Regression Analysis – A Powerful Tool and Riveting Drama

By Tim Chalberg
Pacific Lutheran University

30444 148th Avenue SE
Kent, WA 98042

chalbetl@gmail.com

Introduction

Statistics has asserted itself as an integral field of study in today's world, thanks in large part to regression analysis. Today, statistical analysis, particularly through regressions, can be found almost anywhere a person looks. The idea of a regression is relatively young, first being posed by Francis Galton (1822-1911) in the late 1800s. The preferred method for determining a regression, the method of least squares, owes its origins to Adrien-Marie Legendre (1752-1833) and Carl Friedrich Gauss (1777-1855) in the early 1800s, and its application to statistics to R.A. Fisher (1890-1962) in the 1920s. That hardly tells the whole story though. The story of how modern regression analysis came to be is filled with surprising motives and controversy, making regression arguably as worthy of study for its story as it is for its broad range of applications.

Method of Least Squares

The method of least squares has its own unique story. To start with, it certainly was not devised with regressions in mind, given that it was created over 50 years before the idea of regression was formed. On top of that, there was significant controversy over who should get credit for creating the method. However, before delving into the history of the method of least squares, we shall take a brief look at the method itself.

There are many algorithms for the method of least squares, some more different than others. For a closer look at the ones of Gauss, Laplace (1749-1827), and Yule (1871-1951), refer to the work of Aldrich [1]. Algorithms for the method have become increasingly elegant as mathematical notation has developed, in particular thanks to modern-day matrices. However, the basic premise behind the method of least squares remains the same. We shall illustrate it with the following simple example.

Let $A = \{ a_1, a_2, a_3, \ldots a_n \}$ and $B = \{ b_1, b_2, b_3, \ldots b_n \}$ be ordered sets of distinct, quantifiable observations, both with n elements and $n \in \mathbf{N}$. We assume there is a relation between observations in A and B. So, we assume $a_i x = b_i$ for the ith observation in A and B. This equation can be rewritten as:

$$(1)\ a_i x - b_i = 0$$

In theory, (1) should hold for any pair of corresponding observations in A and B. However, in reality, it does not, thanks to a number of potential sources for errors in observations. So, we re-write (1):

$$(2)\ a_i x - b_i = v_i$$

Here, $v_i \in \mathbf{R}$, and represents the error term associated with A and B's ith observations (so, in this example, there will be a total of n error terms). Note that v_i could be positive or negative, but that v_i^2 will always be a positive value. Thus, $\sum v_i^2$ more accurately describes the error in observations than $\sum v_i$, because large positive and negative error terms will cancel each other in $\sum v_i$, but not in $\sum v_i^2$. This is one way to see that minimizing the *squares* of the error terms for observations is preferred, though this is hardly a formal proof for why the squared error terms are better. Going back to the simple case presented here, the least squares method would express x in (2) in such a way that it minimizes $\sum v_i^2$.[1] Notice that there are virtually no constraints on what x can be, so the method of least squares can be applied to theoretically produce any kind of equation.

The idea of comparing observations and expressing their relation algebraically was around before the method of least squares. Efforts date back to at least the 1700s, and many of

[1] Explanation of the method of least squares influenced by example produced by Aldrich, which can be traced back to Gauss [1, pg 62]

them focused on only linear models [7, pg 239]. By the mid 1700s, Ruggero Boscovich (1711-1787) had proposed estimating error terms in such a way that they summed to zero, and their absolute values were minimized [7, pg 239]. The method of least squares arose out of these investigations, with Adrien-Marie Legendre in 1805 being the first to publish the method, using it to predict the orbits of comets [4, 7, 12]. American Robert Adrain (1775-1843) published a paper on the method in late 1808 or early 1809, and the noted German mathematician Carl Friedrich Gauss published his own work on it in 1809 [7, pg 465].

