Chapter 7

Correlated Cousins: Darwin and Galton

In the early eighteenth century a comfortable English gentleman in the crust of society could enjoy the fruits of England's colonial expansion. He drank the finest tea from China poured into the finest Chinese porcelain teacups. The Chinese adorned their porcelain with hand painted scenes that were exquisite.

Later in the eighteenth century, tea became a standard among not just the upper crust, but an emerging middle class. Thanks to the industrialist Josiah Wedgwood, mass-producible alternatives to Chinese porcelain were available at very reasonable price points. The upper-crust enjoyed their ornate Chinese porcelain, but most of the nation of tea drinkers used cups from one of Wedgwood's many factories.

Despite the loss of America, the blessing of Empire smiled upon the British. Taxes collected from economic activities throughout the empire filled the royal coffers. An ultra-wealthy class profiting from monopolies in trade, shipping and in the case of India, government backed rights to land ownership and taxing authority existed alongside a burgeoning middle class. Unlike their European counterparts, gentlemen resolved their disputes in respected courts and debated in Parliament without taking to the streets.

This nation of gentlemen citizens (woman didn't really count) added much to scientific and industrial progress. The government as well as private donors funded scientific research through the Royal Society. Financiers invested in industrial processes that brought engineering designs to fruition. The British invented the steam engine, and then were the first to apply the steam engine toward powering looms that increased the productivity of textile plants, powering steamships, and finally powering railroads.

When Queen Victoria ascended to the throne (1837) the British were literally sitting on much of the world. Whether from a Wedgwood cup, or an exquisitely painted Chinese porcelain cup, a British gentleman drinking his tea could bastion in self admiration. Their respect of institutions and social civility prevailed and Britain was supreme.

As he drank his tea and admired his civility, did he forget or ignore the violence that his society forced upon others. The British addiction to tea caused an unsustainable trade deficit with China. The Chinese insisted upon payment in silver; this drained the British coffers. One solution, drink less tea. But wait, there's money to be made.

The British triangulated their financial environment and made the following calculation. The East India Company owned land and could force peasants into slavery. Slave labors made it feasible for large scale production of opium. The East India Company sold the opium to distributors who then profited from sales, in silver, to an increasing population of addicts in China. This placed silver in the hands of the British who could then purchase the tea for their population of gentlemen back home. The gentlemen who were civil among themselves turned brutal when encountering others.

And if the English gentlemen were to be honest, they could be brutal to their own as well. As chronicled by Charles Dickens, for many opportunities to climb the social ladder were nonexistent. Serfdom had given way to a new industrial servitude. In British cities slum neighborhoods abounded.

One might be tempted to conclude that British civility was a mere masquerade covering a dog eat dog reality. This conclusion would be equally as unfounded as the gentlemen who ignored his society's shortcomings. The gentlemen's agreements held amongst themselves. Social protocols that tempered destructive behavior held sway. Upward mobility was accessible to many individuals and while a principled moral code did not find universal application, there was broad agreement over right versus wrong and many gentlemen did follow through.

One such individual was Erasmus Darwin, the paternal grandfather of Charles Darwin and Francis Galton. Erasmus was a well known physician and anti-slave proponent. Two generations later Charles upheld his grandfather's egalitarian views, while Francis' perspectives were more ambiguous. Both men are contributors to our body of fundamental scientific knowledge and methods.

7.1 Charles

Charles Darwin's family could easily afford the luxury of hand-made Chinese porcelain; after all, Charles Darwin's grandfather, Erasmus, was one of the wealthiest men in Britain and the grandfather's son, Robert, married into a family of even greater economic and social rank. Robert married Josiah Wedgwood's granddaughter, Sussanah. One wonders which cup Charles drank his tea from. Was it the cup of the privileged, which was certainly within the Darwin family's reach, or was it the cup that enhanced the family privilege?

With hindsight we recognize Darwin's contributions. However, if we had reviewed his resume over his first 22 years of life, we might forecast a very different course. Darwin was a mediocre elementary and high school student. Because he was undisciplined and performed poorly in high school, his father, Robert Darwin, curtailed his high school studies and sent him to medical school under the supervision of Charles' elder brother, Erasmus (named after the grandfather). Erasmus was in his final year of medical school at Edinburgh.

The change of venue had no impact upon Charles' undisciplined nature. Once Erasmus graduated, Charles during his second year, had a freer hand to socialize and spend his father's generous stipend as he pleased. It all ended with Charles declaring his disinterest in a medical profession and dropping out after his second year.

The father attempted a reset guiding Charles toward the life of a clergyman. Robert Darwin passed away in 1848, eleven years prior to Charles most famous publication *On the Origin of Species by Means of Natural Selection*. He never lived to see the irony; Charles' work became the center of a controversy between science and religion that continues to this date.

