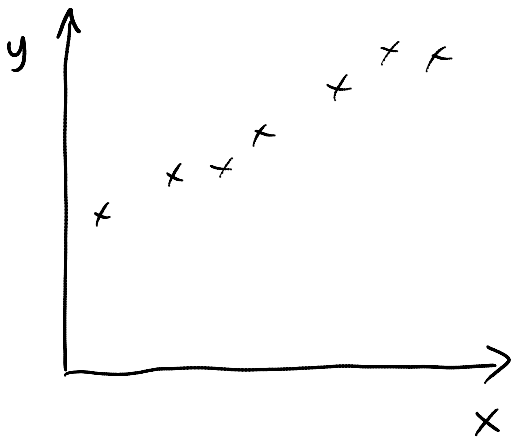
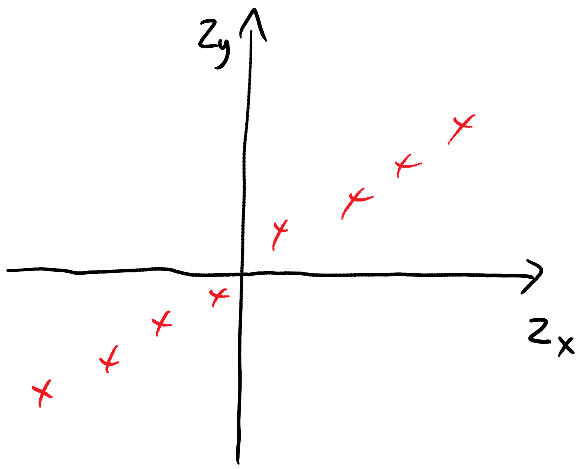
Section 13.1, cont'd

Tuesday, June 11, 2013
11:47 AM

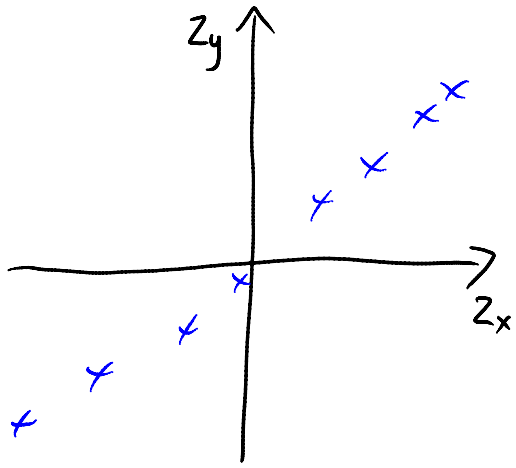


↓ translate all data points into their associated z-scores

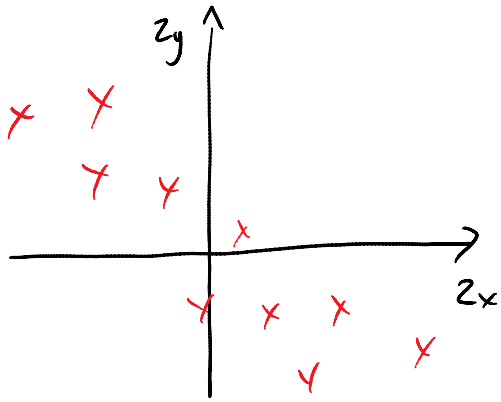


note: plot now centred about the origin

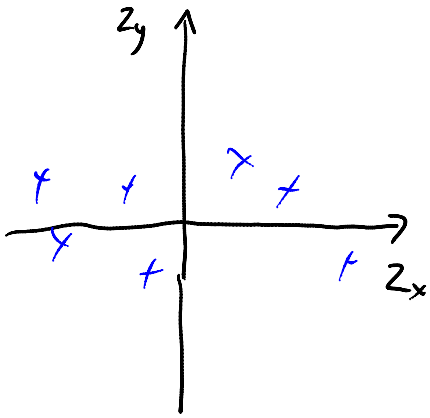
so we can get measure of the "goodness-of-fit"



← the product $z_x z_y$ is all or nearly all positive



← the product $z_x z_y$ for each point is nearly all negative



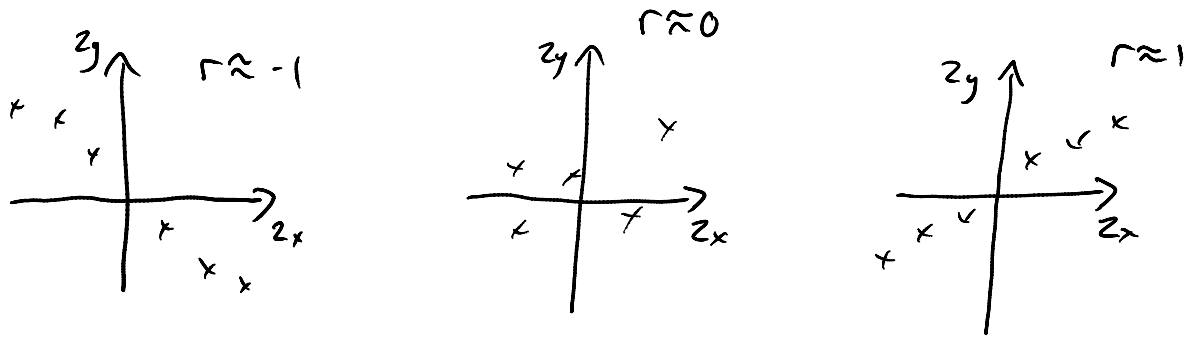
← the product $z_x z_y$ is a mixture of positive and negative
- so the sum will be near zero

correlation coefficient

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

← note: I want you to interpret this, not calculate it

and r runs from -1 to 1



note: outliers can make a big difference!
but you can always report r with and without that point
(but need to either perform a statistical test or consult statistician before throwing the point out)

---

so why do so many stats packages calculate R^2 and not r?

R^2 ← coefficient of determination

R^2 gives the fraction of the data's variation accounted for by the model

if $R^2 = 0.76$, then 76% of variation in the data is accounted for by the fit and the rest is either scatter or some association not in the fit

so how do you calculate R^2? $R^2 = (r)^2$

↑
correlation
coefficient