

Anscombe

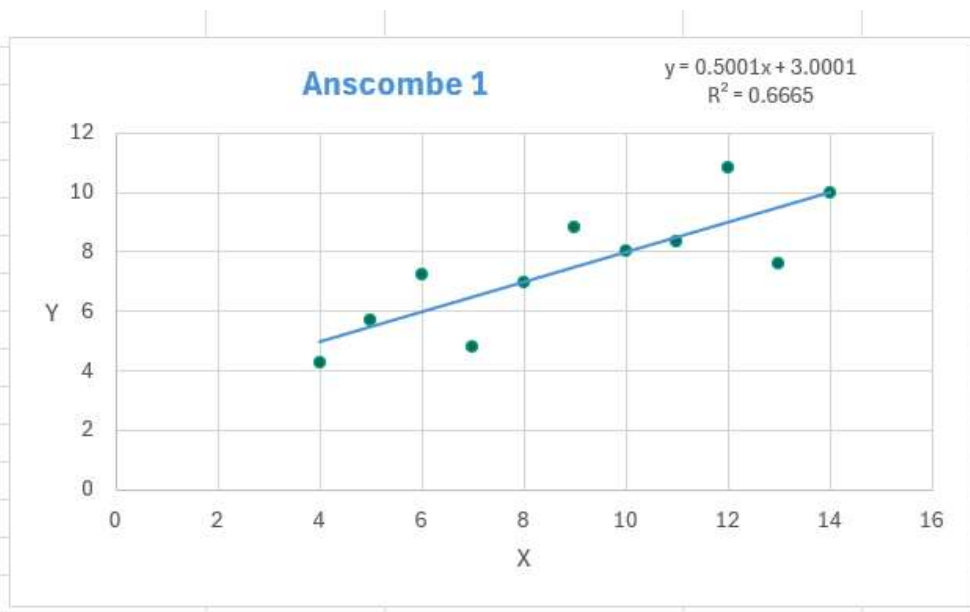
In 1973 creëerde F.J. Anscombe een kwartet van (hypothetische) gegevensverzamelingen ("datasets") met dezelfde statistische kentallen, maar die zeer verschillen van elkaar. Hij illustreerde hiermee dat het van groot belang is om de gegevens grafisch te analyseren vooraleer een verdere statistische analyse uit te voeren.

[Anscombe, F.J.] Graphs in Statistical Analysis. The American Statistician. Vol. 27 No. 1 (February 1973): 19

Elke dataset bestaat uit 11 puntenkoppels (x_i, y_i) , maar bij slechts één ervan is het zinvol om een lineaire regressie uit te voeren, nl. bij kwartet 1.

