Partitioned multivariate normal distributions

-

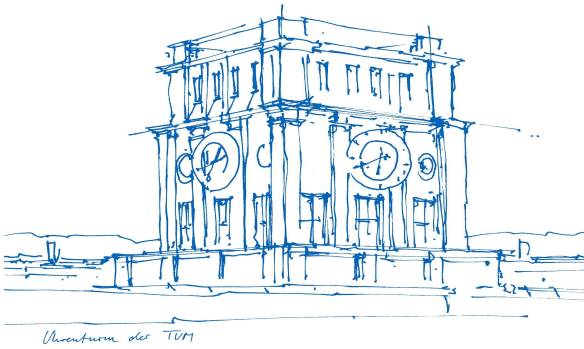
Marginalizing, Conditioning and Regression

Julien Gagneur

Prof. for Computational Molecular Medicine
Technical University of Munich

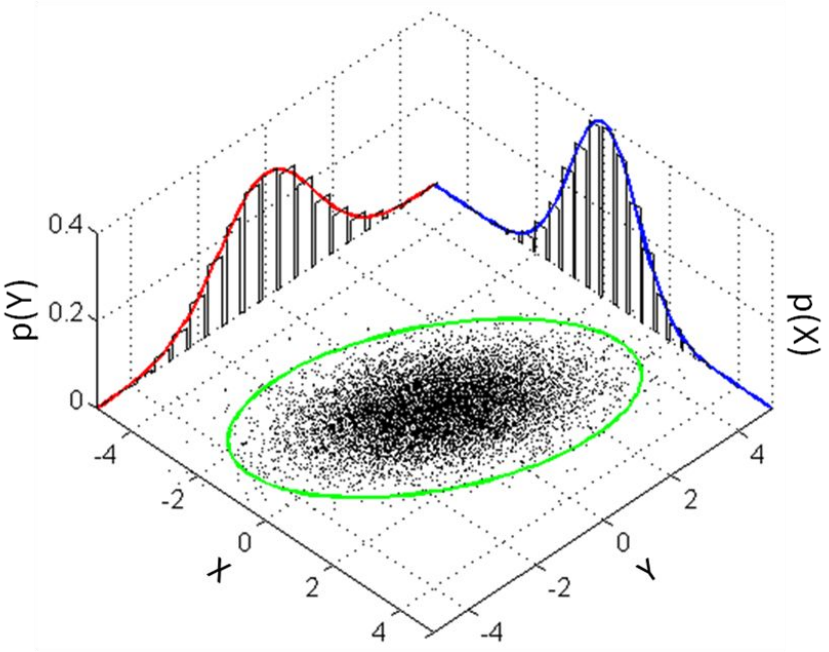
www.gagneurlab.in.tum.de

To understand the genetic basis of gene regulation and its implication in diseases



Motivation

- The MVNs is multivariate distribution across *p* random variables
- Random variables may come in different subsets:
  - Observed vs. latent variables
  - Target or response variable vs. predictive features
- We then consider partitioned MVNs



Multivariate Normal Distribution (source: Wikipedia)

Notations

Given a MVN $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. We consider a partition of the p variables into two sets, leading to

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad (32)$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \quad (33)$$

Marginalizing and conditioning

The marginal distribution is a MVN with the following simple form:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \quad (34)$$

The conditional distribution is also a MVN:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}\right) \quad (35)$$

where

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \quad (36)$$

Conditioning: A thought experiment

- We're looking at the performance of several athletes throwing the javelin twice. We assume the athletes have different average performances distributing in a Gaussian way. We assume further that each throw deviates from the athlete average performance identically, independently and normally. Under these assumptions, the distances of two throws would distribute as a MVN.

