Regression toward the mean: a fresh look at an old story

Back in time, when I took a statistics course from Professor G., I encountered regression toward the mean for the first time.[a] I did not understand the topic, so I raised my hand in class and asked whether the phenomenon is telling us about reality or it is just math. Professor G. seemed puzzled for a moment and then gave an answer which I don't recall. Presumably, its content was not memorable enough.

From time to time I returned to this interesting topic but have not grasped it fully, until my collaborator (my son) cleared up for me key parts of the story several years ago. More recently, I read two more relevant articles[1,2] and decided it was time to make it as clear as possible in my mind and maybe in your mind, too. In particular, I would like to connect the explanation of regression toward the mean with causality, indeterminism, and measurement – three good sources of confusion.

Mr. Smith and indeterminism

Let X_1 be a variable at time 1, say weight, and let's consider Mr. Smith whose weight happened to be 100kg at that time. To understand regression toward the mean, we first need to assume that Smith's weight at time 1 is a variable whose values are distributed about an unknown mean.

On first thought, that assumption might seem strange. Where is this distribution coming from? Why are there other possible values of Smith's weight at time 1 besides 100kg? What exactly is the meaning of this set of unobserved values?

Indeterministic causation offers a clear answer. Smith's actual weight at time 1 (X_1=100) is one realization of all the weights Smith could have had at time 1 – given the causes of weight. Although other weights remain theoretical, each of them could have been realized with a certain probability.

According to indeterminism, the realized values of causal variables behind X_1 create a tendency of X_1 to take each of its values.[3] For instance, if the causes of X_1 are variables *A*, *B*, and *C* (Figure 1) and their realized values were "aa", 121, and 0.7, respectively, we may write a tendency function:[3]

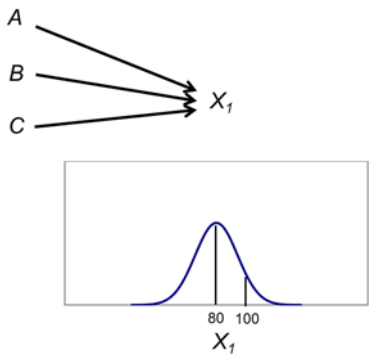
$$T(X_1=x_1 \mid A=\text{aa}\ ,\ B=121,\ C=0.7)$$

The various tendencies for all possible values of X_1 – conditional on *A*=aa , *B*=121, *C*=0.7 – translate to a probability distribution, a probability density function (PDF) of X_1 (Figure 1). The mean of that distribution is by no means the "true weight" of Smith at time 1; it is just the expected value of the distribution, E(X_1), say 80kg. For instance, the probability that Smith's weight would fall in the range 60kg-100kg might have been 70%, whereas the probability of it falling in the range >100kg might have been 15%. Formally speaking, Smith's weight at time 1 (100kg) is realization of a random variable, X_1, given the realized values of its causes (which are also random variables.)

A thought experiment (on reality)

Next, let's play a thought experiment. Suppose we were able to replay once the causes of Smith's weight such that all of them took exactly the same value they actually took before time 1. What value will Smith's weight take in the replay? Since indeterminism (probabilistic causation) is involved, it will likely be a different value, not 100kg. And although we have no idea which value it would be, we can deduce the following: *values closer to the mean than 100 (between 80 and 100) are more likely to be taken than values farther from the mean than 100.* Just look at the bell-shaped PDF. It is telling us exactly what I wrote in italics: the area between 80kg and 100kg is much larger (higher probability) than the tail area to the right of 100kg (lower probability).

Figure 1.



[a] The term "regression *to* the mean", used by many writers, creates the false impression that something has regressed "all the way" to the mean value.

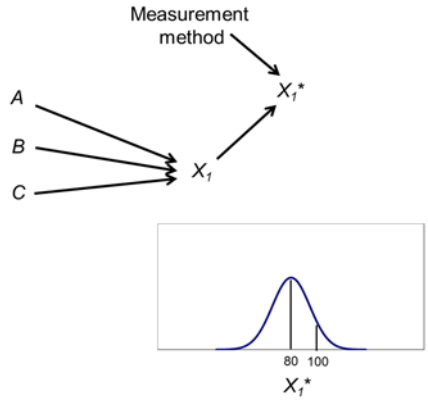
A thought experiment (on measurement)

In the empirical world we never truly know Smith's realized weight. All that we record is some measurement of X_1. Therefore, we need to consider X_1*, the measured variable, whose proximal, typical causes are X_1 and the measurement method (Figure 2). Different methods (scales, conditions, etc.) are expected to yield different values of X_1*.

