Francis Galton and regression to the mean

Galton founded many concepts in statistics, among them correlation, quartile, and percentile. Here **Stephen Senn** examines one of Galton's most important statistical legacies – one that is at once so trivial that it is blindingly obvious, and so deep that many scientists spend their whole career being fooled by it.

Galton was born into a wealthy family. The youngest of nine children, he appears to have been a precocious child – in support of which his biographer cites the following letter from young Galton, dated February 15th, 1827, to one of his sisters:

> My dear Adèle,
>
> I am four years old and can read any English book. I can say all the Latin Substantives and adjectives and active words besides 52 lines of Latin poetry. I can cast up any sum in addition and multiply by
>
> 2,3,4,5,6,7,8,(9),10,(11)
>
> I can also say the pence table, I read French a little and I know the clock.
>
> Francis Galton[1]

Apparently Galton was also a truthful child, since, having written the letter, he had realised that what he had claimed about the numbers 9 and 11 was not quite true and had tried to obliterate them. And before you get too impressed, his birthday was February 16th so he was very nearly five!

Galton's later progress in education was not quite so smooth. He dabbled in medicine and then read mathematics at Cambridge, but eventually had to take a pass degree. In fact he subsequently freely acknowledged his weakness in formal mathematics, but this weakness was compensated by an exceptional ability to understand the meaning of data. Galton was a brilliant natural statistician.

Many words in our statistical lexicon were coined by Galton. For example, *correlation* and *deviate* are due to him, as is *regression*, and he was the originator of terms and concepts such as *quartile*, *decile* and *percentile*, and of the use of *median* as the midpoint of a distribution[2]. Of course, words have a way of developing a life of their own, so that, unfortunately *decile* is increasingly being applied to mean *tenth*. There are, pretty obviously, ten tenths of a distribution, but there are, slightly less obviously, only nine deciles, since the deciles are the boundaries between the tenths. To use *decile* to mean *tenth* – as when, for example, speaking of students "in the top decile" (according to their examination marks) – is not only pompous but also wrong and means that yet another word will eventually have to be invented to perform the function that Galton created *decile* to fulfil.

Galton compared the height of children to that of their parents. He found that adult children are closer to average height than their parents are.

To take another example, we no longer use the term *regression* in quite the way Galton did. We now usually reserve it for the fitting of linear relationships. In Galton's usage regression was a phenomenon of bivariate distributions – those involving two variables - and something he discovered through his studies of heritability. However, the use of regression in Galton's sense does survive in the phrase *regression to the mean* – a powerful phenomenon it is the purpose of this article to explain.

Galton first noticed it in connection with his genetic study of the size of seeds, but it is perhaps his 1886 study of human height[3] that really caught the Victorian imagination. Galton had compared the height of adult children to the heights of their parents.

For this purpose he had multiplied the heights of female children by 1.08. For as he put it:

> 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.

It is interesting to note, incidentally, that he also considered whether 1.07 or 1.09 might not be a better factor to use, but remarked:

> The final result is not of a kind to be affected by these minute details, for it happened that, owing to a mistaken direction, the computer to whom I first entrusted the figures used a somewhat

