THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 3, 227–231
http://dx.doi.org/10.1080/00031305.2015.1087876

Taylor & Francis
Taylor & Francis Group

HISTORY CORNER

On Galton's Change From "Reversion" to "Regression"

Prakash Gorroochurn

ABSTRACT
Galton's first work on regression probably led him to think of it as a unidirectional, genetic process, which he called "reversion." A subsequent experiment on family heights made him realize that the phenomenon was symmetric and nongenetic. Galton then abandoned "reversion" in favor of "regression." Final confirmation was provided through Dickson's mathematical analysis and Galton's examination of height data on brothers.

ARTICLE HISTORY
Received April 2015
Revised August 2015

KEYWORDS
Ancestral type; Atavism; Imperfect correlation

1. Introduction

Regression, as we know it today, was born from Galton's investigations into the laws of heredity. The phenomenon that Galton discovered is best described in his own words:

> …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 … (Galton 1886)

That Galton came to this conclusion almost single-handedly and not by drawing on the contributions from his predecessors is testimony to his genius. The various experiments and analyses that Galton performed before he reached his conclusion have been well described in works such as Cowan (1972); MacKenzie (1978); Porter (1986); Stigler (1986, 1989); and Bulmer (2003). However, what is often not properly discussed is that Galton at first very probably did not understand regression as we know it today. He first called the phenomenon "reversion" (indeed the symbol r was first used by Galton to signify "the coefficient of reversion" (Pearson 1930, p. 9), which was a genetic process well known to both him and his contemporaries. One of his first discoveries was not that there *was* a regression effect, but rather that the reversion phenomenon he had observed and had assumed would occur was operating in a linear fashion. Galton also thought the phenomenon was a *unidirectional* process operating on offspring from remote ancestors (Gorroochurn to appear, Chap. 2). The realization that something other than a unidirectional genetic process was going on, however, soon came about when he found that reversion was also occurring on parents from their offspring. At this stage, Galton made the decision to change "reversion" to "regression." Galton confirmed his hypothesis through the mathematical analysis of J. Hamilton Dickson and later through the examination of height data on brothers. The phenomenon was not genetic reversion as he had at first thought, but a nongenetic, purely statistical phenomenon that could operate in either direction.

2. The 1877 Paper: Reversion or "ATAVISM"

Galton's (1877) groundbreaking paper, "Typical Laws of Heredity" dealt with a problem that had preoccupied him for a while, indeed since his 1869 book *Hereditary Genius* (Galton 1869): Why do the characteristics (mean and variance) of a hereditary attribute (such as height) from an isolated human population remain constant from generation to generation? To explain the constancy in attributes, Galton invoked *reversion*, also known as atavism, which is the *genetic process* by which an individual resembles a grandparent or more distant ancestor with respect to some trait not possessed by the parents. This can happen, for example, if a recessive and previously suppressed trait reappears through the combination of two recessive alleles in a genotype. Alternatively, the process of recombination can give rise to a unique constellation of genes resulting in a long suppressed character to reappear (these two mechanisms were not known to Galton as Mendelism was yet to be rediscovered in 1900). Atavism is thus *reversion to ancestral type* and was well known by Galton's contemporaries, including Darwin (1859, p. 14), who, in fact, first proposed it. This genetic process is quite different from the purely statistical phenomenon that Galton soon discovered and at first identified with reversion.

There is undeniable evidence that Galton believed that atavism *was* the process that would revert offspring's traits to those of their distant ancestors. Thus, back in 1865, he made the following statement:

> Lastly, though the talent and character of both of the parents might, in any particular case, be of a remarkably noble order, and thoroughly congenial, yet they would necessarily have such mongrel antecedents that it would be absurd to expect their children to invariably equal them in their natural endowments. The law of atavism prevents it. (Galton 1865, p. 319).

