“Transmuting” women into men: Galton’s data on human stature

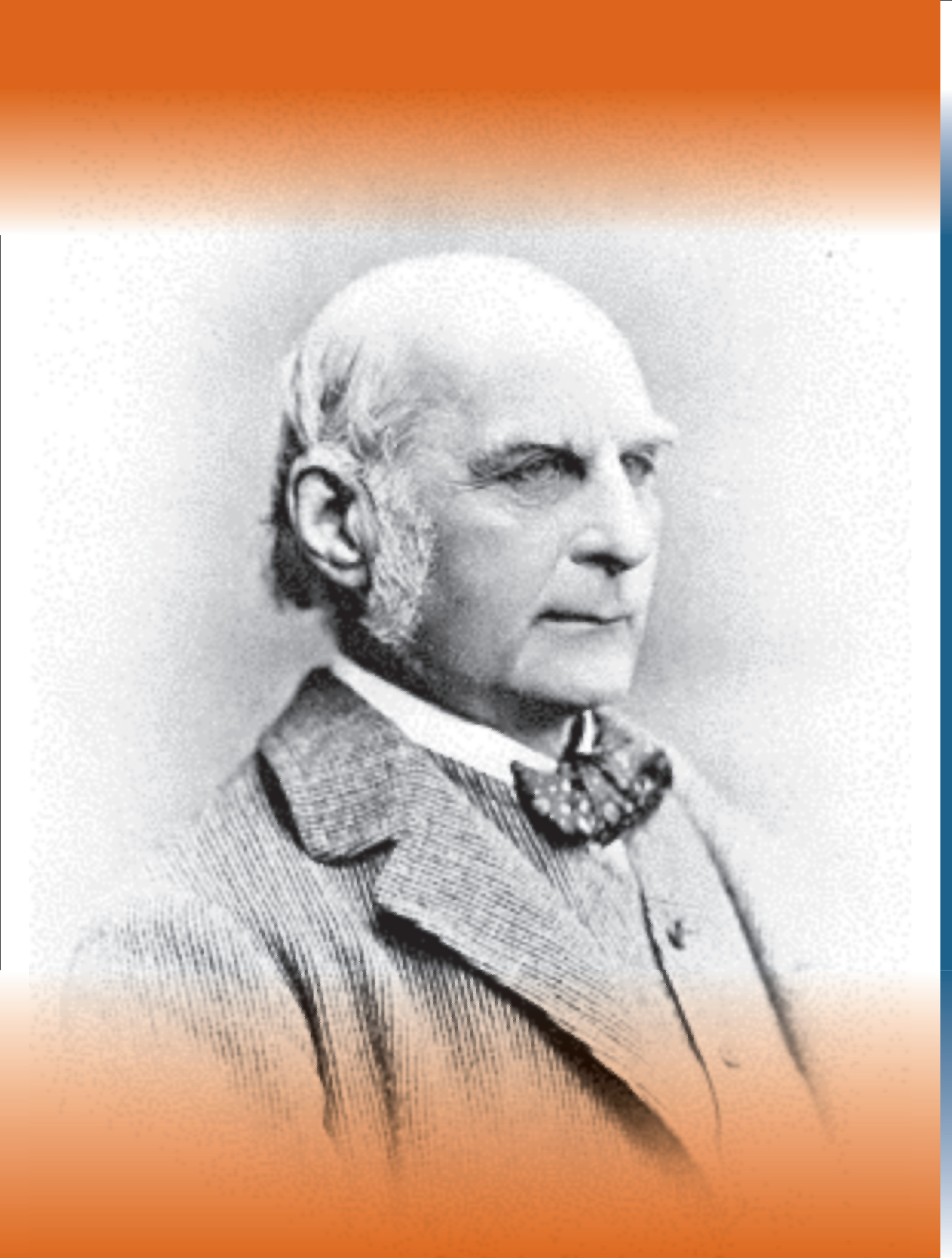
SUMMARY

The first two regression lines, and the first correlations, were calculated by Francis Galton, in his work on heredity in sweet-peas and in humans. When ‘regressing’ the heights of adult children on those of their parents, Galton had to deal with the fact that men are generally taller than women— but without modern-day statistical tools such as multiple regression and partial correlation. This poster uses the family data on stature, which we obtained directly from Galton’s notebooks, to

(a) compare the sharpness of his methods, relative to modern-day ones, for dealing with this complication;

(b) estimate the additional familial component of variance in stature beyond that contributed by the parental heights.

In keeping with Galton’s plea for “a manuscript library of original data”, these historical and pedagogically-valuable data are now available to the statistical community as digital photographs and as a dataset ready for further analyses.



***Sir Francis Galton**, F.R.S.*
1822-1911

Data Collection

Galton tried several times to collect data on family stature, but “tried in vain for a long and weary time to obtain it in sufficient abundance.”

In 1884, Galton “made an offer of prizes for Family Records, which was largely responded to, and furnished me last year with what I wanted.” In particular, he noted that “I especially guarded myself against making any allusion to this particular inquiry in my prospectus, lest a bias should be given to the returns.” In all, records were received from 205 families.

FLY-LEAF

TO

THE RECORD OF FAMILY FACULTIES.

