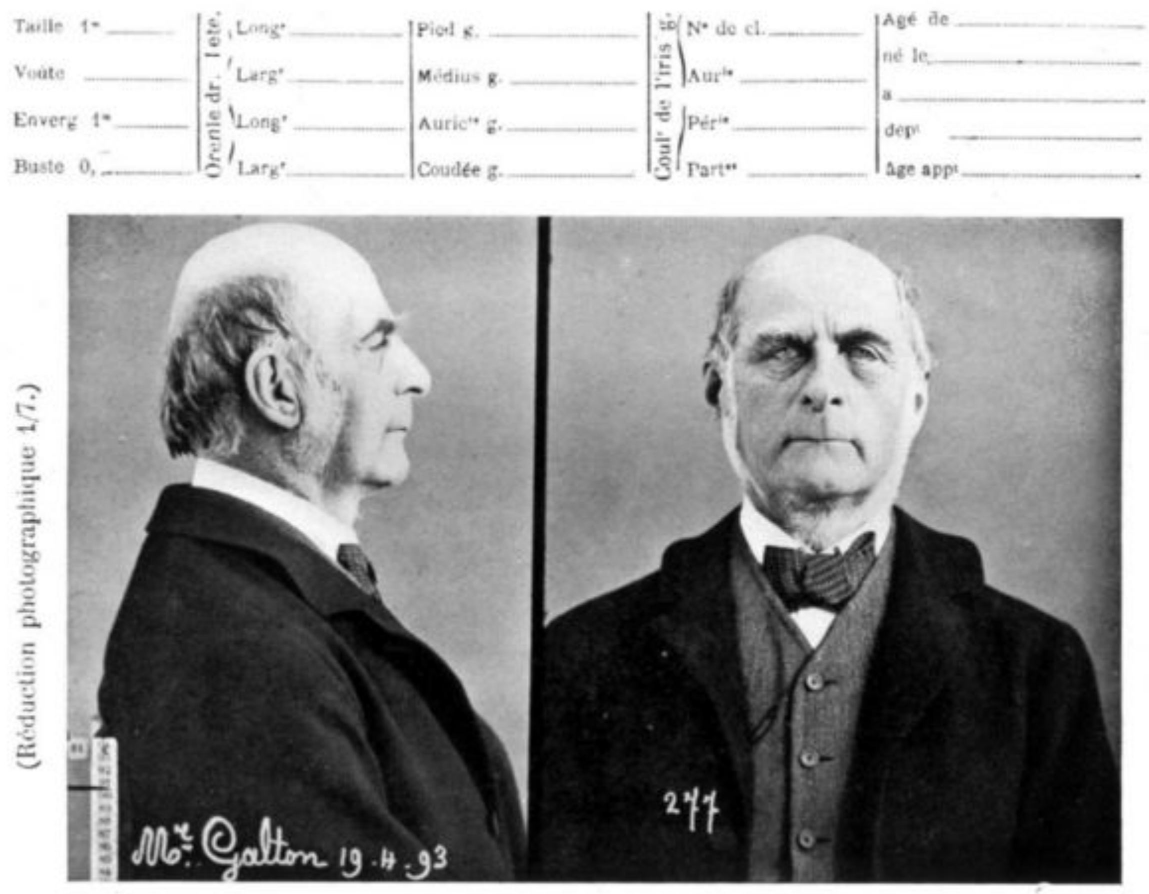


Regression Analysis: Model building, fitting and criticism
Statistics 201b

Regression and its uses

Suppose we have a **response** Y (also known as an output or a dependent variable) and **predictors** $X_1, \ldots, X_p$ (also known as inputs or independent variables) -- We find several goals lurking behind the banner of regression modeling

1. We might want to **examine the relationship between inputs and outputs** -- Do they tend to vary together? What does the “structure” of the relationship look like? Which inputs are “important”?

2. We are often interested in **making predictions** -- Given a new set of predictor values $X_1^*, \ldots, X_p^*$, what can we say about an unseen Y*? How accurate can we be?

3. Regressions can also be little more than **slightly sophisticated descriptive statistics**, providing us with summaries that we interpret like we would the sample mean or standard deviation

4. Finally, regression tools often serve as **a building block for more advanced methodologies** -- Smoothing by local polynomials, for example, involves fitting lots of regression models “locally”, while iteratively fitting weighted regressions is at the heart of the standard computations for generalized linear models

Regression and its uses

We actually find tension between the first two bullets on the previous slide -- Models that are good for prediction are often not the most transparent from a data modeling or data analytic perspective

Put another way, the price of interpretability might be diminished predictive power, especially when we start to look at more modern tools that depend on averages of large number of models (boosting) or willfully amplify the "dimensionality" of the problem (support vector machines)

Breiman's piece on the "two cultures" is still instructive reading...

Statistical Science
2001, Vol. 16, No. 3, 199–231

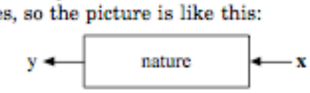
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

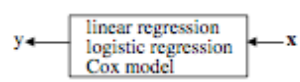
The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

$$\text{response variables} = f(\text{predictor variables}, \text{random noise, parameters})$$

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-4735 (e-mail: leo@stat.berkeley.edu).

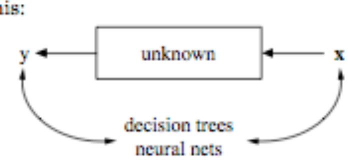
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;

Some history

Yes, you knew it was coming -- It’s not possible for me to present material without at least the barest hint of context or history

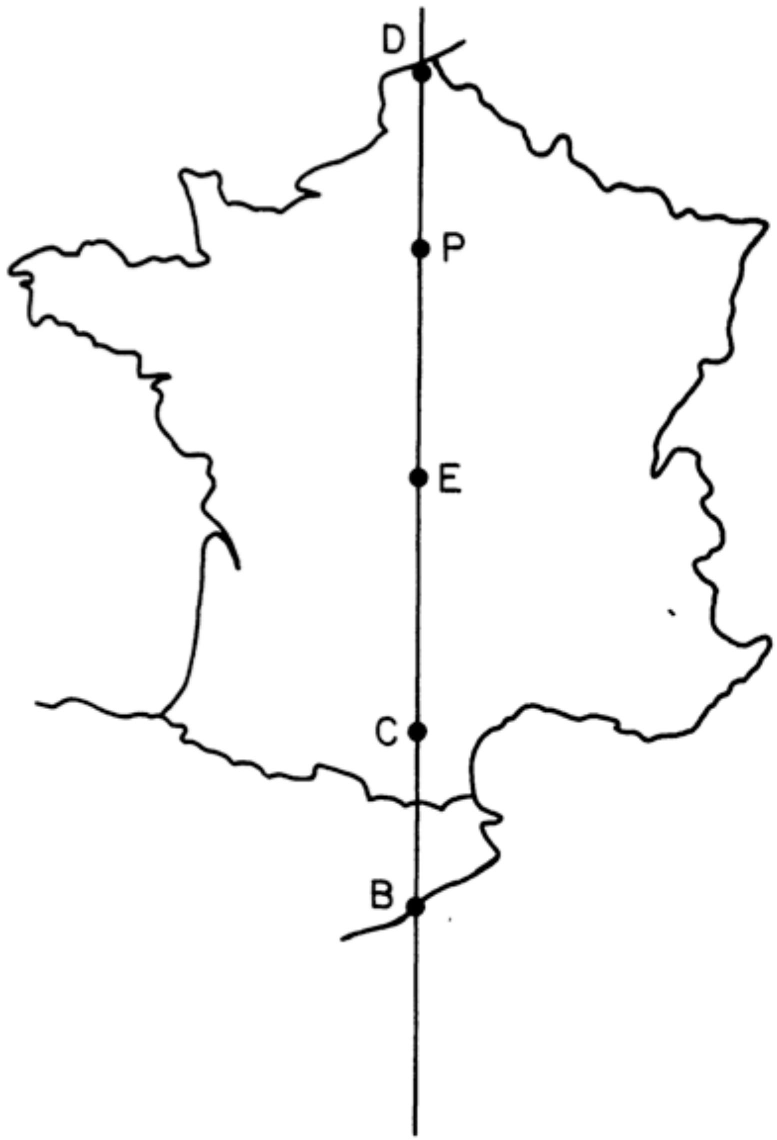
So let’s have a look at the “two cultures” (more or less) through two major advances that produced the technology we’ll come to know as “regression”...

The contested origins of least squares

Stephen Stigler, a well-known statistician who writes extensively on the history of our field, begins a 1981 article on least squares with the sentence “The **most famous priority dispute** in the history of statistics is that between Gauss and Legendre, over the discover of the method of least squares.”

Legendre is undisputedly the first to publish on the subject, laying out the whole method in an article in 1805 -- **Gauss claimed to have used the method** since 1795 and that it was behind is computations of the “Meridian arc” published in 1799

In that paper, Gauss used a famous data set collected **to define the first meter** -- In 1793 the French had decided to base the new metric system upon a unit, the meter, equal to one 10,000,000th part of the meridian quadrant, the distance from the north pole to the equator along a parallel of latitude passing through Paris...


D
P
E
C
B

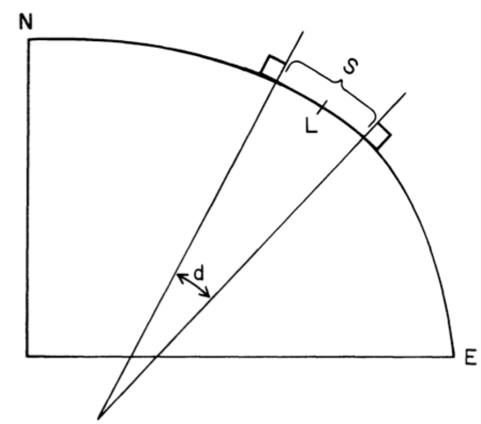

N
S
L
d
E

TABLE 1.