Galton's (1877) in his paper explained that he resorted to experiments with sweet peas to answer his questions. He sorted a large number of sweet pea seeds into seven equally spaced size (weight) classes and sent each of his friends seven packets, each containing 10 seeds of a given class size. The seeds from the offspring were then collected and sent back to Galton. From his analysis of the seed results, Galton made the following two key conclusions:

1. For a given parental class size, the size of the filial seeds was normally distributed, with the same probable error e_p within each class (i.e., the same family variability).

CONTACT Prakash Gorroochurn pg2113@columbia.edu Associate Professor, Department of Biostatistics, Columbia University, New York, NY 10032.

2. Reversion followed "the simplest possible law," being a linear function of the deviation from the grand mean (M). Thus, if the parent size in a given class has mean $M + ke_p$ $(k = 0, \pm 1, \pm 2, \pm 3)$, then the corresponding filial size will have mean $M + k\rho\, e_p$, where, as we shall see below, ρ $(0 < \rho < 1)$ is the fractional coefficient of reversion.

Notice that, in the above, Galton did not say that he has *discovered* reversion, the latter genetic process having already been assumed to be operational; rather Galton stated that it was *linear*.

In an Appendix to the 1877 paper, Galton gave the algebraic conditions necessary for stability in population variability in terms of the modulus c of a distribution (the modulus of a normal distribution was historically used to represent $\sigma\sqrt{2}$). Galton first considered the case of two parents each of whom could be productive, but later turned to the case of simple descent. It is the latter case that we shall describe here. First, Galton wrote the distribution of the "amount of deviation" x in a present population as

$$y = \frac{1}{c\sqrt{\pi}} \exp\left(-\frac{x^2}{c^2}\right).$$

Second, "reversion is expressed by a simple fractional coefficient of the deviation," which we shall denote by ρ $(0 < \rho < 1)$ (Galton himself used the symbol r, but we shall reserve the latter for the *sample* correlation coefficient). Then, in the "reverted parentages,"

$$y = \frac{1}{\rho c\sqrt{\pi}} \exp\left(-\frac{x^2}{\rho^2 c^2}\right).$$

Galton then denoted the modulus of the present population by c_1 and that of the "reverted parentages" by c_2, so that

$$c_2 = \rho c_1. \tag{1}$$

The next step for Galton was to consider the variation of the number of progeny for a given parental class, that is, the family variability.

Family variability was shown by experiment to follow the law of deviation, its modulus, which we will write as v, being the same for all classes. Therefore, the amount of deviation of anyone of the offspring from the mean of his race is due to the combination of two influences—the deviation of his "reverted" parentage and his own family variability; both of which follow the law of deviation. This is obviously an instance of the well-known law of the "sum of two fallible measures" (Airy, "Theory of Errors," §43 (*Galton's footnote*) (Galton 1877, p. 533).

Denoting the modulus of the family variability by v, Galton wrote the above law as

$$c_4^2 = c_2^2 + v^2, \tag{2}$$

where c_4^2 is the overall modulus of the progeny. Combining Equations (1) and (2),

$$c_4^2 = \rho^2 c_1^2 + v^2.$$

Now, for variability to remain constant across generations, we need $c_1 = c_4$, which implies

$$c_1^2 = \frac{v^2}{1 - \rho^2} \tag{3}$$

(ibidem). It should be noted that Equation (3) expresses the fact that the variability across generations remains constant due to the balancing forces of family variability (which tends to increase spread) and reversion (which tends to reduce spread).

In sum, Galton's (1877) paper is groundbreaking but still not fully satisfactory in the sense of projecting a clear and accurate view of regression. Galton's use of reversion to signify "the tendency of that ideal mean filial type to depart from the parent type, reverting towards what may be roughly and perhaps fairly described as the average ancestral type" shows that, at this stage, he thought of the process as being both *unidirectional* and *genetic*. To his amazement, the contrary turned out to be true, and this was to be shown in his 1885 presidential address to the Anthropology Section of the British Association (Galton 1885a). This is also where the term "regression" was first used by Galton, having realized that something else than atavism was going on. We now turn our attention to this lecture.