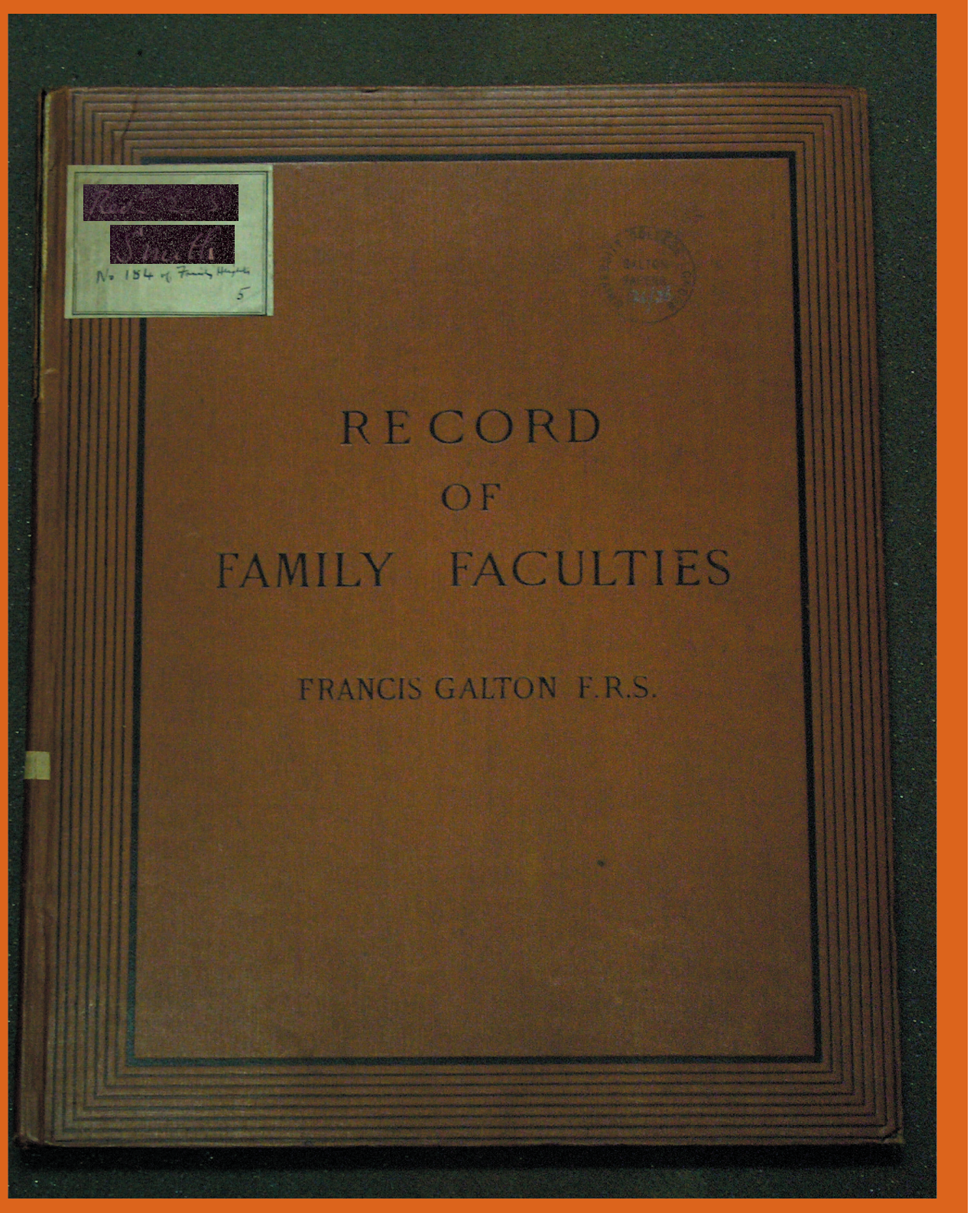
BY

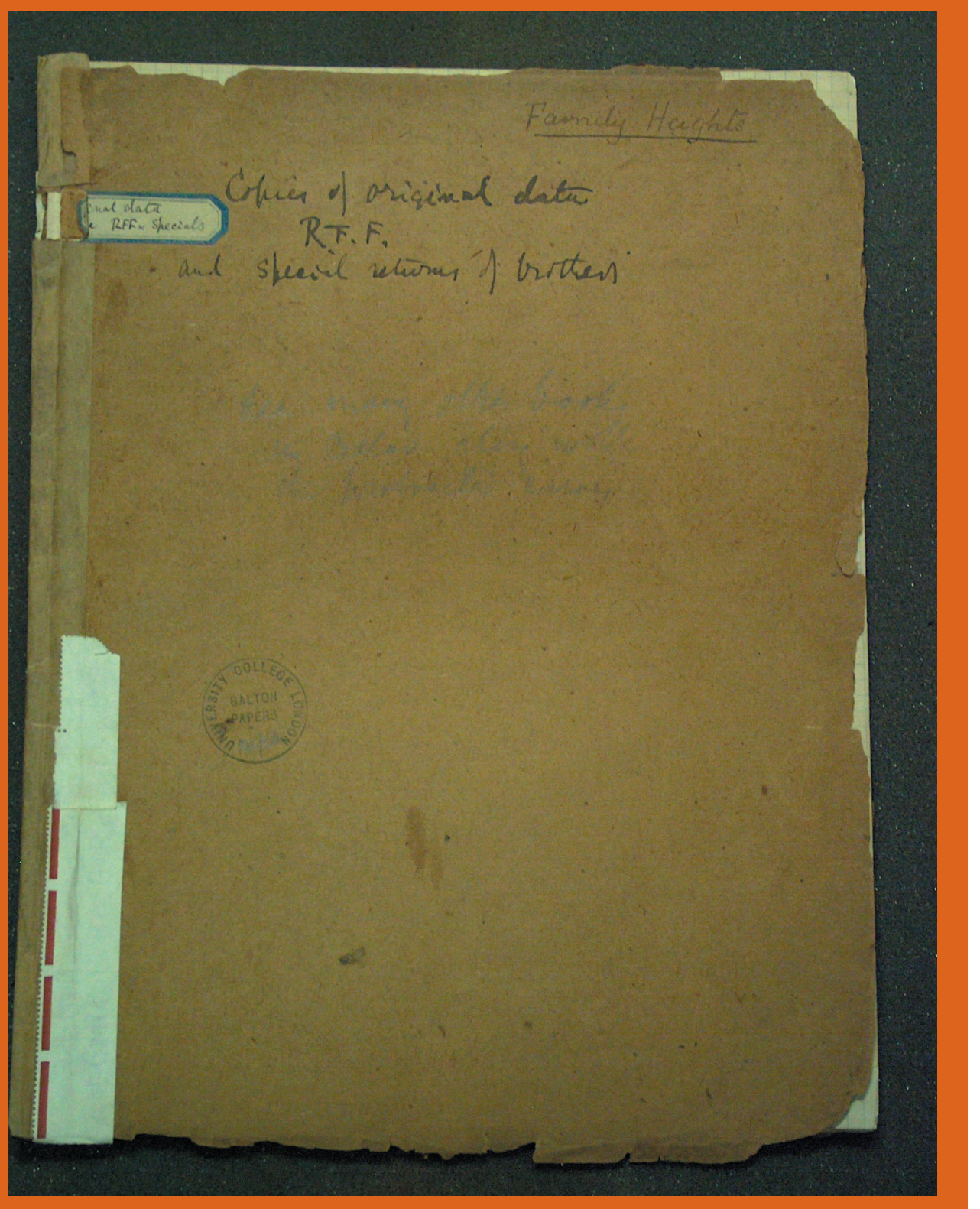
FRANCIS GALTON, F.R.S.

Mr. Francis Galton offers £500 in prizes to those British subjects resident in the United Kingdom who shall furnish him before May 15, 1884, with the best Extracts from their own Family Records.

These Extracts will be treated as confidential documents, to be used for statistical purposes only, the insertion of names of persons and places being required solely as a guarantee of authenticity and to enable Mr. Galton to communicate with the writers in cases where further question may be necessary.

