Journal of Statistics Education

ISSN: (Print) 1069-1898 (Online) Journal homepage: https://www.tandfonline.com/loi/ujse20

Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors

Jeffrey M. Stanton

To cite this article: Jeffrey M. Stanton (2001) Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors, Journal of Statistics Education, 9:3, , DOI: 10.1080/10691898.2001.11910537

To link to this article: https://doi.org/10.1080/10691898.2001.11910537

Journal of Statistics Education

Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors

Jeffrey M. Stanton
Syracuse University

Journal of Statistics Education Volume 9, Number 3 (2001)

Key Words: Correlation; Francis Galton; History of statistics; Karl Pearson.

Abstract

An examination of publications of Sir Francis Galton and Karl Pearson revealed that Galton's work on inherited characteristics of sweet peas led to the initial conceptualization of linear regression. Subsequent efforts by Galton and Pearson brought about the more general techniques of multiple regression and the product-moment correlation coefficient. Modern textbooks typically present and explain correlation prior to introducing prediction problems and the application of linear regression. This paper presents a brief history of how Galton originally derived and applied linear regression to problems of heredity. This history illustrates additional approaches instructors can use to introduce simple linear regression to students.

1. Introduction

The complete name of the correlation coefficient deceives many students into a belief that Karl Pearson developed this statistical measure himself. Although Pearson did develop a rigorous treatment of the mathematics of the Pearson Product Moment Correlation (PPMC), it was the imagination of Sir Francis Galton that originally conceived modern notions of correlation and regression. Galton, a cousin of Charles Darwin and an accomplished 19th century scientist in his own right, has often been criticized in this century for his promotion of "eugenics" (planned breeding of humans; see, for example, Paul (1995). Historians have also suggested that his cousin's lasting fame unfairly overshadowed the substantial scientific contributions Galton made to biology, psychology and applied statistics (see, for example, FitzPatrick 1960). Galton's fascination with genetics and heredity provided the initial inspiration that led to regression and the PPMC.

The thoughts that prompted the development of the PPMC began with a then vexing problem of heredity -- understanding how strongly the characteristics of one generation of living things manifested in the following generation. Galton initially approached this problem by examining characteristics of the sweet pea plant. He chose the sweet pea because that species could self-fertilize; daughter plants express genetic variations from mother plants without contribution from a second parent. This characteristic eliminated, or at least postponed, having to deal with the problem of statistically assessing genetic contributions from multiple sources. Galton's first insights about regression sprang from a two-dimensional diagram plotting the sizes of

daughter peas against the sizes of mother peas. As described below, Galton used this representation of his data to illustrate basic foundations of what statisticians still call regression. The generalization of these efforts into the product-moment correlation and the more complex multiple regression came much later. Current textbooks of behavioral science statistics typically reverse this order: the PPMC is presented first and linear regression is covered later. Many instructors may also feel more comfortable starting with correlation and building up to regression.

The present paper provides historical background and illustrative examples that statistics instructors may find useful in introducing these concepts to college level classes in applied statistics. By briefly tracing the historical development of regression and correlation, this paper shows how introductory statistics instructors can use engaging and historically accurate examples to introduce regression and correlation to students. A number of articles concerning the teaching of regression and correlation indicate that students often have difficulty understanding these concepts and the connection between them (see, for example, Williams 1975; Duke 1978; Karylowski 1985; Goldstein and Strube 1995;). The present article provides new ideas for instruction based on the historical origins of these statistical techniques.