3. The 1885 Presidential Lecture: Regression's First Appearance

In 1885, Galton was to make another major breakthrough following his initial 1877 paper on reversion. In the presidential lecture to the Anthropology Section of the British Association, Galton admitted the following:

> …I was then [in 1877] blind to what I now perceive to be the simple explanation of the phenomenon … (Galton 1885a, p. 507).

Having now realized an important fact he had been "blind to," Galton was poised to make a departure from his original, less accurate "reversion" to the new, more accurate "regression." He thus stated that his previous experiments on sweet peas showed that "the mean filial *regression* (author's italics) toward mediocrity was directly proportional to the parental deviation from it." Moreover, he now not only showed that regression is symmetric in nature but also gave the correct mechanism for it. (Galton was correct in explaining regression from one generation to the other. However, Galton also made an argument for perpetual regression, which turned out to be erroneous. For more details, see Bulmer 2003, p. 285).

Galton explained that he had taken pains to collect the heights of parents and adult children in 205 families. He then multiplied the height of each female parent by 1.08 so as to make the male and female parents comparable and "no objection grounded on the sexual difference of stature need be raised." He then took the average for each pair of parents to obtain a "mid-parent" height. From these, Galton calculated the median for both mid-parents (X) and offspring (Y) as 68.25 inch and the probable errors as $e_{p,X} = 1.2$ and $e_{p,Y} = 1.7$, respectively. (The probable error of a random variable X is simply its semi-interquartile range, i.e., for a symmetric distribution the value e_p such that $\Pr\{|X\text{-}m| < e_p\} = 0.5$. At one time, it was a popular measure of variability but was replaced by the standard deviation, a term coined by Pearson (1894). It can be shown that, when X is normally distributed, $e_p = 0.675$ SD.) Next, for each mid-parental height, Galton calculated the median height of the adult children and plotted the latter median heights against the mid-parental heights. The line is fitted by inspection, and

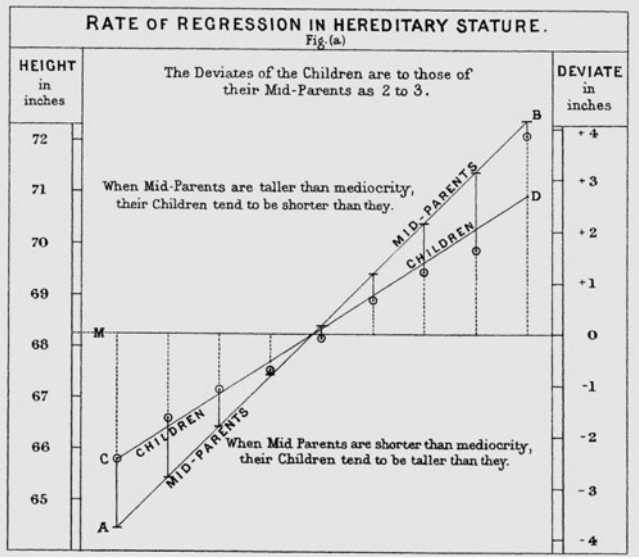

Figure 1. Galton's two regression lines (taken from Galton 1885b).

Galton obtains a slope of 2/3. Galton was now ready to define his law of regression for these data:

> It is that the height-deviate of the offspring is, on the average, two-thirds of the height-deviate of its mid-parentage (Galton 1885a, p. 508).

In modern notation, we can write this as

$$E(Y|X) - M = \frac{2}{3}(X - M)$$
$$E(Y|X) = \frac{2}{3}X + \frac{1}{3}M,$$

where the overall population mean M has been assumed to be constant.