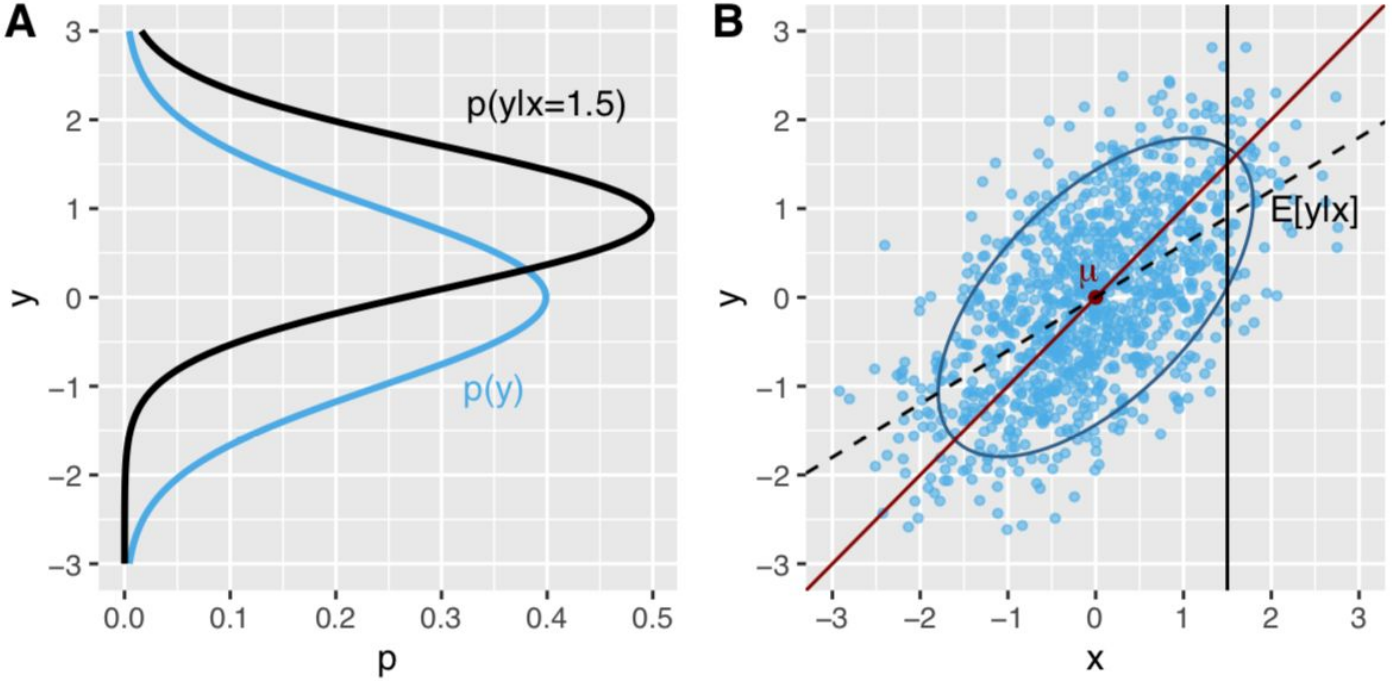
- Question: The average throw across the population of athletes is 40 m. Alice throws the javelin. It lands at 50 m. Is the expected distance of Alice at her next throw...

  1) < 50 m,

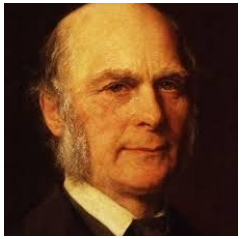
  2) = 50 m, or

  3) >50 m?

Conditioning with MVNs

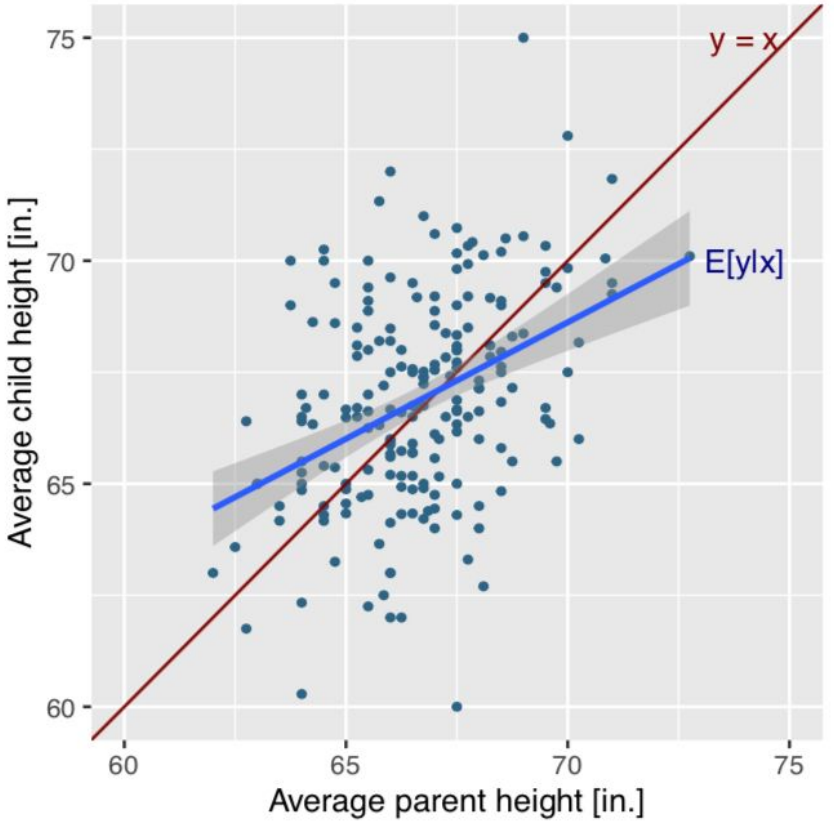


Regression towards the mean in human stature



Sir Francis Galton
1822-1911

- Parents exceptionally tall have children that are tall, but not as much
- The effect is proportional to the deviation from the population mean