2. Galton's Early Considerations of Regression

Besides his role as a colleague of Galton's and a researcher in Galton's laboratory, Karl Pearson also became Galton's biographer after the latter's death in 1911 (Pearson 1922). In his four-volume biography of Galton, Pearson described the genesis of the discovery of the regression slope (Pearson 1930). In 1875, Galton had distributed packets of sweet pea seeds to seven friends; each friend received seeds of uniform weight (also see Galton 1894), but there was substantial variation across different packets. Galton's friends harvested seeds from the new generations of plants and returned them to him (see Appendix A). Galton plotted the weights of the daughter seeds against the weights of the mother seeds. Galton realized that the median weights of daughter seeds from a particular size of mother seed approximately described a straight line with positive slope less than 1.0:

> "Thus he naturally reached a straight regression line, and the constant variability for all arrays of one character for a given character of a second. It was, perhaps, best for the progress of the correlational calculus that this simple special case should be promulgated first; it is so easily grasped by the beginner." (Pearson 1930, p. 5)

The simple, special case that Pearson referred to is, of course, both the roughly equivalent variability of the two measures and their identical units of measurement. Figure 1 uses a simple, invented data set to illustrate Galton's earliest findings. The parent sweet pea size on the *X*-axis and the offspring sweet pea size on the *Y*-axis have approximately equal variability. Thus, the slope of the line connecting the means of the different columns of points is equivalent both to the regression slope and the correlation coefficient. For Galton's purposes, any slope smaller than 1.0 indicated regression to the mean for that generation of peas. The phenomenon of regression to the mean is illustrated by the configuration of points: The *y*-coordinates of most of the points in Figure 1 are closer to the horizontal offspring mean than their *x*-coordinates are to the vertical parent mean. Galton's first documented study of this type suggested a slope of 0.33 (obtained through careful inspection of his scatterplots), which indicated to him that extremely large or small mother seeds typically generated substantially less extreme daughter seeds. This finding is, of course, prototypical of regression to the mean: For many variables, natural processes work to "dampen" extreme outliers and bring them closer to their respective means.