Galton now took the next step that he unfortunately had not ventured into in his sweet pea experiments. Having plotted offspring median heights (Y) against the mid-parental heights (X), Galton next plotted the median mid-parental heights against the offspring heights, and discovered that a similar reversion effect acts on the mid-parental heights (see Figure 1).

That is, not only were offspring more "mediocre" than their mid-parents but mid-parents were also more "mediocre" than their offspring! This single observation is enough to put a serious dent in the hypothesis that genetic reversion to ancestral types was the process responsible for the observations that Galton had initially made. Galton correctly assessed the new discovery and gave the correct explanation for the regression effect:

> The explanation of it is as follows. The child inherits partly from his parents, partly from his ancestry. Speaking generally, the further his genealogy goes back, the more numerous and varied will his ancestry become, until they cease to differ from any equally numerous sample taken at haphazard from the race at large. Their mean stature will then be the same as that of the race; in other words, it will be mediocre. Or, to put the same fact into another form, the most probable value of the mid-ancestral deviates in any remote generation is zero.
>
> For the moment let us confine our attention to the remote ancestry and to the mid-parentages, and ignore the intermediate generations. The combination of the zero of the ancestry with the deviate of the mid-parentage, is that of nothing with something, and the result resembles that of pouring a uniform proportion of pure water into a vessel of wine. It dilutes the wine to a constant fraction of its original alcoholic strength, whatever that strength may have been (Galton 1885a, p. 508).

Indeed, the above explanation is one of the most intuitive ways of understanding the regression effect: In general, suppose a first measurement X is made on a given subject, followed by a second measurement Y. Assume X is exceptionally high. As long as X and Y are imperfectly correlated, X can be thought to be made up of two components:

1. The first component, which is usually extreme and is expected to remain extreme.
2. The second component, which is not extreme and is expected to remain near the center of the distribution.

The first measurement X is extreme because both components are high. However, for the second measurement Y, the first component is expected to remain high, but the second component is expected to be near the center. Hence, the average value of Y will be less extreme than X and closer to the center of the distribution (e.g., Wallis and Roberts 1956, p. 61; Stigler 1997).

Notice that the above explanation does not require any biological, economic, or other force to be present for regression to occur. The regression effect (or regression to the mean) is a purely statistical artifact arising from imperfect correlation between X and Y (of course, the concept of correlation was not known to Galton when he first discovered the regression effect). As such, it is also symmetric in the sense that the same reasoning as above can made to argue that, for a given extreme Y, the value of X is expected to be less extreme and closer to the center.

4. Final Confirmation: Dickson's Analysis and Data on Brothers

The irrefutable confirmation of Galton's hypothesis on the symmetric and purely statistical nature of regression was provided by the mathematician J. Hamilton Dickson and Galton's examination of height data on brothers. Galton used a sheet of squared paper and entered frequencies on it (Figure 2). He noticed that, when identical values (total values of each cell) were joined by line segments, a series of concentric and similar ellipses were formed (e.g., see Friendly and Denis 2005). Galton then reported the observations he made:

> Their common center 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 [Figure 2] 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 1885b, p. 255)

These observations clearly leaned in favor of Galton's hypothesis that regression occurred both forwards and backwards, and did not require a genetic force to be operational. However, an