French arc measurements, from Allgemeine Geographische Ephemeriden, **4,** 1799, *page* xxxv. *The number* 76545.74 *is a misprint; the correct number is* 76145.74. *The table gives the length of four consecutive segments of the meridian arc through Paris, both in modules* S *(one module* ≅ 12.78 *feet) and degrees* d *of latitude (determined by astronomical observation). The latitude of the midpoint* L *of each arc segment is also given.*

	Modules S	Degrees d	Midpoint L
Dunkirk to Pantheon	62472.59	2.18910	49° 56′ 30″
Pantheon to Evaux	76545.74	2.66868	47° 30′ 46″
Evaux to Carcassone	84424.55	2.96336	44° 41′ 48″
Carcassone to Barcelona	52749.48	1.85266	42° 17′ 20″
Totals	275792.36	9.67380	

Least squares

The relationships between the variables in question (arc length, latitude, and meridian quadrant) are all nonlinear -- But for short arc lengths, a simple approximation holds

$$\mathsf{a} = (\mathsf{S}/\mathsf{d}) = \alpha + \beta \sin^2 \mathsf{L}$$

Having found values for α and β , one can estimate the meridian quadrant via

$$\text{meridian quadrant } = 90(\beta + \alpha/2)$$

Label the four data points in the previous table

$$(\mathsf{a}_1, \mathsf{L}_1), (\mathsf{a}_2, \mathsf{L}_2), (\mathsf{a}_3, \mathsf{L}_3) \text{ and } (\mathsf{a}_4, \mathsf{L}_4)$$

and apply **the method of least squares** -- That is, we identify values for α and β such that the sum of squared errors is a minimum

$$\sum_{\mathsf{i}=1}^{4} (\mathsf{a}_\mathsf{i} - \alpha - \beta \sin^2 \mathsf{L}_\mathsf{i})^2$$

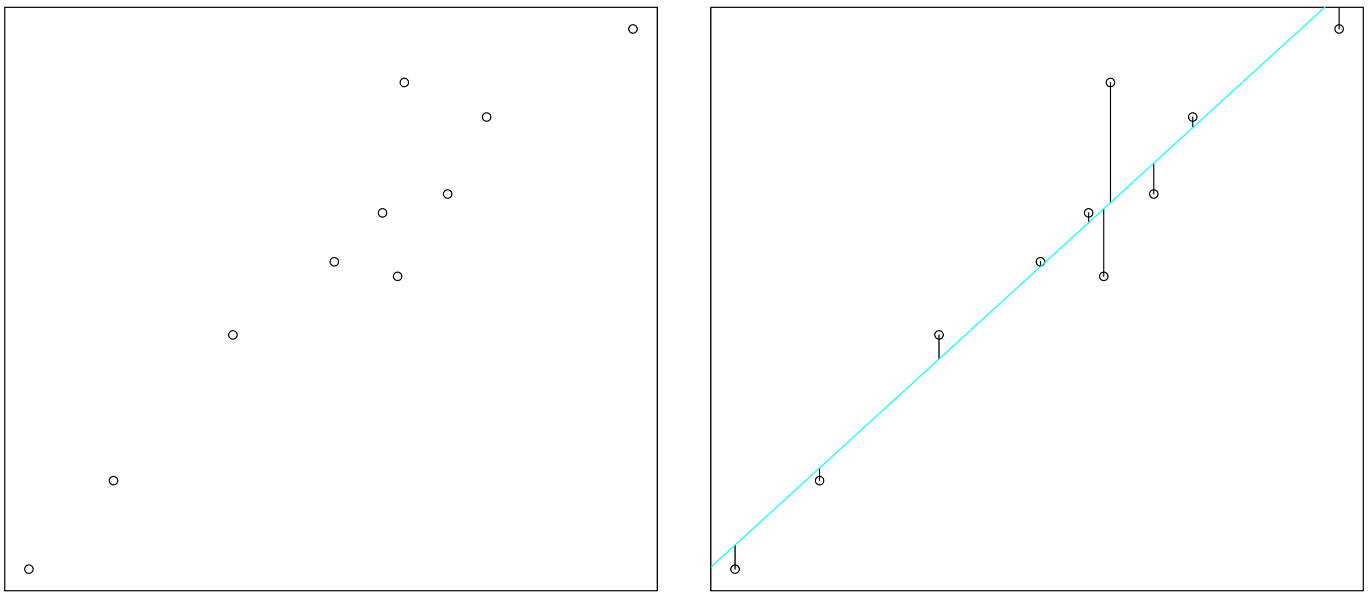
Least squares

Given a set of predictor-response pairs $(x_1, y_1), \ldots, (x_n, y_n)$, we can write the ordinary least squares (OLS) criterion (as opposed to a weighted version that we'll get to) as

$$\mathrm{argmin}_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Least squares

Graphically, in this simple case, we are doing nothing more than hypothesizing a linear relationship between the x and y variables and choosing that line that minimizes the (vertical) errors between model and data



Least squares

Given a set of predictor-response pairs $(x_1, y_1), \dots, (x_n, y_n)$, we can write the ordinary least squares (OLS) criterion (as opposed to a weighted version that we'll get to) as

$$\text{argmin}_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

By now you are all familiar with the idea that we can minimize this quantity by simply taking derivatives with respect to the parameters

$$\frac{\partial}{\partial \alpha} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)$$

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = -2 \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i)$$

Least squares

Setting these to zero yields the so-called normal equations

$$\alpha + \beta\,\overline{\mathsf{x}} = \overline{\mathsf{y}} \quad \text{and} \quad \alpha\,\overline{\mathsf{x}} + \beta\,\overline{\mathsf{x}^2} = \overline{\mathsf{xy}}$$

or in matrix form

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

After a little (familiar?) algebra, we can rewrite the expression in the form

$$\frac{y - \overline{y}}{\mathrm{sd}(y)} = r\,\frac{x - \overline{x}}{\mathrm{sd}(x)}$$

where r is the usual correlation coefficient (we write it this way because we'll see it again in a moment)

Gauss and least squares

Stigler attempts to reproduce Gauss's calculations, but cannot given the simple linearization (and a couple not-so-simple linearizations) on the previous slide

Ultimately, he reckons that because Gauss was a mathematician and not a statistician, he might have derived a more elaborate expansion -- No matter what form was used, **Stigler seems convinced that something like least squares was required**

Gauss eventually publishes on least squares in 1809, and his account of the method is much more complete than Legendre's -- **Linking the method to probability and providing computational approaches**

Gauss and least squares

My intention in bringing up this example is that least squares, as a method, has been around for a long time -- Numerical analysts often use the procedure for **fitting curves to data**, whether the underlying functional form is known (or an approximation is known as in the last example) or not (using a flexible basis of, say, polynomials or piecewise polynomials)

We will see many examples of similar problems in statistics -- **"Smoothing" procedures borrow tools directly from approximation theory** (from polynomials to smooth, piecewise polynomials to wavelets)

But statistics brings with it an emphasis on **model interpretation, model assessment, and the formal incorporation of uncertainty** -- It's interesting to compare how these same tools are used by statisticians and numerical analysts and how the analysis shifts as they pass across disciplinary boundaries

Galton and regression

While least squares, as a method, was developed by several people at around the same time (often ideas are “in the air”), regression as we have come to understand it, was almost entirely the work of one man

Stigler writes “Few conceptual advances in statistics can be as unequivocally associated with a single individual. Least squares, the central limit theorem, the chi-squared test -- all of these were realized as the culmination of many years of exploration by many people. Regression too came as the culmination of many years’ work, but in this case **it was the repeated efforts of one individual**.”

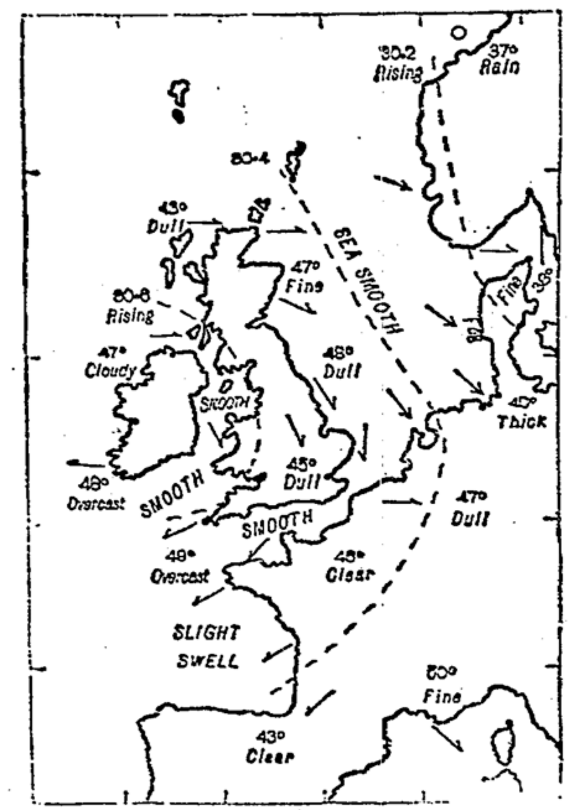
Galton and regression

Francis Galton (1822-1911) was at various points in his career an inventor, an anthropologist, a geographer, a meteorologist, a statistician and even **a tropical explorer** -- The latter gig paid quite well as his book “The art of travel” was a best seller

Among his many innovations, was **the first modern weather map**, appearing in The Times in 1875 -- To draw it, Galton requested data from meteorological stations across Europe

He also developed the use of **fingerprints as a means of identification** -- This work is just one small part of his larger interest how human characteristics (physical or even mental) varied across populations

WEATHER CHART, MARCH 31, 1875.



