Regression to the Mean

Dominic Klyve*

December 3, 2019

1 Introduction

As with so many words in statistics, *regression* can mean multiple things. First-year statistics students often learn the procedure of *linear regression* to find the line of best fit on a scatterplot. It turns out that this procedure is just one special case of *general linear regression*, a powerful statistical technique that can be applied to any number of variables.

However, the term has a second meaning—one that is closely related to its original Latin root. To "regress" is to go back, usually to a previous state.[1] In this sense, the term dates back to the late nineteenth century, when Francis Galton (1822–1911) studied the relationship between the heights of adults and the heights of their adult children. He found that if parents were unusually tall, their children were usually shorter than they; similarly, very short parents tended to have children who were taller. It seemed as if some invisible force was pulling the height of successive generations back to the average height of humans. Galton referred to this process, appropriately enough, as *regression to the mean* [Galton, 1886].

Francis Galton was a scientist, a eugenicist, and a half-cousin of Charles Darwin. He seemed to be interested in everything, but his favorite hobby was measuring; he measured anything and everything he could. He measured the lifespan of kings and queens to test whether they lived longer than other people. (They didn't, and since Galton was fairly sure that more people prayed for the king than prayed for average people, he concluded that prayer is ineffective in prolonging life.) He collected weather data and constructed the first weather map, discovering the phenomenon of "anticyclones" in the process. He even created a "beauty map" of Britain by walking through different cities and making secret records of the attractiveness of women he passed by pricking holes in a piece of paper attached to a work glove.

Galton's primary interest wasn't statistics, but heredity. He was very interested in knowing to what extent people's characteristics are determined by what they inherited from their parents. His fondest wish was to study intelligence, but since this was notoriously difficult to measure, he decided instead to study height. If he could determine the extent to which height is inherited, he reasoned, the same might hold for other characteristics of people.

In this project, we will explore the original work in which Galton described regression to the mean [Galton, 1886], and try to understand why it occurs.

*Department of Mathematics, Central Washington University, Ellensburg, WA 98926; dominic.klyve@cwu.edu.

[1] In fact, the two meanings of the word "regression" aren't as different as they first appear, as we shall see in this project.

2 Heights of children, heights of parents

Galton introduced the primary motivation for his work on the first page of his report.

> It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did not tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small.

Task 1 Explain in your own words what Galton had discovered.

Task 2 Give a possible explanation for why the offspring of tall plants are usually shorter than their parents.

Galton was interested in questions concerning the laws of heredity in general, and he was especially interested in human heredity. He thus worked quite hard to gather data from 930 adult children, from 205 sets of parents.[2]

After a careful study of the data (which we will examine below), Galton discovered that not only did the same relationship he had observed between the height of plants and their offspring exist in humans, but that the tendency to be more mediocre than one's parent followed a fairly precise mathematical pattern.

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

[2] The mean number of children per family seems large, and this is not an accident. Galton thought he would get better data if he gathered information only from families with many kids, which might help to smooth out unusual measurements.

Task 3 Let's first try to understand Galton's claim. Assume that the height of the average adult in his study was 69 inches. If a parent were 75 inches tall, how tall would Galton predict the child to be? What if the parent were 66 inches tall?

Task 4 Using this $\frac{2}{3}$ value, write an equation that predicts the height of an adult child, C, in terms of the height of the parent, P. Make sure the values your equation gives match your answers in the previous task.

Task 5 Children seem to be, on average, more "mediocre" (in terms of their height) than their parents; what if we turn the question around? If an adult child is 72 inches tall, how tall would you expect the parent to be? Why?

Before he could do much analysis, Galton needed to find a good way to combine the heights of both an individual's parents into one value. In order to avoid breaking his data into two sets (for males and females), he also needed to find a reasonable way to compare the heights of males (who are taller on average) to the heights of females.

My data consisted of the heights of 930 adult children and of their respective parentages, 205 in number. In every case I transmuted the female statures to their corresponding male equivalents and used them in their transmuted form, so that no objection grounded on the sexual difference of stature need be raised when I speak of averages. The factor I used was 1.08, which is equivalent to adding a little less than one-twelfth to each female height. It differs a very little from the factors employed by other anthropologists, who, moreover, differ a trifle between themselves.

Task 6 Explain what Galton did to his data values in order to account for the different average height of males and females.

Task 7 Do you think Galton's method of doing this was reasonable? Why or why not?

In order to discuss the height of two parents as one value, Galton came up with the idea of the height of the "mid-parent."

> [W]hen dealing with the transmission of stature from parents to children, the average height of the two parents, or, as I prefer to call it, the "mid-parental" height, is all we need care to know about them.

Task 8 Galton didn't specify whether he was taking the mean or the median height of the two parents. Why?

Galton summarized all of his data in one table (Table I, next page), and most of the remainder of his essay (and of this project) is devoted to describing and analyzing it. Take a minute and examine the Table before reading on.