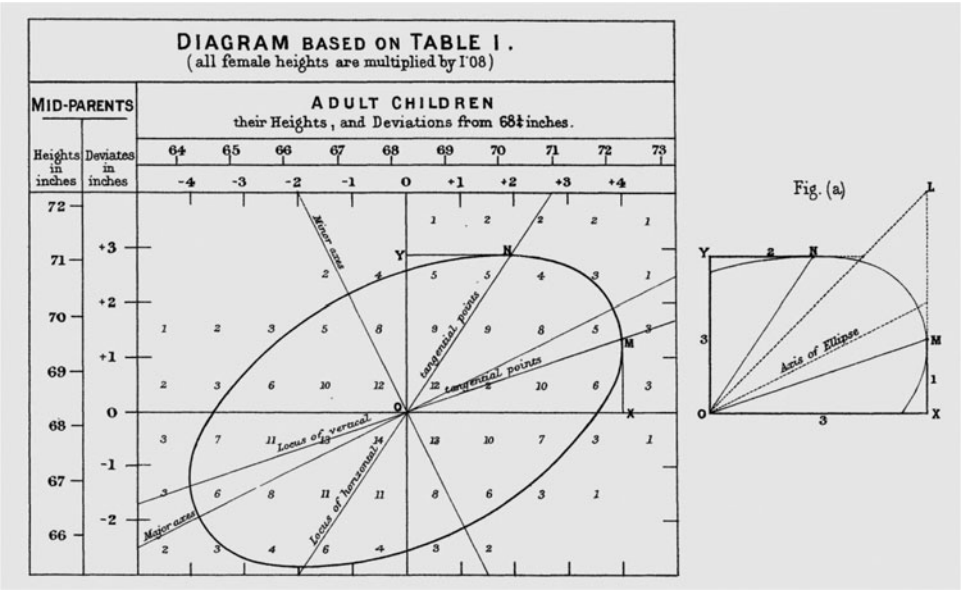

Figure 2. Ellipse generated by Galton by joining points with the same frequency.

irrefutable confirmation needed a proper mathematical analysis, a task that Galton thought was beyond his analytical skills. Therefore, he solicited the help of the able mathematician J. Hamilton Dickson. In modern mathematical language (this is exactly how the problem was phrased: "A point P is capable of moving along a straight line P'OP, making an angle $\tan^{-1}2/3$ with the axis of y, which is drawn through O the mean position of P; the probable error of the projection of P on Oy is 1.22 inch: another point p, whose mean position at any time is P, is capable of moving from P parallel to the axis of x (rectangular coordinates) with a probable error of 1.50 inch. To discuss the 'surface of frequency of p' (Galton 1886), Dickson was provided with the information that $Y \sim N(0, \sigma_Y^2)$ and $X|Y \sim N(\beta_{X|Y}y, \sigma_{X|Y}^2)$, and was asked the following questions:

1. What is the joint density of (X, Y), and what is the shape of the contours of equal probability density?
2. How can the regression coefficient $\beta_{Y|X}$ be calculated?
3. What is the density of Y given X?
4. What is the relationship between $\beta_{Y|X}$ and $\beta_{X|Y}$?

Dickson answered each of the above questions without much trouble, and the solution was published as an Appendix to Galton's (1886) paper "Family Likeliness in Stature." In modern notation, the joint density of X and Y is

$$f_{XY}(x, y) = f_Y(y)\, f_{X|Y}(x|y) \propto \exp\left[-\left\{\frac{y^2}{2\sigma_Y^2} + \frac{(x-\beta_{X|Y}y)^2}{2\sigma_{X|Y}^2}\right\}\right]. \tag{4}$$

(Pearson (1921) expressed his puzzle as to why Galton did not himself derive the joint density of X and Y since he already knew both $f_Y(y)$ and $f_{Y|X}(x|y)$. However, it is unlikely that Galton thought in terms of conditional and marginal distributions.)

To obtain the contours of equal probability, Dickson sets the expression in the above exponent to a constant (say K):

$$\frac{y^2}{\sigma_Y^2} + \frac{(x-\beta_{X|Y}y)^2}{\sigma_{X|Y}^2} = K, \tag{5}$$

which is the equation of a set of ellipses.

