Regression Preamble

Predictive Modeling & Statistical Learning

Gaston Sanchez

So Far ...

So far ...

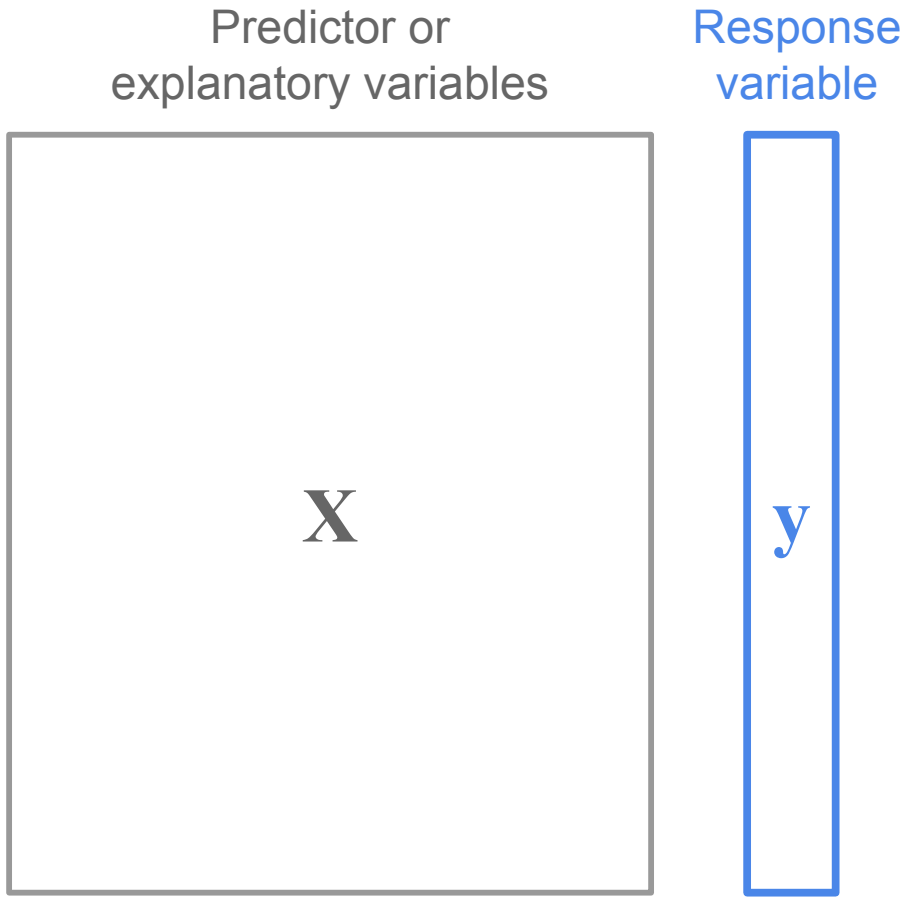
We've talked about PCA and general concepts of matrix algebra that constantly appear in Statistical Learning methods:

- data matrix
- the duality of a data matrix
- sum-of-squares and cross-products SSCP matrices
- EVD
- SVD
- *etc*

Now we switch to *Supervised Learning*

Supervised Methods

Predictive Methods (Supervised)



Supervised Methods

Two flavors

- **Regression**: quantitative response variable
- **Classification**: qualitative response variable

Supervised or Predictive

Build models and procedures for regression and classification tasks, and assess the predictive accuracy of those models and procedures when applied to new data.

Concept of a Model

- Suppose we observe a response Y
- We also observe p different predictors, $X_1, X_2, \ldots, X_p$
- We assume Y is related with $[X_1, \ldots, X_p]$
- The relationship can be written in a general form as

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

We want to use the information/values of variables $X_1, X_2, \ldots, X_p$ to "learn" about the response Y.

Concept of a Model

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

- $f()$ represents the systematic information—the *signal*—that the predictors provide about Y
- ϵ represents an *error* term—the *noise*—that is a catch-all for what we miss with the model

What kind of $f()$?