Galton explained his table this way:

> The meaning of the table is best understood by examples. Thus, out of a total of 928 children who were born to the 205 mid-parents on my list, there were 18 of the height of 69.2 inches (counting to the nearest inch), who were born to mid-parents of the height of 70.5 inches (also counting to the nearest inch). So again there were 25 children of 70.2 inches born to mid-parents of 69.5 inches.

Task 9 Write down three things you notice about the data or the values in this table.

Task 10 Write down at least one question that you have about the table.

One nice property of this table is that we can see the average height of adult children for each different parental height, and we can also do the opposite—for all children of a given height, we can see the average height of their parents.

TABLE I.

Number of Adult Children of various statures born of 205 Mid-parents of various statures.

(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

Note.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Task 11 In Task 5 you tried to the predict the height of a parent, given that the child's height is 72 inches. Use Table I to find the actual value. Is it what you predicted? Why do you think this is?

Galton addressed this question directly. In reference to his "Law of Regression" (that adult children will deviate only $\frac{2}{3}$ as much as their parents from mediocre height), he wrote

> The converse of this law is very far from being its numerical opposite. Because the most probable deviate of the son is only two-thirds that of his mid-parentage, it does not in the least follow that the most probable deviate of the mid-parentage is $\frac{3}{2}$, or $1\frac{1}{2}$ that of the son. . . . It appears from the very same table of observations by which the value of the filial regression was determined when it is read in a different way, namely, in vertical columns instead of in horizontal lines, that the most probable mid-parentage of a man is one that deviates only one-third as much as the man does.

Task 12 Look again at the table. Is it true that the average mid-parent height of children of a given height is more "mediocre" than that of the children?

Task 13 Give a possible explanation for why this seemingly-contradictory result might hold.

The conclusion is this: children of very tall parents are usually tall, but are shorter than their parents. But if we randomly choose a tall child, it's likely that the parents are shorter than the child! Galton himself tried to explain this counter-intuitive result.

> The number of individuals in a population who differ little from mediocrity is so preponderant that it is more frequently the case that an exceptional man is the somewhat exceptional son of rather mediocre parents, than the average son of very exceptional parents.

Now let's try to make sense of Galton's explanation.

Task 14 Look at all the adult children in Table 1 who are 70.2 inches tall.

a. How many such children are there?
b. How many of them have parents that are even taller than they?
c. What property of tall parents might lead to the fact that not many of these children were born to such parents?

Task 15 Now rewrite Galton's explanation in your own words.

Task 16 Check whether your textbook has a section on regression to the mean. If it does, how does its description compare to the one you gave above?

Today the phenomenon that unusual values of almost any random variable are generally followed (and preceded) by more average values is still referred to as "regression to the mean."

3 Regression to the Mean in the Modern World

Now that we have some understanding of this phenomenon, we can sometimes spot errors made by people who haven't yet learned it. In 2016, an article by Jeffery Linder appeared in a medical journal entitled *Primary Care Respiratory Medicine*. This article [Linder, 2016] re-told a story about the dangers of not understanding regression to the mean as follows:

> Giving effective feedback is hard. In *Thinking Fast and Slow* [Kahneman, 2011], Daniel Kahneman describes a time when he was giving a lecture to Israeli fighter pilots about effective training practices. During his talk, Kahneman discussed the well-supported concept that rewarding good performance is more effective than punishing poor performance. After the talk, an incredulous senior instructor confronted Kahneman and said that criticising his trainee pilots for poor execution of aerial manoeuvers worked. The instructor had noticed that when he criticised trainees after they poorly executed a manoeuver, they almost always improved on their next attempt.

Task 17 Let's assume that the instructor was correct, and that trainees who poorly executed a manoeuver almost always improved on their next attempt if they were criticized. Is it reasonable to conclude that the criticism helped? Why or why not?

Linder summarized Kahneman's next step as follows:

> Kahneman saw a teaching opportunity. He drew a chalk target on the floor, had the officers turn their backs to the target and try to hit the target with two successive no-look coin throws. Those familiar with regression to the mean can guess what happened: officers who were far from the target on their first throw generally improved on the second throw; officers who were close to the target on their first throw generally did worse on their second throw. Kahneman showed how easy it is to conflate 'effective feedback'—either positive or negative—with regression to the mean.

Task 18 Suppose a basketball player has one of the best games of her season. What would you expect her performance to be in the next game if her coach praises her? What if her coach criticizes her?

Task 19 If her next game is quite average, what would you say to a sports commentator who claimed that she lost her confidence after her successful run?

Task 20 Give one (or ideally, several) examples from your life in which you could observe regression to the mean. If someone who didn't understand this phenomenon were to observe this regression, what might they wrongly conclude?

Task 21 The journal mentioned above is devoted to doctors who specialize in respiratory medicine. Suggest one or more reasons that these doctors should be aware of the phenomenon of regression.

References

Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Nicholas W Gillham. Sir Francis Galton and the birth of eugenics. *Annual review of genetics*, 35(1): 83–101, 2001.

Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

Jeffrey A Linder. Moving the mean with feedback: insights from behavioural science. *Primary Care Respiratory Medicine*, 26, 2016. doi: 10.1038/npjpcrm.2016.18.

