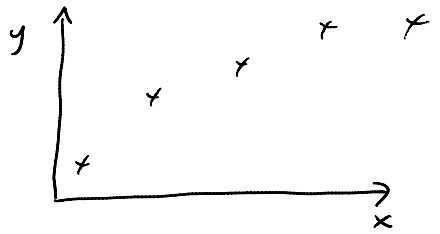
# Section 6.1: Coefficients of Correlation and Determination

Wednesday, June 17, 2015
1:54 PM

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

one common way to display this is called a scatterplot




mark I human eyeball:

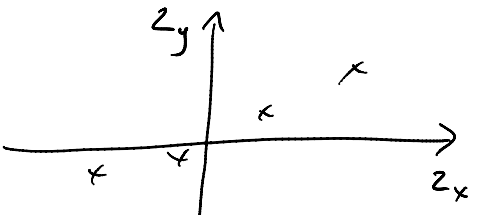
allows us to see (by inspection!) relationships between variables (if any), trends, outliers, etc.

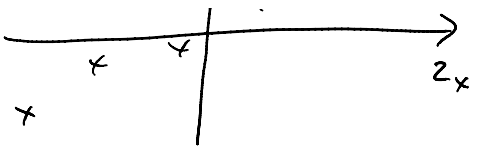
---

consider our scatterplot:


and for all $x$, calculate the mean and std dev
for all $y$, " " " " " " "

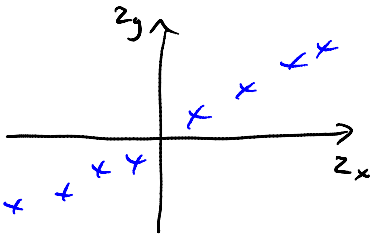
and for each point, replace $x$ by $z_x$
$y$ by $z_y$



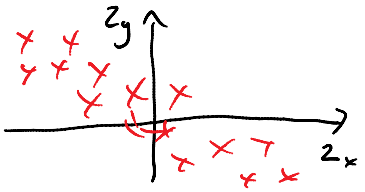



note how the product $z_x z_y$ is positive for points in QI and QIII

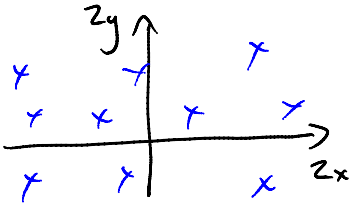
and negative for points in QII and QIV




the sum of $z_x z_y$ is positive




the sum of $z_x z_y$ is negative




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

the correlation coefficient $r = \frac{\sum z_x z_y}{n-1}$

where $-1 \leq r \leq 1$

↑ negative association

↑ positive association

the coefficient of determination:

$R^2 = r^2$ (Why uppercase? no idea)

no idea)

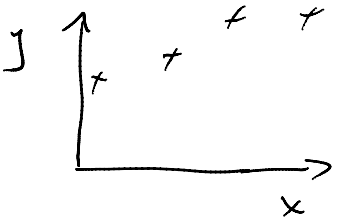
where $0 \leq R^2 \leq 1$

↑ 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

example:




$R^2 = 0.76$

then you could say that "a linear model accounts for 76% of the variability in y"

remainder is the residual