

## Anscombes data

---

Man kan illustrere det meningsløse i påstande som: 67% af datavariationen i  $y$  er forklaret ved  $x$ -variablen - Gennem følgende eksempel og øvelse:

Nedenfor ses fire datasæt præsenteret af statistikeren Francis Anscombe i 1973

$x$	10	8	13	9	11	14	6	4	12	7	5
$y_1$	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
$y_2$	9.14	8.14	8.74	8.77	9.26	8.1	6.13	3.1	9.13	7.26	4.74
$y_3$	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
$x_4$	8	8	8	8	8	8	8	19	8	8	8
$y_4$	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.5	5.56	7.91	6.89

- Udfør lineær regression på fire datasæt  $(x, y_1)$ ,  $(x, y_2)$ ,  $(x, y_3)$  og  $(x_4, y_4)$ , og undersøg, om der kan være tale om en lineær sammenhæng.
- Bestem et residualplot for hver af sammenhængene.
- Bestem  $R^2$  for hver af sammenhængene.
- Beskriv forskellene mellem de fire sammenhænge og deres datasæt.