At 18, Charles undertook theological studies at Cambridge University. Charles admits that much of the time was wasted in frivolity. Charles was a genuinely likable individual who made friends easily. Combine that with a generous pocketbook and you've got a lot of hours drinking beer at the local pub. If God disapproved he did not show his displeasure.

The hours that Charles spent in the pubs were greater than the hours he spent attending classes. The cajoling of his concerned classmates instilled enough discipline so that in 1831, Charles passed his exams and earned a B.A. degree. After which he declared that he had no interest in joining the clergy.

If we were to take a cursory glance at Charles' resume, we might forecast that he would live off his family's wealth and enjoy a leisurely yet inconsequential life. One essential clue along with a somewhat random circumstance would dramatically alter the forecast.

Despite Charles' distaste for structured learning, he had an insatiable intellectual curiosity. He paid no heed to lectures or required course readings, but read vociferously in areas that peeked his curiosity. Mostly, his interests were in geology and studies of nature.

Alongside his readings, Charles would attend more informal discussion sessions in his topics of interest. He was a regular attendee of a weekly discussion group that Professor John Henslow sponsored. Henslow was a professor of botany who was also knowledgeable in the field of geology. Naturally, the discussions centered around geology and natural history.

Henslow saw something in Charles that eluded others. Aside from his weekly discussion group, Henslow organized frequent field trips attended by other accomplished geologists and naturalists. How strange it was that among the accomplished professors there was singularly one student who was also invited to participate. It was even more remarkable that this student was not even formally enrolled in the fields of geology or natural history. Charles was an outlier who made his impression.

Following his graduation in April 1831, Henslow became aware of a naval expedition chartered for the purpose of mapping unmapped territories. In line with the English attitudes of fusing theory with practice, the expedition offered free passage (room and board) on the vessel to any naturalist or geologist who might wish to explore territories that the expedition would survey.

On the economic side, few men could afford an unsalaried position over several years, a circumstance that a fortunate son born into a very wealthy family did not have to worry about. However, by trade nor formal education, Charles was neither a natural historian nor a geologist. Nevertheless Henslow immediately proposed Darwin for the position and with a recommendation from the esteemed professor, Charles was under serious consideration.

The only barrier to the position was Roger Darwin, who opposed. Perhaps Roger Darwin's view of the voyage was that it invited further aimless wandering, not a disciplined environment that his son needed.

But then there was Josiah Wedgwood II, grandson of the industrialist Josiah Wedgwood and sister of Charles' mother, Sussanah Wedgwood Darwin. Charles' mother passed away when Charles was only eight years old. Possibly as a commitment to his departed sister, Josiah watched after Charles. When the opportunity to join the Beagle arose, Josiah intervened on Charles' behalf and convinced Roger to give permission to Charles.

Hemslow and Josiah were right, Roger was wrong. While on firm ground in the protected custody of the finest British educational institutions, Charles stumbled. On the Beagle, freely out in the open seas and able to explore new territories as he pleased, Charles found his footing.

The experiences from his five year journey, 1831 - 1836 constitute the foundations of Charles' works until his death in 1882. Charles was an intrepid adventurer. By foot and by horseback, he traveled thousands of miles through forests and deserts alike. He traveled along seashores and up the peaks of the Andes taking in everything with a power of observation that cannot be taught, but is innate to few individuals.

The world was open to Charles' powers. He saw with his own eyes seashells on peaks 15,000 feet high; evidence of powers that reshape the Earth. He saw with his own eyes coral reefs that rise above the shoreline while building upon the graves of coral ancestors thousands of feet below. He saw with his own eyes the diversity of life upon the Galapagos that branched from distant relatives upon the South American continent. From these observations, Charles was able to deduce theories that would have been inaccessible to men who did not see it with their own eyes.

Below is a Table of Charles' publications that contain direct records from Charles' Beagle expedition.

Table 7.1 lists the major publications by Charles Darwin that contain references to his observations and collections from the voyage of the *HMS Beagle*.

Concerning the list, there are several notable points. First until the publication of *On the origin of Species*, geological works dominate the biological works. Afterwards, there are no geological entries. From his central idea of natural selection, Charles germinates a series of further investigations into biology.

Charles' publication *Voyage of the Beagle* brought him instant fame and recognition as a preeminent scientist. Charles wrote in an engaging style that reads like an action thriller. While the public feasted on the accessible writing style, the scientific community wondered at the novelty of the scientific content.