The dotted lines indicate the gradations of barometric pressure The variations of the temperature are marked by figures, the state of the sea and sky by descriptive words, and the direction of the wind by arrows—barbed and feathered according to its force. ⊙ denotes calm.

Galton and regression

Galton was also half-cousins with Charles Darwin (sharing the same grandfather) and took a strong interest in how physical and mental characteristics move from generation to generation -- **Heredity**

His work on regression started with a book entitled Hereditary Genius from 1869 in which he studied **the way “talent” ran in families** -- The book has lists of famous people and their famous relatives (great scientists and their families, for example)

He noted that there was a rather dramatic reduction in awesomeness as you moved up or down a family tree from the great man in the family (the Bachs or the Bernoullis, say) -- And thought of this as a kind of **regression toward mediocrity**

Galton and regression

In some sense, his work builds on that of Adolphe Quetelet -- Quetelet saw **normal distributions in various aggregate statistics** on human populations

Galton writes “Order in Apparent Chaos -- I know of scarcely anything so apt to impress the imagination as the wonderful cosmic order expressed by the Law of Frequency of Error. The law would have been **personified by the Greeks and deified**, if they had known of it.”

Galton and regression

Relating the normal curve (and the associated central limit theorem) to heredity, however, proved difficult for Galton -- He could not **connect the curve to the transmission abilities** or physical characteristics from one generation to the next, writing

"If the normal curve arose in each generation as the aggregate of a large number of factors operating independently, no one of them overriding or even significant importance, what opportunity was there for a single factor such as parent to have a measurable impact?"

So at first glance, the normal curve that Galton was so fond of in Quetelet's work was at odds with the possibility of "inheritance" -- Galton's solution to the problem would be **the formulation of regression and its link to the bivariate normal distribution**

NATURAL INHERITANCE

BY

FRANCIS GALTON, F.R.S.

AUTHOR OF

"HEREDITARY GENIUS," "INQUIRIES INTO HUMAN FACULTY," ETC.

FIG. 7.

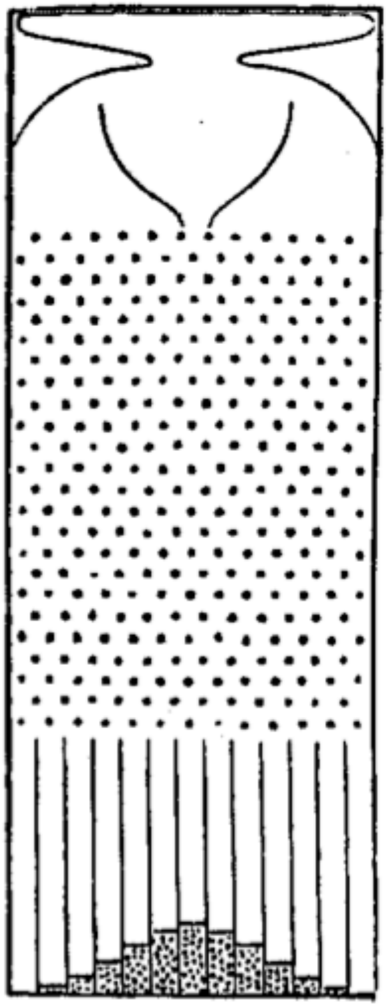


FIG. 8.

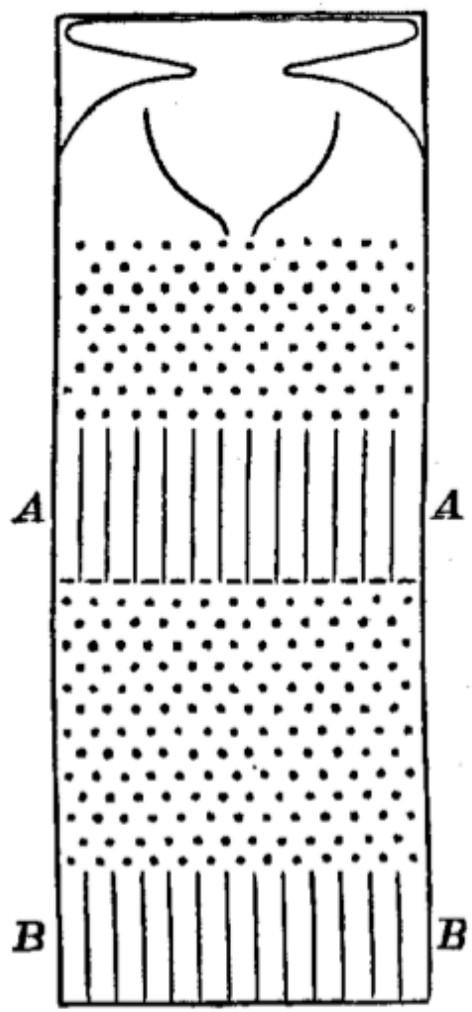
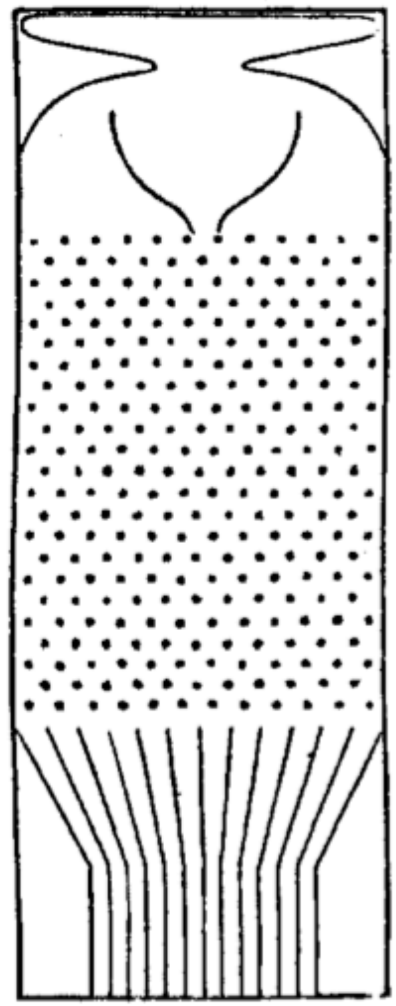


FIG. 9.



Galton and regression

In 1873, Galton had a machine built which he christened **the Quincunx** -- The name comes from the similarity of the pin pattern to the arrangement of fruit trees in English agriculture (quincunxial because it was based on a square of four trees with a fifth in the center)

The machine was originally devised to **illustrate the central limit theorem** and how a number of independent events might add up to produce a normal distribution -- Lead shot were dropped at the top of the machine and piled up according to the binomial coefficients at the bottom

The other panels in the previous slide illustrate a thought experiment by Galton (it's not clear the other devices were ever made) -- The middle region (between the A's) in the central machine, could be closed, **preventing the shot from working their way down the machine**

FIG. 7.

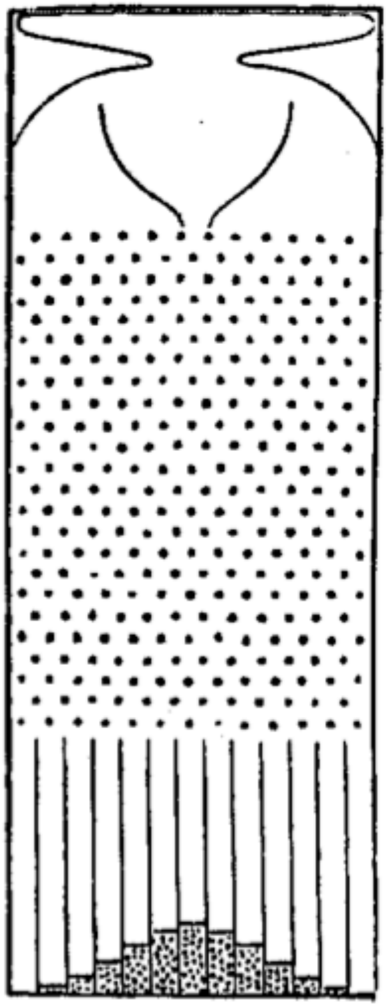


FIG. 8.

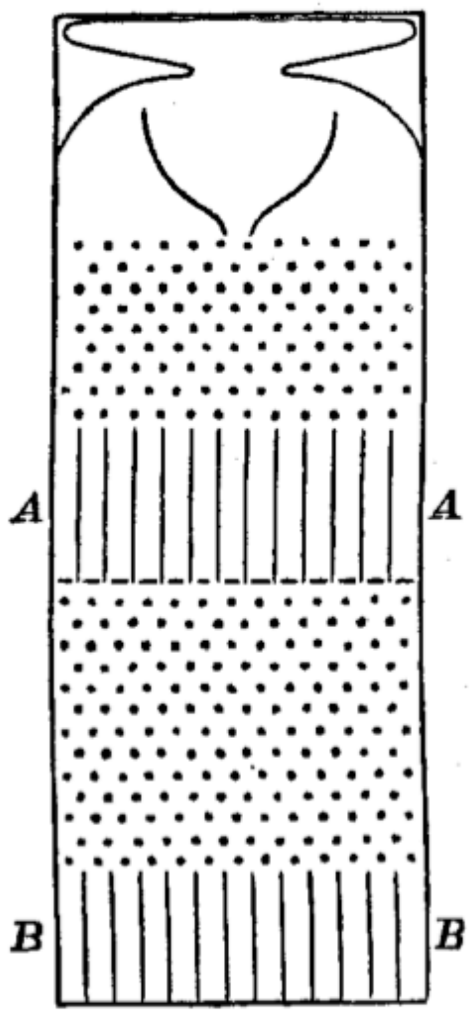
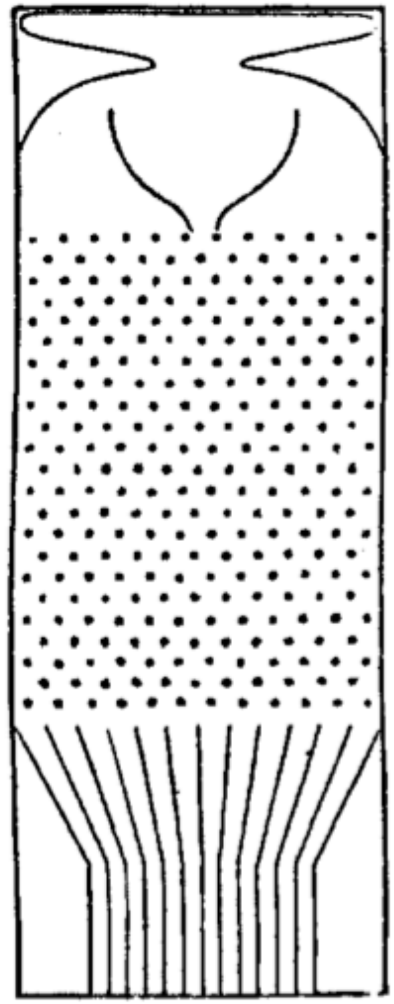


FIG. 9.



