Relational data and linear regression

Viggo Andreasen, Roskilde Universitet

September 9, 2025

BEFORE we start please

- Go to Moodle and find the section for today
- Upload the following files:
  - `Overheads.pdf`
  - `snake.csv`
  - `RegressionForSnakes.ipynb`

    Moodle may show the file in a strange format press Ctr-S to save a copy (on some systems you have to remove a .JSON from the end of the file name, when saving)
- Start Python (so that it is ready - check to see that the files are where Python can find them! and the return to this page)

Relational data

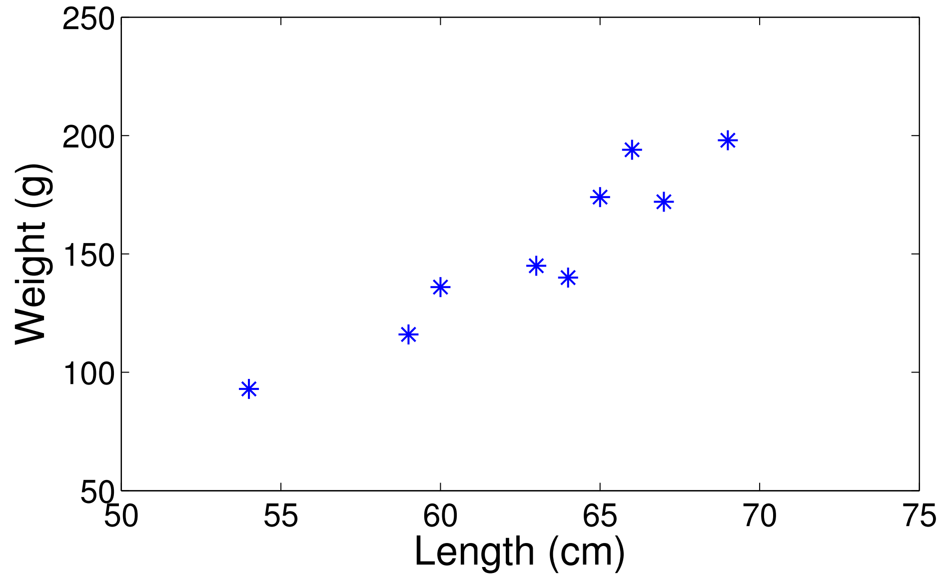
Data where we have more than one measurement for each **observational unit** (each object that has been studied).
Focus on a special situation that occurs very often: For each unit of study we have measured two continuous variables.

Length X (cm)	Weight Y (g)
60	136
69	198
66	194
64	140
54	93
67	172
59	116
65	174
63	145

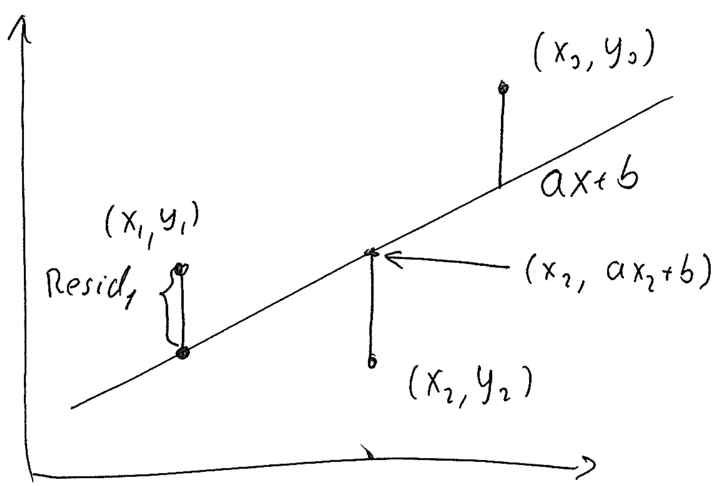
Example: Length and Weight of 9 female adult snakes (species adder/viper, DK: hugorm. *Vipera berus*). Observational unit is a snake.

Scatter plot

- Unit of study (observation)
  $k = 1..n$
  in our case $n = 9$
- x_k length of snake k
- y_k weight of snake k



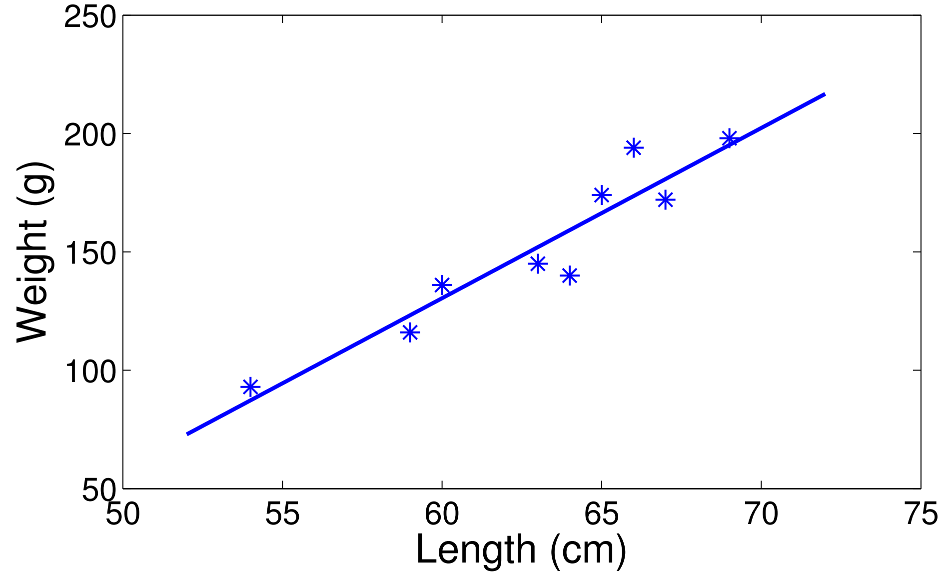
Distance from the line $y = ax + b$ to the data points



