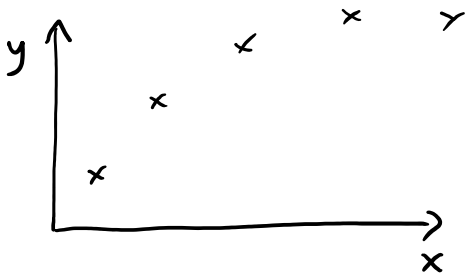
Section 9.1: Coefficients of Correlation and Determination

Tuesday, March 27, 2018 9:42 AM

suppose you have a set of bivariate data (x, y)

one common way to display this data is called a **scatterplot**



it is important that you <u>look</u> at the graph

- allows us to see (by inspection) relationships (associations) between variables (if any), trends, outliers, etc.

---

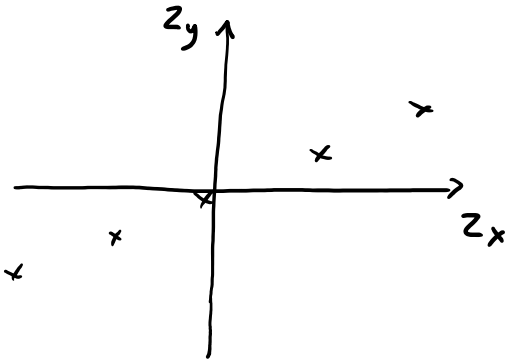
consider our scatterplot

and for all x, calculate mean and standard dev

y " " " " "

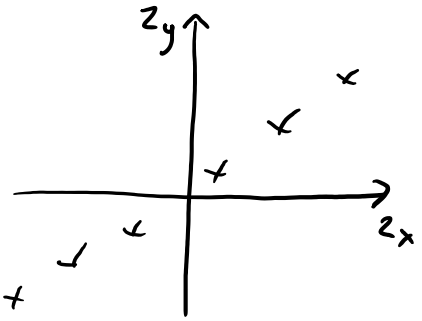
and for each point, replace x by z_x
y by z_y

you get:

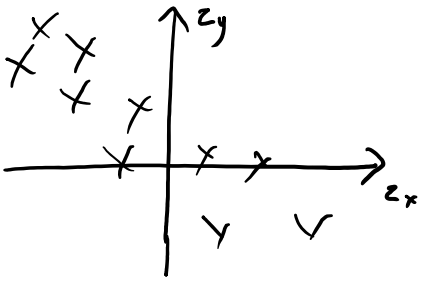


note how the product $z_x z_y$ is positive for points in QI and III

and negative for points in QII and QIV

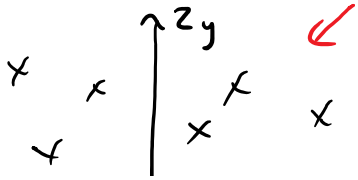


the sum of $z_x z_y$ is positive

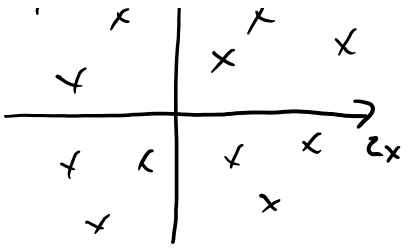


the sum of $z_x z_y$ is negative

in physics "shotgun plot"

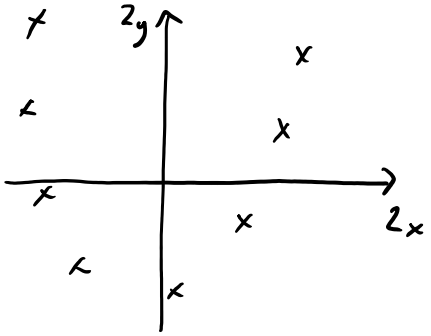


the sum of $z_x z_y$ is close to zero because



the sum of $z_x z_y$ is close to zero because the positives and negatives cancel out

warning:



note: this parabola will also have

$\Sigma z_x z_y$ close to zero

$\Rightarrow$ look at your graph

the correlation coefficient

$$r = \frac{\Sigma z_x z_y}{n-1}$$

(more complicated versions of this equation exist)

where $-1 \leq r \leq 1$

↑ negative association

↑ positive association

the coefficient of determination:

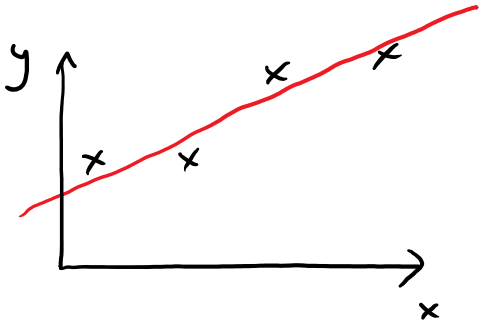
$$R^2 = r^2$$

where $0 \leq R^2 \leq 1$

↑ no association

↑ strong association

the nice thing about R^2 is that it gives the fraction of the data's variation accounted for by the fit



← the data is mostly linear but with a bit of scatter

if $R^2 = 0.76$, then "a linear model accounts for 76% of the variability in y"

→ the remainder is called the residual

next section ↵