Galton and regression

By imagining holding back a portion of the shots, Galton expected to still see a normal distribution at the bottom of the machine, but one with less variation -- As he opened each barrier, t**he shot would deposit themselves according to small normal curves**, adding to the pattern already established

Once all the barriers had been opened, you'd be left with the original normal distribution at the bottom -- Galton, in effect, showed how the normal curve **could be dissected into components** which could be traced back to the location of the shot at A-A level of the device

In effect, he established that a normal mixture of normals is itself normal -- But with this idea in hand, **we see his tables of human measurements in a different light**...

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 simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Table 13 (Special Data).

Relative number of Brothers of various Heights to Men of various Heights, Families of Five Brothers and upwards being excluded.

Heights of the men in inches.	Heights of their brothers in inches.													Total cases.	Medians.
	Below 63	63·5	64·5	65·5	66·5	67·5	68·5	69·5	70·5	71·5	72·5	73·5	Above 74		
74 and above	1	1	...	...	...	...	...	1	1	...	5	3	12	24	
73·5	...	...	...	...	...	1	3	4	8	3	3	2	3	27	
72·5	...	...	...	...	1	1	6	5	9	9	8	3	5	47	71·1
71·5	...	1	...	1	2	8	11	18	14	20	9	4	...	88	70·2
70·5	...	...	1	1	7	19	30	45	36	14	9	8	1	171	69·6
69·5	...	1	2	1	11	20	36	55	44	17	5	4	2	198	69·5
68·5	...	1	5	9	18	38	46	36	30	11	6	3	...	203	68·7
67·5	2	4	8	26	35	38	38	20	18	8	1	1	...	199	67·7
66·5	4	3	10	33	28	35	20	12	7	2	1	...	...	155	67·0
65·5	3	3	15	18	33	36	8	2	1	1	...	...	...	110	66·5
64·5	3	8	12	15	10	8	5	2	1	...	...	...	...	64	65·6
63·5	5	2	8	3	3	4	1	1	...	1	...	...	1	20	
Below 63.........	5	5	3	3	4	2	...	...	...	...	...	...	1	23	
Totals.........	23	29	64	110	152	200	204	201	169	86	47	28	25	1329	

Galton and regression

Looking at these tables, we see the Quincunx at work -- The righthand column labeled “Total number of Adult Children” being the **counts of shot at the A-A level**, while the row marked “Totals” can be thought of as **the distribution one would see at the bottom of the device** when all the barriers are opened and the **individual counts in each row as the corresponding normal curves**

By 1877, Galton was starting to examine these ideas mathematically -- He essentially **discovered the important properties of the bivariate normal distribution** (the bivariate normal had been derived by theorists unknown to Galton, but they did not develop the idea of regression, nor did they attempt to fit it from data as Galton did)

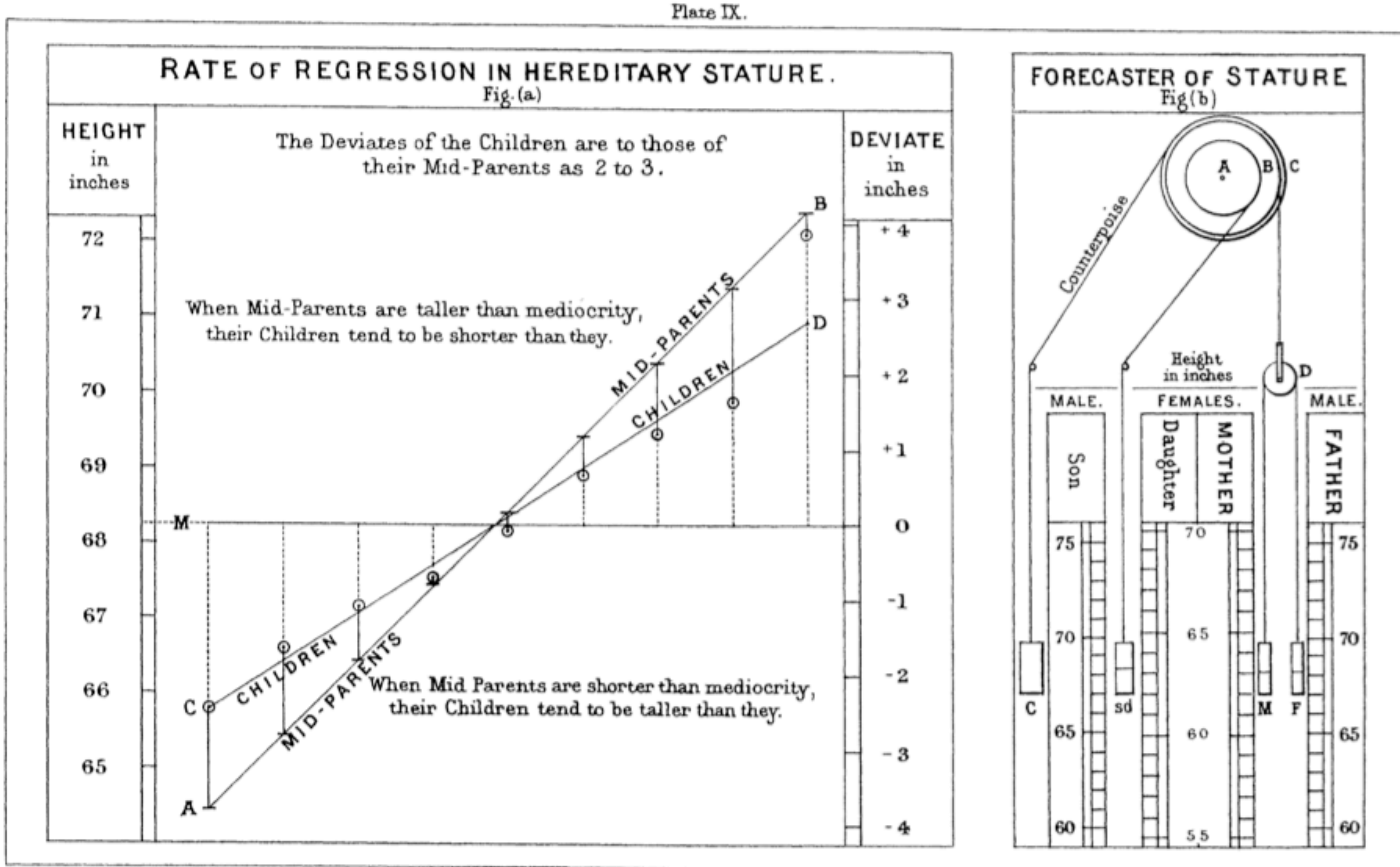
Galton and regression

In his text Natural Inheritance, he approached a table like this by first examining the **heights of the mid-parents** and noted that it appeared to be normal -- He then looked at the **marginal distribution of child heights** and found them to also be normally distributed

He then considered the heights of the children associated with different columns in his table, plotting median values against mid-parental height and finding a straight line (which he fit by eye)

He found that the slope was about 2/3 -- If children were on average as tall as their parents, he'd expect a slope of 1, leading him to coin the phrase "regression toward mediocrity"

Plate IX.



J P & W R Emslie, lith

Galton and regression

What Galton found through essentially geometric means was the following relationship (which we've seen earlier in the lecture)

$$\frac{y - \bar{y}}{sd(y)} = r \frac{x - \bar{x}}{sd(x)}$$

where we might take x to be the heights of mid-parents and y to be the heights of their adult offspring -- The quantity r is the correlation coefficient between x and y (another Galton innovation)

This gives a precise meaning to his phrase "regression to the mean"

FIG. 10.