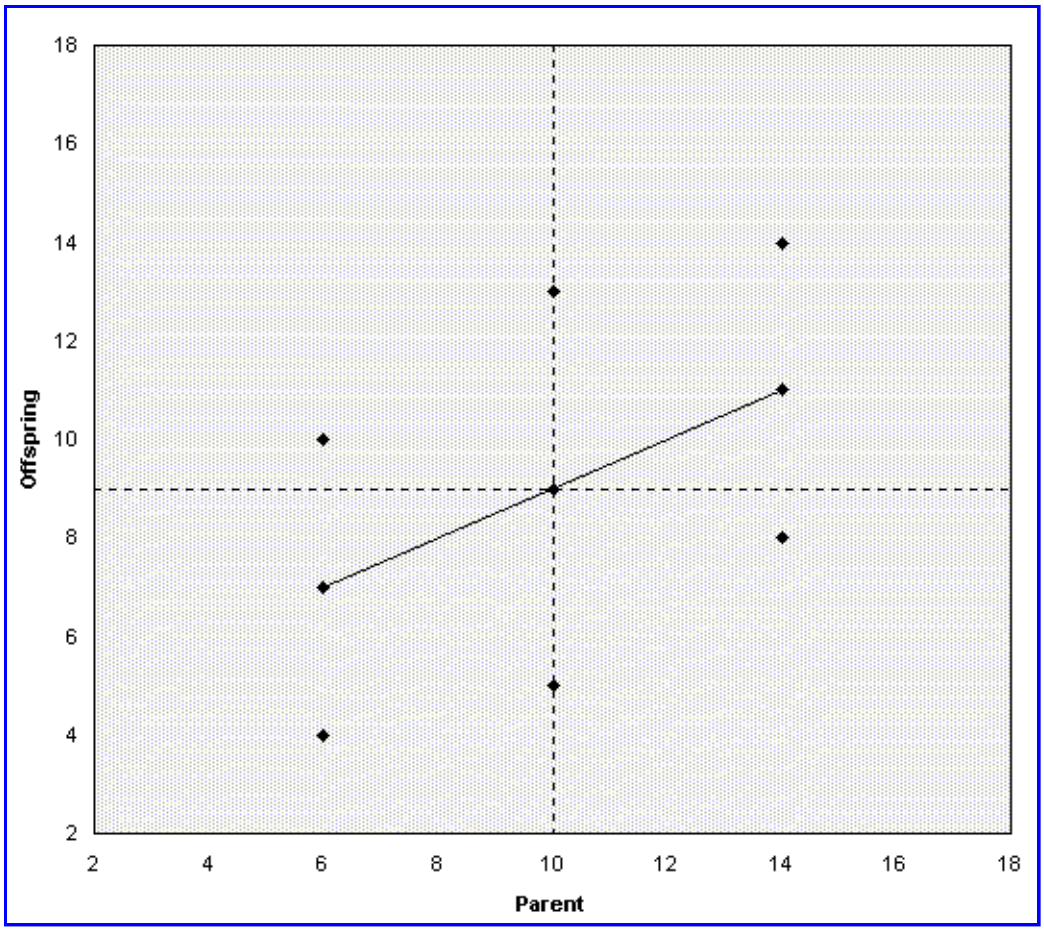

Figure 1. Connecting the means of the individual columns of data provides a crude approximation of the regression line. The slope is exactly 0.50 and the correlation is approximately $r = 0.51$. Many, though not all, of the points are closer to the offspring pea size mean of 9 on the *Y*-axis than to the parental pea size mean of 10 on the *X*-axis. The numeric data appear in Appendix B.

Nonetheless, only a horizontal line would have indicated no heritability in seed size whatsoever, so Galton's finding affirmed his basic assumptions concerning the heritability of "characters." Figure 1, greatly simplified from Galton's original graph, illustrates how a line connecting the means of the columns of data points indicates the degree to which extreme values in the first generation (on the *X*-axis) tend to regress toward the mean of the second generation (on the *Y*-axis). I invented the data points, which are listed in Appendix B, to simplify hand calculation in a classroom setting. In contrast to Figure 1, Galton's original data did not produce a perfectly smooth line, but he was able to draw, by hand, a single line that fit all the data reasonably well (Galton's first regression line was presented at a lecture in 1877; see (Pearson 1930). The slope of this line he designated "*r*" for regression. Only under Pearson's later treatment did *r* come to stand for the correlation coefficient (Pearson 1896).

Galton's progress was both eased and hobbled by his choices for descriptive statistics; he used the median as a measure of central tendency and the semi-interquartile range as a measure of variability. One advantage of these measures lay in the simplicity of obtaining them. Galton was nearly fanatical about graphing and tabulating every available data point. These descriptive values could emerge from an inspection of the resulting figure or table with a minimum of computation. It is understood now that the median and semi-interquartile range do not have the favorable mathematical properties of the mean and standard deviation (for example, they cannot be manipulated using covariance algebra). But Galton was not a sophisticated enough

mathematician to recognize the deficiency. So Galton's progress toward a more general implementation of regression was delayed by his choice of descriptive statistics. In *Natural Inheritance* (Galton 1894), Galton expended a page or two making various arguments about the exact value of the slope of a regression line as calculated with various techniques to estimate the change in *Y* versus the change in *X* on the scatterplot. At that point in time, his efforts lacked the mathematical foundation to derive the slope from the data themselves. As an interesting footnote, in the late 1870s, Galton did not have access to a mechanical calculating machine, whereas Pearson had one for personal use on his desk no later than 1910 (Pearson 1938).

3. Galton's Recognition of the Generality of Regression Slope

Even with his poor choice of descriptive statistics, Galton was able to generalize his work over a variety of heredity problems. He tackled personality temperament, artistic ability, and disease incidence, among others (Galton 1894). The important breakthrough in the process of analyzing these data came from his realization that if the degree of association between two variables was held constant, then the slope of the regression line could be described if the variability of the two measures were known. At that time Galton believed he had estimated a single heredity constant that was generalizable to many or most inherited characteristics. But he wondered why, if such a constant existed, the observed slopes in his parent-child plots varied so much over these characteristics. After noticing differences in variability between generations, he arrived at the idea that the differences in regression slopes that he obtained were solely due to differences in variability between the different sets of measurements.