Charles' most famous geological insight is his explanation for the formation of coral reefs and atoll formation. The reefs mysteriously appear as an empty circular frame. Where is the picture? Why do they emerge from the ocean's bottom unattached to any land mass? Charles' explanation was that the accompanying landmass once was a volcano that formed an island. The volcano became inactive and sank into the ocean under its own weight. As the volcano sank, dying coral gave birth to new generations that sustained themselves at comfortable depths by building upwards upon the calcium deposit graves of their fore-bearers. Modern investigations confirm Charles' explanation with coral sculptures scaling from the depths of up to 2000 feet.

Another notable point is the time it took Charles to clarify his ideas on evolution and natural selection. Charles' Beagle expedition notebooks are pregnant with clues, but the theory gestates for over 23 years before publication. What finally sparked the publication?

One might say it was a message from Charles' intellectual twin who Charles never had heard of. In June of 1858, Charles received a manuscript from an unknown Alfred Wallace (1823-1912) entitled *On the Tendency of Varieties to Depart Indefinitely from the Original Type*. There Charles read the theory of natural selection that had been germinating within his mind for decades.

As with Charles, Wallace coupled keen observation skills with a hands-on see it for yourself approach. Wallace had years of extensive field experience in South America and what is now known as Malaysia and Indonesia. As with Charles, Wallace suffered from, malaria during his field work. As with Charles, Thomas Malthus' work on population growth influenced Wallace. As with Charles, Wallace determined that natural selection produced species differentiation. Unlike Charles, Wallace formalized his ideas in a ready for publication manuscript.

Charles did the honorable thing. He forwarded Wallace's manuscript to Charles Lyell and Joseph Hooker, both prestigious scientists and members of the Linnean Society of London. Charles supported its publication. Both Charles Lyell and Joseph Hooker were aware that Charles Darwin had previously outlined the theory of natural selection in an unpublished 95 page essay that Charles Darwin wrote in 1844.

The August 20, 1858 Volume 3 Issue 9 TextitJournal of the Proceedings of the Linnean Society of London was dedicated to a single topic with three entries. The topic *On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection*. The entries were:

1. An excerpt from Darwin's 1844 essay
2. A letter from Darwin to Asa Gray (1857) explaining his theory
3. Wallace's manuscript.

After this episode, Charles became serious about addressing any reservations that kept him from previously publishing the theory. The result is his most famous book, *On the Origin of Species*. In this book, Charles once again displays his mastery. It is a dive into science that is singular for being both accessible to the general

Table 7.1: Darwin's Publications Referencing the Beagle Expedition

Publication	**Year**	**Focus**
Voyage of the Beagle	1839	Travel and natural history journal documenting Darwin's detailed observations of geology, biology, and indigenous peoples during the expedition.
Zoology of the Voyage of H.M.S. Beagle	1838–1843	Multi-volume work on specimens collected during the Beagle voyage, with contributions by expert zoologists; Darwin wrote introductory sections and habitat notes.
The Structure and Distribution of Coral Reefs	1842	Marine geology; developed the theory of coral reef and atoll formation based on volcanic island subsidence.
Geological Observations on Volcanic Islands	1844	Geology; analyzed volcanic island formations observed during the Beagle voyage.
Geological Observations on South America	1846	Geology; focused on fossil finds, glacial action, and tectonic uplift in South America.
On the Origin of Species	1859	Evolutionary biology; introduced the theory of natural selection, using key examples from Beagle data, especially biogeographical patterns.
The Variation of Animals and Plants under Domestication	1868	Variation and heredity; includes comparative references to wild species observed during the Beagle voyage.
The Descent of Man, and Selection in Relation to Sex	1871	Human evolution and sexual selection; draws on behavioral observations of animals made during the voyage.
The Expression of the Emotions in Man and Animals	1872	Comparative psychology; includes incidental references to behaviors Darwin first recorded on the Beagle.
The Power of Movement in Plants	1880	Plant physiology; occasional references to tropical plants first seen during the expedition.
The Formation of Vegetable Mould through the Action of Worms	1881	Soil science and ecology; may allude to early ecological observations made during the Beagle expedition.

public and inspiring to the scientific community. It was a best seller and at age 50, Charles' fame reached new heights; this time connecting with a younger generation.

In 1876, Charles wrote a confidential letter to his children. Charles' eldest son posthumously published the letter. It is a biographical essay detailing the highlights of his life. After reminiscing about his childhood, the letter focuses on Charles' scientific achievements, publications, and relations with colleagues within the scientific community. As Charles lived upon his family's wealth and had no formal employment, he rarely came into contact with individuals of a younger generation.

Whereas Charles is extremely generous in his mentions of those who mentored him, as well as colleagues with whom he communicated, he mentions not a single individual from a younger generation. The man who authored the most revolutionary scientific ideas of his times and holds high influence across generations of scientists to this age, mentions not a single individual who could carry on with his work.