Evidence suggests that Legendre published on the method of least squares not long after discovering it, and there is also evidence that points to Adrain "discovering" it by reading Legendre's work [7, pg 465]. However, while Gauss unquestionably published on the method after Legendre, he claimed to have discovered it in either 1794 or 1795 [4, 7, 12], and Gauss believed that was enough to claim credit for its discovery.

There are a number of documents that support Gauss devising the method of least squares before Legendre, namely correspondences that Gauss had with others before he published his work in 1809 [7, pp 240-243]. One document that has been studied is a letter Gauss sent to the editor of a journal in 1799, noting some errors he had found in a publication. The most interesting was a miscalculation involving an ellipse, which Gauss claimed to have found after applying "his method" [7, pg 240]. It was hypothesized that Gauss could be referring to the method of least squares [4, 12], and in 1981 Stigler tried to reproduces Gauss's results using Gauss's published least squares method [12]. Stigler was not able to, but did not reject the idea that Gauss may have used the least squares method [12]. However, Celmins reviewed Sigler's work and claimed that the difference in answers was too great to say that Gauss used least squares to find the miscalculations he wrote the editor about [4]. Still, whether Gauss used the method for the ellipse equations in a 1799 paper or not, evidence does suggest that Gauss and Legendre discovered the method of least squares independently, and that Gauss discovered it before Legendre [4, 7, 12].

Legendre wrote to Gauss about the method of least squares, particularly once Gauss had published his work and made it known publicly that he considered the method to be his own. Legendre expressed his views to Gauss quite clearly and respectfully:

> It was with pleasure that I saw that in the course of your meditations you had hit on the same method which I had called the method of least squares in my memoir on comets…I will therefore not conceal from you, Sir, that I felt some regret to see that in citing my memoir p. 221 you say *principium nostrum quo jam inde ab anno* 1795 *usi sumus* etc…In Mathematics it often happens that one discovers the same things that have been discovered by other and which are well known; this has happened to me a number of times, but I have never mentioned it and I have never called *principium nostrum* a principle with someone else had published before me. You have treasures enough of your own, Sir, to have no need to envy anyone…[2]

Legendre certainly respected Gauss, but also felt that publishing the method of least squares first made him the rightful discoverer, regardless of when Gauss discovered the method.

Interestingly, Gauss never considered the method of least squares to be a major discovery, but continued to fight for credit by calling it his own method his entire life [7, pg 248]. Regardless, Gauss did generalize the method considerably in what he published in 1809 by providing an algorithm to calculate estimates, and also linking the method to the calculus of probabilities [12, pg 472]. His algorithm has evolved in to modern-day Gaussian elimination [1, pg 61]. As for Legendre, though his name was not attached to any of his work, the term "method of least squares" comes directly what he called the method in his 1805 publication of his work [4, pg 123]. So, in a way, both of their contributions have been credited.

[2] Excerpt from a letter to Gauss from Legendre dated May 31, 1809, and as quoted by Plackett [7, pg 243]

In the end, the method of least squares was significant when first discovered thanks to the significance of predicting the orbits of celestial bodies. In conjunction with Kepler's laws, the method was powerful and applicable. Eventually, the method of least squares would become as powerful and applicable in statistics.

Regression and Correlation

The next major step towards regression analysis as we know it today came from well outside of mathematics, and well after the discovery of the least squares method. In fact, it came from a person even on the fringe of science at his time, Francis Galton. Though never known as an able mathematician, Galton ran into many mathematical theories of probability and distribution, and had a practical understanding of the ideas [9]. More importantly, Galton had a unique passion for measuring [9], and interest in heredity. The two came together as Galton discovered the ideas of regression and correlation.

By the time Galton was researching heredity, normal distributions and the central limit theorem were known, but relations between a parent generation and its offspring had not been examined closely. Galton was bothered by the central limit theorem, because he felt certain dominant traits were passed from parents to offspring, and he did not see how that could be true if the theorem was correct [11, pg 75]. Galton would have to somehow investigate the issue.