In modern terms, this principle can be illustrated by assuming a constant correlation coefficient but varying the standard deviations of the two variables involved. Figures 2A, B, and C display the scatterplots of three closely related data sets. The correlation in each data set is identical at $r = 0.64$ (note that 0.66 was one of the basic heritability constants derived by Galton). In Figure 2A the standard deviation of *Y* is the same as the standard deviation of *X*. In Figure 2B the standard deviation of *Y* is smaller than the standard deviation of *X*. Finally, in Figure 2C the standard deviation of *Y* is larger than the standard deviation of *X*. The data, which appear in Appendix B, were designed to center both variables on zero and to simplify hand calculations for classroom use. The scales of measurement are arbitrary. Anchoring a straight edge or ruler at the origin, one can estimate the slope of the line of best fit: the line becomes shallower as the variability of *Y* decreases relative to the variability of *X* and steeper as variability of *Y* increases relative to the variability of *X*.

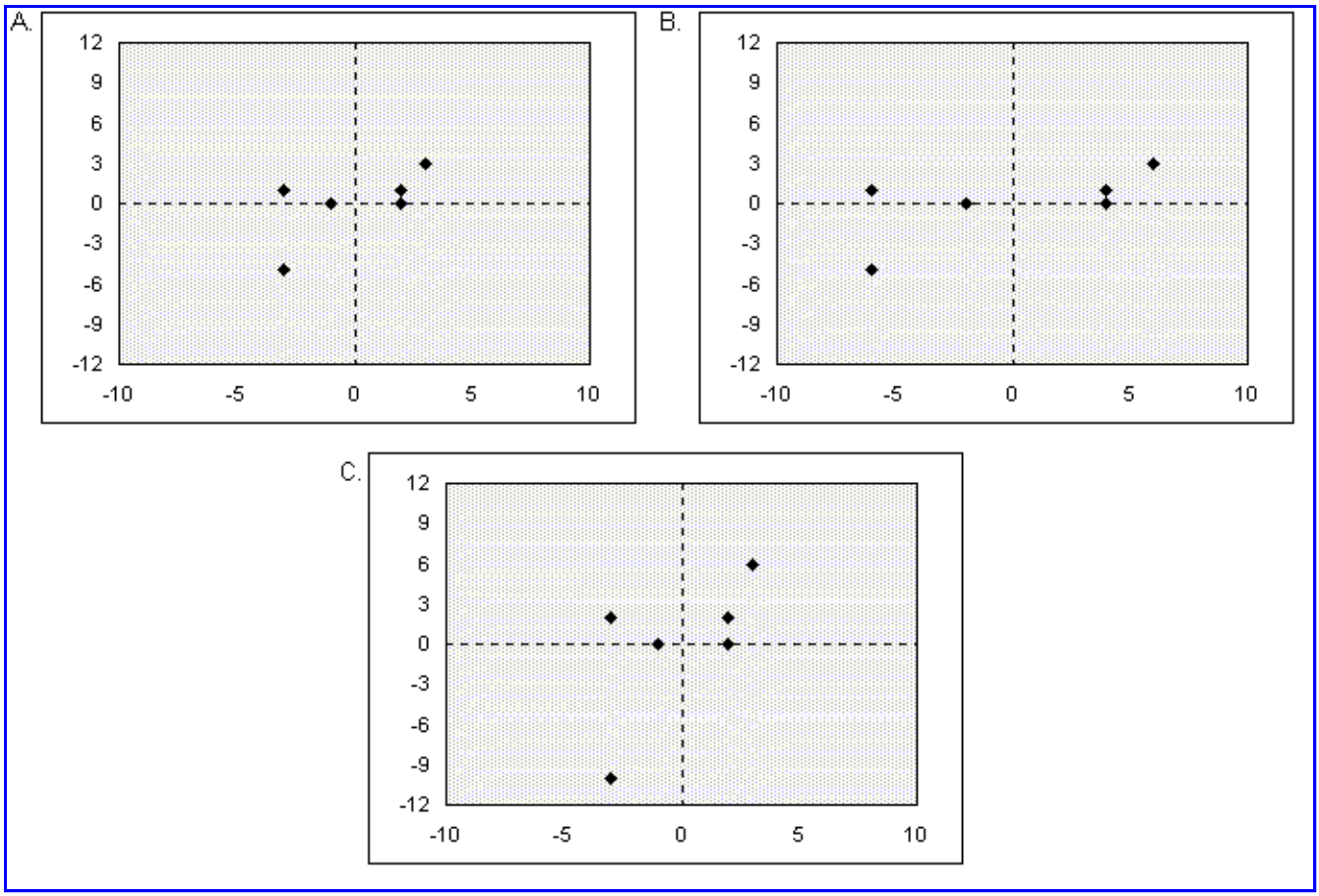

Figure 2. Correlation remains constant even though the slope of the line changes as a result of differences in variability between the two variables. In all three graphs, the correlation is identical at approximately $r = 0.64$. In 2A the variability in X is identical to the variability in Y. In 2B the variability in X was magnified to be larger than the variability in Y. In 2C the variability in Y was magnified to be larger than the variability in X. The numeric data appear in Appendix B.

Although he represented it with different symbols, Galton had at this point recognized that the rudimentary regression equation $y = r (S_y / S_x) x$ described the relationship between his paired variables. Galton used an estimated value of r, because he did not know how to calculate it. The expression (S_y / S_x) was, in effect, a correction factor that allowed him to adjust the observed slope based on the variability of each measure. Modern work on genetics and inheritance has not supported Galton's belief in a universal heritability constant (that is, a fixed correlation of characteristics across generations), but this mistaken assumption gave him the impetus to advance the conceptual division between correlation and regression. As represented in Figure 2, the visible squashing and stretching of the X and Y distributions across the three figures does not hide the evident relationship between the two variables. The degree of association between the two variables remains constant even though the slope of the line is different in all three graphs. Galton correctly recognized that the ratio of variabilities of the two measures was a key factor in determining the slope of the regression line.