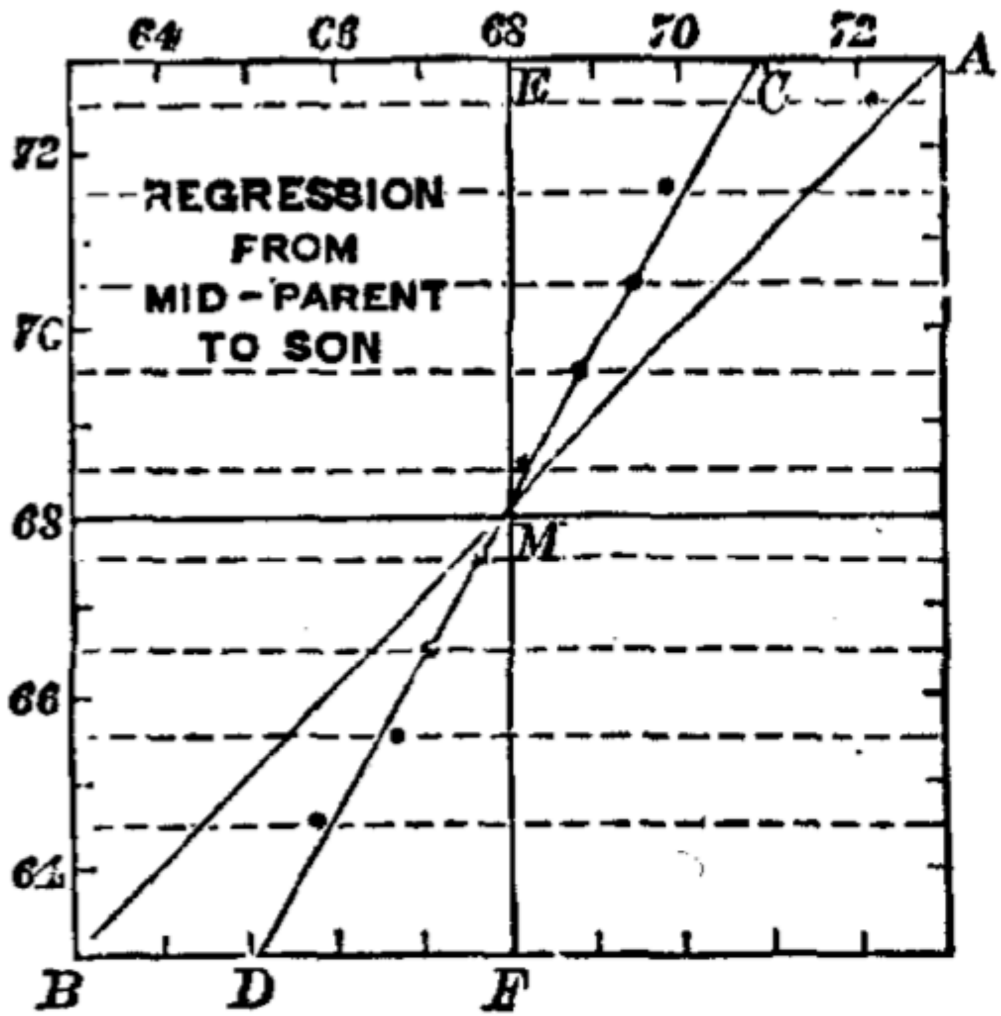
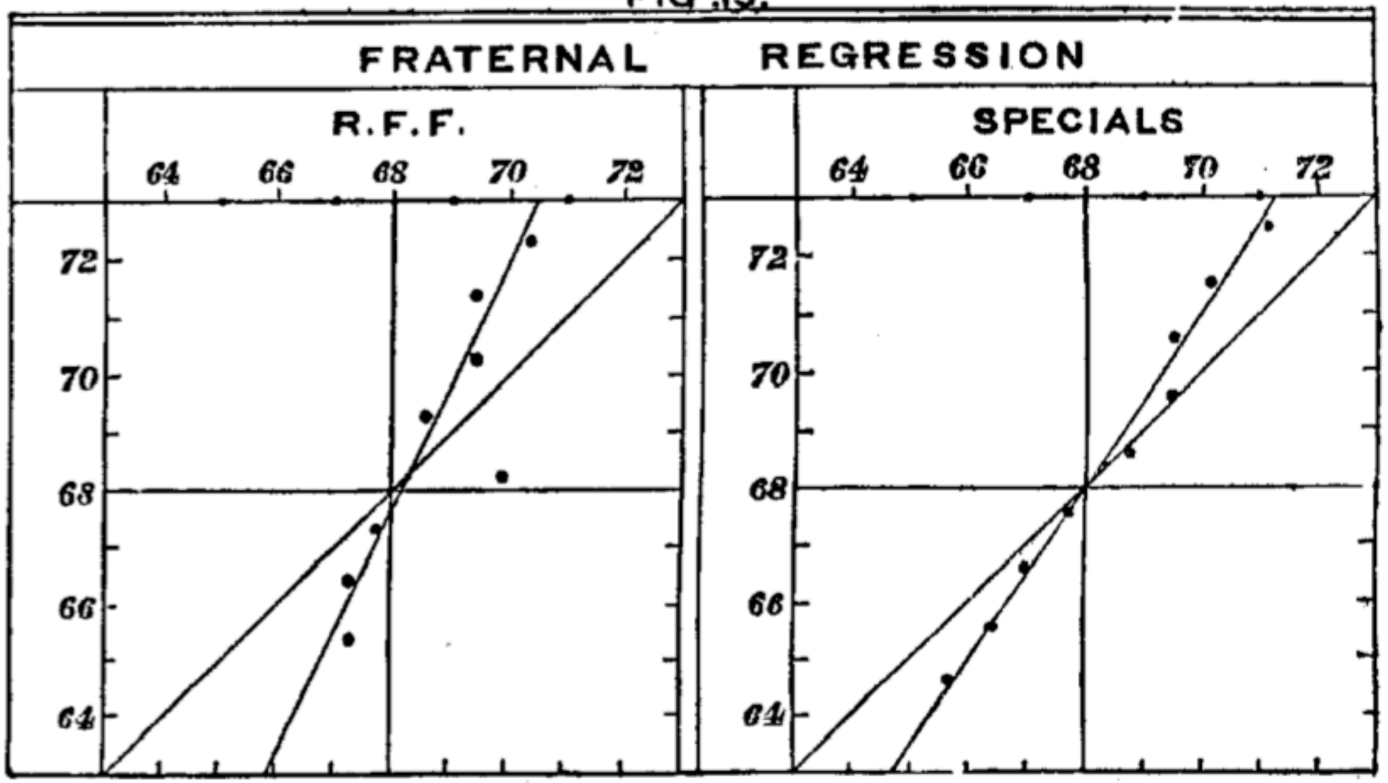


FIG. 13.


FRATERNAL REGRESSION
R.F.F.
SPECIALS
64
66
68
70
72
64
66
68
70
72
72
70
68
66
64
72
70
68
66
64

Galton and regression

Galton also noticed, however, that **a similar kind of regression happened in reverse** -- That is, that if you transposed the table, you'd find a slope of 1/3 relating the average mid-parent's height to that of their children

He surmised that the regression effect was more a fact about **the bivariate normal distribution** than anything else -- This is a lesson that many researchers have failed to appreciate even now

Here's Galton -- Notice that he's not content to just invent regression, but he also exhibits one of the first (if not the first) bivariate kernel density estimate!

I found it hard at first to catch the full significance of the entries in the table, which had curious relations that were very interesting to investigate. They came out distinctly when I "smoothed" the entries by writing at each intersection of a horizontal column with a vertical one, the sum of the entries in the four adjacent squares, and using these to work upon. I then noticed (see [fig. 6.6]) that lines drawn through entries of the same value formed a series of concentric and similar ellipses. Their common centre lay at the intersection of the vertical and horizontal lines, that corresponded to 68.25 inches. Their axes were similarly inclined. The points where each ellipse in succession was touched by a horizontal tangent, lay in a straight line inclined to the vertical in the ratio of 2/3; those where they were touched by a vertical tangent lay in a straight line inclined to the horizontal in the ratio of 1/3. These ratios confirm the values of average regression already obtained by a different method, of 2/3 from mid-parent to offspring, and of 1/3 from offspring to mid-parent, because it will be obvious on studying [fig. 6.6] that the point where each horizontal line in succession is touched by an ellipse, the greatest value in that line must appear at the point of contact. The same is true in respect to the vertical lines. These and other relations were evidently a subject for mathematical analysis and verification. (Galton 1885c, 254–255)

Galton and regression

To complete this story, Galton enlisted the help of a mathematician, Hamilton Dickson -- The problem he wanted solved was the following

Suppose x and y are expressed as deviations from the mean and that x is normal with mean zero and standard deviation Q_x

Also suppose that conditional on a fixed value of x, y is also normal with mean $\beta_{y|x}$ and standard deviation $Q_{y|x}$

What is the joint density of x and y and, in particular, are the contours of equal probability elliptical?

What is the conditional distribution of x given y, and in particular, what is the relation between the two regression coefficients?

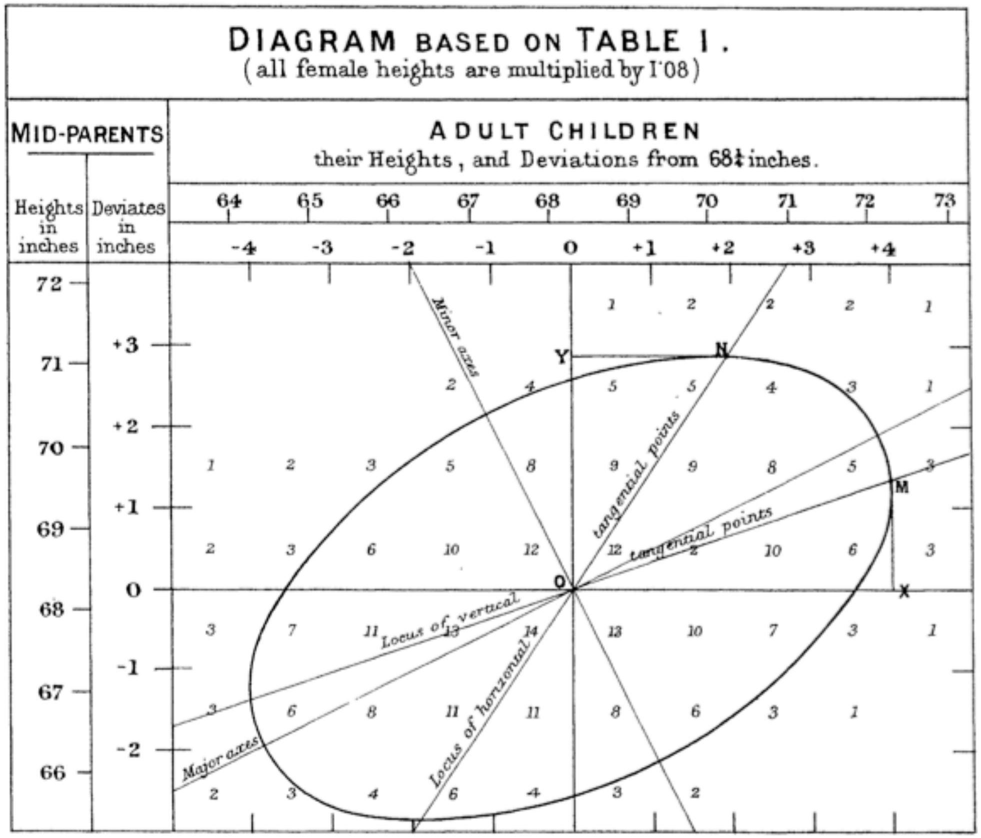
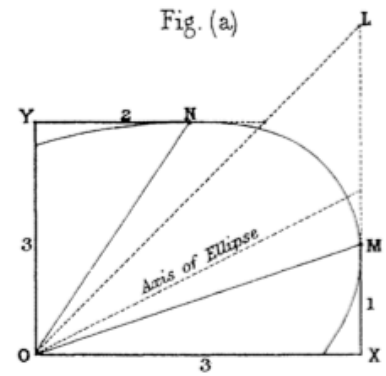