- For point k
  $\text{Residual}_k = y_k - (ax_k + b)$
- Total distance

$$s(a,b) = \sqrt{\frac{1}{n-2}\sum_{k=1}^{n}(\text{Residual}_k)^2} = \sqrt{\frac{1}{n-2}\sum_{k=1}^{n}\left(y_k - (ax_k + b)\right)^2}$$

Best fitting line



- Set $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$

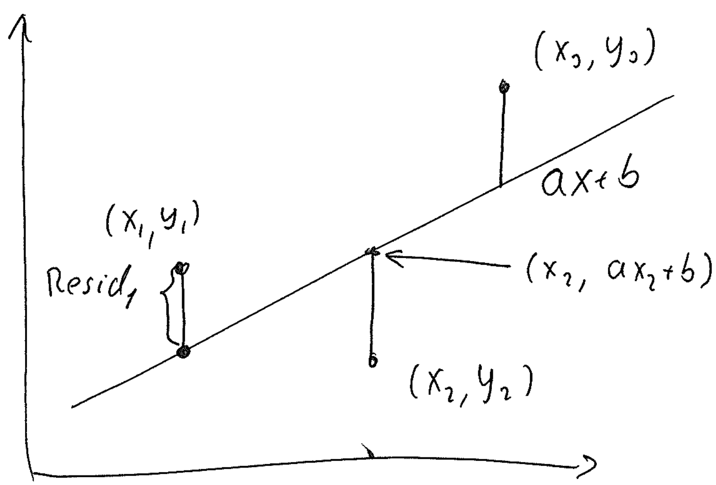
  and $\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$
- Compute $a_1 = \sum(x_k - \bar{x})(y_k - \bar{y})$ and $a_2 = \sum(x_k - \bar{x})^2$
- $a = a_1/a_2$ and $b = \bar{y} - a\bar{x}$
- Don't memorize the formula!
  In Python
  `b,a=polyfit(x,y,1)`
  gives a and b

The hard way:

- $\frac{1}{n-1}a_2$ is the *variance of* x.
  In Python `x.var()`
- $\frac{1}{n-1}a_1$ is *covariance of* x *and* y.
  `x.cov(y)`
- $a =$ `x.cov(y)/x.var()`

Compute and draw line

- We that we know a and b we can compute the line as $y_{\text{line}} = ax + b$
- The residuals are the difference between observed value and predicted value:
  $\text{Resid}_k = y_k - (ax_k + b)$ at the point (x_k, y_k)



Assumptions

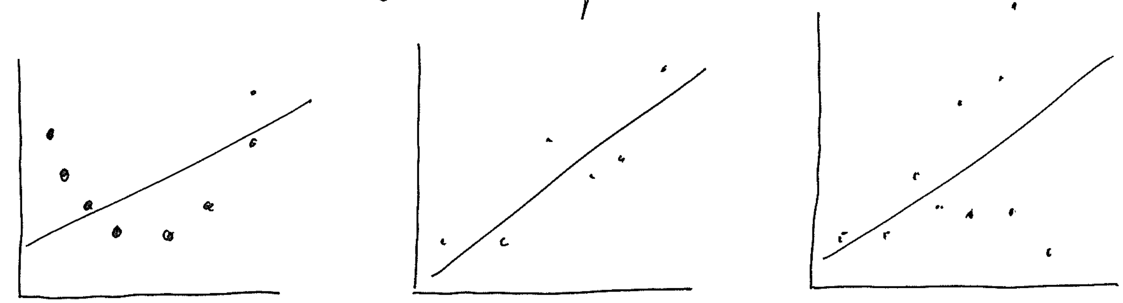
- The uncertainty/variation is associated with the measurement of y_k (not with x_k).
- We minimize the variation in data (in the y_k) that we cannot explain by our model (our line).
- The remaining variation gives the *standard deviation* around the line