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

- In "classic" statistics, $f()$ takes the form of a function (with parameters to be estimated)
- Within statistical learning, $f()$ is more open-ended
- It can also take the form of an algorithm
- Sometimes $f()$ is a *black box*

Concept of a Model

Instead of $f()$:

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

You could also replace it with machine():

$$Y = \text{machine}(X_1, X_2, \ldots, X_p) + \epsilon$$

A couple of comments on $f()$

- No credible data scientist would ever claim that a given procedure will provide $f(X_1, X_2, \ldots, X_p)$ *exactly* capturing Y
- The data will be an imperfect reflection of $f(X_1, X_2, \ldots, X_p)$ because of ϵ
- ϵ is unobservable, and it cannot be removed from Y in order to obtain $f(X_1, X_2, \ldots, X_p)$

Conditional Distribution of $Y|X_1, X_2, \ldots, X_p$

As we shall see, the expression

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon$$

as well as the main idea behind Regressoin, focuses on the distribution of the response Y with respect to the predictors $X_1, X_2, \ldots, X_p$.

Regression Goal

To understand as far as possible with the available data how the conditional distribution of some response Y varies across subpopulations determined by the possible values of the predictor or predictors

Cook and Weisberg, 1999

Modeling for what?

Goal Tradeoff

Understanding -vs- Prediction

Modeling Goals

Understanding / Description Goal

A statistical model typically aims to provide a certain comprehension of the data and the mechanism that generated them through a parsimonious representation of a random phenomenon.

Prediction Goal

Sometimes also, a statistical model seeks to Predict new observations with "good" accuracy.

In this course we'll give priority to the prediction goal

Predictive Modeling

The Process of developing a mathematical tool or model that generates an accurate prediction.

Kuhn and Johnson, 2013

Model Performance

- From the predictive modeling standpoint, a "good" model is one which gives accurate predictions.
- By *predictions* we mean predictions of new data.
- Therefore we focus on the generalization ability of the model to predict unobserved data.
- This involves finding measure(s) of accuracy for predictions.

Some Important Comments

- Regression analysis seeks to characterize conditional distributions.
- Overall aim in regression analysis is to examine the distribution of $Y|X$
- Many features of $Y|X$ can be examined but the conditional mean $\bar{Y}|X$ is usually the focus.
- You are free to choose whatever procedures seem to be the most useful (e.g. graphs, linear models, nonlinear, parameteric, non-parameteric, etc).

Some Important Comments

- There is nothing in regression analysis that requires regression be formulated as a "causal model".
- There is nothing in regression analysis that requires statistical inference (e.g. test of hypotheses, confidence intervals).
- Regression analysis can be used to describe relationships but also to predict.

Historical Background

Introduction

In order to better understand the "modern" side of predictive models, it is worth talking about the origins of the Regression framework, and its consolidation during the first half of the 20th century.

Overview

- The method of Least Squares was discovered (or invented) in the early 1800's
- The idea of *Regression* appeared around 1885
- The connection between LS and Regression appeared in 1897