To obtain the required regression coefficient $\beta_{Y|X}$, first Equation (5) is differentiated:

$$\frac{ydy/dx}{\sigma_Y^2} + \frac{(x-\beta_{X|Y}y)(1-\beta_{X|Y}dy/dx)}{\sigma_{X|Y}^2} = 0,$$

so that

$$\frac{y}{\sigma_Y^2 dy} + \frac{(x-\beta_{X|Y}y)(dx-\beta_{X|Y}dy)}{\sigma_{X|Y}^2} = 0.$$

By setting the coefficient of dy to zero, tangents to the ellipse in Equation (5) parallel to the y-axis can be obtained and these intersect the ellipse at points lying on the line OM (see Figure 2) with the following equation:

$$\frac{y}{\sigma_Y^2} - \frac{(x-\beta_{X|Y}y)\,\beta_{X|Y}}{\sigma_{X|Y}^2} = 0,$$

or

$$y = \frac{\beta_{X|Y}\sigma_Y^2}{\sigma_{X|Y}^2 + \beta_{X|Y}^2\sigma_Y^2}x.$$

Thus,

$$\beta_{Y|X} = \frac{\beta_{X|Y}\sigma_Y^2}{\sigma_{X|Y}^2 + \beta_{X|Y}^2\sigma_Y^2}. \tag{6}$$

To obtain the conditional density $f_{Y|X}(y|x)$ with the aid of Equation (6), the exponent in Equation (4) can be rewritten as

$$\begin{aligned}
&\frac{y^2}{2\sigma_Y^2} + \frac{\left(x - \beta_{X|Y} y\right)^2}{2\sigma_{X|Y}^2} \\
&= \frac{x^2 + \left(\dfrac{\beta_{X|Y}^2 \sigma_Y^2 + \sigma_{X|Y}^2}{\sigma_Y^2}\right)\left(y^2 - \dfrac{2x\beta_{X|Y}}{\beta_{X|Y}^2 + \sigma_{X|Y}^2/\sigma_Y^2} y\right)}{2\sigma_{X|Y}^2} \\
&= \frac{x^2}{2\sigma_{X|Y}^2/\left(1 - \beta_{X|Y}\beta_{Y|X}\right)} + \frac{\left(y - \beta_{Y|X} x\right)^2}{2\sigma_{X|Y}^2 \beta_{Y|X}/\beta_{X|Y}},
\end{aligned}$$

which gives the two results:

$$\begin{aligned}
X &\sim N\left[0,\ \frac{\sigma_{X|Y}^2}{1 - \beta_{X|Y}\beta_{Y|X}}\right], \\
Y|X &\sim N\left[\beta_{Y|X} x,\ \sigma_{X|Y}^2 \frac{\beta_{Y|X}}{\beta_{X|Y}}\right]. \qquad (7)
\end{aligned}$$

(It can be shown that $\beta_{Y/X}\beta_{X/Y} = \rho^2$, the square of the correlation coefficient between X and Y.)

Dickson also gave the relationship between $\beta_{Y|X}$ and $\beta_{X|Y}$ through Equation (6).

Galton was clearly elated when he received Dickson's analysis, since it confirmed all his empirical results. As Galton had suspected, with only the three pieces of information σ_Y^2, $\beta_{X|Y}$, and $\sigma_{X|Y}^2$ (together with the normal distribution assumption), the elliptical contours he had observed could be constructed. Moreover, given that there was linear regression of X on Y, Dickson had not only shown that there was also linear regression of Y on X but also had been able to deduce coefficient for regression $(\beta_{Y|X})$. These were powerful results confirming Galton's hypothesis that regression was both symmetrical and intrinsically statistical (e.g., Denis 2001; Maraun, Gabriel, and Martin 2011).

It is also very likely that Galton's appreciation of the consequence of symmetry was to be further consolidated, in addition to Dickson's analysis, through his later examination of the data on brothers (Galton 1889, p. 210; Stigler 1986, p. 290). The data were obtained from returns of 295 families and comprised 783 brothers in all. From these, Galton constructed a table showing the distributions of heights of brothers of men with a given height. The resulting table turned out to be symmetric, reflecting the symmetric nature of regression (although, as Stigler (1986, p. 293) has pointed out, the degree of symmetry was exaggerated due to Galton erroneously counting each pair of brothers twice).