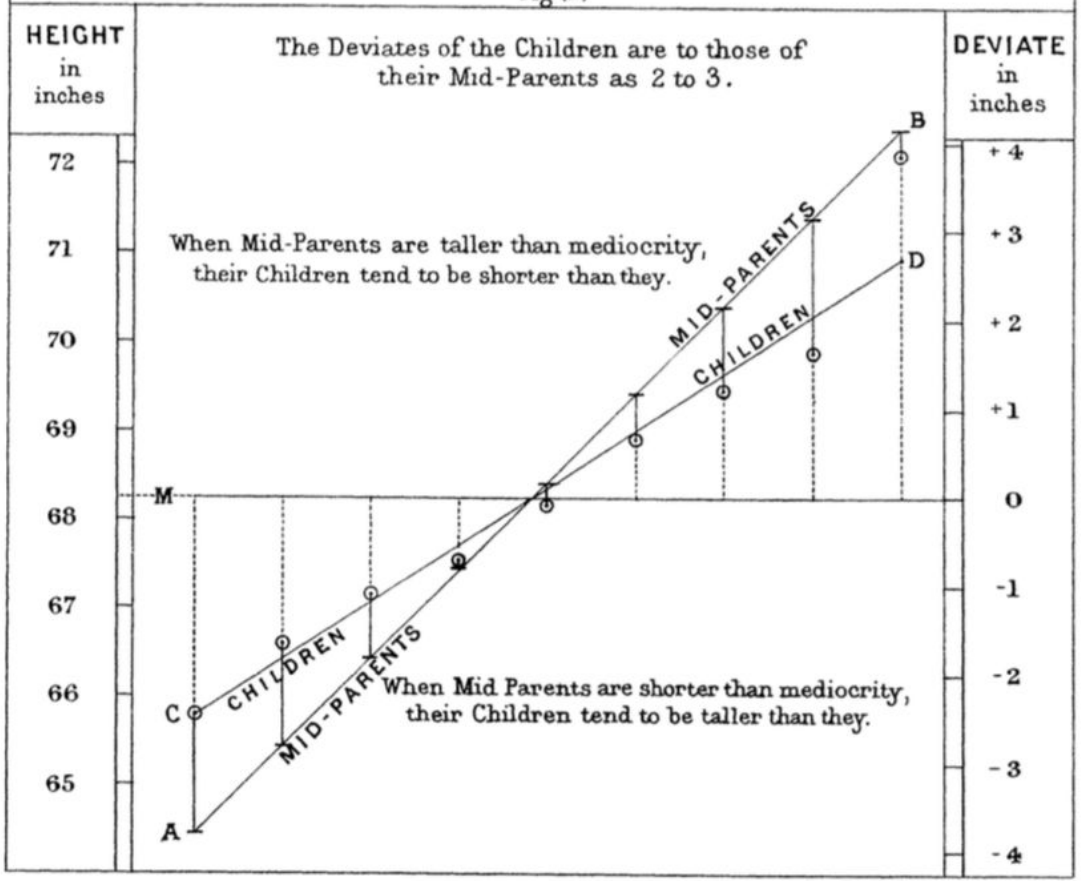


Galton, Journal of the Anthropological Institute of Great Britain and Ireland, 1886
Data at http://www.randomservices.org/random/data/Galton.html

Plate IX.

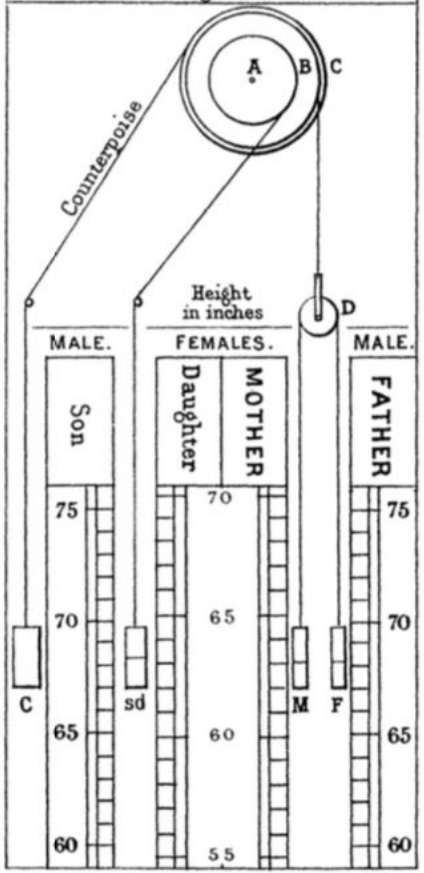
RATE OF REGRESSION IN HEREDITARY STATURE.