In Figure 2, the value 100kg is the value of X_1*, and the PDF describes the probability distribution of X_1*, not X_1. Accordingly, our thought experiment of the replay should consider X_1 retaking its (unknown) value *and* exact replication of the measurement method.

This is still a thought experiment, though. We can't be certain that any repeated measurement is indeed an *exact* replication. Things might have changed – slighly or not slightly: the calibration, the temperature, the weight itself… Keep in mind that exact replication is a theoretical idea that is never synonymous with actual replication.

Figure 2.



Indeterministic causation, which easily explained the PDF of X_1, readily accommodates the PDF of X_1*, too (Figure 2). The measurement method is not conceptually different from X_1, a proximal cause of X_1*. It is just another causal variable whose value (the actual method) shows up in the tendency function of X_1* along with the realized value of X_1:

$$T(X_1^*=100\text{kg} \mid X_1 = x_1, \textit{Measurement Method}=m)$$

Preferring determinism?

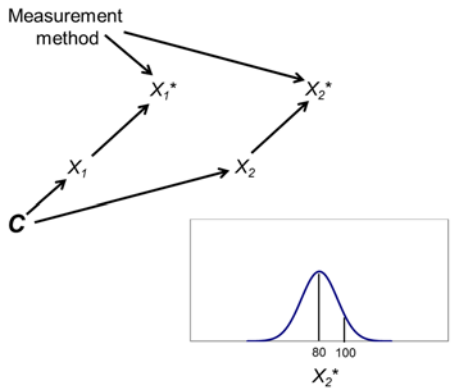
Deterministic causation encounters a problem with both Figure 1 and Figure 2 because tendencies and "probabilistic realization" do not exist in a deterministic universe. According to determinism, our thought experiments should always generate the same value of Smith's weight: 100kg. No PDF is in sight (other than bizarre posterior Bayesian probabilities).

So how do authors who cling to determinism (knowingly or unknowingly) handle the conundrum? How do they explain the origin of the PDF which is essential to explaining regression toward the mean? A little deception. The first thought experiment (Figure 1) is simply ignored as the story is completely theoretical; the true realized weight remains unknown forever. To save the second thought experiment, they offer a supplement called "random measurement error", which supposedly explains the origin of the PDF of X_1*. There is one problem, however: randomness has no logical place in deterministic causation. Random error belongs in determinism about as well as an electrical outlet belongs in water.

A second time point

Let's set aside the ailments of determinism, and return to Mr. Smith at some later time point, time 2 (Figure 3).

Figure 3.



To simplify, I depicted ***C*** (vector notation) as a substitute for all the causes of X_1 and X_2 other than the measurement method. I also made an important causal assumption – no arrow from X_1 to X_2 – which will be discussed later. What can we say about the likely value of X_2* for Mr. Smith?

If the values of variables in ***C*** and the measurement method have not changed, the variables X_1* and X_2* are conditionally independent and identically distributed. Therefore, the PDF of X_2* should be identical to the PDF of X_1*, and "sampling" from the

PDF at time 2 (Figure 3) is equivalent to "re-sampling" from the PDF at time 1 (Figure 2).

Recall the measurement of Smith's weight at time 1 (X_1*=100kg). What would you expect to observe if you were able to re-measure Smith's weight at the same time? We already answered that question in a thought experiment: a value closer to the mean. And what would you expect to observe when you actually measure his weight again at time 2? Exactly the same thing. It's the same PDF at both times! *Values closer to the mean than 100 (between 80 and 100) are more likely to be observed at time 2 than values farther from the mean than 100*. Since Smith's measured weight at time 1 was 100kg (much larger than a mean of 80kg), his measured weight at time 2 is likely to be smaller than 100kg. That's regression toward the mean in the empirical world.

And in general: if you observe someone whose value of X_1* is "far enough" from the mean of X_1* in either direction, his value of X_2*, the second measurement, is likely to be closer to that mean.[b] Or the other way around: if you observe someone whose value of X_2* is "far enough" from the mean of X_2*, his value of X_1*, the first measurement, is likely to have been closer to that mean. In each case you are simply sampling again *from the same PDF* (same mean; same distribution). To sum up: if the time point variables are conditionally independent, and the PDF is bell-shaped, and you happen to hit a tail value at one time – you are likely to hit a closer-to-the-mean value at another time.