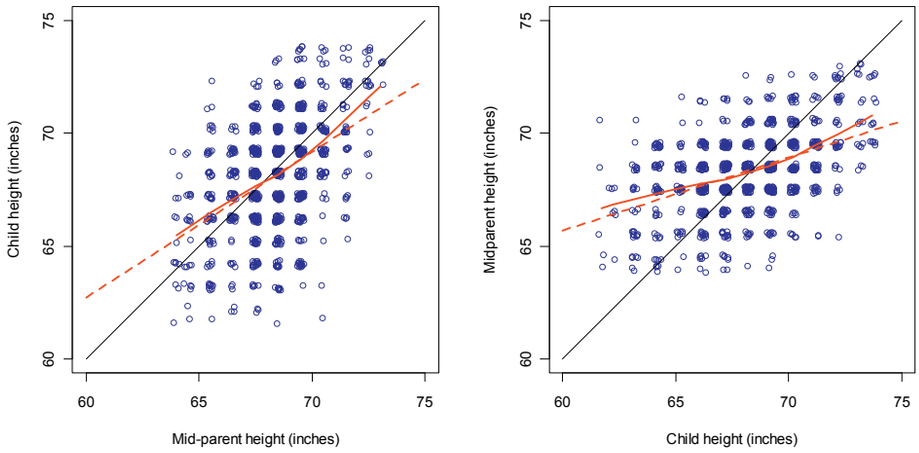

Figure 1. Galton's height data: two scatterplots showing the regression phenomenon (drawn from data listed at http://www.math.uah.edu/stat/data/Galton.txt)

different factor, yet the result came out closely the same.

The year being 1886 the computer in question was, of course, a human and not an electronic assistant! The more interesting point, however, is that Galton is describing what we would now call robustness in statistics – and, simultaneously, provides an early example of what is now recognised as a general scientific phenomenon: scientists never seem to fail the robustness checks they report. It is interesting to speculate why.

Galton's data consisted of 928 adult children and 205 "parentages" – that is to say, father-and-mother couples. (The mean number of children per couple was thus just over 4.5 – families were larger in those days.) He represented the height of parents using a single statistic, the "mid-parent", this being the mean of the height of the father and of his wife's height multiplied by 1.08. Of course, as previously noted, for the female children the heights were also multiplied by 1.08. For the male children they were unadjusted.

Figure 1 is a modern graphical representation of Galton's data. Galton had grouped his results by intervals of 1 inch, and in consequence, if a given child's recorded height were plotted against its recorded mid-parent height, many points would be superimposed on top of each other. I have added a small amount of "jitter" in either dimension to separate the points, which are shown in blue. The data are plotted in two ways: child against mid-parent on the left and mid-parent against child on the right. The thin solid black diagonal line in each case is the line of equality. If a point lies on this line then child and mid-parent were identical in height. Also shown in red in each case are two different approaches one might use to predicting "output" from "input". The dashed line is the least squares fit, what we now (thanks to Galton) call a *regression line*. The thick red line is a more local fit, in fact a so-called *LOWESS* (or locally weighted scatterplot smoothing) line. The point about either of these two approaches – irrespective of whether we predict child from mid-parent or vice versa – is that the line that is produced is less steep than the line of exact equality. The consequence is that we may expect that an adult child is closer to average height than its parents – but also, paradoxically, that parents are closer to average height than is their child.

The first part we might expect. The second may seem absurd – but is just as true. I will say it again: a tall child will have parents, on average, less tall than himself. This particular point is both deep and trivial. It is deep because the first time that students encounter it (I can still remember my own reaction) they assume that it is wrong; its truth is well hidden. Once understood, however, it becomes so obvious that one is amazed at how regularly it is overlooked. It is a point not about genetics but about statistics.

In fact I am confident that at this stage I can divide my readers into two: those who will claim that I am wasting their time repeating a hackneyed truth, and those who will say that I have so far failed in anything that I have said to show that the hackney in question is a genuine carriage. So I will say farewell to the members of the first group and address myself to the second – but before I say goodbye to the first I will ask them one question. Do you think that there is good evidence that the placebo effect is genuine? If so, stick around for a while because I will try and show you that you (and ten thousand physicians with you) are wrong. What this has to do with Francis Galton will be revealed in due course.