It is odd, for there is one individual of a younger generation that was so inspired by *On the Origin of Species* that he changed his entire career path. The individual wrote glowing letters to Charles describing an intense emotional effect that the book had upon him. Charles most definitely read the letter. The two even collaborated on a set of experiments to determine whether or not inherited characteristics were transferred to sex organs through the bloodstream.

How did Charles, one of the most famous scientists of his age and certainly a man who was sought out by many, become aware of this particular individual? Beyond that, why when he preferred to work in solitude did Charles agree upon a scientific collaboration? And why does Charles give no mention of the individual in his autobiographical letter?

7.2 Francis

Francis Galton never had personal economic anxieties that are common to many throughout history. Born into the same privileged bloodline as Charles Darwin, Francis was the son of Violetta Darwin and Samuel Tertius Galton, a wealthy banker. His grandfather was Erasmus Darwin—poet, physician, philosopher, and the very same grandfather that Charles Darwin claimed. From the cradle, Francis had the liberty to explore life on his own terms. And explore he did.

As a child, Francis exhibited a kind of precocious brilliance that seemed to mark him for greatness. By age five he was reading and writing Latin. By eight, he could recite large portions of Shakespeare and classical texts. And by adolescence, he was already a polymath-in-the-making. In contrast to Charles, whose early years did not reveal his latent brilliance, Galton's early life seemed to presage an extraordinary destiny.

But early bloomers are not guaranteed smooth trajectories. When Francis entered Cambridge, the weight of expectation was heavy, and for the first time his intellectual wanderings were constrained by formal learning. He initially pursued medicine like Charles, but quickly became disenchanted. He switched to mathematics, but when his father died suddenly in 1844, Francis—midway through his studies—suffered a nervous breakdown. He graduated, but without the distinction once expected.

Though untethered from a traditional career path, Francis never lacked direction—only that his compass pointed in unusual ways. One of his boldest decisions was to explore southern Africa. In 1850, inspired by the travels of David Livingstone and others, Francis set out for what was then called Damaraland—present-day Namibia. His aim was ambitious: to explore unknown interior regions, chart unmapped territory, assess the geography and meteorology of the region, and test himself against the limits of endurance and danger.

The idea was Francis' own. Although he sought and received support from the Royal Geographical Society, which provided a small grant and logistical endorsements, the vast majority of the expedition's financial support was like the idea, Francis' own. His ability to fund the project without hesitation was emblematic of his life. Francis, like Charles, never required employment. He lived off inherited wealth and was therefore free to pursue science with no external obligations or institutional expectations. While most scientists of his time were tethered to universities or the clergy, Francis built a life of inquiry entirely outside conventional academic pathways.

Over the course of his expedition, Francis traveled more than 4,500 miles—by ox-wagon, horseback, and on foot. He navigated challenging terrain, from searing deserts to mountainous ridges, and recorded his observations with meticulous detail. He charted previously unrecorded areas of Damaraland and Ovamboland and created some of the first accurate maps of the region. These maps were later shared with the Royal Geographical Society, earning him its prestigious *Founder's Medal*.

Francis' journey was not without peril. His expedition encountered repeated threats: sandstorms, food shortages, river crossings that nearly drowned pack animals, and tense standoffs with local groups. On one occasion, he faced what he called a "critical situation" with a group of Herero tribesmen. The tribesman were armed, suspicious, and had a well deserved reputation of savagery having marauded defenseless villages. Francis had to rely on a mix of negotiation, calm demeanor, and a strategic show of strength to de-escalate tensions. Dangers went beyond interactions with suspicious tribesman. Illness plagued his crew. At one point, Francis was incapacitated for days by fever, with only minimally trained assistants to care for him. Yet he pressed on.

True to form, Francis measured everything: barometric pressure, temperature, wind velocity, soil types, and human physiognomy. His notebooks were filled with observations and data—part travelogue, part scientific report, part self-experiment. He brought back not just maps, but a methodology for scientific travel that emphasized precision, measurement, and reproducibility.

Despite never holding an academic post, Francis embedded himself in Britain's scientific elite. He became a Fellow of the Royal Geographical Society and, later, of the Royal Society itself. He regularly corresponded with leading figures such as Charles Darwin, Thomas Huxley, Karl Pearson, and Herbert Spencer. Though independent, he was never isolated. He built his scientific network through letters, personal visits, and attendance at key societies and clubs in London—environments where his wit, ideas, and pedigree earned him attention and respect.