In the last two paragraphs I wrote "likely" repeatedly, not "certainly", and I assumed that neither the values of ***C*** nor the measurement method have changed (implying conditional independence). If that's not the case – say, Smith added 1,000 calories to his daily diet in between – then Smith's weight at time 2 may no longer be considered as "sampled" from his PDF at time 1. Having lost the theoretical basis of regression toward the mean, we can no longer deduce that his weight at time 2 is likely to be smaller than 100kg. I am sure you would have agreed even before reading about regression toward the mean…

What about $X_1 \rightarrow X_2$?

No one knows whether causation is deterministic or indeterministic; it is an axiomatic choice for every methodologist – with implications.[c] Likewise, no one knows how causation works, but we have to start with a set of so-called assumptions, which are actually axioms. One such axiom states that $X_1 \rightarrow X_2$: a natural variable at one time is a cause of that variable at any future time.[4] Rejection of the axiom, although certainly permissible, carries undesirable consequences with respect to continuity properties.

Notice that the axiom does not claim anything about the magnitude of the effect, which might depend on the variable and the time interval. In some cases the effect may be extremely strong, say, time 2 is very close to time 1, whereas in others it may be close to null. Regardless, we ought to add the arrow $X_1 \rightarrow X_2$ (Figure 4).

Figure 4.

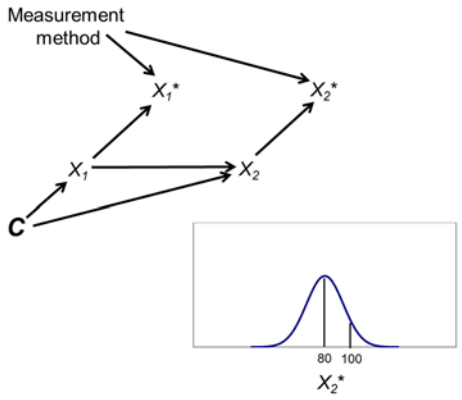


Figure 4 reveals that X_1 and X_2 are never fully conditionally independent because $X_1 \rightarrow X_2$. Their PDFs may be "very similar", but not identical. Strictly speaking, the condition for regression toward the mean does not hold for these unknown variables.

It does not strictly hold for their measured counterparts, X_1* and X_2*, either. The arrow $X_1 \rightarrow X_2$ also account for dependence between the measured variables through the path $X_1^* \leftarrow X_1 \rightarrow X_2 \rightarrow X_2^*$. Notice that this path is not blocked in our thought experiments about X_1* and X_2* because the realized values of X_1 and X_2 are likely different when $X_1 \rightarrow X_2$ non-deterministically. Might the effect $X_1 \rightarrow X_2$ be strong to the point of identical values? Could X_1 and X_2 share the same PDF because they are perfectly correlated? Such suggestions of exactness try to sneak in deterministic causation through the back door. Remember: If you let them in, there is no story to tell about the origin of the PDFs.

Are we far enough already?

Writers on regression toward the mean usually use terms such as "far from the mean", "tail

[b] "his" is used generically instead of "her", "him/her", "their", or other variations.

[c] Interestingly, many deterministic writers ignore problematic epistemological implications of their choice.

observations“, “extremely low”, “unusually high”, and the like. That makes life deceptively easy, so let’s make it more difficult.

Look at the next series of PDFs (Figures 5a, 5b, 5c) in which I shifted the observed value farther and farther from the mean. When are we far enough to expect regression toward the mean of another observation? If the probability area to the left of the observation (bound at the mean) is identical to the area to the right (left-to-right ratio 0.25:0.25), a repeated observation is as likely to be closer to the mean (on the same side) as it is to be farther from the mean. But if the ratio of the two areas is 0.26:0.24, regression toward the mean is likely! But only trivially more likely than not…

“More likely to be closer to the mean” means nothing interesting without quantification of how much more likely. The farther we move observation A toward the tail, from an even split of the probability area, the more likely observation B will fall closer to the mean than observation A. Still, it might end up being only a little closer.

Figure 5a.

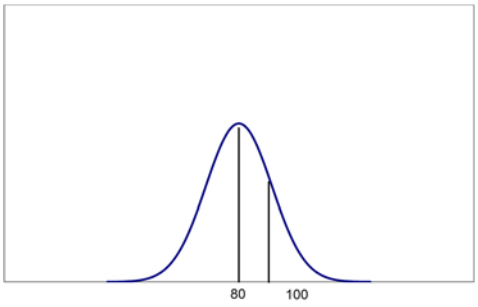


Figure 5b.