The value of the Extracts will be estimated by the degree in which they seem likely to facilitate the scientific investigations described in the preface to the *Record of Family Faculties*.[1]





Biometrika, Then and Now

In 1901, Galton helped launch Biometrika, with the following wish:

“(…) This journal, it is hoped, will justify its existence by supplying these requirements either directly or indirectly. I hope moreover that some means may be found, through its efforts, of forming a manuscript library of original data. Experience has shown the advantage of occasionally rediscussing statistical conclusions, by starting from the same documents as their author. **I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of his data in some place where it should be accessible, under reasonable restrictions, to those who desire to verify his work.**” (Vol 1, pp 7-10, 1901)

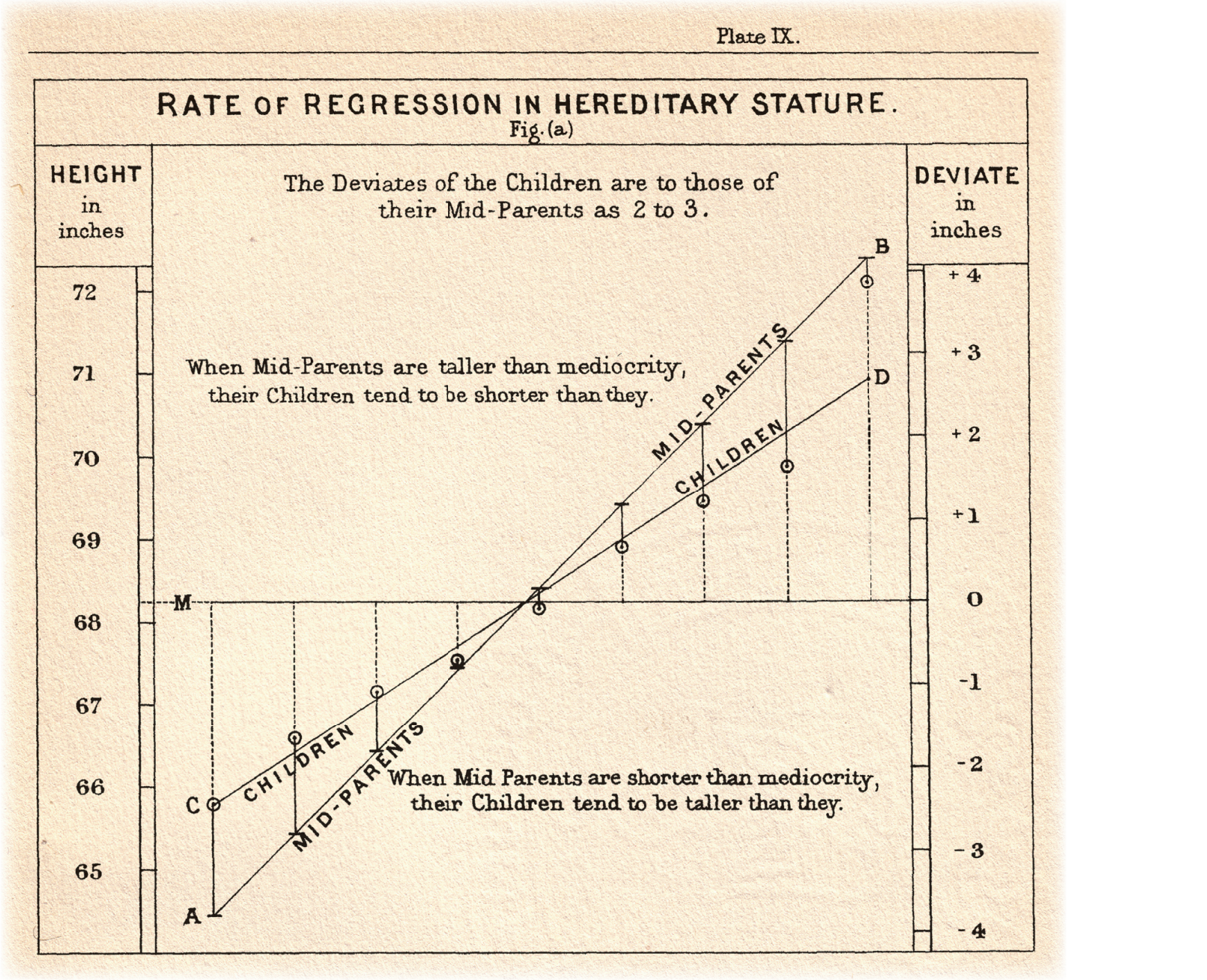
Galton, with his flair for the technological, would have welcomed the internet, ‘computers’ that follow instructions, and digital photography. He would also have been pleased that, with the approval of University College London, digital photographs of the pages of his notebook of heights, along with an electronic copy of the numbers they contain, and some other related photographs, are available online at **www.epi.mcgill.ca/hanley/galton**.

See also the article, “‘Transmuting’ women into men: Galton’s family data on human stature,” by James Hanley, in the August 2004 edition of **The American Statistician**.

www.epi.mcgill.ca/hanley/galton/

Galton and Regression: An Introduction and Background

Galton defined *regression* as a reversion of a characteristic measured in offspring, *away* from the mean value of the *same* characteristic in *their own* parents, and *towards* the mean value in *all* parents/offspring. In his “regression line” (see Figure below), “the Deviates of the Children are to those of their Mid-Parents as 2 to 3” implying that “When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they”, and conversely.



The contours of equal frequency in the two-way frequency table (see right) led Galton to the correlation coefficient of the bivariate Gaussian distribution. From these, Karl Pearson developed a full treatment of correlation, multiple and partial. Pearson’s early work relied on these family data, which “Mr. Galton, with his accustomed generosity”, had placed at Pearson’s disposal.