Collecting data for investigation and analysis came naturally to Francis Galton, thanks to a unique passion for counting or measuring virtually anything. For instance, Galton wrote the following account about a visit to South Africa, where he apparently ran into some attractive women:

> I have dexterously even without the knowledge of the parties concerned, resorted to actual measurement…I sat at a distance with my sextant, and as the ladies turned themselves about, as women always do, to be admired, I surveyed them in

> every way and subsequently measured the distance of the spot where they stood- worked out and tabulated the results at my leisure.[3]

Galton made interesting observations domestically as well. There are accounts of him building a "ticker" that he used to count how many times different people fidgeted in meetings [9, pg 510]. Galton simply seemed to have a peculiar fascination with counting and measuring things. He soon applied counting and measurement to something never examined quantitatively before, heredity.

In 1865, Galton published a paper titled "Hereditary Talent and Character." He investigated in the paper biographical dictionaries by counting how many people in the dictionaries were related. He concluded that the rate of relatives in the dictionaries was higher than the rate of relatives in the general population, so he argued that intelligence on some level is inherited. It was Galton's first formal step in advancing heredity studies. By the late 1860s, Galton was aware of Charles Darwin's (1809-1882) evolution theories, and the two began to correspond, even doing some experiments together. At this point, Galton was working feverishly on his theories about heredity, and he continued to correspond with Darwin while writing papers. Between all the writing, Galton designed the experiment that led him to discover regression.[4]

By the early 1870s, Galton was still highly interested in heredity, and had turned his attention to obtaining data about school boys to analyze. He hoped to compare heights and weights of school boys to their parents to determine the role heredity played in height and weight. Meanwhile, Galton also was in the midst of an experiment with sweet peas. He took and measured each sweet pea seed before planting, and then measured the plants that grew, repeating the process to gather information about multiple sweet pea generations. By 1875, he had

[3] Francis Galton in a letter to his brother, as quoted in Schwartz Cohen [9, pg 510]. Schwartz Cohen quoted the passage from *The Life of Galton* by Karl Pearson.
[4] The information in this paragraph is all from Schwartz Cohen [9, pp 510-516]

sufficient data to analyze both the school boys and the sweet peas. He was analyzing them separately, but at some point realized that he could apply the ideas of deviation he was using with the school boys to his analysis of the sweet pea plants. Galton published his work on the school boys in 1876, and then his work on the sweet peas in 1877. Though Galton was more interested in human heredity, the work he released about the sweet peas was much more significant.[5]

When Galton discovered that the theories he was using with human heredity could be applied to explain the sweet pea data, he realized how general his ideas on heredity were. So, inferences could be made between the sweet peas and humans, and vice versa. What he noticed in the sweet peas is that larger parents did produce larger offspring, but the offspring tended to be smaller than their parents. On the other hand, smaller plants had smaller offspring, but the offspring tended to be larger than their parents.[6] Galton plotted the size of parents versus the size of their offspring to calculate a "reversion line," though he later renamed it a "regression line" [9, pg 520]. The idea of regressing to the mean had been born. Equally important to Galton was that he could finally explain deviation from parent to offspring without violating the central limit theorem, and he even created the Quincunx to demonstrate how it worked [11, pp 74-75]. Galton's ideas finally all came together once he discovered regression. In 1890, he further generalized his work on regression to present the idea of correlation for the first time [11], though it seems that our modern understanding of correlation is more general than Galton's [11, pg 76].