Fig. (a)



FORECASTER OF STATURE

Fig (b)

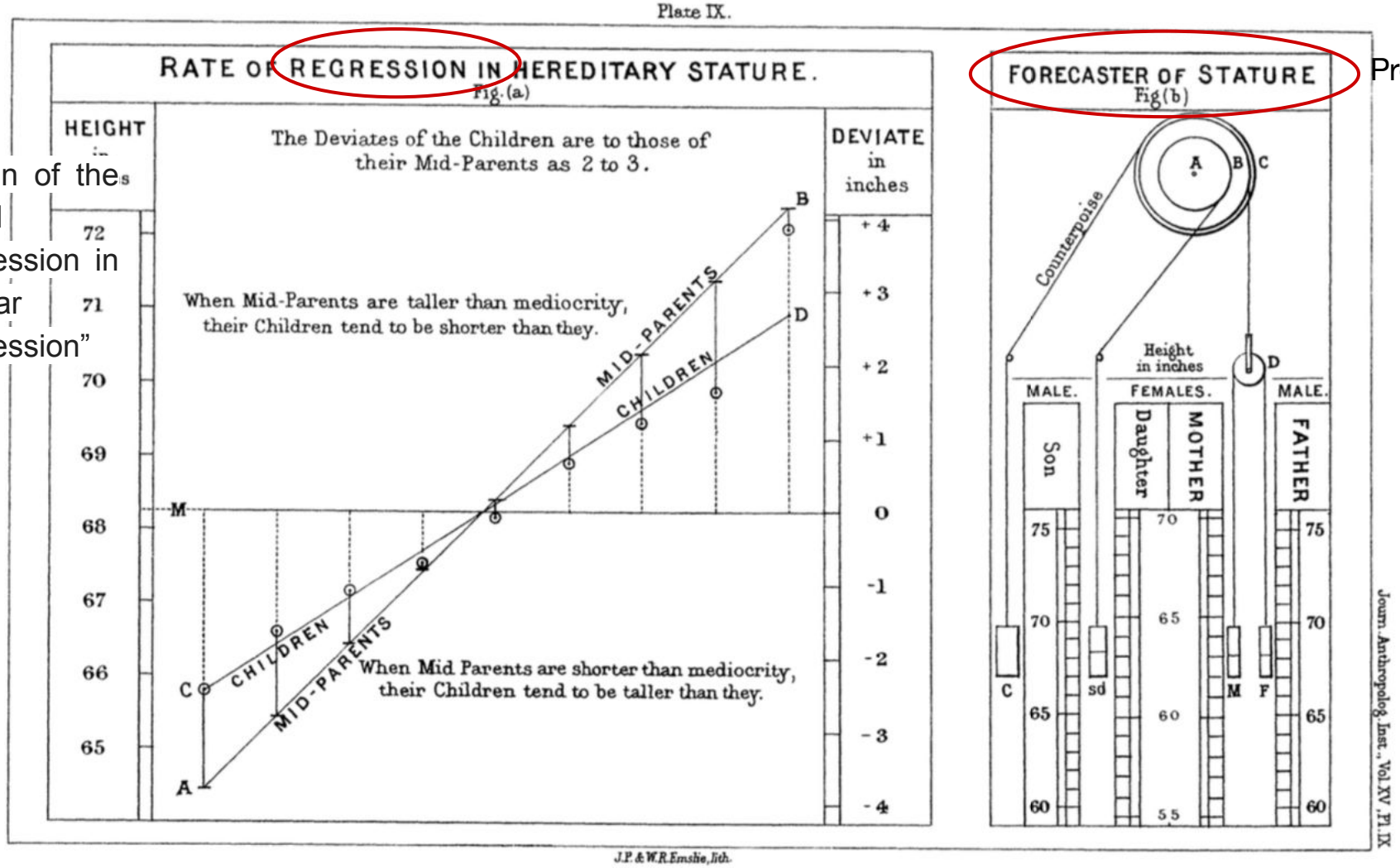


Journ. Anthropolog. Inst., Vol. XV. Pl. IX

J.P. & W.R. Emslie, lith.

Galton
1886

Origin of the word regression in "linear regression"



Predictor

Galton 1886