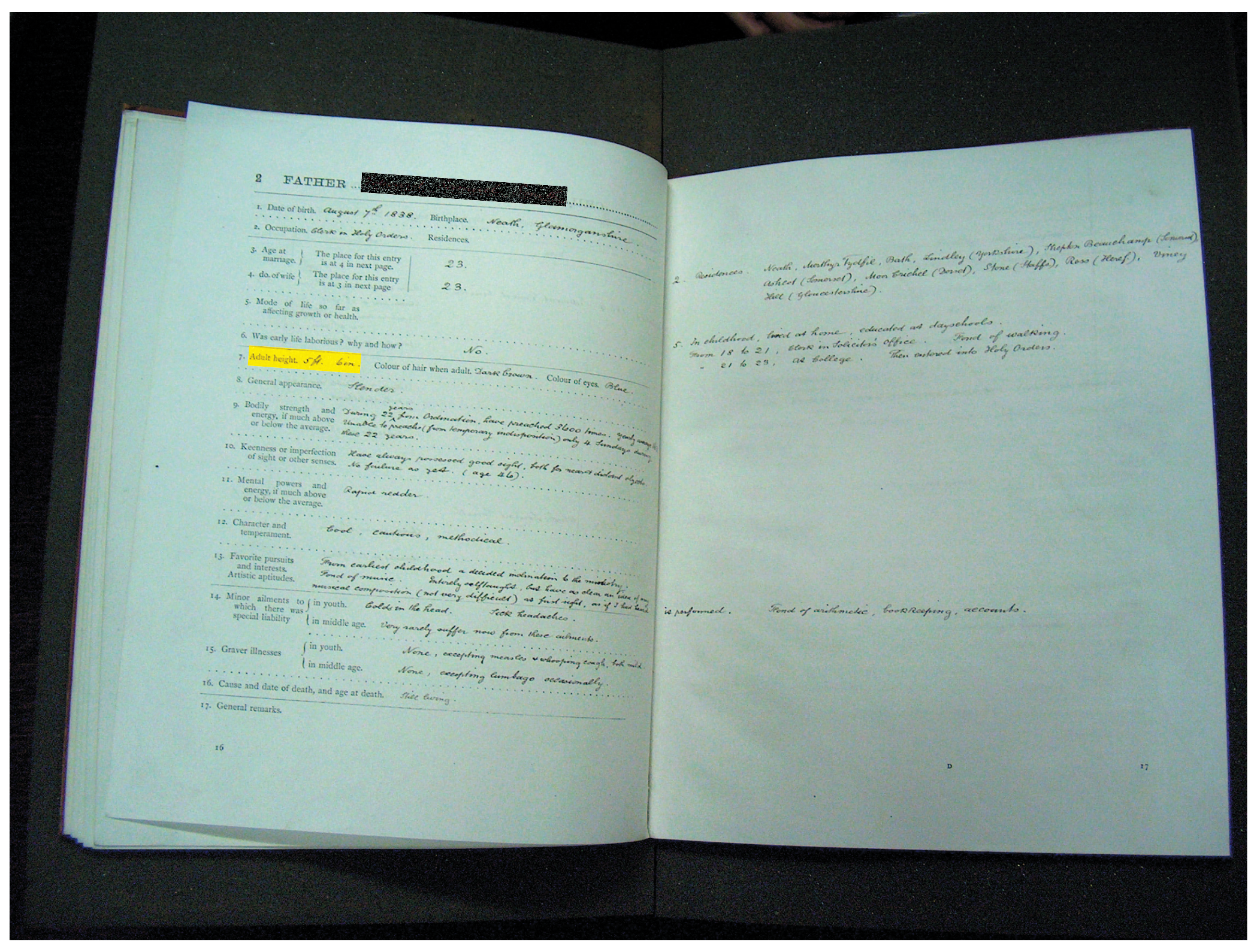
account for the results, contrasted in their values but concurrent in their significance, that are derived from comparisons between different groups of the returns.

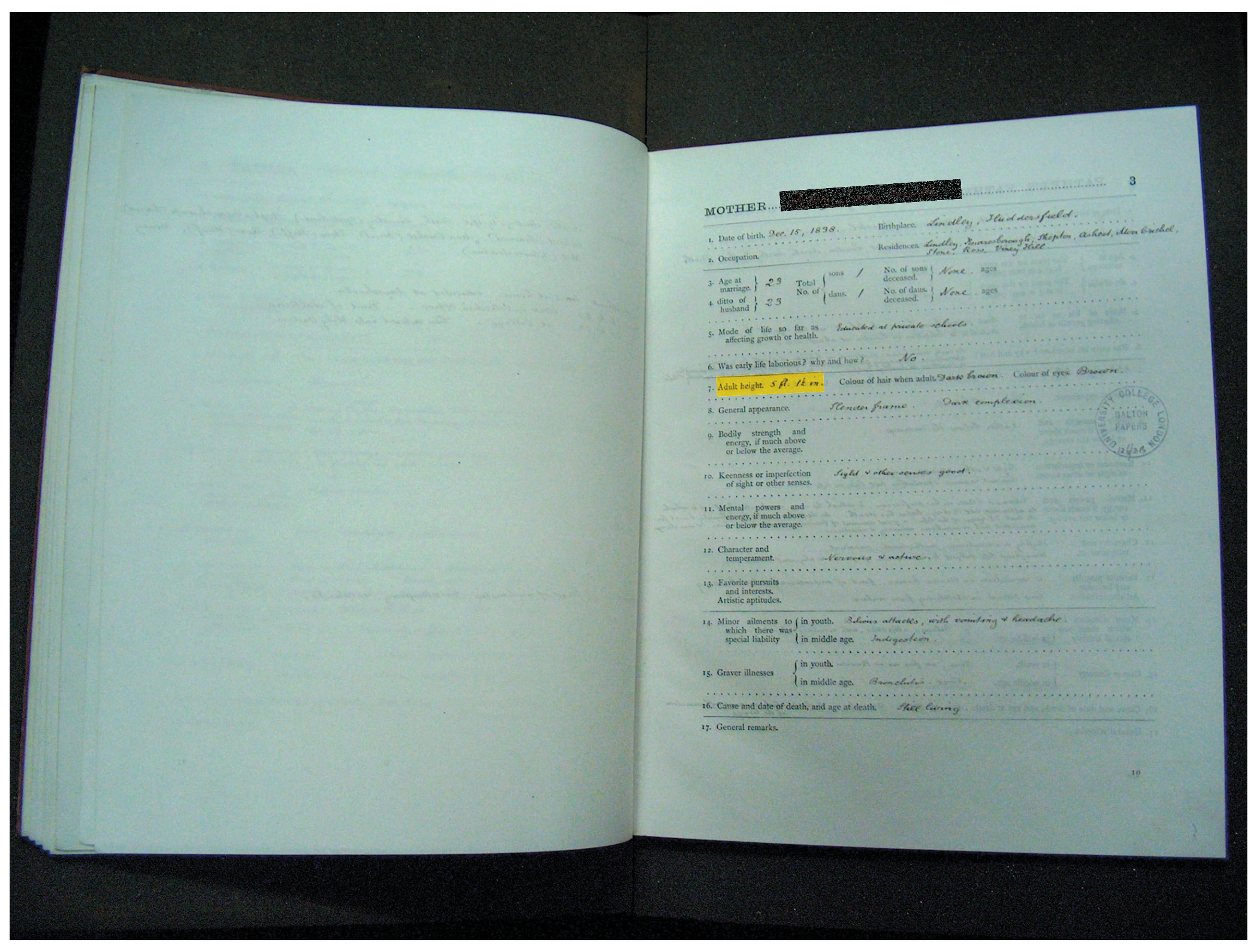
An analysis of the Records fully confirms and goes far beyond the conclusions I obtained from the seeds. It gives the numerical value of the regression towards mediocrity in the case of human stature, as from 1 to $\frac{2}{3}$ with unexpected coherence and precision [*see* Plate IX, fig. (*a*)], and it supplies me with the class of facts I wanted to investigate—the degrees of family likeness in different degrees of kinship, and the steps through which special family peculiarities become merged into the typical characteristics of the race at large.