---

[5] The information in this paragraph is all from Schwartz Cohen [9, pp 517-518]

[6] The information in this paragraph is all from Schwartz Cohen [9, pg 518]

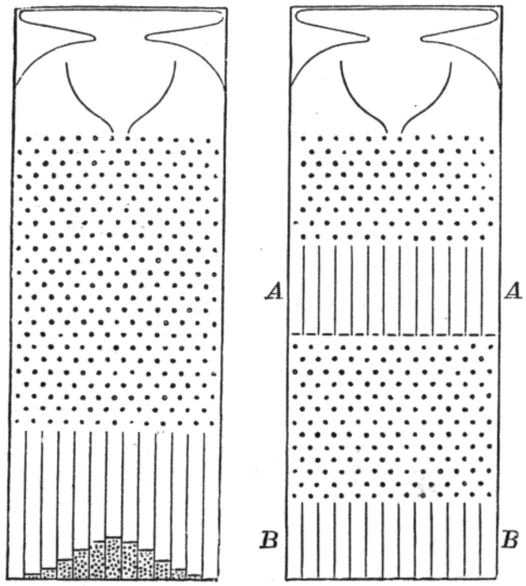

Figure 1: Galton's Quincunx[13]

Shot was fed through the top, and can be seen in the quincunx on the left at the bottom in a normal distribution. The quincunx on the right stops the shot in the middle, where it will already be in a normal distribution. Once released, it still forms a normal distribution at the bottom.

Galton's contributions to statistics are remarkable. He did a great deal for heredity by making observations not made before, and a great deal for statistics by noticing generalizations he could make in his investigations of heredity. However, while Galton's intense interest in heredity was his greatest strength, the reason he was drawn to heredity in the first place is what ultimately makes him a controversial figure. Galton had no intrinsic interest in advancing statistics, or laying the foundations for genetic research. He wanted to create a better society, and saw heredity as the answer [9]. As such, he hoped to uncover the laws of heredity, and then have human reproduction controlled based on them in an effort to maximize human evolution [9, pg 527]. In other words, Galton was the first promoter of eugenics, and as such remains a controversial historic figure. Still, ethical views of eugenics not withstanding, the discoveries Galton made with the hopes of creating a eugenic society were significant in many ways, particularly for the development of regression analysis.

It All Comes Together – Modern Regression Analysis

In the wake of Galton's work, a number of great statisticians emerged, and the field of statistics in developed rapidly. However, Karl Pearson (1857-1936) and R.A. Fisher in particular stand out from the crowd. Pearson further developed Galton's ideas, while R.A. Fisher challenged Pearson. The two became rivals and had some well-documented disputes, adding a dramatic backdrop to the development of modern regression analysis.

Karl Pearson unquestionably followed in Francis Galton's footsteps. A more able mathematician than Galton, Pearson clarified many of Galton's statistical ideas [3, pg 394], while also further generalizing them. Galton's work dealt only with normal distributions and linear relations, but Pearson hoped to extend it to other distributions and relations. In particular, Pearson is noted for developing the idea of non-normal distributions, which he felt were signs of natural selection at work [8, pp. 237-238]. Pearson typically referred to the non-normal distributions as "skew," and his studies led him to begin studying correlation curves, instead of correlation lines [3, pg 394]. Pearson's initial investigations laid the groundwork for further advancements, particularly with non-linear statistical analysis.

However, perhaps Pearson's greatest contribution to statistics was the zeal with which he followed Galton. The two men had mutual respect for each other that shaped much of Pearson's life. In 1901, Galton, Pearson, and W.F.R. Weldon created *Biometrika*, a journal focused on statistics, and Pearson was the editor [3, pg 8]. Then, in 1906, Galton established and funded a eugenic laboratory, with an "ostensibly reluctant" Pearson as supervisor [8, pg 278]. Once Galton passed away, Pearson was selected as the first Galton Professor of Eugenics, as Galton's will requested [8, pg 279]. Clearly, Galton respected Pearson, and Pearson was certainly an advocate for Galton. Perhaps Pearson's greatest sign of respect for Galton is the impressive biography of Galton he wrote called *The Life of Galton*. All in all, Pearson rose to prominence on the shoulders of Francis Galton.

