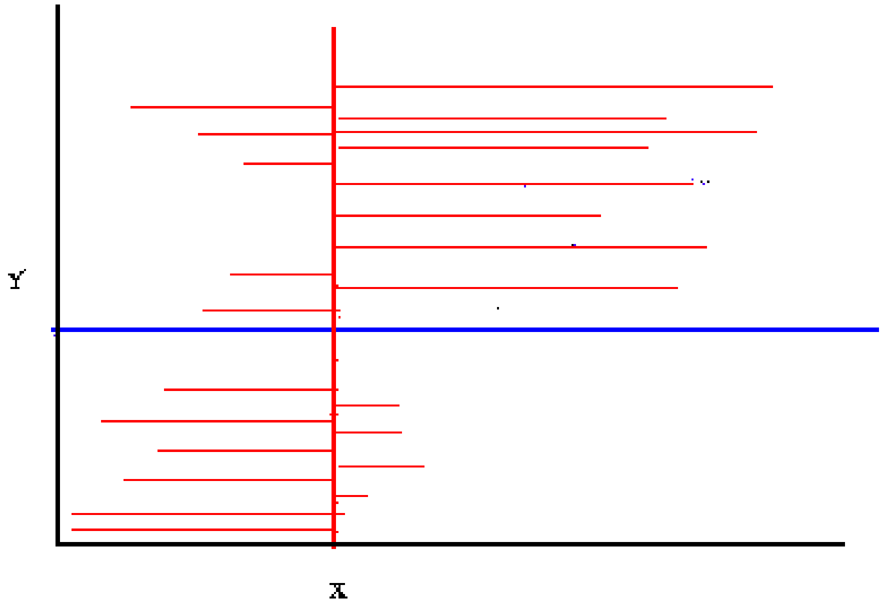
Fig 5: Scatterplot with regression line of Y on X



Students easily understood that the pole stops rotating when all the tensions balance out (green line) although I'm not sure if the mathematics of elastic is exactly the same for minimising sums of squares. After that it's easy to explain that the green line actually represents the regression line of y on x.

I then built up a new diagram by drawing horizontal lines from each data point to mean X:

Fig 6: Scatterplot with horizontal lines from points to mean X

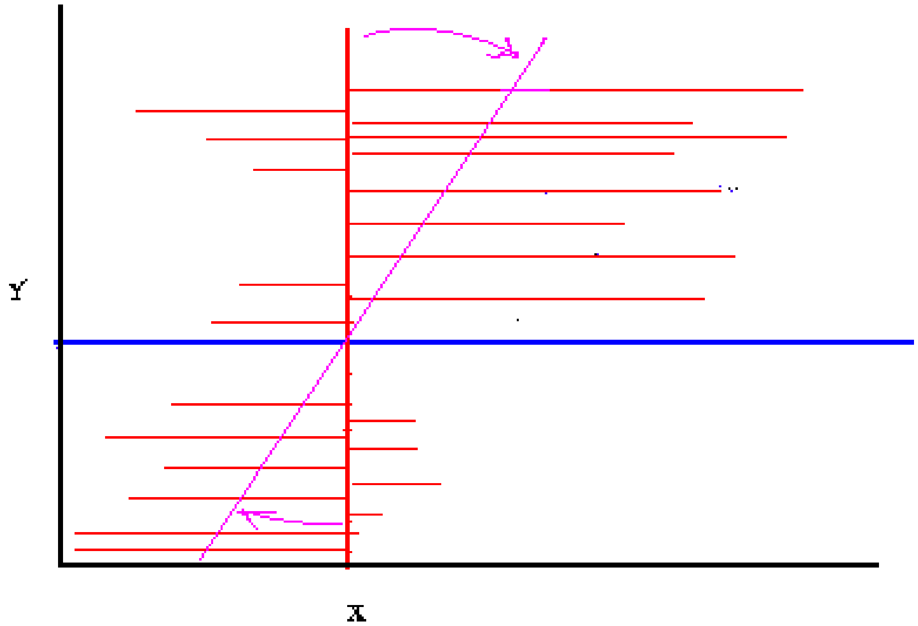


By this time there was chalk-dust everywhere.

Now think of the vertical red line as a rigid pole free to rotate around the intersection of the means. If you hold the pole steady and imagine the horizontal red lines as elastic bands attached to each data point, What happens when you let go of the pole?

Yup, right again!

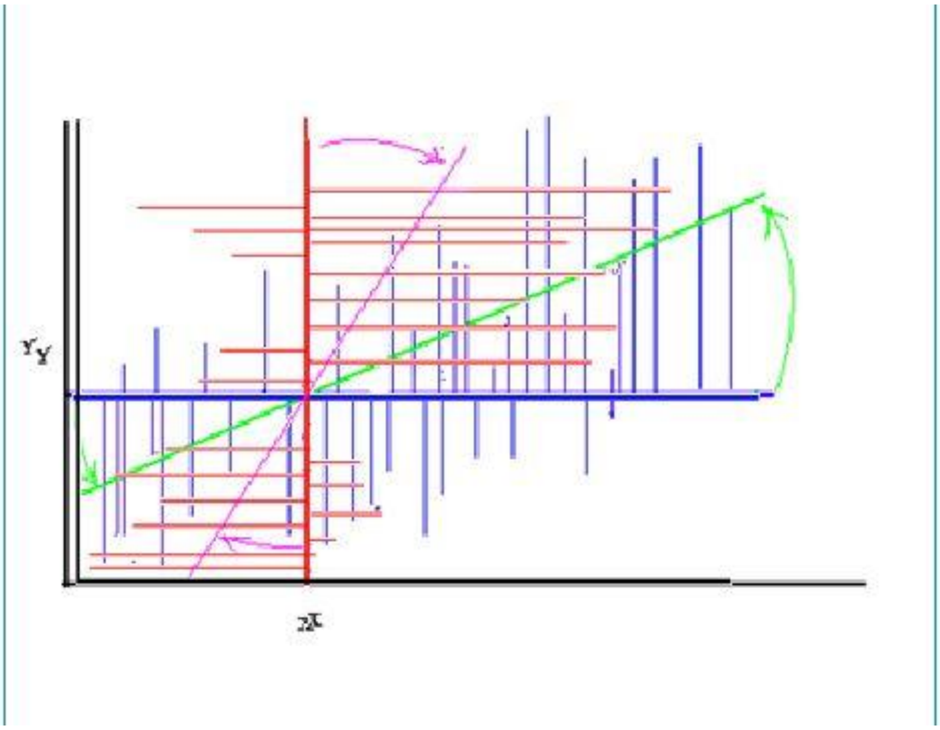
Fig 7: Scatterplot with regression line of Y on X



So the pink line must be the regression line of X on Y then?

If you overlay two diagrams with the green and purple lines:

Fig 8: Scatterplot showing angle between regression lines of Y on X and X on Y
(Overlay of Figs 5 and 7 made with [Serif] by Terry Blom on his mobile whilst walking the dog!)



. . from which you can explain correlation in terms of the cosine of the angle between the two regression lines.

Cosine 90° is 0, cosine 0° is 1, *et voilà*!

Postscript

Looking back I should perhaps have made the X and Y axes themselves the starting points for the lines (rods), forced them to remain horizontal and vertical, and used the same elastic band analogy to see where they ended up (hopefully the two means). I assume they would stabilise at the mean of X and Y respectively, with the vertical one representing mean X and a horizontal one mean Y? If so they could then be pinned through the intersection of mean X and mean Y and allowed to rotate, producing two regression lines, Y on X and X on Y, the cosine of the angle between them being Pearson's r.