Stephen M Stigler. *The seven pillars of statistical wisdom*. Harvard University Press, 2016.

Instructor Notes for "Regression to the Mean"

Dominic Klyve[3]

November 16, 2019

This set of notes accompanies the Primary Source Project "Regression to the Mean" written as part of the TRIUMPHS project (see end of notes for details about TRIUMPHS).

PSP Content: Topics and Goals

Regression to the mean—the phenomenon that given a series of measurements, unusual values are usually followed by more average values—seems not to be a standard topic in many modern statistics classes. Yet the phenomenon is statistical in nature, and is certainly important in understanding parts of the modern world. Moreover, the concept is easy to understand, and a reasonable explanation exists for why it occurs. This mini-Primary Source Project has the goal of introducing students to the concept and explaining why it occurs via the work of Francis Galton, who first described (and named) the phenomenon.

Useful Background

Francis Galton pioneered the study of inheritance, and the tools that he developed to track changes in traits between generations helped to usher in modern statistics. It's worth noting that his work is somewhat controversial today. While rather progressive in his day, Galton's work on intelligence led to some very racist work in psychology purporting to demonstrate racial differences in intelligence. Moreover, some of the ideals of the Nazi party in Germany (and elsewhere) can be traced to him, as the person who coined the word "eugenics." This project does not touch on any of these topics, but instructors should be aware of these issues, in case a student asks about them (see [Gillham, 2001] for more about Galton and eugenics).

Student Prerequisites

This project has no formal prerequisites. Indeed, it offers an introduction to regression to the mean written by a scientist who assumed his readers would not have mathematical or statistical training beyond basic algebra.

Commentary on PSP Design and Individual Tasks

This PSP has only three sections: a one-page introduction, a five-page section exploring Galton's work, and an optional one-page section that explores the application of that work to understanding certain modern reasoning errors. It is designed to be completed in about one 50-minute class day.

[3]Department of Mathematics, Central Washington University, Ellensburg, WA 98926, dominic.klyve@cwu.edu.

- Task 3 has an easy solution. Since, in the first case, the 75-inch parent is 6 inches taller than average, their child is expected to be $\frac{2}{3}(6) = 4$ inches taller than average, or 73 inches. A student without strong quantitative skills, however, may try to set up and solve this problem in a more complicated way. Instructors should consider whether they have a preference for how their students approach this problem and act accordingly.
- Task 5 looks fairly trivial, and students will likely have no problem solving it. However, it is quite important. The "obvious" solution turns out not to be correct—the linear equation that predicts children's height as a function of their parents' height cannot be used to make predictions in the opposite direction. This paradox is central to Galton's work, as students will discover in later tasks.
- Tasks 6–7, on the other hand, are not important for the underlying statistics. They exist only to make interpreting Galton's tables easier later.

Suggestions for Classroom Implementation

Because regression to the mean is fairly independent of most topics taught in a first (or second!) statistics course, I give this project to students after a discussion of sampling distributions, but before hypothesis testing. Indeed, the project assumes that students have not seen hypothesis testing formally described.

The PSP includes several open-ended discussion questions, and lends itself well to group work. I suggest assigning groups of three students (or letting students choose their own, as your classroom culture warrants). The schedule given below is based on a 50-minute class period.

Sample Implementation Schedule

- **Day 0:** Introduce the project. Assign the Introduction and the first part of Section 2 (including Tasks 1–3) as homework. This will allow students to do the longest part of the reading at their own pace, saving the shorter readings and more active tasks for group work time.
- **Day 1:** Students can work through all (or almost all) of Section 2 of the PSP in a 50-minute class period.
- **Optional Day 2: Instructors who wish to do so could spend 15–25 minutes on whole-class or small-group discussion of select tasks in Section 3.**

  Homework: Assign formally-written solutions to all tasks.

Other Recommendations for Further Reading

A marvelous description of Galton's paper, together with useful historical context, can be found in Stephen Stigler's *Seven Pillars of Statistical Wisdom* [Stigler, 2016]. (In fact, I strongly recommend this entire book to any teacher of statistics. It's short but packed with powerful insight into statistics.) Instructors looking for a variety of useful suggestions for teaching statistics, including the incorporation of projects into the curriculum, are also encouraged to consult the 2016 GAISE report:

> R Carver, M Everson, J Gabrosek, N Horton, R Lock, M Mocko, and B Wood. *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016.* American Statistical Association. Available at `www.amstat.org/education/gaise`.

Acknowledgments

The development of this student project has been partially supported by the TRansforming Instruction in Undergraduate Mathematics via Primary Historical Sources (TRIUMPHS) Program with funding from the National Science Foundation's Improving Undergraduate STEM Education Program under grant number 1524098. Any opinions, findings, and conclusions or recommendations expressed in this project are those of the author and do not necessarily represent the views of the National Science Foundation. For more information about TRIUMPHS, visit `https://blogs.ursinus.edu/triumphs/`.