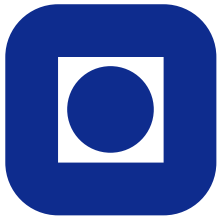


NTNU
Norwegian University of
Science and Technology

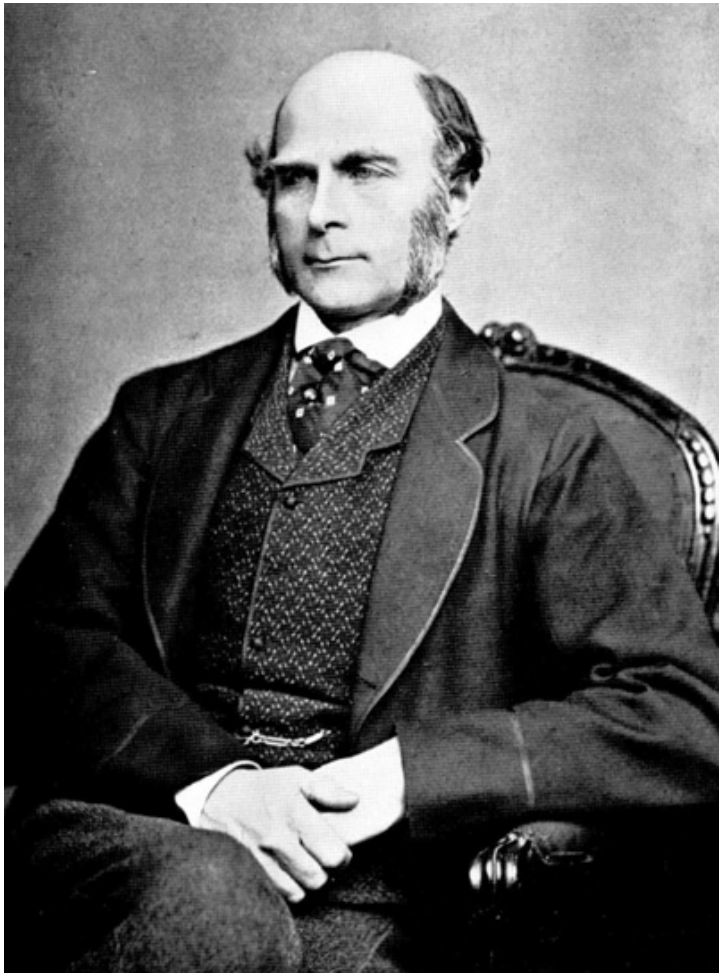
TMA4267 Linear Statistical Models V2014 (2)

Least squares [1.2], bivariate normal [1.5]