Piqued by livestock breeding, Francis developed an interest in inherited traits. But it was Charles' *On the Origin of Species* that electrified his mind and transitioned the interest to an obsession. Francis read the book shortly after its publication in 1859 and immediately wrote to Charles with glowing praise. In the letter, Francis declared that the book had struck him "like lightening," and that he was overcome by its explanatory power. "I shall never be easy," he wrote, "until I have followed it out." This letter began a sustained correspondence between the two cousins, linking Charles' biological theory with Francis' growing obsession: the measurement and inheritance of human traits.

Francis quickly turned his energy to testing ideas that followed from those of Charles. In the 1860s, he developed a hypothesis—based on the idea of pangenesis, which Charles himself had proposed—that hereditary information might be carried in the blood. According to this model, the body's cells released minute particles (called "gemmules") into the bloodstream, which then transmitted traits to the reproductive organs and thereby to the offspring.

To test this, Francis designed an experiment: he transfused blood from one breed of rabbit into another, hoping that traits from the donor might be inherited by the offspring of the recipient. Charles was intrigued and supported the plan. Over several years, Francis performed extensive rabbit transfusions. If blood did indeed carry hereditary material, then offspring of the transfused animals should have shown mixed traits.

They did not. The results were entirely negative. No offspring bore traits of the donor rabbits. Galton had disproved his own hypothesis—and, inadvertently, Darwin's model of pangenesis. Though Charles accepted the failure with scientific grace, he privately found it disappointing. Francis, for his part, published the negative results without consulting Charles.

The publication held negative consequences for Charles' theory of pangenesis; subsequent correspondences from Charles to Francis had a chillier tone than prior correspondences. Charles continued to respect Francis' brilliance, but became more cautious in engaging with his speculative theories. Francis, increasingly independent, moved further into uncharted terrain—laying the groundwork for bio-statistics, quantitative psychology, and eugenics.

From the 1860s onward, Galton turned increasingly to questions of heredity and measurement, laying the foundation for entire fields that did not yet exist. His enduring contributions lie not only in the concepts he introduced but in the statistical methods he developed to support them.

Selected Publications of Francis Galton (Emphasis on Biostatistics and Measurement):

- *Narrative of an Explorer in Tropical South Africa* (1853) – Account of Galton's African travels; includes geographic, meteorological, and ethnographic observations.
- *Hereditary Genius* (1869) – Argued that intellectual abilities run in families; laid groundwork for study of behavioral heredity.
- *English Men of Science: Their Nature and Nurture* (1874) – Survey-based study examining the relative influence of environment and heredity.
- *Statistics of Mental Imagery* (1880) – Pioneering study in psychology using statistical and experimental methods to examine individual cognitive differences.
- *Inquiries into Human Faculty and Its Development* (1883) – Introduced the term "eugenics"; covered sensory acuity, memory, and reaction time.
- *Natural Inheritance* (1889) – Systematized Galton's biometric theories; introduced the concepts of regression to the mean and correlation.
- *Fingerprints* (1892) – First scientific classification of fingerprints; became foundational in forensic science.
- *Biometrika* (founded 1901, with Karl Pearson) – Although not a single work, Galton was instrumental in launching this journal to formalize the field of biostatistics.

Francis died in 1911, aged 88. In his will, he endowed a chair in eugenics at University College London. His intellectual descendants would carry his ideas forward—some into statistical and psychological sciences, others into dark chapters of history that Francis never imagined.

7.3 Correlation: The Partnership

From the modern day experience of science, it is obvious that statistics would be central to Galton's investigation of inherited characteristics. But no such obvious notions existed in Francis' times. The field of statistics was in its infancy. There wasn't the language or methodology to address such questions as: what is the relationship between physical, mental, and character traits of parents and offspring? This was not an obstacle for Galton. Working with the mathematician, Karl Pearson, together they would create the language and methods that are core to modern day statistics. Before presenting Galton and Pearson's contributions, let's examine the state of the art of statistics in his day.

In the mid-nineteenth century, statistics was largely a discipline of counts and averages. Governments collected census data, insurers calculated life tables, and astronomers studied errors of measurement. The prevailing model was that of “the law of large numbers”—the idea, formalized by Jacob Bernoulli and refined by Laplace, that regularity emerges from chance when observations are numerous. Adolphe Quetelet, a Belgian polymath, had gone further in the 1830s by applying the mathematics of probability to human beings. He introduced the notion of *l’homme moyen*—the “average man”—whose characteristics, he argued, followed the bell curve of the normal distribution. Quetelet measured traits like chest circumference of soldiers and showed that human variation could be treated mathematically, just like astronomical errors.

Yet Quetelet’s vision, though influential, remained descriptive rather than relational. He focused on distributions within a population, not on the connections between individuals across generations. The language of random variables, as we use it today, did not yet exist. Probability was applied to dice and planets, to accidents and averages, but rarely to heredity or psychology. The leap from viewing traits as fixed attributes to viewing them as quantities subject to probabilistic variation had only just begun.