4. Pearson's Mathematical Development of Correlation and Regression

In 1896, Pearson published his first rigorous treatment of correlation and regression in the *Philosophical Transactions of the Royal Society of London* . In this paper, Pearson credited Bravais (1846) with ascertaining the initial mathematical formulae for correlation. Pearson noted that Bravais happened upon the product-moment (that is, the "moment" or mean of a set of products) method for calculating the correlation coefficient but failed to prove that this provided the best fit to the data. Using an advanced statistical proof (involving a Taylor expansion), Pearson demonstrated that optimum values of both the regression slope and the correlation coefficient could be calculated from the product-moment, $\sum xy/n$, where x and y are deviations of observed values from their respective means and n is the number of pairs. For example, in linear regression, if the slope is calculated from the product-moment, then the observed x values predict the observed y values with the minimum possible sum of squared errors of prediction, $\sum (y - \hat{y})^2$.

A simpler proof than Pearson's for the product-moment method appeared in Ghiselli (1981). Although neither Pearson's nor Ghiselli's proof is likely to enhance the flow of a typical introductory statistics class, a simple numerical re-creation of Ghiselli's proof can illuminate the important point about the optimal prediction of y from x. Such an example appears in Table 1. This table uses pairs of deviation scores to help demonstrate that the expression $\hat{y} = b\,x$ minimizes squared prediction errors when b is calculated as the product-moment. The first three columns demonstrate the calculation of b as the mean (moment) of the products of the deviation scores. Both x and y vectors are centered on 0 so that each score is itself a deviation score. The rightmost four columns of Table 1 show that adding a small offset to the value of b or subtracting a small offset from b enlarges the sum of the squared errors of prediction. Both Ghiselli's and Pearson's proofs demonstrate more generally that any departure from the product-moment worsens the prediction of y from x.

Table 1. Numeric Example Demonstrating that Adding or Subtracting an Offset from the Product-Moment Worsens the Prediction of *Y* from *X*.