Least Squares

LS origins

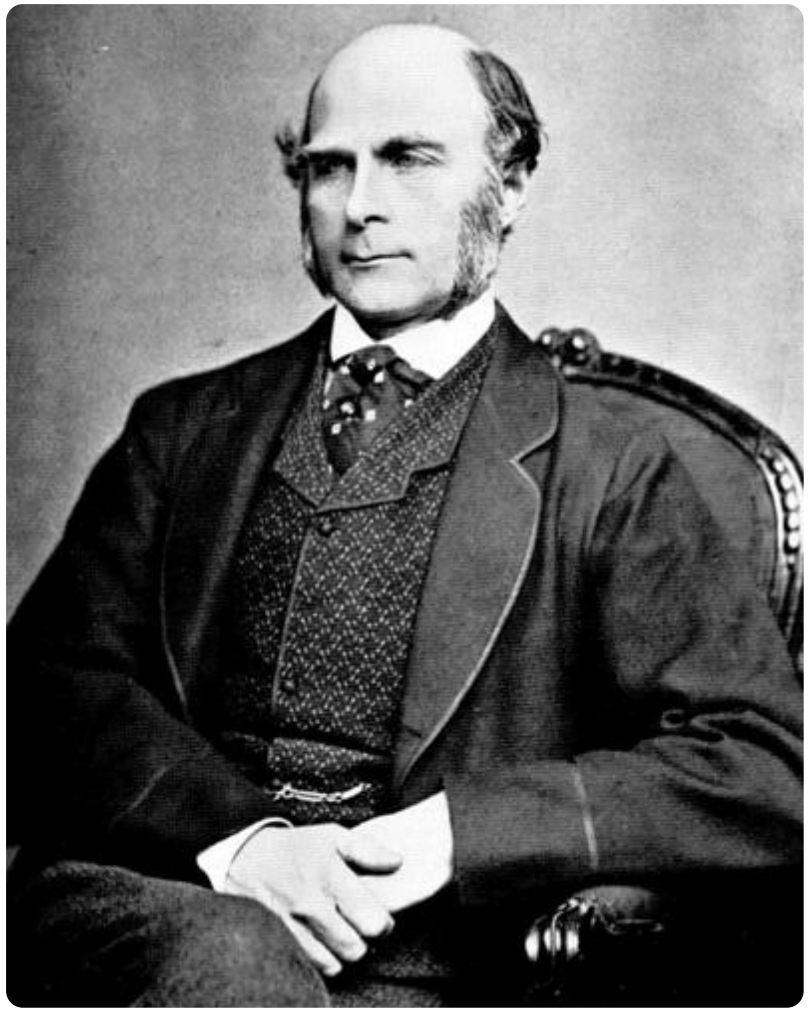
- Least Squares owes its origins to astronomy.
- Adrien-Marie Legendre's 1805 pioneering work on the determination of the orbits of planets.
- Carl Friedrich Gauss stated in 1809 that he had been using LS since 1795, but could not prove his claim with documented evidence.
- Legendre -vs- Gauss confrontation about who was the discoverer of Least Squares.

LS origins

- Within a few years, Gauss and Pierre Simon Laplace added a probability component (curve to describe the error distribution).
- Gauss deviced an elimination algorithm to compute LS estimates.
- Once introduced, LS caught on immediately in Astronomy and Geodetics.
- But it took 80 yers for these ideas to be applied to other disciplines.

Regression Ideas

Sir Francis Galton (1822 - 1911)



Galtons motto: "Count wherever you can"