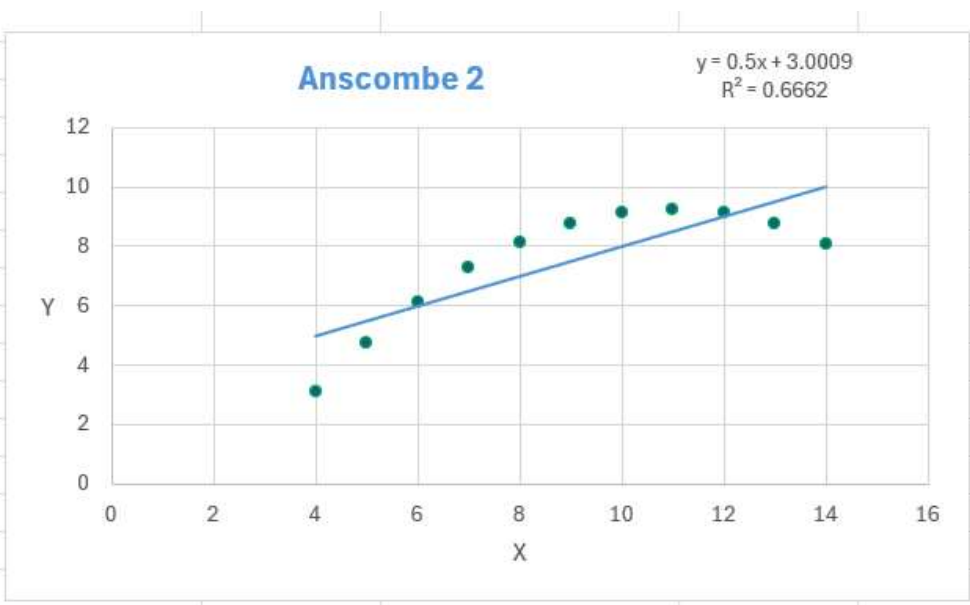
Kwartet 1

Xgegevens	Ygegevens1
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



Kwartet 2

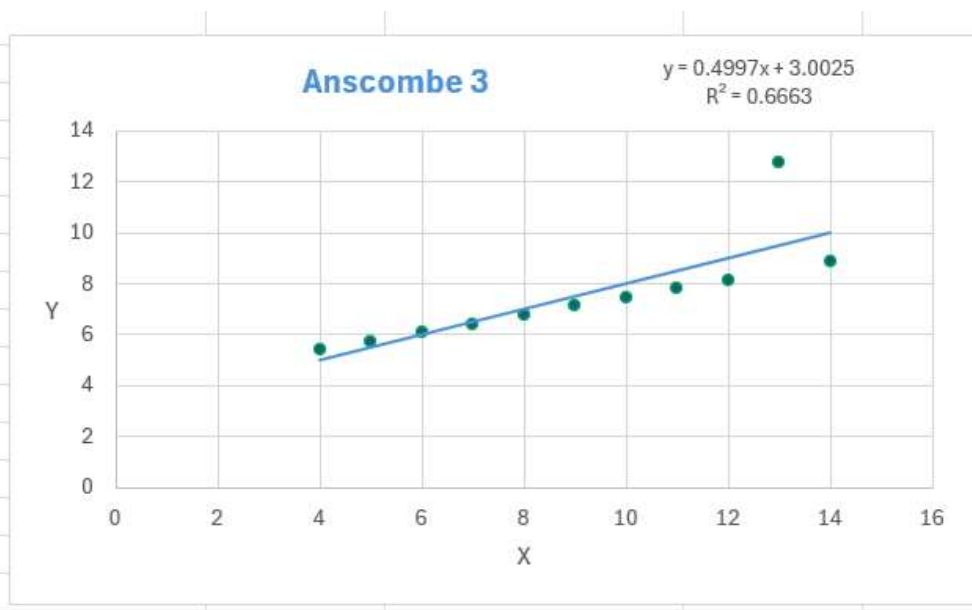
Xgegevens	Ygegevens2
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74



Hier is grafisch duidelijk dat een lineaire regressie niet zinvol is. Het lijkt eerder te gaan om een kwadratische regressie of een veeltermregressie met een hogere graad dan 2.

Kwartet 3

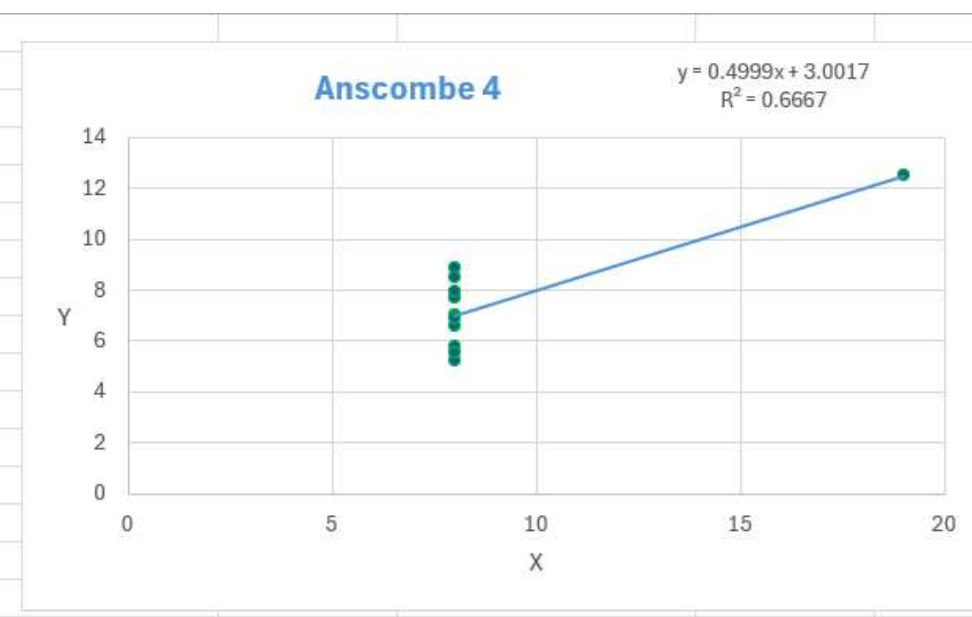
Xgegevens	Ygegevens3
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73



Hier is duidelijk een uitbijter ("outlier").

Kwartet 4

Xgegevens4	Ygegevens4
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89



Dit is een zeer rare situatie met een uitbijter die de hele zaak domineert.

Merk op dat in de figuren hierboven de regressielijn zich beperkt tot een lijnstuk, in tegenstelling tot GeoGebra waar die getoond wordt voor alle reële x-waarden, terwijl statistisch gezien alleen de x-waarden tussen de kleinste x en de grootste x qua gegeven punten zinvol zijn.

Michael Borchers heeft een leuk [GeoGebra bestand voor dit kwartet](#).