Deviation Data		Product	No Offset (Optimal)		Positive Offset		Negative Offset	
x	y	x*y	2.0x[b]	Squared Errors[c]	2.1x	Squared Errors	1.9x	Squared Errors
-1	-3	3	-2	1	-2.1	0.81	-1.9	1.21
-1	-2	2	-2	0	-2.1	0.01	-1.9	0.01
-0.5	-1	0.5	-1	0	-1.05	0.025	-0.95	0.025
-0.5	-1	0.5	-1	0	-1.05	0.025	-0.95	0.025
0.5	1	0.5	1	0	1.05	0.025	0.95	0.025
0.5	3	1.5	1	4	1.05	3.8025	0.95	4.2025

	2	3	6	4	1	4.2	1.44	3.8	0.64
Sum	0	0	14		6		6.1375		6.1375
Mean	0	0	2.0[a]						

[a] The mean of the deviation cross-products, that is, the product-moment.

[b] This column uses 2.0 as the value of the slope in the equation $\hat{y} = 2.0\,x$. The raw product-moment value can be used here as the slope without correcting for the variability in x because the variance of x was programmed to be exactly equal to 1. Note that the general equation for regression slope is $\sum(X - \bar{X})(Y - \bar{Y}) / n\sigma_x^2$.

[c] The squared errors of prediction $(\hat{y} - y)^2$

---

5. A Note on Multiple Regression

Galton realized soon after he had collected and analyzed his sweet pea data that the generations prior to the immediate parents could also influence individual characteristics (Pearson 1930). He even noticed that certain characteristics occasionally skipped one or more generations; a man may appear more similar to his grandfather than to his father in certain respects. In an 1898 paper to the journal *Nature* (cited in Pearson 1930), Galton published a clever diagram that partitioned a unit square into successively smaller squares, where each square represented the ever diminishing influence of previous generations of ancestors on the present individual. A modified and simplified version of the diagram appears in Figure 3. Within each row, smaller and smaller divisions represented each ancestor. Traveling backward through time, each generation had only half as much influence on the present individual as the generation succeeding it.

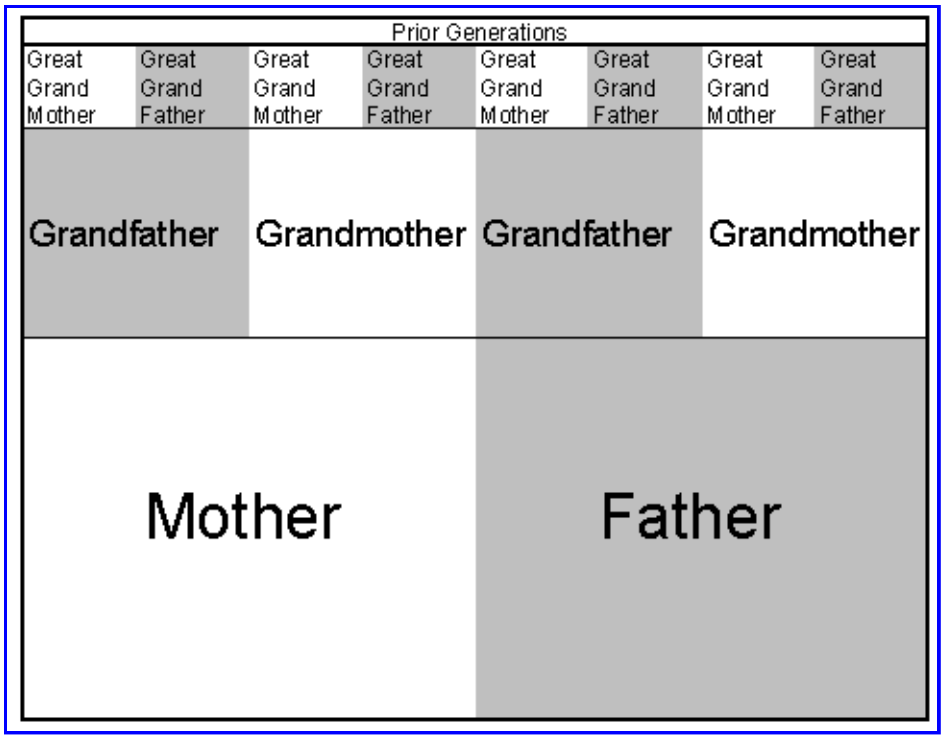

Figure 3. Simplified diagram showing influences of three preceding generations of ancestors on the present day individual.