Pearson's loyalty to Galton was most visible with his work in the laboratory and as editor of *Biometrika*. Pearson, by many accounts, was a great supervisor in the eugenic laboratory. He was supportive and energetic. He not only hired women but recruited them, though perhaps partly because he could pay women less. However, Pearson demanded great loyalty in the laboratory too, especially to the "ideas of our Founder," as Pearson would refer to Galton as. Whenever someone deviated from the path, Pearson would dismiss them quickly and coldly. Not surprisingly, this led to some tension and hostility between him and the people he dismissed. Pearson's attitude also guided his editing of *Biometrika*, which resulted in hostility from many outside the laboratory as well.[7]

The most noteworthy dissenter was R.A. Fisher. Pearson and Fisher quarreled quite publicly and vigorously, partly due to their personalities, and partly due to their beliefs. However, they did not quarrel at first.

Fisher read *Biometrika*, and first submitted work to it in 1915. In 1916, Fisher sent a noteworthy hand-written note in response to an article by Kristine Smith. Smith's article had argued that solving a regression so that the chi-squared statistic is a minimum was the most accurate method possible. Simply put, Fisher did not agree, saying that the chi-squared method is arbitrary, and suggesting the Gaussian method (the method of least squares) would be better. Fisher laid out his argument, but it was hardly sufficient for Pearson. He responded honestly and strongly, pointing out flaws in Fisher's logic. Pearson defended the chi-squared method presented in the paper (which is not surprising since Pearson created chi squares [8, pg 255]), and essentially told Fisher that the only way he could win the argument would be by somehow proving the Gaussian method is more accurate.[8]

In 1919, Fisher began work on his response to Pearson. He was looking at two formulas for estimating the standard deviation of a normal distribution, including one related to the

---

[7] Information in this paragraph from Porter [8, pp 266-276]
[8] Information in this paragraph from Stigler [10, pp 38-40]

Gaussian method. Guided by an idea he had read in a paper by Eddington (1882-1944), Fisher made a startling discovery. The Gaussian method not only had a smaller standard deviation, it contained all information provided by the sample population.[9]

The method of least squares had already been applied in statistics, but it was just one of several methods being used [2]. Fisher's work indicated that the least squares method was superior to them all [2, pg 407]. What had started as a response to Pearson's criticisms and challenge had ballooned into an entire new theory on regression. Fisher's version of regression joined the Pearson understanding with the method of least squares to form a powerful new reinterpretation [2, pg 401]. Fisher published his new theory in 1922, though it was largely ignored due his dense writing style and complicated math [10, pp 44-45]. The most noteworthy part of the article at first was where Fisher claimed that Pearson made an error with the degrees of freedom of the chi-squared statistic when parameters were estimated [10, pg 45].

It seemed inevitable that Fisher and Pearson would engage in an ugly dispute. Even though Pearson's challenge largely sparked Fisher to discover his new theory, Fisher did not acknowledge Pearson in any manner [10, pg 45]. On top of that, Pearson was hardly pleased that Fisher had questioned the degrees of freedom with his chi-squared statistic, and so he published several responses asserting that his original formula was still correct [8, pg 256]. The clash of their theories would serve as the backdrop for their verbal attacks against each other for the next 20 years. In fact, Pearson's very last article, written in 1936 just months before his death, was a detailed 25-page account of the argument he had carried on with Fisher for 20 years, as if none of the mostly published dispute was known [10, pg 46]. Pearson also never accepted Fisher's reinterpretation of regression as an answer to the challenge he posed to Fisher in 1916 [10, pg 45]. Pearson simply would not drop the long-standing quarrel.

