Why r^2 is meaningless

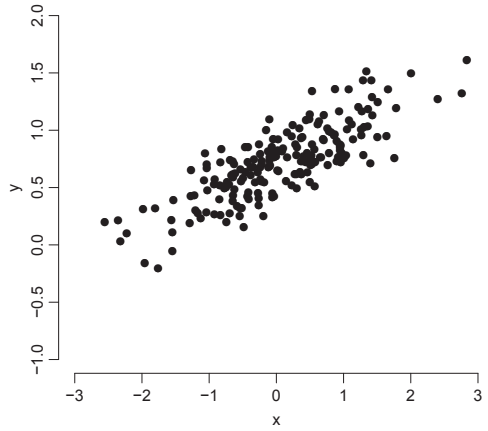
Bendix Carstensen

Steno Diabetes Center, Denmark

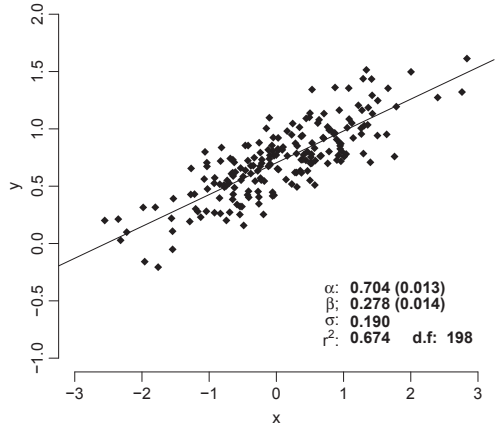
& Department of Biostatistics, University of Copenhagen

bxc@steno.dk http://BendixCarstensen.com

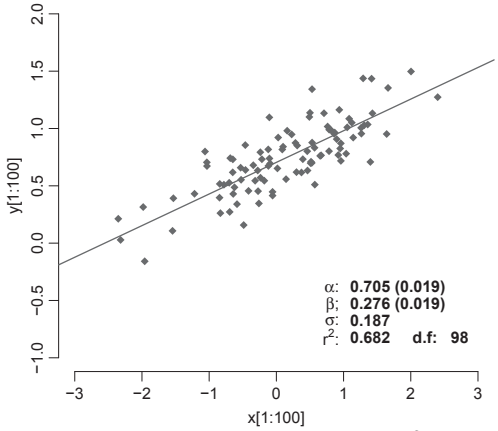
A dataset



Regression analysis

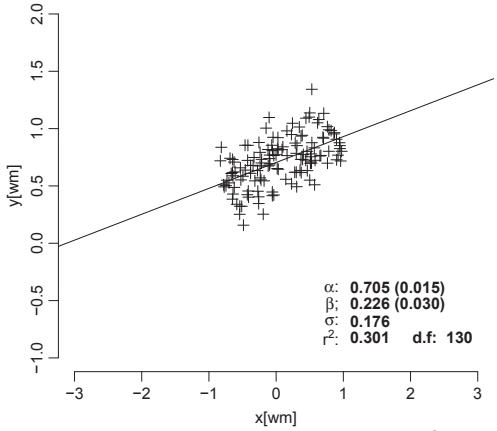


A 50% random sample



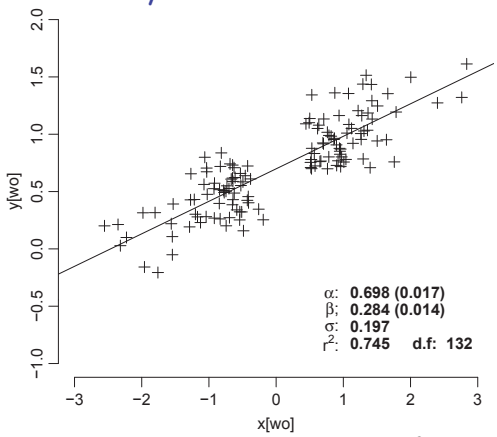
Same estimates of α, β and σ and **same** r^2

The middle 2/3 of the data



Same estimates of α, β and σ but **smaller** r^2

The outer 2/3 of the data



Same estimates of α, β and σ but **larger** r^2

What r^2 is and is not

- r^2 is related to the **population** analysed:
  It is the fraction of the **population** variation in y which explained by x.
- r^2 does **not** convey **any** information on the **size** of the relationship. The relationship is judged from the estimates of α and β:
  Is the effect clinically relevant?.
- r^2 does **not** convey any information on the **precision of predictions**. This is contained in the residual variation, σ. A 95% prediction interval for given $x = x_0$ is:

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta} \times x_0 \pm 1.96\hat{\sigma}$$

(disregarding the estimation error in α and β).

Moral:

- The clinically relevant parameters α, β and σ are the same no matter how the population is sampled.
- They reflect the relationship between y and x.
- r^2 involves the population distribution, which is alien to the relationship between y and x.
- Hence, r^2 is mathematical mumbo-jumbo where the link to subject matter relevance has been obscured by mixing in the distribution of y in the study population.