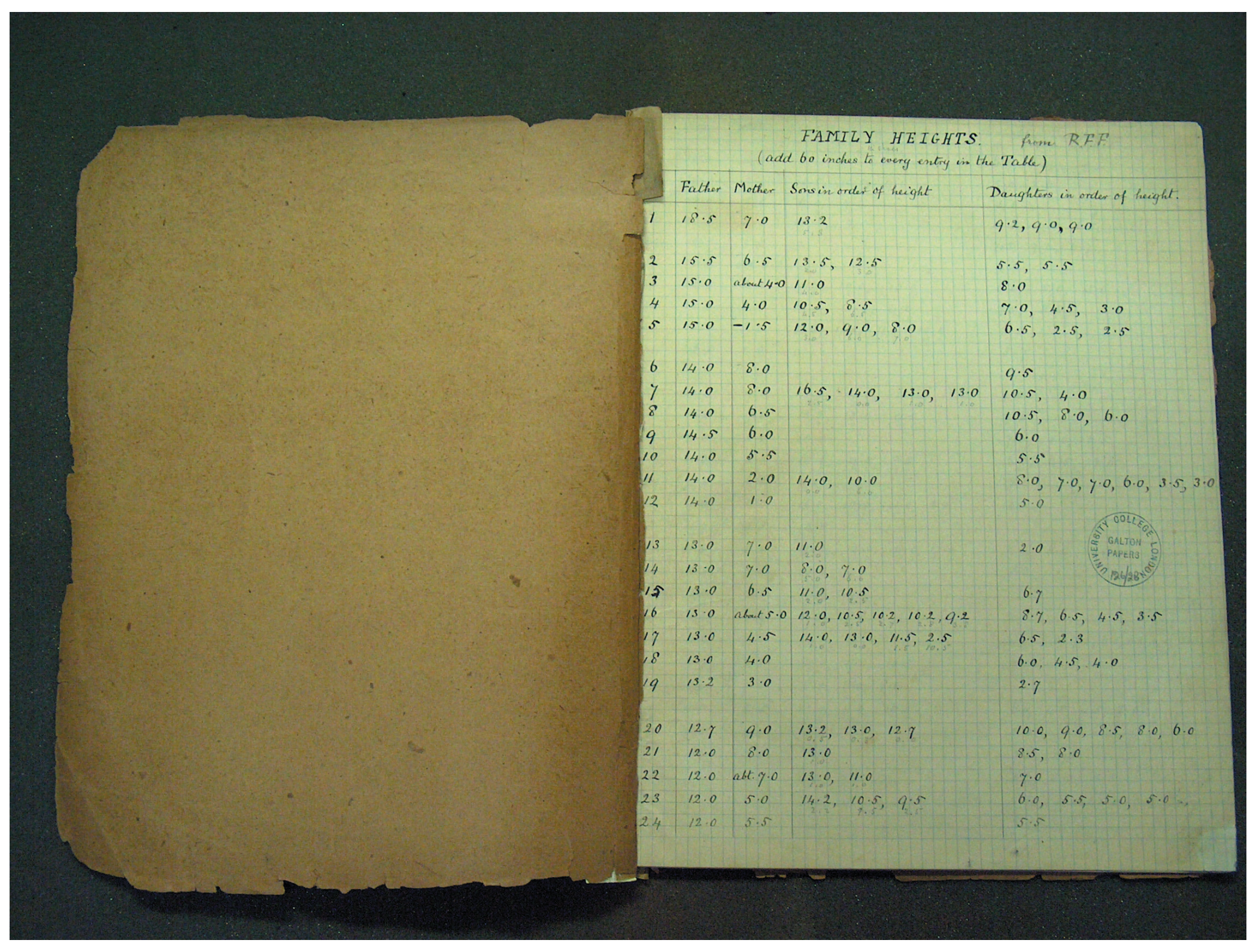
My data consisted of the heights of 930 adult children and of their respective parentages, 205 in number. In every case I transmuted the female statures to their corresponding male equivalents and used them in their transmuted form, so that no objection grounded on the sexual difference of stature need be raised when I speak of averages. The factor I used was 1·08, which is equivalent to adding a little less than one-twelfth to each female height. It differs a very little from the factors employed by other anthropologists, who, moreover, differ a trifle between themselves; anyhow, it suits my data better than 1·07 or 1·09. The final result is not of a kind to be affected by these minute details, for it happened that, owing to a mistaken direction, the computer to whom I first entrusted the figures used a somewhat different factor, yet the result came out closely the same.

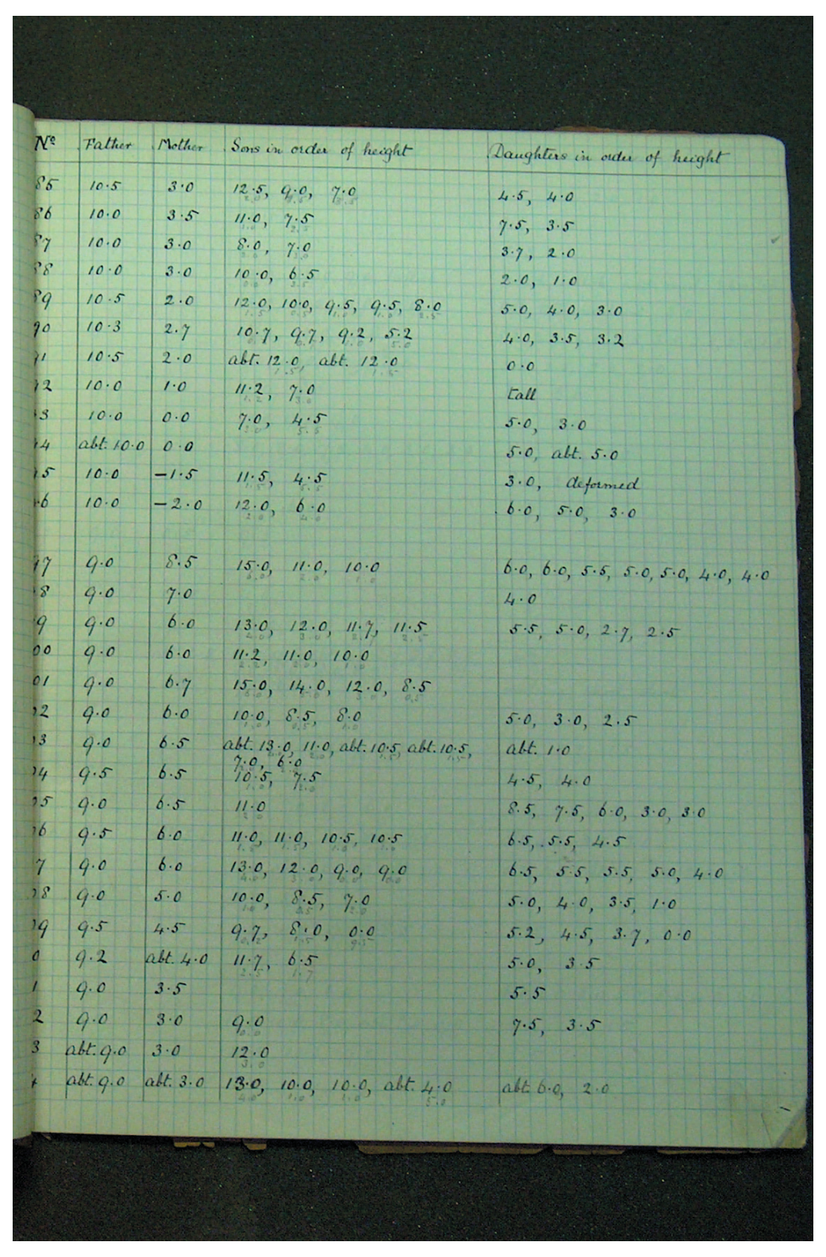
I shall now explain with fulness why I chose stature for the