5. Conclusions

Galton's first encounter with the regression phenomenon led him to believe that reversion or atavism was really the process operating. This process is both unidirectional and genetic. The subsequent realization that reversion was actually symmetric made him realize that something else was going on, namely, a purely statistical phenomenon arising from the imperfect correlation between two measurements. Galton then decided to change "reversion" to "regression." Of course, Galton cannot be blamed for at first having been "blind" to the correct explanation of the regression phenomenon, as he himself admitted. Galton was a pioneer in the true sense of the word and, as such, it would be quite unfair to expect anything more than what he had already achieved.

References

Bulmer, M. (2003), *Francis Galton: Pioneer of Heredity and Biometry*, Baltimore, MD: Johns Hopkins Press. [227,228]

Cowan, R. S. (1972), "Francis Galton's Statistical Ideas: The Influence of Eugenics," *Isis*, 63, 509–528. [227]

Darwin, C. (1859), *On the Origin of Species by Means of Natural Selection*, London: John Murray. [227]

Denis, D. (2001), "The Origins of Correlation and Regression: Francis Galton or Auguste Bravais and the Error Theorists?" *History and Philosophy of Psychology Bulletin*, 13, 36–44. [231]

Friendly, M., and Denis, D. (2005), "The Early Origins and Development of the Scatterplot," *Journal of the History of the Behavioral Sciences*, 41, 103–130. [229]

Galton, F. (1865), "Hereditary Talent and Character," *Macmillan's Magazine*, 12, 157–166 (Part I), 318–327 (Part II). [227]

——— (1869), *Hereditary Genius*, London: Macmillan (Reprinted 1979, London: Friedmann). [227]

——— (1877), "Typical Laws of Heredity," *Nature* 15, 492–495, 512–514, 532–533. (Also in *Proceedings of the Royal Institution* 8, 282–301). [227,228]

——— (1885a), "Presidential Address, Section H, Anthropology," *Nature*, 32, 507–510 (Also published in (1885) *British Association Reports* 55, 1206–1214). [228,229]

——— (1885b), "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute*, 15, 246–263. [229]

——— (1886), "Family Likeness in Stature," *Proceedings of the Royal Society of London*, 40, 42–73. (Appendix by J. D. Hamilton Dickson, 63–66). [227,230]

——— (1889), *Natural Inheritance*, London: Macmillan. [231]

Gorroochurn, P. (to appear), *Classic Topics on the History of Modern Mathematical Statistics: From Laplace to Modern Recent Times*, Hoboken, NJ: Wiley. [227]

MacKenzie, D. A. (1978), *Statistics in Britain, 1865–1930*, Edinburgh, UK: Edinburgh University Press. [227]

Maraun, M. D. , Gabriel, S., and Martin, J. (2011), "The Mythologization of Regression Towards the Mean," *Theory & Psychology*, 21, 762–784. [231]

Pearson, K. (1894), "Contributions to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society of London*, 185, 71–110. [228]

——— (1921), "Notes on the History of Correlation," *Biometrika*, 13, 25–45. [230]

——— (1930), *The Life, Letters and Labours of Francis Galton (Part IIIA)*, Cambridge, UK: Cambridge University Press. [227]

Porter, T. M. (1986), *The Rise of Statistical Thinking, 1820–1900*, Princeton, NJ: Princeton University Press. [227]

Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge, MA: Harvard University Press. [227,231]

——— (1989), "Francis Galton's Account of the Invention of Correlation," *Statistical Science*, 4, 73–79. [227]

——— (1997), "Regression Towards the Mean, Historically Considered," *Statistical Methods in Medical Research*, 6, 103–114. [229]

Wallis, W. A. , and Roberts, H. V. (1956), *Statistics: A New Approach*, Glencoe, IL: The Free Press. [229]