So let us leave Francis Galton for the moment and consider another example, this time a simulated one. Figure 2 shows simulated values in diastolic blood pressure (DBP) for a group of 1000 individuals measured on two occasions: at baseline and at outcome. ("Outcome" simply means "some time later"; they have not received any medical treatment between the two occasions.) If your blood pressure is high, you are *hypertensive*. Using a common but arbitrary definition of hypertension as a diastolic pressure of 95 mmHg or more, the subjects have been classified as consistently hypertensive (red diamonds) consistently normotensive (blue circles) or inconsistent – hypertensive on one occasion, normal on the other (orange stars). The distributions at outcome and baseline are very similar, with means close to 90 mmHg and a spread that can be defined by that Galtonian statistic, the *inter-quartile range*, as being close to 11 mmHg on either occasion. In other words, what the picture shows is a population that all in all has not changed over time although, since the *correlation* (to use another Galtonian term) is just under 0.8 and therefore less than 1, there is, of course, variability over time for many individuals. Some have increased their blood pressure between the measurements, some have reduced it.

However, in the setting of many clinical trials, Figure 2 is not a figure we would see. The reason is simple: we would not follow up individuals who were observed to be normotensive at baseline. If you are "healthy", we would not bother to call you back for the second test. In-

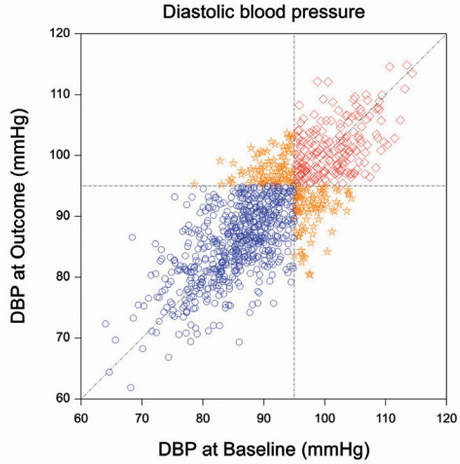

Figure 2. Simulated diastolic blood pressure for 1000 patients measured on two occasions – blue circles, normotensive on both occasions; red diamonds, hypertensive on both occasions; orange stars, inconsistent

stead what doctors and medics see is the picture given in Figure 3. Of the 1000 subjects seen at baseline, 285 had DBP values in excess of 95 mmHg. We did not bother to call the other 715 back; we concentrated instead on those we deemed to have a medical problem – and those are the only ones shown in the figure. And if now we compare the outcome values of those 285 subjects we have left to the values they showed at baseline, we will find that mean DBP seems to have gone down. At outcome it is more than 2 mmHg lower than it was at baseline. What we have just observed is what Francis Galton called *regression to the mean*.

There has been an apparent spontaneous improvement in blood pressure. Apparently many patients who were hypertensive at baseline became normotensive. It is important to understand here that this observed "improvement" is a consequence of this stupid (but very common) way of looking at the data. It arises because of the way we select the values. What is missing because of our selection method is bad news. We can only see patients who remain hypertensive or who become normotensive. We left out the patients who were normotensive but became hypertensive. They are shown in Figure 4. If we had their data they would correct the misleading picture in Figure 3, but the way we have gone about our study means that we will not see their outcome values.

Regression to the mean is a consequence of the observation that, on average, extremes do not survive. In our height example, extremely tall parents tend to have children who are taller than average and extremely small parents tend to have children who are smaller than average, but in both cases the children tend to be closer to the average than were their parents. If that were not the case the distribution of height would have to get wider over time. Of course there can be changes in such distributions over time and it is the case that people are taller now than in Galton's day, but this is a separate phenomenon in addition to regression to the mean.

However, regression to the mean is not restricted to height nor even to genetics. It can occur anywhere where repeated measurements are taken.

Does it happen that scientists get fooled by Galton's regression to the mean? All the time! Right this moment all over the world in dozens of disciplines, scientists are fooling themselves either by not having a control group, which would also show the regression effect, or, if they do have a control group, by concentrating on the differences within groups between outcome and baseline rather than the differences between groups at outcome. It is regression to the mean that is a very plausible explanation for the placebo effect, since entry into clinical trials is usually only by virtue of an extreme baseline value. This does not matter as long as you compare the treated group to the placebo group, since both groups will regress to the mean. It does mean, however, that you have to be very careful before claiming that any improvement *in the placebo group* is due to the healing hands of the physician or psychological expectancy.