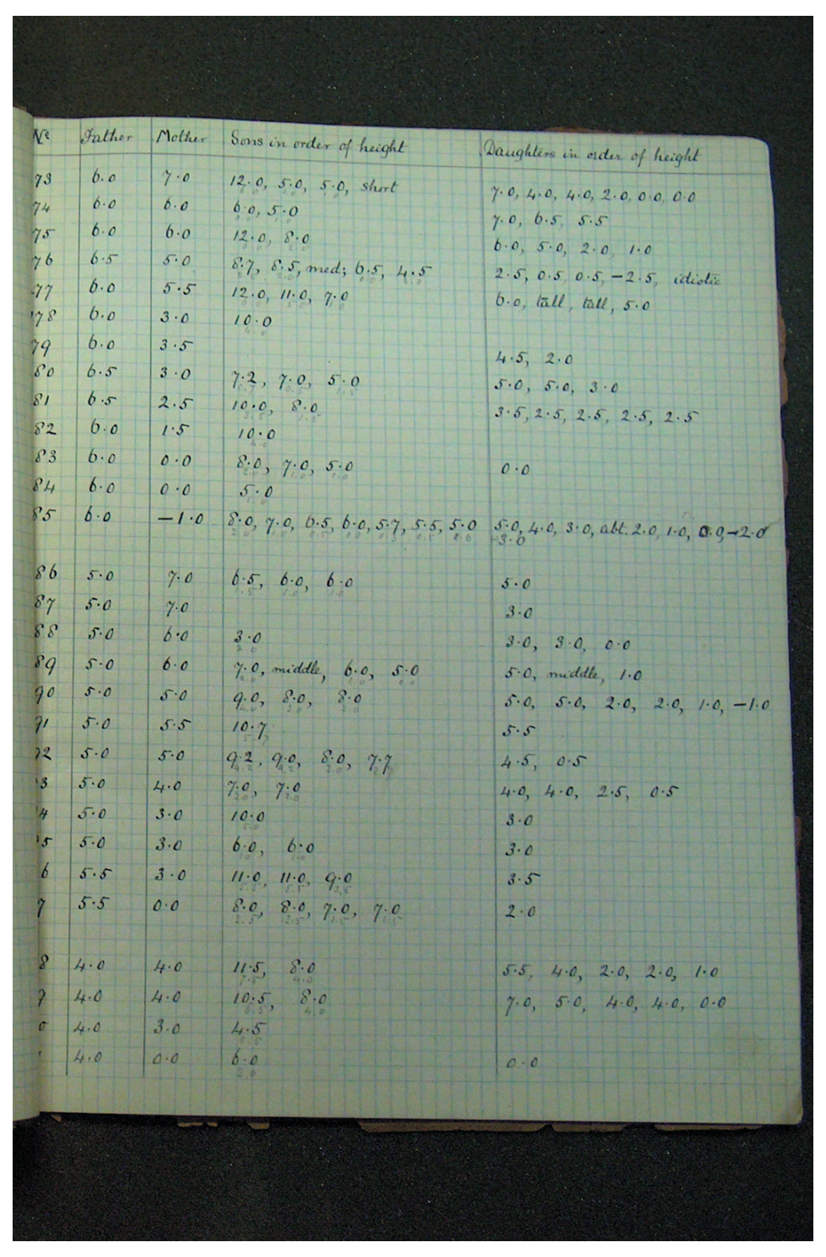
Record of Family Faculties and the Galton Notebooks