This was the intellectual climate into which Galton stepped. Building on Quetelet’s notion of human measurements as distributed across a population, Galton asked the next question: how are those distributions linked between parents and children? If Quetelet’s “average man” embodied the population at a moment in time, Galton wanted to uncover the laws that governed how traits moved through time, across generations. For this, description was not enough. A new statistical tool was required—something to capture not just variation, but association.

Galton began with height. He gathered data from hundreds of families, measuring the heights of fathers, mothers, and children. To make comparisons simpler, he averaged the heights of the two parents into a single number he called the “mid-parent” value. He then plotted the child’s height on the vertical axis of a graph and the mid-parent’s height on the horizontal axis. Each family became a single point on the page.

The pattern was clear: taller parents tended to have taller children, and shorter parents tended to have shorter children. The points drifted upward together. But the relationship was not exact. Very tall parents often had children who were still tall but not quite as extreme. Very short parents had children who were somewhat taller than themselves. On average, the children’s heights “regressed” toward the middle of the population. Using terminology that would be considered toxic today, Galton called this tendency regression toward mediocrity—today we call it regression to the mean.

Mathematically, Galton expressed his ideas as follows.

$$\Delta y = a(\Delta x)$$

where

- Δy is the difference between the height of the offspring and the mean height of all offspring in the sample
- Δx is the difference between the mid-parent height and the mean mid-parent height
- a is a multiplier (slope) estimated by Galton

By hand fitting a line to data, Galton estimated the value of a as $a = 2/3$. The value, being less than one, indicates that the offspring are closer to the mean than the mid-parent are to their mean. Galton’s interaction with Pearson lead Pearson to generalize Galton’s regression equation and rather than eyeballing a regression line, provide formulas that optimally fit the line to the data.

The regression equation is:

$$y = ax + b$$

where

- y is a random variable such as the height of offspring
- x is a random variable such as the height of parents
- a and b are regression parameters that are chosen to minimize the difference between the regression line given by $y = ax + b$ and the actual data.

While Pearson independently provided a calculation method along with the resulting formulas for the parameters a and b, Gauss and Legendre had solved this problem nearly a century before Pearson.

While appearing different, Galton's expression and the regression expression are different only by a shift in the coordinate system. Galton centers his data so that the mean of Δx and the mean of Δy lie at the origin. The general regression equation allows for uncentered data, yielding the intercept parameter, b. For both equations, the scaling parameter, a has the same value.

From this work, a further question emerged: how strongly do two variables—like parent and child heights—move together? Galton addressed the question with expressible ideas that he coined correlation.

Galton visualized correlation by examining how tightly data points clustered around the regression line. If most points lay close to the line, he inferred a strong linear relationship; if widely scattered, the linear relationship was weak. The sign of the slope indicated whether the variables moved together (positive) or in opposition (negative). This intuitive approach laid the groundwork for Pearson's formal correlation coefficient.

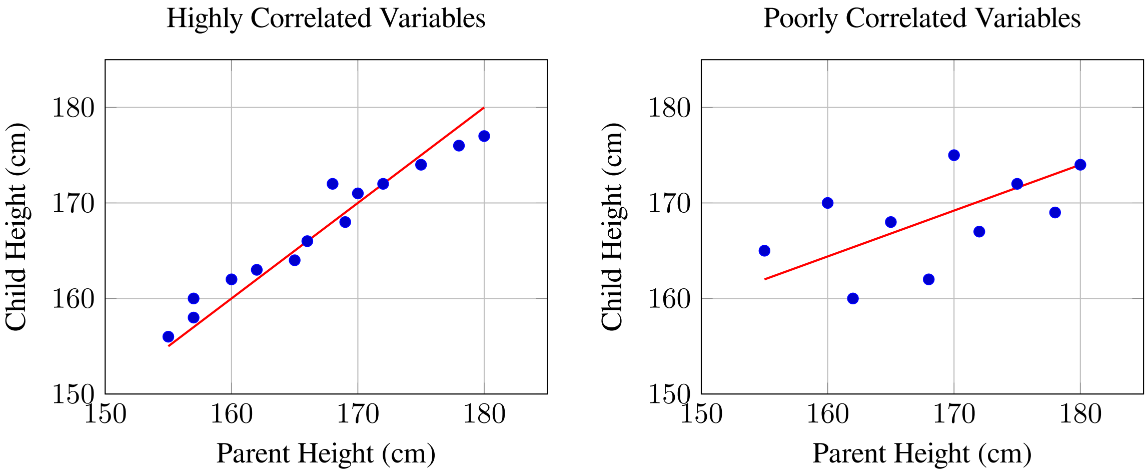

Figure 7.1: Left: Highly correlated data with regression line. Right: Poorly correlated data with regression line. The red line represents the regression line for each dataset.