To prove *that* would require a three-arm trial: an active group, a placebo group and a group given nothing at all. Then all three groups would have the same regression to the mean improvement and differences between the placebo and the open arm could be judged to be due to a true placebo effect. Not surprisingly, very few such trials have been run. However, analysis of those that have been run suggests that only in the area of pain control do we have reliable evidence of a placebo effect[4,5].

But regression to the mean is not just limited to clinical trials. Did you choose dangerous road intersections in your region for corrective engineering work based on their record of traffic accidents? Did you fail to have a control group of similar black spots that went untreated? Are you going to judge efficacy of your intervention by comparing before and after? Then you should know that Francis Galton's regression to the mean predicts that sacrificing a chicken on such black spots can be shown to be effective by the methods you have chosen[6]. Did you give failing students a remedial class and did they improve again when tested? Are you sure that subsequence means consequence? What have you overlooked?

Do scientists get fooled by Galton's regression to the mean? All the time!

A Victorian eccentric who died 100 years ago, although no great shakes as a mathematician, made an important discovery of a phenomenon that is so trivial that all should be capable of learning it and so deep that many scientists spend their whole career being fooled by it.

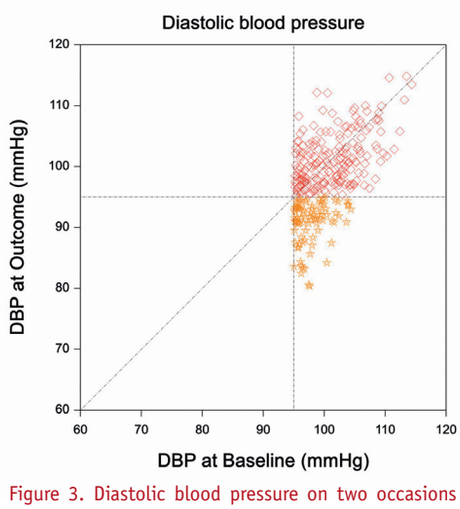

Figure 3. Diastolic blood pressure on two occasions for patients observed to be hypertensive at baseline

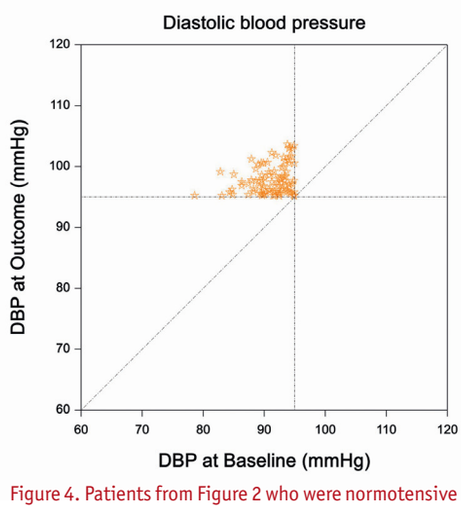

Figure 4. Patients from Figure 2 who were normotensive at baseline but hypertensive at outcome

References

1. Forrest, D.W. (1974) *Francis Galton: The Life and Work of a Victorian Genius*. London: Paul Elek.
2. David, H.A. (1995) First (questionable) occurrence of common terms in mathematical statistics. *American Statistician*, **49**, 121–133.
3. Galton, F. (1886) Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246–263.
4. Hrobjartsson, A. and Gotzsche, P.C. (2001) Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine*, **344**, 1594–1602.
5. Kienle, G.S. and Kiene, H. (1997) The powerful placebo effect: fact or fiction? *Journal of Clinical Epidemiology*, **50**, 1311–1318.
6. Senn, S.J. (2003) *Dicing with Death*. Cambridge: Cambridge University Press.

Stephen Senn is a professor of statistics at Glasgow University. His book *Dicing with Death* is published by Cambridge University Press and covers (among other matters) Francis Galton and the history of regression to the mean.