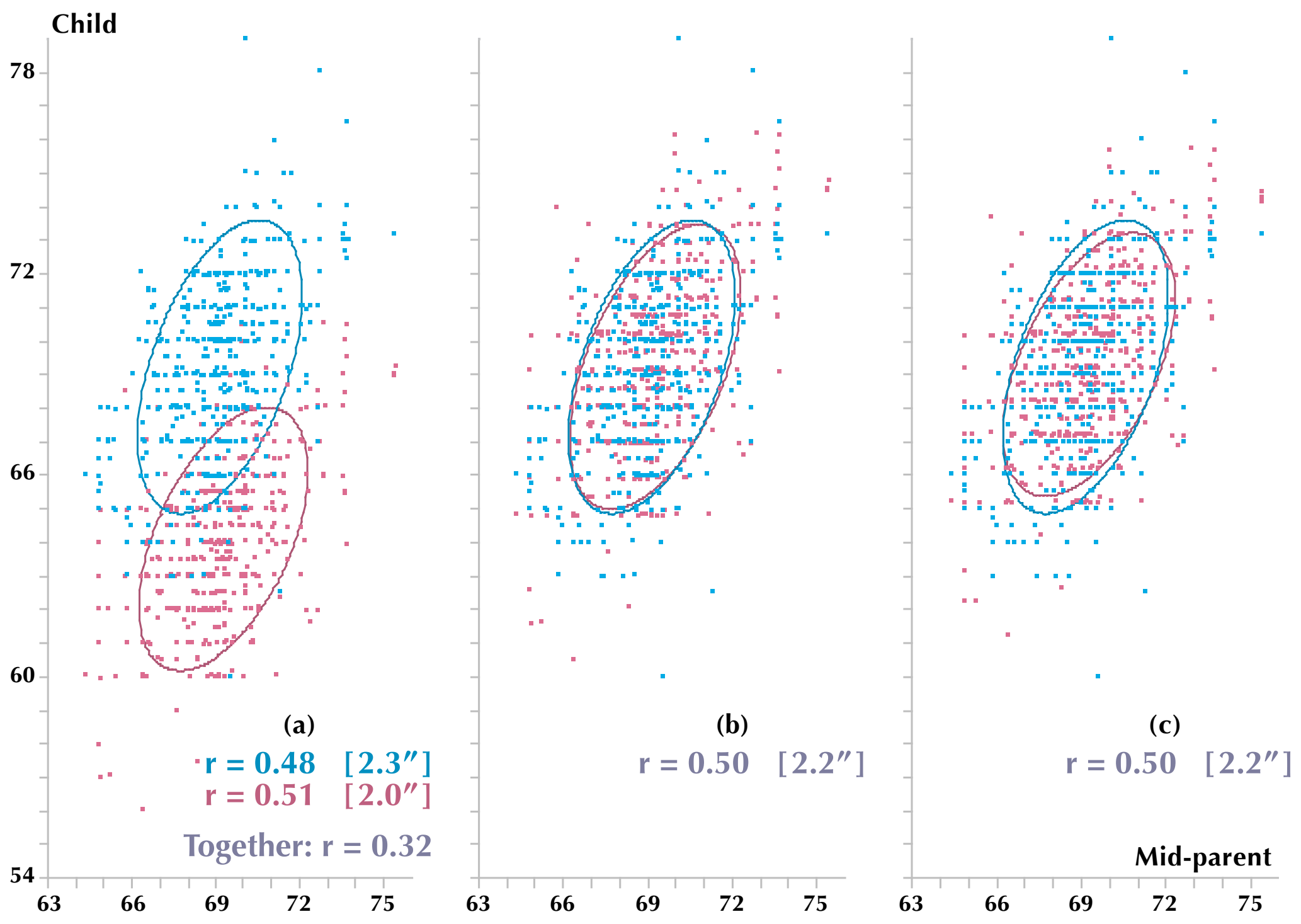






ANALYSIS 1

“Transmuting” of Female Heights



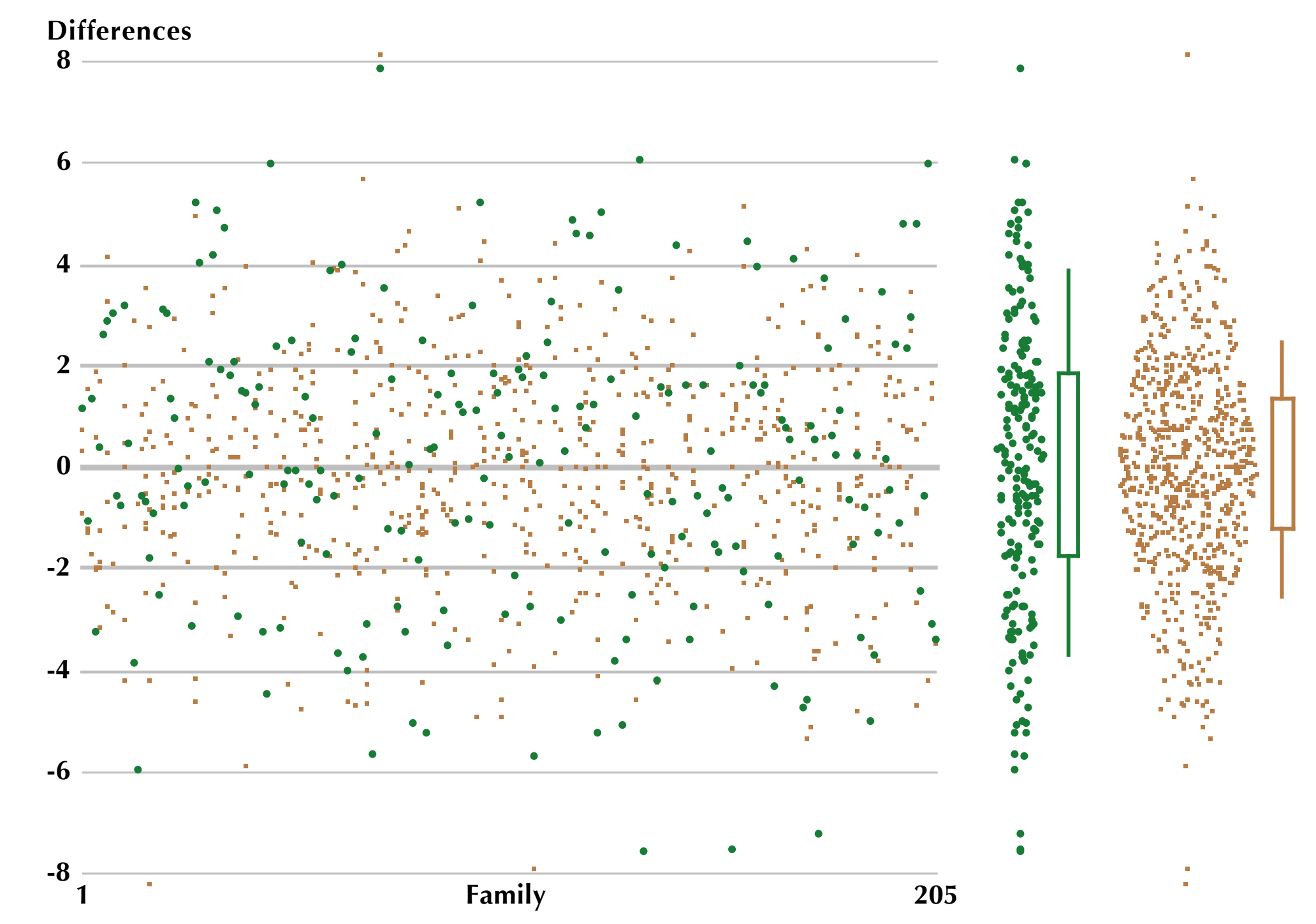
Heights (in inches) of adult children in relation to their mid-parent height. **(a)** each daughter’s height ‘as is’ **(b)** daughter’s height multiplied by 1.08 **(c)** 5.2 inches added to daughter’s height. Daughters’ heights are shown in pink, and sons’ in blue, symbols. Ellipses (75%) are drawn based on the observed means and covariances.

In all three panels, and in analyses for “Do Residuals Segregate along Family Lines?”, the mid-parent height is calculated as (father’s height + 1.08 x mother’s height) / 2.

[Average Residual, in inches]

ANALYSIS 2

Do Residuals Segregate along Family Lines?



Distribution of within- and between-family residuals from simple linear regression, after daughters’ heights have been multiplied by 1.08, of offspring height on mid-parent height. Families listed left to right, in same order as in Galton’s notebook.

Larger green dot: the average residual for a family, multiplied by the square root of the number of offspring in the family, so as to put all 205 averages on the same scale. *Smaller brown dot*: orthogonal difference of within-family residuals (729 in all, from 172 families with two or more offspring). Marginal distributions shown on right. Boxplots show the 10th, 25th, 75th and 90th percentiles. **ICC = 19 %**

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.