Pearson formalized Galton's description of correlation into a formula with the following properties.

- $-1 \leq \text{Corr}(X,Y) \leq 1$ where Corr(X,Y) is the correlation between the variables X and Y
- if no clear linear relation between the variables exists, $\text{Corr}(X,Y)$ is close to zero. In this case the regression line provides little information
- If the regression line provides an excellent fit for the data then $|\text{Corr}(X,Y)|$ is close to 1.

- The sign of Corr(X,Y) is the same as the sign of the regression coefficient a.
- Corr(X-q,Y-p) is the same for all constants q and p. The choice of constants only shifts the coordinate system without affecting the fitness of the regression line to the data.

7.4 Galton and Pearson, The Data Scientists

This section presents a case study exploring hereditary influence on height. The case study follows Galton's work with some modifications. Our modifications allow us to make observations that go beyond Galton's demonstration of regression to the mean.

Define the problem.

Determine relationships between the heights of parents and their fully grown offspring.

Propose an input-output parametric model of the system.

$$y = ax + b$$

where

- y is the modified height of offspring
- x is the average of the modified height of the parents
- a and b are regression parameters that are chosen to fit the data.

The purpose of modifying heights is to assure that men and women are given equal weight on the model's outcome. To accomplish this, there is a multiplier on all female heights so that their average corresponds to the average height of the men.

The adjustment allows for a standardization of heights so that differences between the model and actual results are not attributable to the known difference in female and male heights.

We use the same multiplier that Galton applied, 1.08, however, whereas we apply the multiplier to all females, Galton applied it only to the mothers, not to daughters.

Identify the required data.

For this exercise, we adopt Galton's dataset containing the fathers height, mothers height, offspring's height, and offspring's sex having 934 adult offspring. [1]

Collect and organize data as inputs and outputs.

- Inputs, the average height of the parents, with a maternal height modifier
- Outputs, the heights of the offspring with a daughter height multiplier.

Below is a small sample of modified data from Galton's actual dataset.

[1] The dataset is from Kaggle's website. This is slightly larger than the 928 observations in Galton's original table, likely due to later re-tabulations by Pearson that introduced minor duplications or rounding adjustments.

Input Average Modified Parent Height	Output Modified Child Height
69.6	70.7
73.60	72.7
69.5	70.0
67.7	69.5
65.9	64.0
70.0	68.0
70.4	69.6
71.1	70.7
68.2	68.6
69.8	74.0

Table 7.2: Sample of Galton's Modified Data

Define a metric that quantifies the error between model predictions and observed outputs.

For the purpose of parametrization, we use the least squares metric of Gauss and Legendre. Recall, that the least square method minimizes the sum of the square of all errors between the regression line and the actual data points.

Apply an optimization routine to adjust the parameters and minimize the error.

We apply the least squares method of Gauss, Legendre, and Peterson. The result is $a = 0.7134$ and $b = 19.9120$. Figure 6.2 below gives a plot of the data against the regression line.

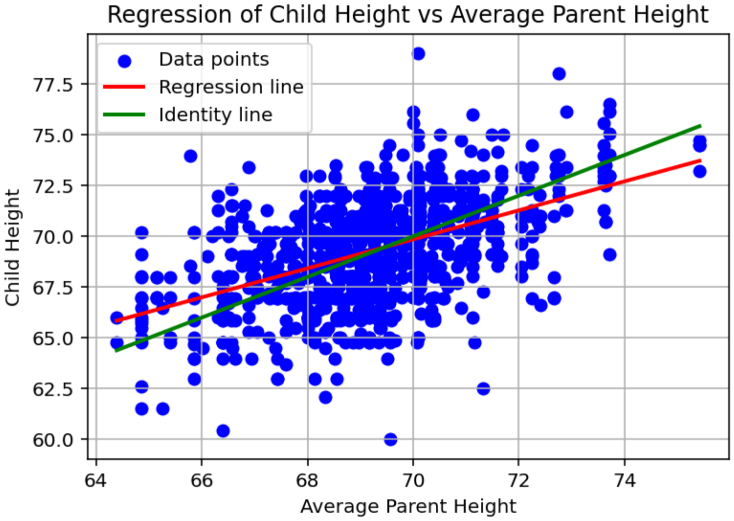

Figure 7.2

The correlation between the adult's average height and the children's height is corr$= 0.4969$.