With the realization of this diminishing effect, Galton had hit upon the kernel of the idea of multiple regression. A characteristic or variable may be influenced not simply by a single important cause but by a multitude of causes of greater and lesser importance. Some of these causes may even overlap with one another (that is, the predictors themselves are intercorrelated). In later publications Galton listed some mathematical formulae that captured this same basic idea, but he was never able to develop a complete mathematical treatment of the matter:

> "The somewhat complicated mathematics of multiple correlation, with its repeated appeals to the geometrical notions of hyperspace, remained a closed chamber to him." (Pearson 1930, p. 21)

Nonetheless, Galton's conceptualization of the multiple influences of progenitors on characteristics of the present day individual was entirely parallel to the modern conception of multiple regression. As with simple linear regression and the correlation coefficient, Galton laid the imaginative groundwork that Pearson later developed into a rigorous mathematical treatment. Pearson's subsequent work included further development of multiple regression as well as innovative progress on other statistics such as chi-square (Pearson 1938).

6. Instructional Ideas

The previous discussion of the development of regression and the correlation coefficient provides a historical account of aspects of scientific and mathematical progress in the late 1800s and early 1900s. In conjunction with the graphical and numerical examples presented, classroom use of some of these historical examples may improve comprehension and sustain student interest by showing the various problems that Galton, Pearson, and others struggled with and solved as they worked out the techniques that are so widely used today. The historical account describes a series of small conceptual steps that an instructor could package

into one or more lectures for an introductory statistics class.

Such a series might begin with a brief overview of the heredity problems that Galton was considering. The simulated sweet pea data from Figure 1 illustrate the basic concepts of plotting data in columns, regression to the mean, and hand fitting a line to the data using the means of the columns. Galton's actual sweet pea data are summarized in Appendix A, and readers may request a copy of the complete data set from the author. The complete data set could be analyzed and plotted for classroom use with any statistical program. The instructor might also wish to plot an additional example of total regression to the mean to show that a horizontal line signifies the absence of an association between X and Y.

Next, students can ascertain why the data behind Figure 1 are a unique case by examining Figure 2. The instructor can emphasize how the level of association between the two variables remains constant even though the slope changes across the three graphs. In much of the heredity data Galton collected, differences in variability were common from generation to generation. Differences in variability between the two variables influence the slope of the regression line but not the level of association between the variables. Assuming that r is known or can be estimated, the slope can be calculated by multiplying r by (S_y / S_x).

Next, the instructor can introduce Pearson's approach to the calculation of r by first demonstrating that the product-moment -- the mean of the cross products of the deviations of X and Y -- provides the most accurate prediction of y scores from x scores as exemplified by the calculations in Table 1. To bring students full circle to the modern notation of the regression equation, $Y = b\,X + a$, the instructor can show that the value for r, adjusted by multiplying by the expression (S_y / S_x), provides the formula for b, the regression slope. For the general case where variables have means other than zero, instructors can demonstrate how to obtain the y-intercept from the means of X and Y.

Although introductory statistics courses do not usually cover the mathematical basis for multiple regression, instructors may wish to discuss the basic concepts using the hereditary example described in the previous section. In particular, the example and the figure illustrate how different predictors contribute in combination but to different degrees to the prediction of an outcome variable. Such a discussion could complete a student's introduction to the basic concepts of correlation and regression.

References

Bravais, A. (1846), "Analyse Mathematique sur les Probabilites des Erreurs de Situation d'un Point," *Memoires par divers Savans*, 9, 255-332.

Duke, J. D. (1978), "Tables to Help Students Grasp Size Differences in Simple Correlations," *Teaching of Psychology*, 5, 219-221.

FitzPatrick, P. J. (1960), "Leading British Statisticians of the Nineteenth Century," *Journal of the American Statistical Association*, 55, 38-70.