Fig. (a)



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

Galton and regression

In his response, Dickson **derived the bivariate normal distribution** and the associated marginals and conditionals -- For simplicity, let X and Y have two standard normal distributions with correlation

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right]$$

Then, after a little algebra, the conditional density of Y given X=x is just

$$\begin{aligned} f(y|x) &= \frac{f(x,y)}{f_x(x)} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{1}{2}\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right)^2\right] \end{aligned}$$

which we recognize as a normal with mean ρx and standard deviation $1-\rho^2$

Galton and regression

Despite the tremendous influence Galton had on the practice and (indirectly) theory of statistics, it’s worth asking **why he was so concerned with heredity?**

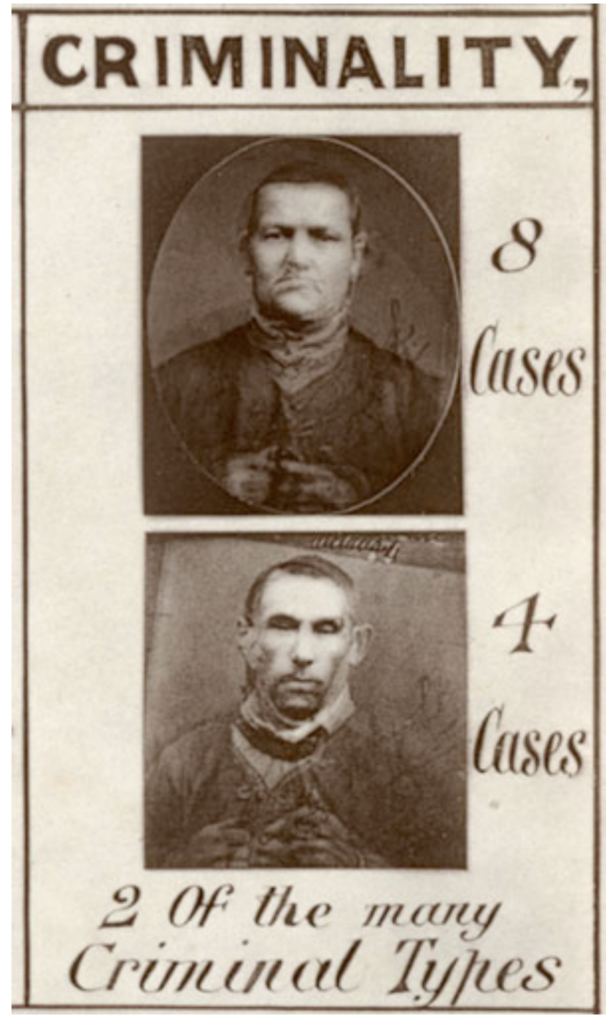
Tables of heights seem innocent, tracking familial “eminence” is maybe less so, but his photographic work...

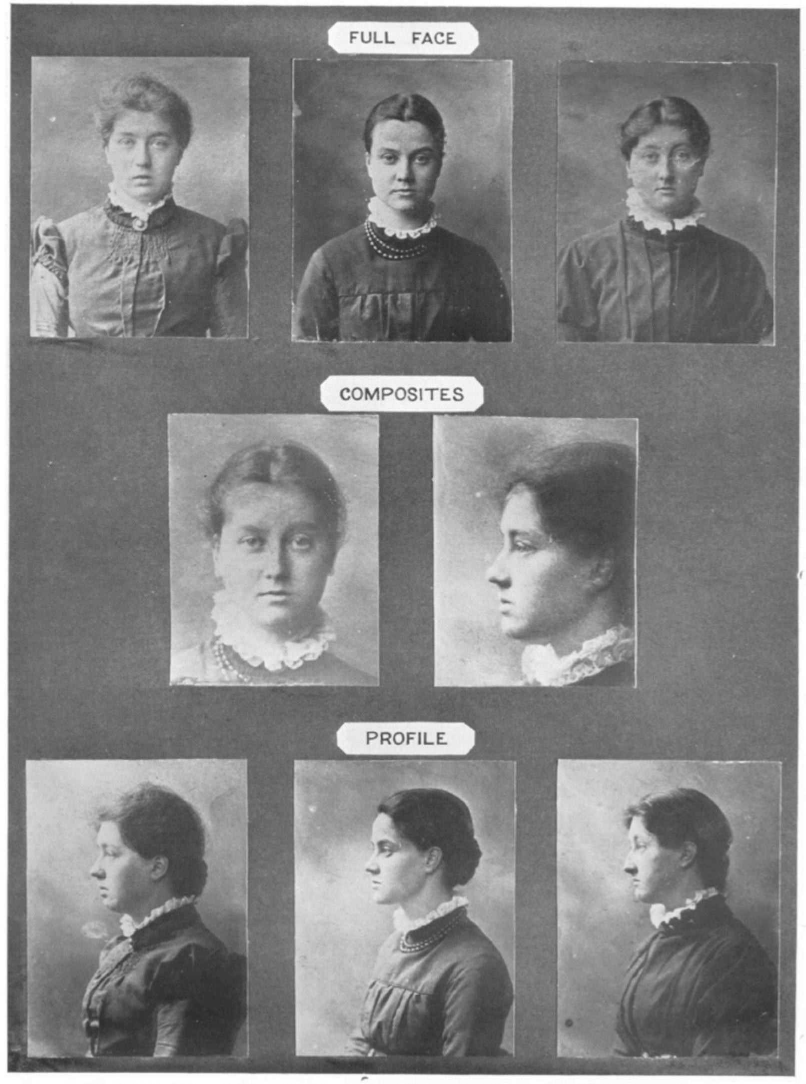
Galton

As we've seen, Galton was deeply committed to the idea of **the normal curve as an important force in nature** and (as with Quetelet) thought the mean value had particular importance as **an indicator of "type"**

Quetelet was more extreme than Galton, however, in that he believed deviations from the mean were more like small errors, and **regarded the mean as something perfect or ideal**

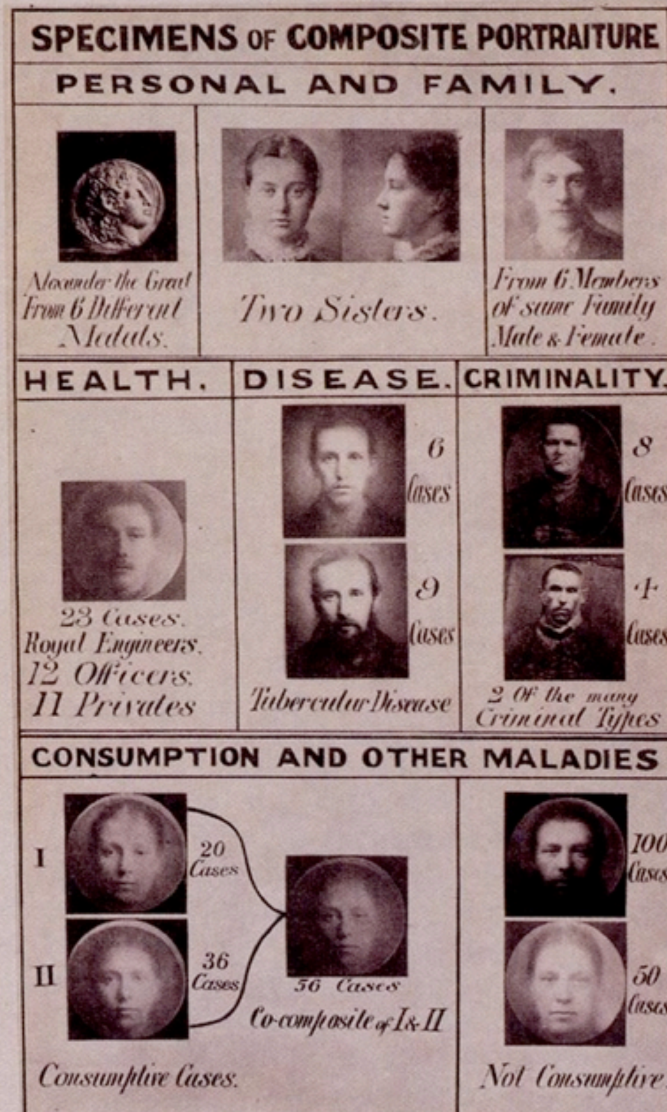
For Galton, these types were stable from generation to generation -- You can see this in his work on fingerprints or even in his **composite photography**





Portraits of three Sisters, full face and profile, with the corresponding Composites.