About Galton

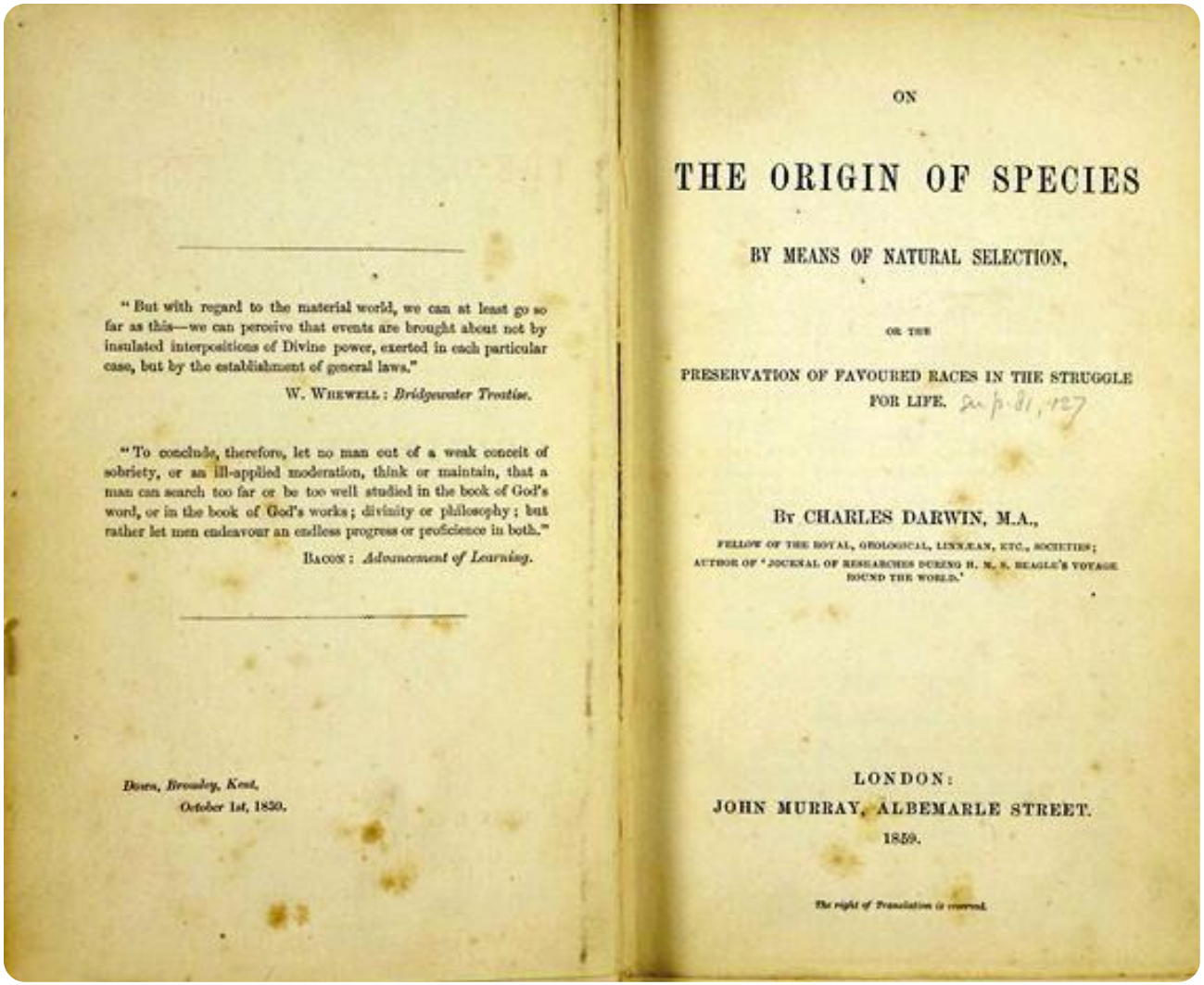
- Victorian Polymath
- 1859 - Reads *On the Origin of the Species* by Charles Darwin (Galton's half-cousin)
- 1869 - Publishes **Hereditary Genius** (book)
- 1883 - Coined the term **Eugenics**
- 1886 - **Regression towards Mediocrity in Hereditary Stature** (paper)

Galton's terminology

Some terms coined by Galton

- Percentile
- Decile
- Quartile, IQR
- Deviate (later called "deviations")
- Normal curve (before it was called the Error curve)
- Correlation
- Regression

1859 Darwin's "On the Origin of Species"



"But with regard to the material world, we can at least go so far as this—we can perceive that events are brought about not by insulated interpositions of Divine power, exerted in each particular case, but by the establishment of general laws."

W. Whewell: *Bridgewater Treatise.*

"To conclude, therefore, let no man out of a weak conceit of sobriety, or an ill-applied moderation, think or maintain, that a man can search too far or be too well studied in the book of God's word, or in the book of God's works; divinity or philosophy; but rather let men endeavour an endless progress or proficience in both."

Bacon: *Advancement of Learning.*

Down, Bromley, Kent,
October 1st, 1859.

ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

Thesis: Evolution by Natural Selection

Galton's research topics and ideas

- Improve humanity by carefully designed selection
- How to quantify effects of heredity?
- Inheritance of mental abilities
- Theory: intelligence, morality, civism, etc are determined by heredity

1870s Sweet pea seeds experiment

Measure weight of parents and offspring

1870s Sweet pea seeds experiment

Measure weight of parents and offspring

- Smaller parent seeds produce offspring that were larger.
- Larger parent seeds produce offspring that were smaller.

ANTHROPOMETRIC LABORATORY

For the measurement in various ways of Human Form and Faculty.