Mette Langaas

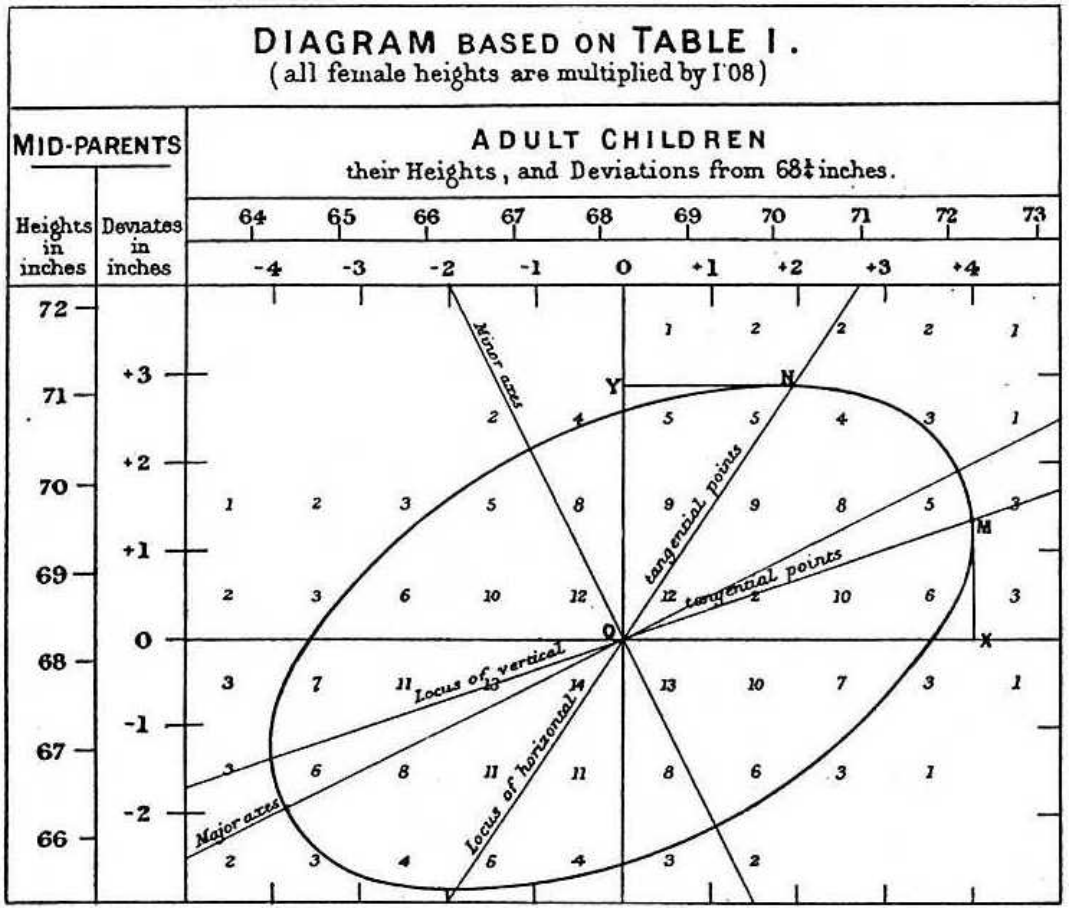
To be lectured: January 7, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

Sir Francis Galton



http://en.wikipedia.org/wiki/File:Francis_Galton_1850s.jpg

Child adult height vs midparent height



http://en.wikipedia.org/wiki/File:Galtons_correlation_diagram_1875.jpg

ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

[With Plates IX and X.]

This memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those

Galtons observations

Galton observed that

- tall parents often got tall children, but the tallest partents on average got shorter children than themselves
- short parents got often short children, but the shortest parents on average got taller children than themselves

Regression towards the mean - a principle stating that of related measurements, and selecting those where the first measurement is either higher or lower than the average, the expected value of the second is closer to the mean than the observed value of the first.

Least squares [1.2]

Coefficients (a,b) found to to minimize sum of squared differences $y_i - a - bx_i$ for linear model

$$y_i = a + bx$$

$i = 1, \ldots, n$. The normal equations:

$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0$$

Least squares

$$
\begin{aligned}
b &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{s_{xy}}{S_{xx}} \\
a &= \bar{y} - b\bar{x}
\end{aligned}
$$

The bivariate normal distribution

$$f(x,y) = ce^{-\frac{1}{2}Q(x,y)}$$

$$c = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

$$Q(x,y) = \frac{1}{(1-\rho^2)}[(\frac{x-\mu_X}{\sigma_X})^2 + (\frac{y-\mu_Y}{\sigma_Y})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y})]$$

Computational trick

$$Q(x,y)(1-\rho^2) = [(\frac{y-\mu_Y}{\sigma_Y}) - \rho(\frac{x-\mu_X}{\sigma_X})]^2 + (1-\rho^2)(\frac{x-\mu_X}{\sigma_X})$$

$$f(x,y) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp[(\frac{x-\mu_X}{\sigma_X})^2]$$

$$\cdot \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_Y} \exp[(\frac{y-c_X}{\sqrt{1-\rho^2}\sigma_y})^2]$$

$$c_x = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$$

Findings

— The marginal distributions $f_X(x)$ and $f_Y(y)$ are both normal, and with parameters (μ_X, σ_X^2) and (μ_Y, σ_Y^2), respectively.

3 questions

Write down the answers on the piece of paper handed out - and put in the box at the exit.

1. How many percent of the lecture did you fully understand?
2. Divide the population version of the regression slope with the correlation coefficient - what do you get?
3. Which result did you like the most?