PREVALENT TYPES OF FEATURES AMONG MEN CONVICTED OF LARCENY (WITHOUT VIOLENCE)

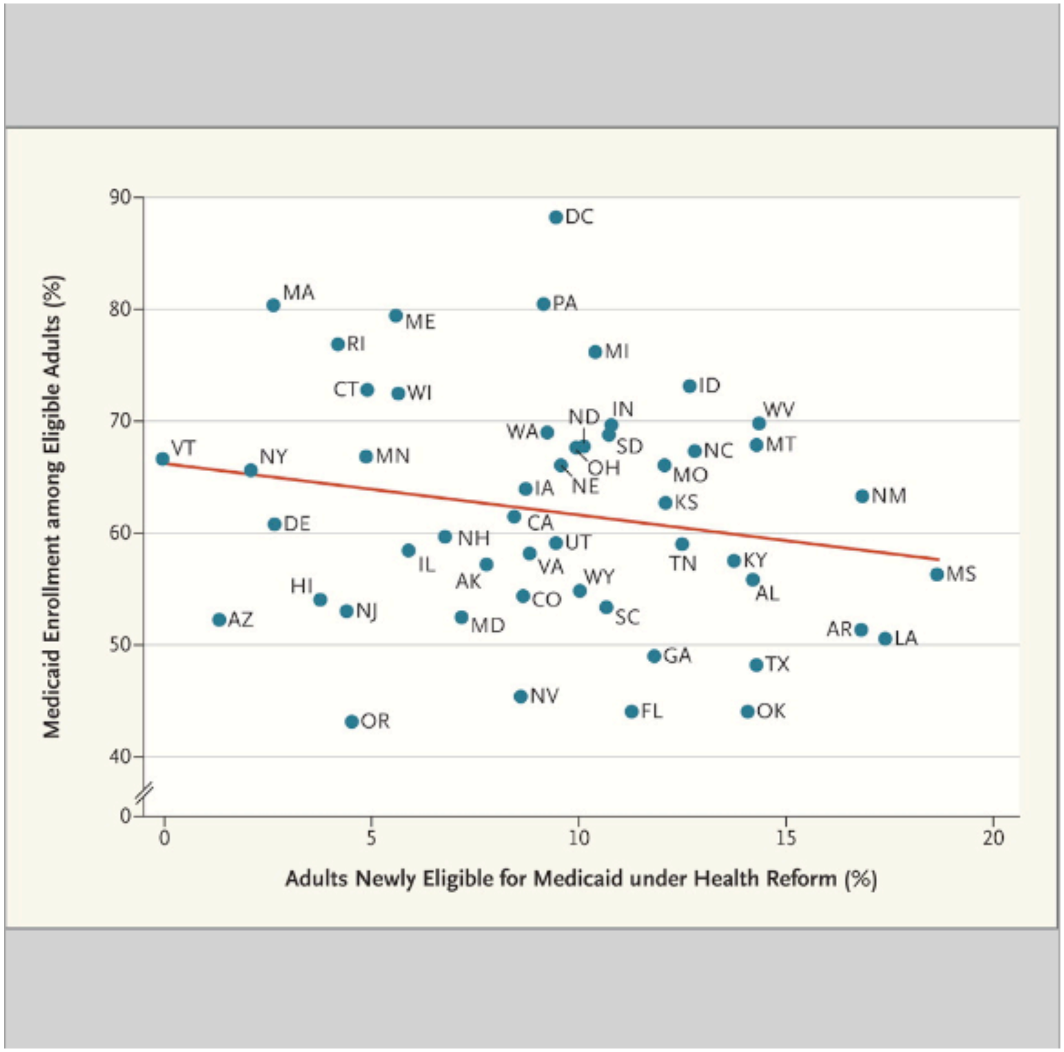

SPECIMENS OF COMPOSITE PORTRAITURE
PERSONAL AND FAMILY.
Alexander the Great From 6 Different Medals.
Two Sisters.
From 6 Members of same Family Male & Female.
HEALTH.
DISEASE.
CRIMINALITY.
23 Cases. Royal Engineers. 12 Officers. 11 Privates
6 Cases
9 Cases
Tubercular Disease
8 Cases
4 Cases
2 Of the many Criminal Types
CONSUMPTION AND OTHER MALADIES
I
20 Cases
II
36 Cases
56 Cases
Co-composite of I & II
Consumptive Cases.
100 Cases
50 Cases
Not Consumptive.

Regression today

Regression has become a powerful tool in many quantitative disciplines -- In many cases, a regression model acts as a kind of social probe, providing researchers with a glimpse into the workings of some larger phenomnenon

OK, that's generous. **It's also a highly abused tool**, one for which the elegant mathematics breaks down rather quickly once you hit modern practice -- Researchers often choose between many competing models, often through exhaustive searchers; data go missing and some form of imputation is often required; the underlying functional form is rarely linear and must also be estimated...

But here's what regression looks like in other fields...



Medicaid Enrollment among Currently Eligible Adults (2007 through 2009) and Percentage of Adults Who Will Become Eligible in 2014 under Health Care Reform, by State.

The population sample was restricted to eligible adults with no other form of health insurance; noncitizens were excluded from the analysis. Results are based on an analysis of data from the Current Population Survey of 2007 through 2009. The red line shows the regression equation: Enrollment=0.660.46×Newly Eligible (P=0.17).

Table 3. Regression Equations Predicting Sales, Number of Customers, Market Share, and Relative Profitability with Racial and Gender Diversity and Other Characteristics of Establishments

	Model 1	Model 2	Model 3	Model 4
Independent Variables	Sales	Customers	Market Share	Profitability
Constant	4.998***	61545.4	3.403***	3.363***
Racial diversity	.093***	433.86***	.007**	.006*
Gender diversity	.028**	195.642**	.001	.005**
Proprietorship	–.821	–370.78	–.232*	–.161
Partnership	.663	–6454.6	–.017	.256
Public corporation	–.109	7376.29	.214*	.202
Private corporation	–1.484**	–8748.7*	.008	.019
Company size	.00001*	.352*	.000**	.000**
Establishment size	.00001**	.119	.000	.000
Organization age	.013**	44.813	.001	.001
Agriculture	–1.942	4188.66	–.206	–.033
Mining	.739	–28856*	–.168	–.264
Construction	–.967	–875.7	–.152	–.036
Transportation/communications	–.052	1498.75	.119	.226
Wholesale trade	.008	–16383**	.136	–.064
Retail trade	–1.183**	7209.83*	.08	–.087
F. I. R. E.	–.683	–8335.4	–.212	.085
Business services	–1.49*	1552.28	.204*	.112
Personal services	–1.566*	1480.03**	.423**	–.001
Entertainment	–4.708**	–1504.5	–.191	–.076
Professional services	–.615	–13539**	.138	–.023
North	2.196***	23143.9***	–.06	–.039
Midwest	2.616***	14968.3***	–.023	–.073
South	1.82***	21152.8***	–.055	.059
R^2	.165***	.155***	.075**	.064**
N	506	506	469	484

Notes: Coefficients are unstandardized. For the dummy (binary) variable coefficients, significance levels refer to the difference between the omitted dummy variable category and the coefficient for the given category.
* $p < .1$; ** $p < .05$; *** $p < .01$.

Table 2. Percent Approving Same-Sex Marriage Ban; OLS Estimates, U.S. Counties

Independent Variable	1	2	3	4
Percent Women Not Working in Labor Force	.152**	.170**	.075*	.090*
Occupational Sex Segregation	16.310***	18.048***	8.216***	9.559***
Percent Households Married with Children	.219*	.322***	.156**	.217***
Percent Same-Sex Households	−2.038	−2.237	−5.428***	−5.956***
Percent Unmarried Opposite-Sex Households	−1.682***	−1.720***	−1.314***	−1.210***
Residential Instability	−.033	−.024	−.083**	−.066*
Percent Homes Not Owner Occupied	−.076*	−.084*	−.079***	−.109***
Crime Rate		.099***		.022*
Percent Production or Construction Occupations	.194***	.234***	.067**	.063*
Percent Professional Occupations	−.178*	−.100	−.118*	−.119*
Percent Self-Employed	−.100	−.093	−.198***	−.234***
Median Family Income ($1,000s)	−.138***	−.045	−.166***	−.148***
Percent Receiving Public Assistance	−.054	.037	.125	.207
Mean Years of Education	−1.328*	−2.611***	−2.040***	−2.606***
Percent Enrolled in College	−.060	.020	−.185**	−.139
Population Density (logged)	−.111	−.332	.001	.040
Percent Urban	.029**	.011	.015**	.011
Republican Voting (percent Bush 2000)	.350***	.381***	.287***	.317***
Percent Evangelical	.180***	.179***	.031***	.020
Percent Catholic	.063***	.064***	−.025**	−.035**
Median Age	−.155	−.022	−.314***	−.247**
Percent African American	.092***	.085***	−.006	.009
Percent Latino	.043*	.024	−.108***	−.118***
LGBT Organizations	−.114	.974	−2.096***	−1.576**
Civil Rights Organizations	−.007	−.004	.073	.113
Antidiscrimination Legislation	−3.283***	−2.797***	−2.656***	−1.918*
Alabama			−6.679***	−6.132***
Arizona			−23.814***	−23.878***
Arkansas			−9.454***	−9.085***
California			−2.079*	−1.867*
Colorado			−13.112***	−12.478***
Florida			−7.906***	−8.263***
Georgia			−2.787***	−3.307***
Idaho			−17.638***	−17.591***
Kansas			−5.857***	−6.051***
Kentucky			−7.623***	−9.111***
Louisiana			−4.771***	−5.492***
Michigan			−15.107***	−15.211***
Mississippi			−2.865***	−1.619
Missouri			−5.502***	−5.583***
Montana			−8.701***	−8.893***
Nebraska			−5.916***	−5.754***
Nevada			−3.886***	−3.964***
North Dakota			−4.316***	−4.057***
Ohio			−13.621***	−13.853***
Oklahoma			−8.002***	−7.835***
Oregon			−10.999***	−10.574***
South Carolina			−3.390***	−3.947***
South Dakota			−30.373***	−29.254***
Tennessee			−1.166	−.920
Utah			−18.009***	−18.697***
Virginia			−17.326***	−17.351***
Wisconsin			−12.236***	−11.792***
Texas (basis of comparison)				
Number of Observations	2,231	1,602	2,231	1,602
R-Square	.713	.735	.909	.915

* $p < .05$; ** $p < .01$; *** $p < .001$.

