Understanding regression to the mean

Jorma Jormakka

Regression to the mean is a phenomenon that has to happen in case the probability distribution of any property, like for instance IQ, stays stable when a new generation arrives. (If some new environmental factors, like the invention of writing or book, appear, the probability distribution need to stay stable, but usually the distribution is stable. If genes are not changing in the population and nothing much is happening, the distribution must be stable over generations.)

Let us take the normal distribution. It has the probability distribution function

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}x^2}$$

where x denotes the property. Let this property be the IQ score. The value $p(x)$ gives the fraction of the population having the score x (but you have to integrate e.g. over scores $\le x$ to get a meaningful number. You might imagine that if two people with the IQ score x make children, the IQ scores y of the children would be given by the normal distribution

$$p(y)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(y-x)^2}$$

but unfortunately this is not possible. If this were so, then the IQ score distribution of the next generation would be

$$p'(y)=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}x^2}\,\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(y-x)^2}dx=\frac{1}{\sqrt{2\pi}\sqrt{2}\sigma}e^{-\frac{1}{2(\sqrt{2}\sigma)^2}y^2}\ .$$

That is a normal distribution, but the variance has grown to twice the size. If we want that the next generation has the distribution

$$p'(y)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}y^2}$$

i.e, the same as the parent generation, we must assume that the fraction of people of intelligence score x have children with the probability

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{1}{2\sigma'^2}x^2}\quad\text{where }\sigma'\le\sigma$$

that is, high IQ people have fewer children than their numbers would indicate, and that the IQ distribution of the children is

$$p(y)=\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{1}{2\sigma'^2}(y-bx)^2}\quad\text{where } b=\sqrt{\frac{\sigma^2}{\sigma'^2}-1}\,.$$

The average IQ score of the children of a person with the score x is clearly bx from this new distribution $p(y)$. This solution does give you the correct stable distribution

$$p'(y)=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{1}{2\sigma'^2}x^2}\,\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{1}{2\sigma'^2}(y-bx)^2}dx=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}y^2}\ .$$

The proof is not difficult but I will not write it here. Basically, find a solution such that $\int p(x)dx=1$ and $\int p(y)dy=1$ and $p(x)$ does not depend on y, and the functions are normal.

Notice that if $b=1$, then $\sigma'=\sigma/\sqrt{2}$. We cannot set $b=1$ and $\sigma'=\sigma$ in a stable distribution. The number bx (here $0\le b\le 1$) is the regression to the mean and it comes from the children inheriting only a part of the intelligence of their parents. This part includes the genetic part, but also has some environmental part. The standard deviation $\sigma' \leq \sigma$ is smaller than the typical 15, because children are slightly more similar in their IQ to parents than the general population.

We can estimate σ' from IQ measurements. I found once one paper where the average IQ of fathers was 139=2.6σ in the highest class and the IQ of their sons was 120=1.333σ. This gives $b = bx / x = 1.333 / 2.6$ and solving σ' gives

$$\sigma' = \sigma\left(\sqrt{b^2 + 1}\right)^{-1} = 13.35 .$$

I am not so happy with phenomenological formulae, but there is a phenomenological (heuristic) formula for regression to the mean: the breeder's equation. In this equation you estimate the average score for children by using heritability. Thus, if heritability H is e.g, 0.6, then you estimate that if parent's IQ scores were 39 points above the population average, then the children average is 0.6*39=23 points above the population average. In fact, $b = H$, but I prefer to motivate it with the stability of the distribution rather than by heritability.

What the breeder's equation does not tell you is that the probability distribution is stable only if higher IQ people make fewer children. Thus, there is no reason to be worried of this situation, which quite commonly prevails, and must prevail.