Entered from the Science Collection of the S. Kensington Museum.

This laboratory is established by Mr. Francis Galton for the following purposes:—

1. For the use of those who desire to be accurately measured in many ways, either to obtain timely warning of remediable faults in development, or to learn their powers.

2. For keeping a methodical register of the principal measurements of each person, of which he may at any future time obtain a copy under reasonable restrictions. His initials and date of birth will be entered in the register, but not his name. The names are indexed in a separate book.

3. For supplying information on the methods, practice, and uses of human measurement.

4. For anthropometric experiment and research, and for obtaining data for statistical discussion.

Charges for making the principal measurements:
THREEPENCE each, to those who are already on the Register.
FOURPENCE each, to those who are not:— one page of the Register will thenceforward be assigned to them, and a few extra measurements will be made, chiefly for future identification.

The Superintendent is charged with the control of the laboratory and with determining in each case, which, if any, of the extra measurements may be made, and under what conditions.

H & W. Brown, Printers, 20 Fulham Road, S.W.

The Galton Collection at UCL

The Galton Laboratory began life as the Anthropometric Laboratory which was part of the London International Health Exhibition of 1885. Visitors to the Exhibition were tested with a battery of machines many of which Galton had devised himself and paid a fee for a copy of their measurements and other data. Over 9,000 people contributed to the exercise, and the data gathered were not properly analyzed until the 1920s/30s. Following its success at the Exhibition, Galton established a permanent home for the Anthropometric Laboratory at the South Kensington Museum (which was renamed the Victoria & Albert Museum). Again, so much data was gathered that it was not until advantages in computer technology in the 1980s that any appropriate statistical analysis was done of these.

Galton's lab, South Kensington, 1884-5

Regression Ideas

ANTHROPOLOGICAL MISCELLANEA.

1886

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

Galton's reversion and regression

Reversion: reversio, -onis (go back to)

Regression: regressio, -onis (to return)

Galton's Height Data

- Height data
- 205 pairs of parents
- 928 adult children
- This was a breakthrough:
  - scatterplot with same variables (very novel at that time)
  - *mid-parent* idea
  - summarizing relationship with a line