[9] All information in the above paragraph from Stigler [10, pg 42]

R.A. Fisher was a strong personality in his own right though. He had no issues publicizing his own accomplishments, and rarely gave credit to others that had inspired or influenced his work. Ironically, Fisher heavily criticized Pearson for his lack of interest and appreciation for past mathematicians; yet Fisher only recognized Gauss, the man whom he had taken the method of least squares from, in one small paper in 1812 [2, pg 414]. Fisher was hardly interested in softening his views towards Pearson either, even once Pearson had passed away. In 1945, almost a decade after Pearson's death, Fisher was asked to write a short summary of Pearson's life for the *Dictionary of National Biography* [5, 10]. Even then, Fisher was stingy, and after repeated revisions and requests to make changes by the editor, he ended up not writing the summary that was published [5].

Fisher had a great mathematical mind, but it is hard to separate his advances from the challenge posed by Pearson. The two men developed many of the theories and methods still used in statistics today, especially through their development of regression, even as they rather bitterly attacked each other..

Conclusion

Statistical analysis plays a key role in today's world, with regressions being applied to data in all sorts of fields. It seems fitting that an analytic tool equally suited for sports, economics, biology, psychology, and other areas, originally developed from studying orbits of comets and planets, and early attempts to quantify heredity. The story of how regressions came to be is full of surprises and drama, which along with its remarkably broad range of applications makes it a topic unlike most any other.

Bibliography

1. Aldrich, John. “Doing Least Squares: Perspectives from Gauss and Yule.” *International Statistical Review*, Vol. 66, No. 1 (April 1998): 61-81 (accessed via JSTOR 24 Jan 2009)

2. Aldrich, John. “Fisher and Regression.” *Statistical Science*, Vol. 20, No. 4 (Nov 2005): 401-417 (accessed via JSTOR 24 Jan 2009)

3. Blyth, Stephen. “Karl Pearson and the Correlation Curve.” *International Statistical Review*, Vol. 62, No. 3 (Dec 1994): 393-403 (accessed via JSTOR 24 Jan 2009)

4. Celmins, Aivars. “The Method of Gauss in 1799.” *Statistical Science*, Vol. 13, No. 2 (May 1998): 123-135 (accessed via JSTOR 17 Jan 2009)

5. Edwards, A.W.F. “R.A. Fisher on Karl Pearson.” *Notes and Records of the Royal Society of London*, Vol. 48, No. 1 (Jan 1994): 97- 106 (accessed via JSTOR 24 Jan 2009)

6. Inman, Henry F. “Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange from *Nature*.” *The American Statistician*, Vol. 48, No. 1 (Feb 1994): 2-11 (accessed via JSTOR 24 Jan 2009)

7. Plackett, R.L. “Studies in the History of Probability and Statistics. XXIX: The Discovery of the Method of Least Squares.” *Biometrika*, Vol. 59, No. 2 (Aug 1972): 239-251 (accessed via JSTOR 17 Jan 2009)

8. Porter, Theodore M. *Karl Pearson : The Scientific Life in a Statistical Age*. New York: Princeton UP, 2004.

9. Schwartz Cohen, Ruth. “Francis Galton's Statistical Ideas: The Influence of Eugenics.” *Isis*, Vol. 63, No. 4 (Dec 1972): 509-528 (accessed via JSTOR 24 Jan 2009)

10. Stigler, Stephen. “Fisher in 1921.” *Statistical Science*, Vol. 20, No. 1 (Feb 2005): 32-49 (accessed via JSTOR 24 Jan 2009)

11. Stigler, Stephen M. “Francis Galton's Account of the Invention of Correlation.” *Statistical Science*, Vol. 4, No. 2 (May 1989): 73-79 (accessed via JSTOR 24 Jan 2009)

12. Stigler, Stephen M. “Gauss and the Invention of Least Squares.” *The Annals of Statistics*, Vol. 9, No. 3 (May 1981): 465-74. (accessed via JSTOR 17 Jan 2009)

13. *Wikimedia Commons*. http://commons.wikimedia.org/wiki/File:Quincunx_(Galton_Box)_-Galton_1889_diagram.png (accessed 14 Mar 2009)