Two questions led me to pursue these same raw data which Galton placed at Pearson’s disposal:

1. **How would today’s statisticians deal with the fact that men are generally taller than women?**

“Partialing out” the “effect” of sex; or “adjusting for sex in a regression model”, is conceptually like *adding* so many inches to the height of each female, or subtracting this amount for each male. In Galton’s analysis, “All female heights were *multiplied by 1.08*”; i.e., he “transmuted” them. I wished to test whether Galton’s ‘proportional’ scaling is a more biologically appropriate adjustment than the purely additive one. i.e., *whether Galton’s multiplicative model is sharper than today’s additive model? i.e., despite stronger computers and user-friendly statistical procedures, would modern-day data-analysts find weaker correlations than Galton did?*

2. **To what extent do the deviates from the regression line segregate further by family?**

Galton’s two-way frequency table did not identify which children with the same mid-parental height belonged to which families. Among children with the same mid-parental height, to what extent do their deviates from the regression line segregate further by family, and how might we show this familial variation graphically?

The 205 Families

	Min	Max	Sum	Mean
NUMBERS OF...				
Sons	0	10	487	2.4
Daughters	0	9	476	2.3
Sons + Daughters	1	15	963	4.7
NUMBERS FOR WHOM HEIGHT REPORTED AS A NUMBER...				
Sons	0	10	481	2.3
Daughters	0	9	453	2.2
Sons + Daughters	1	15	934	4.6

PRELIMINARY ANALYSIS

Role of Stature in Marriage Selection

Table 9.

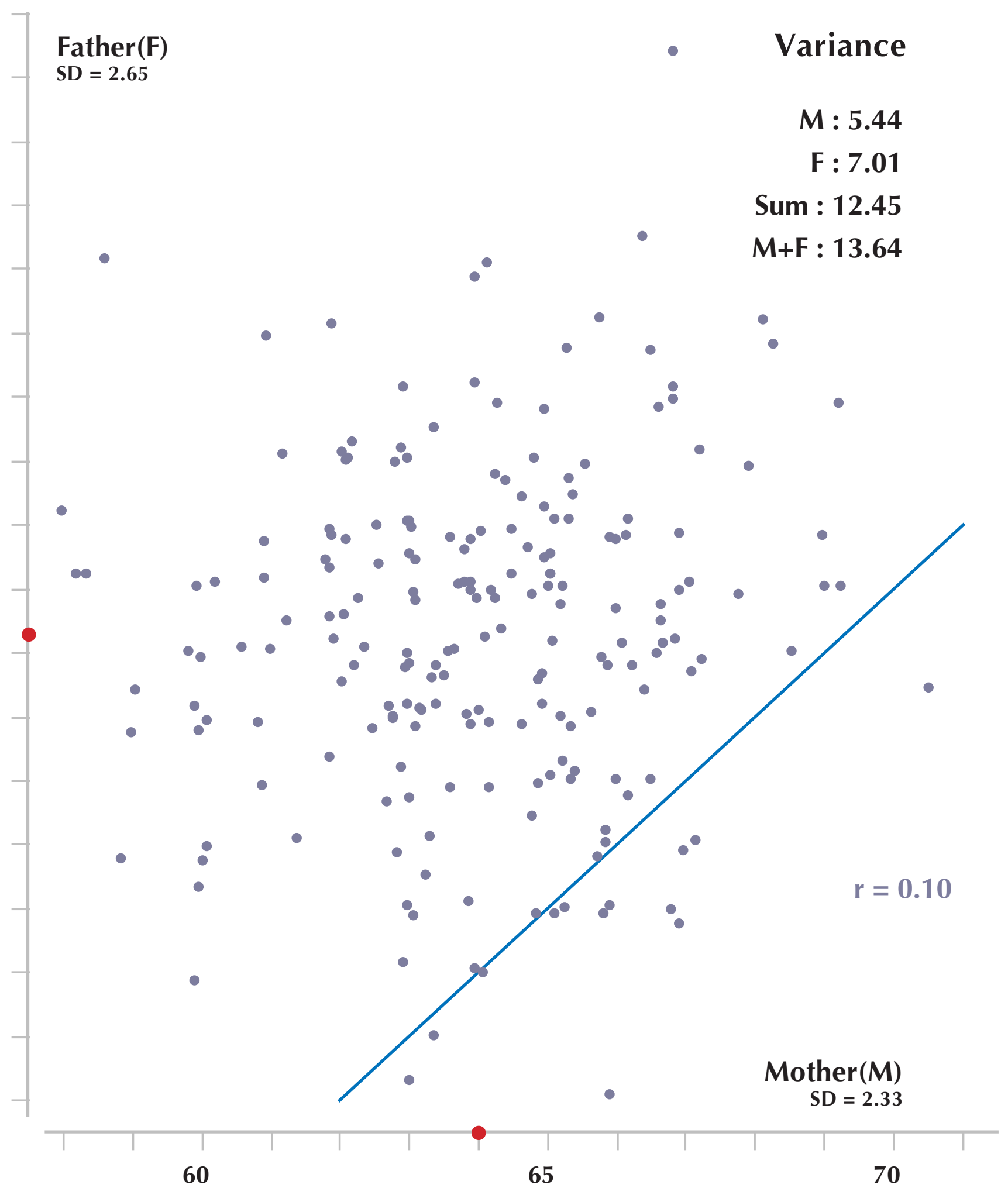
Marriage Selection in respect to Stature.

S., t. 12 cases.	M., t. 20 cases.	T., t. 18 cases.
S., m. 25 cases.	M., m. 51 cases.	T., m. 28 cases.
S., s. 9 cases.	M., s. 28 cases.	T., s. 14 cases.

Short and tall, 12 + 14 = 32 cases.
Short and short, 9 } = 27 cases.
Tall and tall, 18 }

We may therefore regard the married folk as couples picked out of the general population at haphazard when applying the law of probabilities to heredity of stature.

T,M,S = Tall/Medium/Short men; t,m,s = tall/medium/short women.



Acknowledgements

Stephen Stigler directed me to the Galton Papers at University College London (UCL). My current chair, Rebecca Fuhrer, introduced me to Beverly Shipley, a post-graduate student at UCL, who located the material in March 2001. Because of the frail condition of the notebook, photocopying was not permitted, so she first transcribed the family heights onto paper, and later from there into a spreadsheet. In February 2003, I requested and obtained permission for Colin Hanley and Louise Koo to digitally photograph the material.

This work was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada.
The photographs are published with the permission of the Director of Library Services of University College London.

McGill

James A. Hanley
Department of Epidemiology and Biostatistics
McGill University

Address 1020, Pine Avenue West,
Montréal (Québec) H3A 1A2 • CANADA
Telephone +1 (514) 398-6270
Facsimile +1 (514) 398-4503
E-mail James.Hanley@McGill.CA
Web Page http://www.epi.mcgill.ca/hanley/

Design by Colin J. Hanley • Colin.Hanley@mail.McGill.CA