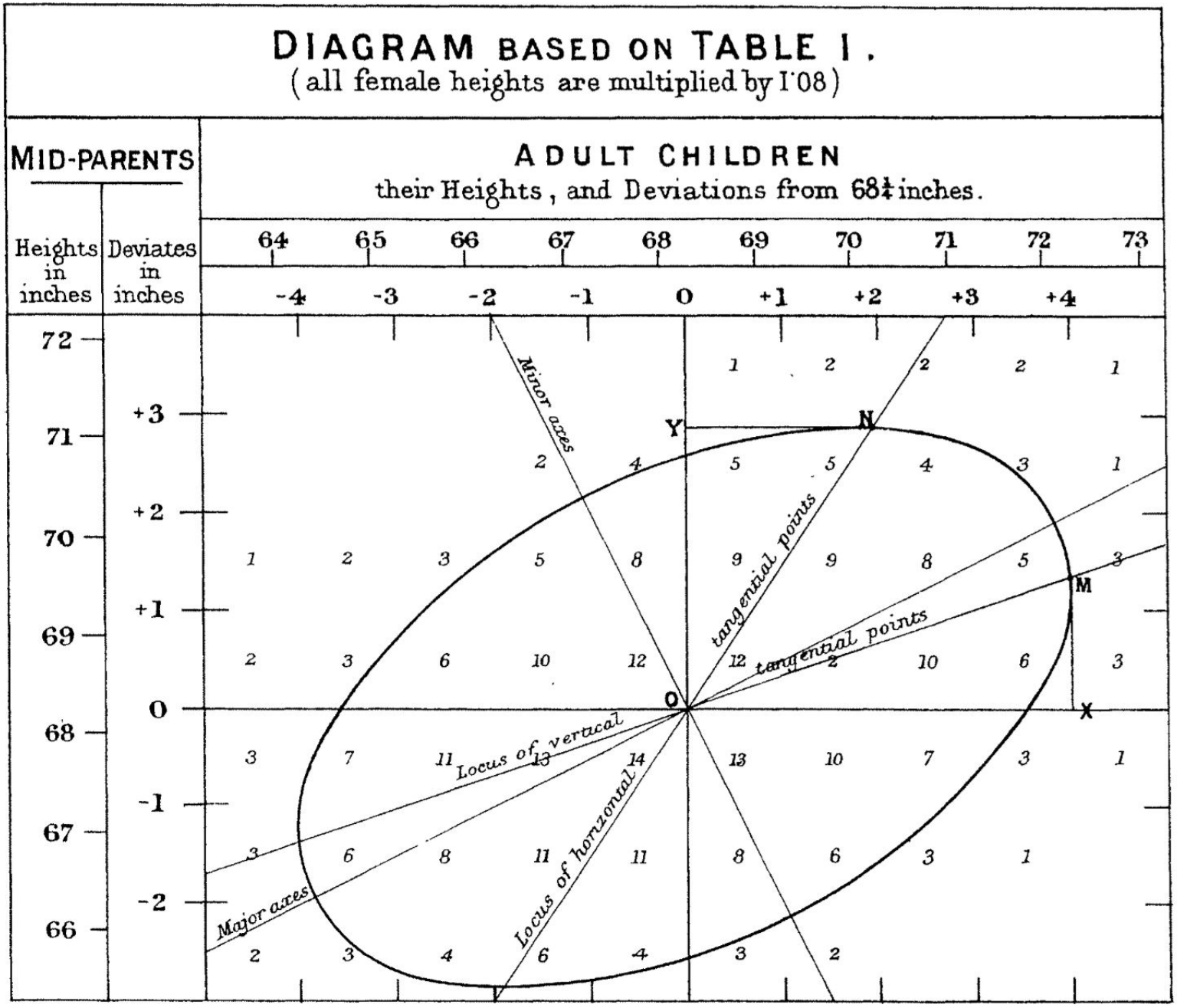
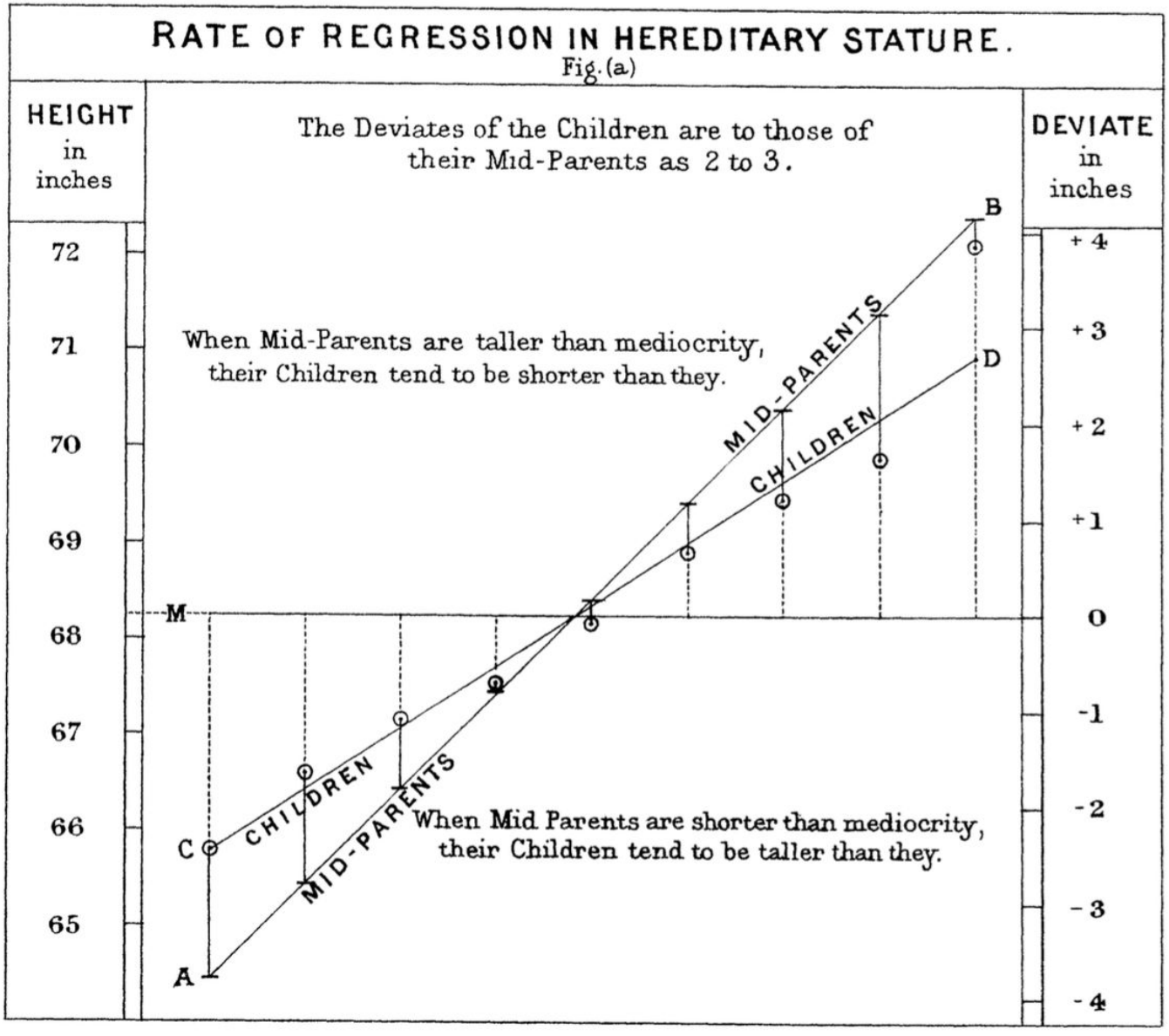