  $$s(a,b) = \sqrt{\frac{1}{n-2}\sum_{k=1}^{n}\left(y_k - (ax_k + b)\right)^2}$$

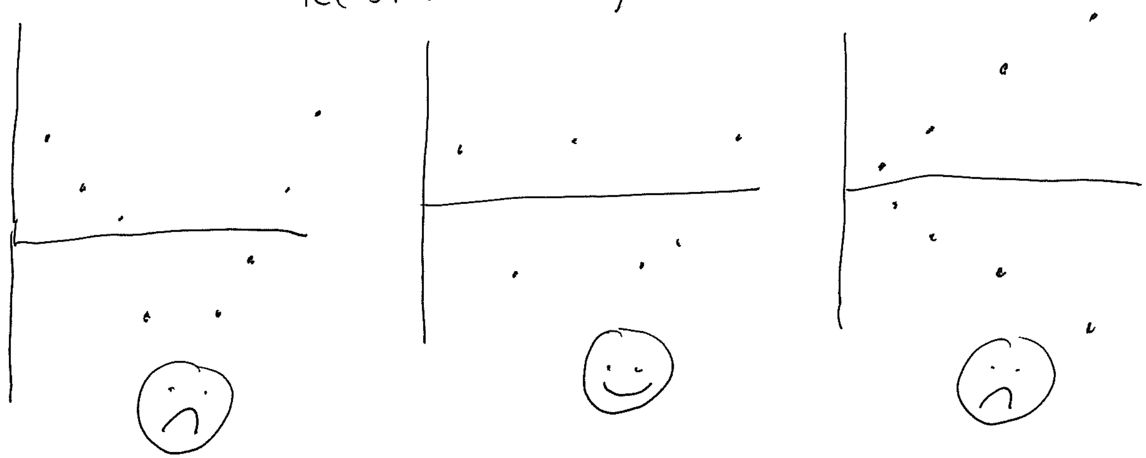
- All y_k are subject to the same amount of variation. In particular the variation does not depend on x_k in a systematic way. This assumption can (and should) be checked. See next overhead.
- Even if the regression line fits very well, one cannot conclude that there is a *causal relationship*.

Analysis of residuals

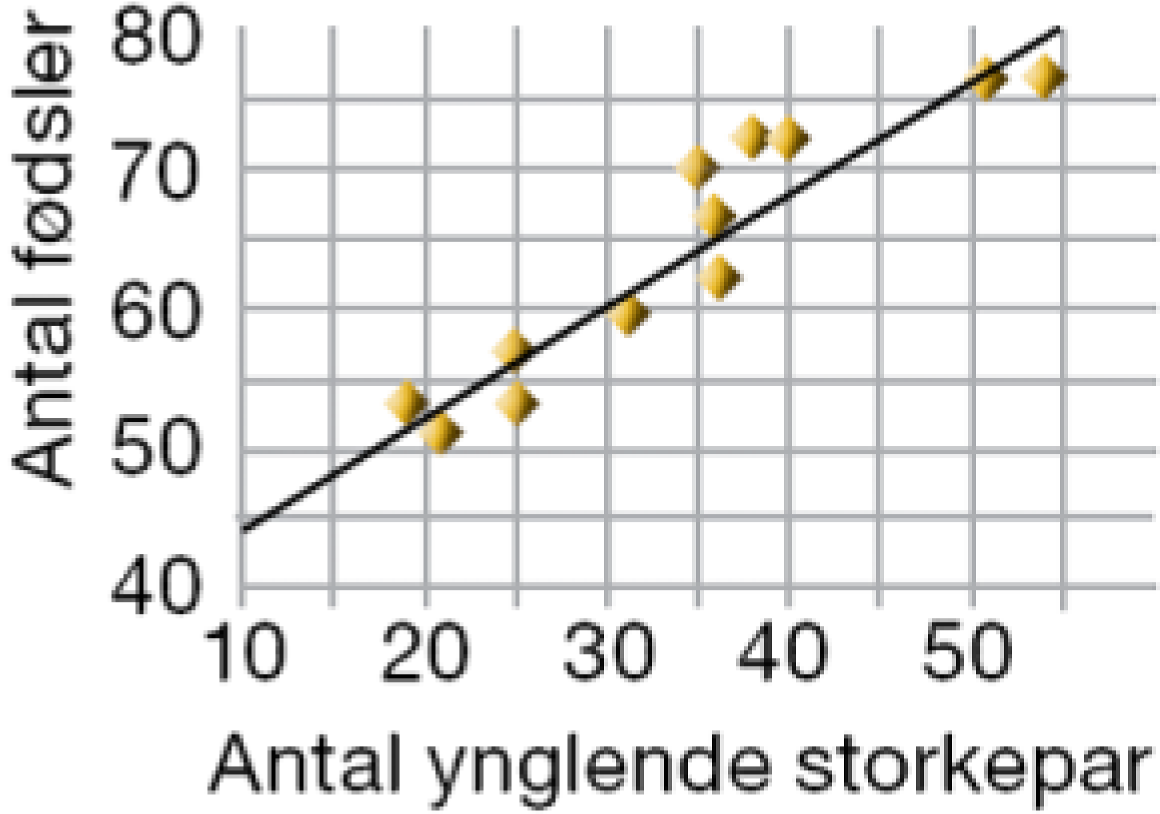
Scatter plot



Residual plot



Causality



Antal fødsler og antal ynglende storkepar i Danmark.