Discussion of Results

This section uses the data from Galton's paper *Regression towards Mediocrity in Hereditary Stature*. As previously noted, this paper credits Galton with the concept of regression toward the mean. Examining the scatter plot reveals trends that Galton notes.

- Parents with a short or tall stature are more likely to produce children of the same stature. One observes this by the upper drift of the data points. The positive regression coefficient, $a = 0.7134$ and the positive correlation confirm this trend.
- Regression toward the mean is apparent. The identity line (green) displays the hypothetical case in which the children's heights are equal to their parents. The portion of the observations where parents' heights are relatively short, less than 67 inches, lies predominantly above the green line; children are predominantly taller than their parents. Alternatively, the portion of the observations where parents' heights are relatively tall, above 72 inches, lies predominantly below the green line; children are predominantly shorter than their parents. This trend causes the regression coefficient $a = 0.7134$ to be less than one.

Aside from the usual observations, there are other points of note.

- The observations are not clustered tightly around the regression line. This means that the regression equation in which parent's average height is the input, is a partial indicator of a child's height. The correlation coefficient corr$= 0.4969$ is not close to one indicating a moderate linear relation between the input and output variables.
- The deviations from the mean are considerable. Indeed the two extreme offspring cases, minimum child height of 60 inches and maximum child height of 79 inches stem from observations in which the average height of the parents is close to the mean of all averages. A further investigation of the fathers' and mothers' heights reveals that all of the parents' heights yielding these extremes are of near average height. It is noteworthy that these extreme child cases are the most extreme of all observations, parent and child. While there is a trend of regression toward the mean, the data also reveals a natural variance in which average parents may produce extreme offspring.

Validate results against additional data.

Galton's dataset has too few observations to validate the conclusions.

7.5 Final Thoughts

Galton and Pearson's work illustrates a transition in the usage of data. Previous works utilize data to parametrize a model that in turn addresses a specific issue. Galton does not gather data to parametrize a model, instead he uses the model to infer and confirm facts about the data. The data is central and the model is useful only if it reveals information about the data.

Galton and Pearson rediscovered Gauss' and Legendre's least squares methodology and used it to determine the regression line. Then they pressed on further. Addressing the need to determine relations between separate sets of observations Galton proposed and Pearson formalized the concept of correlation.

Correlation lies at the heart of ChatGPT. ChatGPT finds relationships between words and uses its findings to construct texts. The chapter Neural Networks and the Connectors along with the chapter The Chat, explore a neural networks architecture and programming methodology that allows ChatGPT to uncover the necessary relationships that yield expressive texts.

Before moving on to those chapters, the book looks at Henry Ford's use of data science. While previous chapters describe academic applications, Henry Ford is noteworthy for applications that built the industrial economy. This is in the list of top ten ideas that have transformed society. An argument could be made that it is at the top of the list.

7.6 Summary Poem: Cousins of Measure

In England's age of steam and grace,
Where porcelain cups and empire's face
Reflected order, wealth, and tea—
Two cousins dreamed of destiny.

Charles sailed far through wind and storm,
To watch life's endless shapes transform;
He saw in shells on mountain stone
The whisper: *change is nature's own*.

His voyage turned to measured thought,
Each leaf and beak with meaning fraught;
He mapped the unseen threads that bind
The random world to ordered kind.

Francis, child of equal line,
Measured bodies, hearts, and mind.
His gift was not of coral reefs,
But of numbers drawn from human griefs.

He charted traits from sire to son,
The rise and fall of everyone.
In parent heights and offspring's span,
He saw the curve that shapes all man.

The scatter plots began to speak—
A slope, a pull, the mild and meek;
Extremes would fade, the mean return—
And *regression* was the name he'd earn.

Through data dense and insight keen,
He sought the pattern in between.
The random cloud became a line,
Where chance and cause could intertwine.

From this new view of nature's plan,
He joined with Pearson, hand in hand—
To teach the world that numbers tell
The truths that words alone can't spell.

They built a language, pure and clean,
Where symbols bridged the seen unseen;

Correlation gave it tone and tune,
Regression sang beneath the moon.

No longer muse or mystic's art—
Now knowledge born from measured part.
Each dataset a hidden song,
Each graph a proof of right or wrong.

And thus began a modern creed:
To let the numbers speak, not plead.
Galton asked, and Pearson found—
How reason lives in the counted ground.

So from their quills and careful sight
Came data's dawn, a new-found light;
The world, once read by faith or art,
Was now revealed by scatter's heart.

Two cousins—one who sailed, one who scaled—
Their legacies entwined, unveiled.
Through ink and graph, both sought the same—
To map the truth behind the name.

And in their wake, we trace their aim:
From chance to chart, from hunch to scheme;
Where models learn and measures frame—
This is the modern data dream.