DIAGRAM BASED ON TABLE I.
(all female heights are multiplied by 1·08)
MID-PARENTS
Heights in inches
Deviates in inches
ADULT CHILDREN
their Heights, and Deviations from 68¼ inches.
64 65 66 67 68 69 70 71 72 73
-4 -3 -2 -1 0 +1 +2 +3 +4
72 71 70 69 68 67 66
+3 +2 +1 0 -1 -2
Minor axes
tangential points
tangential points
Locus of vertical
Locus of horizontal
Major axes
Y
N
M
X
O


RATE OF REGRESSION IN HEREDITARY STATURE.
Fig. (a)
HEIGHT in inches
DEVIATE in inches
The Deviates of the Children are to those of their Mid-Parents as 2 to 3.
When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they.
When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.
MID-PARENTS
CHILDREN
CHILDREN
MID-PARENTS
A
B
C
D
M
72
71
70
69
68
67
66
65
+4
+3
+2
+1
0
-1
-2
-3
-4

Galton's "Regression" Meaning

- Children's heights tend to "revert" to the average height of the population rather than diverting from it
- Future generations of offspring who are taller than average are not progressively taller than their respective parents
- and parents who are shorter than average do not beget successively smaller children

Karl Pearson and Francis Galton

Some Comments

- Galton (and Edgeworth and Pearson) failed to connect least squares to regression
- It was George Udny Yule (1897) who showed the connection of LS with regression
- Assuming errors in regression following a Normal distribution, could be replaced by an assumption that the variables were linearly related

References

- **Statistical Modeling: The Two Cultures** by Leo Breiman (2001). *Statistical Science, Vol 16 (3), 199-231*.
- **Models for Understanding versus Models for Prediction** by Gilbert Saporta (2008). COMPSTAT 2008. Physica-Verlag.
- **Applied Predictive Modeling** by Kuhn and Johnson (2013). Springer.
- **Statistical Learning from a Regression Perspective** by Richard Berk (2008). Springer.
- **Applied Regression Including Computing and Graphics** by Cook and Weisberg (1999). Springer.
- **Modern Multivariate Statistical Techniques** by Julian Izenman (2008). Springer.
- **Modern Regression Methods** by Thomas Ryan (1997).