Table 2. OLS Estimates of Effect of Selected Measures of Residential Segregation on Log of Total Foreclosures

Variables	Dissimilarity Index B	Dissimilarity Index SE	Isolation Index B	Isolation Index SE
Index of Segregation				
African Americans	3.718**	.725	2.122**	.619
Hispanics	−.773	.596	.080	.656
Asians	−2.080*	.920	−2.161	1.636
Control Variables				
Housing Starts Ratio	2.980**	.960	3.067**	1.077
Wharton Land Use Index	.250**	.082	.272**	.096
Change in Housing Price Index	.082**	.024	.092**	.029
CRA-Covered Lending Share	−1.295	.912	−.810	1.061
Subprime Loan Share	3.022*	1.353	4.310**	1.581
MSA Credit Score Index	−.015*	.007	−.016*	.007
Log of Population	1.008**	.089	1.013**	.093
Percent with College Degree	−1.341	1.315	−.997	1.459
Log Median Household Income	.253	.509	.340	.515
Percent with Second Mortgage	.751	3.687	.225	4.350
Percent Workforce Unionized	−.025**	.011	−.022*	.011
Unemployment Rate	−.010	.064	.012	.071
Change in Unemployment Rate	.245**	.052	.213**	.063
Age of Housing Stock	.004	.012	.014	.013
Region				
Midwest	.434*	.200	.631**	.200
South	.042	.257	.081	.296
West	.463	.384	.679	.436
Coastal MSA	−.053	.123	.070	.133
Borders Rio Grande	−1.030**	.370	−1.054**	.380
Constant	1.960	7.557	.979	8.150
R^2	.91		.90	
Joint F-Test for Region	3.35*		7.97**	
Joint F-Test for Segregation	10.48**		6.28**	

Note: $N = 99$. Robust standard errors. Model also includes percent black, percent Hispanic, and percent Asian.
$^*p < .05$; $^{**}p < .01$ (two-tailed tests).

Regression

In the two major examples from this lecture (Gauss and Galton), we see two approaches to regression -- One based on **a loss-function** (the least squares criterion) and one that involves **a conditioning argument** using a formal data model (the bivariate normal)

Let’s see if we can bring these two approaches into agreement...

Regression

Let's recall a little probability and take Y to be a random variable representing our output, and $X = (X_1, \ldots, X_p)$ a random vector denoting our inputs -- Suppose we want to find a function h(X) for predicting values of Y

This seems to require some criterion to judge how well a function h is doing -- Let L(Y,h(X)) represent **a loss function that penalizes bad predictions**

For convenience, we start this quarter with squared error loss or simply

$$L(Y, h(X)) = (Y - h(X))^2$$

and define the expected (squared) prediction error to be

$$E\ L(Y, h(X)) = E(Y - h(X))^2 = \int [y - h(x)]^2 f(x, y)\, dx\, dy$$

Regression

We can rewrite this expression, conditioning on X to yield

$$\mathsf{E}\,\mathsf{L}(\mathsf{Y}, \mathsf{h}(\mathsf{X})) = \mathsf{E}_{\mathsf{X}} \mathsf{E}_{\mathsf{Y}|\mathsf{X}}([\mathsf{Y} - \mathsf{h}(\mathsf{x})]^2 | \mathsf{X})$$

which we can then consider solving pointwise

$$\mathsf{h}(\mathsf{x}) = \mathrm{argmin}_{\mathsf{z}}\, \mathsf{E}_{\mathsf{Y}|\mathsf{X}}([\mathsf{Y} - \mathsf{z}]^2 | \mathsf{X} = \mathsf{x})$$

This produces the simple conditional mean $\mathsf{h}(\mathsf{x}) = \mathsf{E}(\mathsf{Y}|\mathsf{X} = \mathsf{x})$ -- So, under squared error loss, the conditional mean is the best prediction of Y at any point X=x

Regression

In the case of a bivariate normal distribution, this conditional expectation is, in fact, linear -- There are certainly plenty of other situations in which an assumption of linearity is (at best) an approximation (all smooth functions h looking linear in small neighborhoods)

Regression, then, has come to stand for a suite of algorithms that attempt to estimate the mean (or some centrality parameter) of an output conditional on one or more input variables

Modern regression

As an analysis tool, the practice of regression seems to have undergone a massive shift in the 1970s -- Writing to some of the big names publishing at the time (Cook, Allen, Weisberg), this seems to be due in part to the shifting nature of computing

It was also noted that an interest in regression was “in the air” as it was the hot topic of the decade (What is the hot topic today? What’s “in the air today?)

TECHNOMETRICS

INDEX TO CONTENTS, Volume 12, 1970

TECHNOMETRICS

INDEX TO CONTENTS, Volume 13, 1971

TECHNOMETRICS

INDEX TO CONTENTS, Volume 15, 1973

Practicalities

OK so that wasn’t quite what I wanted, but in the 1970s you have the following innovations appearing

1. Diagnostic tools (leave-one-out measures, influence, Cook’s distance)
2. “Automatic” criterion for variable selection (Cp, AIC, BIC)
3. Simulation techniques for inference (the bootstrap)
4. Computational schemes for subset selection (leaps and bounds, say)
5. Computational tools for fitting (SVD, QR decomposition -- well, mid1960s)
6. Biased or penalized estimates (ridge regression)
7. Alternate loss functions (robust regression, quantile regression)
8. Flexible modeling (local polynomials, global B-splines, smoothing splines)
9. New data types (generalized linear models)
10. The Bayesian linear model

Practicalities

Since the 1970s, regression has continued to flourish with new advances in nonparametric methods (wavelets, averaged or boosted predictors, kernel methods), new approaches to penalties (the lasso, say) and an explosion in Bayesian tools

We intend to cover all of this during the quarter!

To start

Let's go back to our general framework -- We have an output or response Y and inputs or predictors $X_1, \ldots, X_p$, both of which we can think of as random variables (although given most of you are coming from 201a, you can think of the X's as deterministic)

We can express a stochastic relationship between inputs and outputs with the formula

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

where we assume the error term is independent of the predictors, have mean zero and constant (finite) variance σ^2

To start

Suppose we are now given data from this model -- That is, we have n data pairs $(x_1, y_1), \ldots, (x_n, y_n)$ where (with a slight abuse of notation) $x_i = (x_{i1}, \ldots, x_{ip})$

The simple linear model (regression model) relating inputs to outputs is then

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

For the moment, we'll assume that if the model has an intercept, it is represented by one of the p predictors -- It's a boring predictor that's simply 1 for each data pair

To start

We determine estimates for the unknown parameters $\beta_1, \ldots, \beta_p$ and σ^2 via ordinary least squares -- That is, we want to minimize the quantity

$$\sum (y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

Taking partial derivatives as we had for the case of simple regression (and now with p > 1, we use the term multiple regression), we can derive the so-called normal equations

$$\begin{aligned}
\beta_1 \sum x_{i1}^2 + \beta_2 \sum x_{i1}x_{i2} + \cdots + \beta_2 \sum x_{i1}x_{ip} &= \sum y_i x_{i1} \\
\beta_1 \sum x_{i1}x_{i2} + \beta_2 \sum x_{i2}^2 + \beta_3 \sum x_{i2}x_{i3} + \cdots + \beta_2 \sum x_{i1}x_{ip} &= \sum y_i x_{i2} \\
&\vdots \\
\beta_1 \sum x_{i1}x_{ip} + \cdots + \beta_{p-1} \sum x_{i,p-1}x_{ip} + \beta_p \sum x_{ip}^2 &= \sum y_i x_{ip}
\end{aligned}$$

To start

While this seems tedious, we can again appeal to a matrix formulation -- Let X denote the so-called design matrix and y the matrix of responses

$$M = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1,p-1} & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2,p-1} & x_{2p} \\ \vdots & & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{n,p-1} & x_{np} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Then, we can (and this is not news to any of you I am certain) rewrite the normal equations in the form

$$M^t M\beta = M^t y$$

where we have collected our regression coefficients into the vector $\beta = (\beta_1, \ldots, \beta_p)^t$

To start

Now, assuming the matrix M^tM is invertible (it's time to dust off your linear algebra books!) we can form an estimate of our regression coefficients using the (symbolic!) manipulation

$$\widehat{\beta} = (M^tM)^{-1}M^ty$$

Similarly, the estimated conditional mean for the ith data point is simply

$$\widehat{y}_i = \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip}$$

which we can write in matrix notation as

$$\begin{aligned}
\widehat{y} &= M\widehat{\beta} \\
&= M(M^tM)^{-1}M^ty \\
&= Hy
\end{aligned}$$

To start

The matrix H is known as the hat matrix (for the obvious reason that it carries our observed data into an estimate of the associated conditional means, in effect placing a “hat” on y)

We can derive some properties of H somewhat easily -- For example, H is symmetric (easy) and it’s idempotent

$$H^2 = HH = M(M^tM)^{-1}M^tM(M^tM)^{-1}M^t = M(M^tM)^{-1}M^t = H$$

We can compute the residuals from our fit $\widehat{\epsilon}_i = y_i - \widehat{y}_i$ as $\widehat{\epsilon} = (I - H)y$, so that the residual sum of squares can be written as

$$\sum_i \widehat{\epsilon}_i^2 = \widehat{\epsilon}^t\widehat{\epsilon} = y^t(I-H)^t(I-H)y = y^t(I-H)y$$