Galton, F. (1894), *Natural Inheritance* (5th ed.), New York: Macmillan and Company.

Ghiselli, E. E. (1981), *Measurement Theory for the Behavioral Sciences*, San Francisco: W. H. Freeman.

Goldstein, M. D., and Strube, M. J. (1995), "Understanding Correlations: Two Computer Exercises," *Teaching of Psychology*, 22, 205-206.

Karylowski, J. (1985), "Regression Toward the Mean Effect: No Statistical Background Required," *Teaching of Psychology*, 12, 229-230.

Paul, D. B. (1995), *Controlling Human Heredity, 1865 to the Present*, Atlantic Highlands, N.J.: Humanities Press.

Pearson, E. S. (1938), *Mathematical Statistics and Data Analysis* (2nd ed.), Belmont, CA: Duxbury.

Pearson, K. (1896), "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia," *Philosophical Transactions of the Royal Society of London*, 187, 253-318.

Pearson, K. (1922), *Francis Galton: A Centenary Appreciation*, Cambridge University Press.

Pearson, K. (1930), *The Life, Letters and Labors of Francis Galton*, Cambridge University Press.

Williams, R. H. (1975), "A New Method for Teaching Multiple Regression to Behavioral Science Students," *Teaching of Psychology*, 2, 76-78.

Appendix A - Galton's Sweet Pea Data

In Appendix D of *Natural Inheritance*, Galton (1894) provided Table 2 (p. 226), which contained a list of frequencies of daughter seeds of various sizes organized in rows according to the size of their parent seeds. I generated raw data for $n = 700$ cases from this table. A simple linear regression analysis on these data using parent seed size to predict filial seed size produced almost exactly the same slope value that Galton reported. A summary of the generated data appears in the accompanying Table 2:

Table 2. Raw Data on Diameters of Parent and Daughter Seeds Generated from Galton (1894) Table 2 in *Natural Inheritance*.

Diameter of Parent Seed (0.01 inch)	Diameter of Daughter Seed (0.01 inch)	Frequency
21.00	14.67	22
21.00	15.67	8
21.00	16.67	10
21.00	17.67	18
21.00	18.67	21
21.00	19.67	13
21.00	20.67	6
21.00	22.67	2

20.00	14.66	23
20.00	15.66	10
20.00	16.66	12
20.00	17.66	17
20.00	18.66	20
20.00	19.66	13
20.00	20.66	3
20.00	22.66	2
19.00	14.07	35
19.00	15.07	16
19.00	16.07	12
19.00	17.07	13
19.00	18.07	11
19.00	19.07	10
19.00	20.07	2
19.00	22.07	1
18.00	14.35	34
18.00	15.35	12
18.00	16.35	13
18.00	17.35	17
18.00	18.35	16
18.00	19.35	6
18.00	20.35	2
17.00	13.92	37
17.00	14.92	16
17.00	15.92	13
17.00	16.92	16
17.00	17.92	13
17.00		

	18.92	4
17.00	19.92	1
16.00	14.28	34
16.00	15.28	15
16.00	16.28	18
16.00	17.28	16
16.00	18.28	13
16.00	19.28	3
16.00	20.28	1
15.00	13.77	46
15.00	14.77	14
15.00	15.77	9
15.00	16.77	11
15.00	17.77	14
15.00	18.77	4
15.00	19.77	2

Appendix B

Figure 1: {(6, 4), (6, 7), (6, 10), (10, 5), (10, 9), (10, 13), (14, 8), (14, 11), (14, 14)}

Figure 2A: {(-3, -5), (-3, 1), (-1, 0), (2, 0), (2, 1), (3, 3)}

Figure 2B: {(-6, -5), (-6, 1), (-2, 0), (4, 0), (4, 1), (6, 3)}

Figure 2C: {(-3, -10), (-3, 2), (-1, 0), (2, 0), (2, 2), (3, 6)}

Jeffrey M. Stanton
School of Information Studies
Center for Science and Technology
Syracuse University
Syracuse, NY 13244-4100
jmstanto@syr.edu