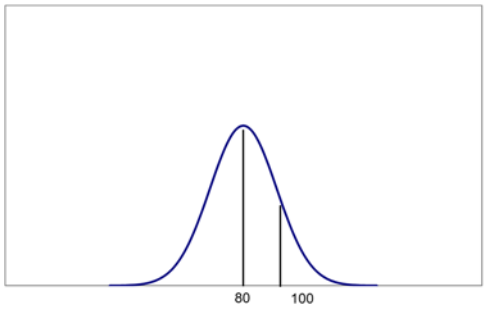
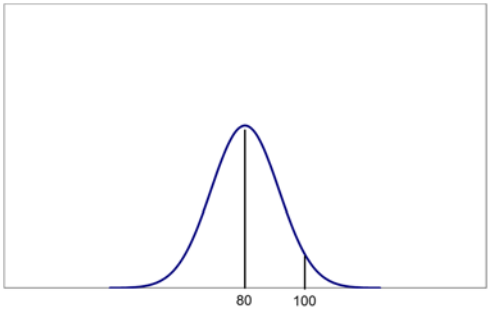


Figure 5c.



Lastly, it is not enough to be “far enough” from the mean. Ideally, the PDFs should be bell-shaped. Imagine a PDF bound on the left at zero, a mean of 1, and most of the probability area located between 1 and 100. If you hit a zero value on the first measurement, are you expecting a second measurement to be “closer to the mean” (between 0 and 1)? No.

Sampling from a finite population

Consider some finite population in which Smith is counted. At time 1, some people in that population have measured values of weight that happen to be high, either because they are “far enough” from their means, just like Smith’s weight, or because their means are high. We’ll call them the “heavy weight” group. Obviously, the mean weight of this group is higher than, and far from, the population mean. Let’s assume it is 110kg as compared with a population mean of 90kg. Now, think about a bell-shaped PDF for that population, the origin of which are the individual PDFs.[d] And think about a thought experiment in which we replay the causes of measured weight in the “heavy” group (taking their realized values). What would be the mean weight of this group in the replay? Of course, it will likely be smaller than 110kg, likely closer to the population mean of 90kg, because some members of this group were “far enough” from their individual means.

Imagine, next, a PDF for the population at time 2, identical to the PDF at time 1 (because the PDFs of individuals have not changed). What would likely happen when you observe again that heavy weight group at time 2? Exactly what was likely to have been observed in our thought experiment. The mean

[d] Reference 1 contains a nice illustration of the population PDF superimposed on individual PDFs.

weight of the heavy weight group at time 2 is likely to be smaller than 110kg, closer to the population mean of 90kg. That's regression of an "extreme" subpopulation mean toward the population mean. Again, all the conditions we discussed for individual observations must hold in this population.

Parent versus child: who is taller?

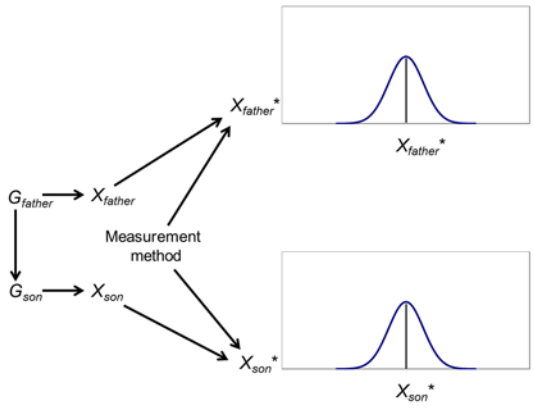
Tall parents are likely to have shorter children. Short parents are likely to have taller children. Tall children are likely to have shorter parents. Short children are likely to have taller parents.

The short paragraph above roughly summarizes Galton's 1886 paper,[5] a historical example of regression toward the mean you can find in almost every writing on this topic. The height of an (adult) child is often less extreme than an extreme average height of the parents, and the average height of parents is often less extreme than an extreme height of their child. What you will not find, I think, is a causal diagram that explains why regression toward the mean applies to an example where the PDFs pertain to *different* people, such as parent and child. We understand, for example, why (and when) Smith is likely to be measured taller at age 50 than at age 30 – if he was "short enough" at age 30 ("far enough" from the mean). But why would Smith's father likely to be measured taller than Smith due to regression toward the mean?

It is quite simple. All that we need to assume is that some genes are the causes of height and that those genes were transferred to Smith from his father (Figure 5).[e] Under this assumption the height variables X_{son} and X_{father} are conditionally independent and identically distributed. Therefore, sampling from one is like sampling from the other. If you hit an extreme value of height for the son, you are likely to hit a closer-to-the-mean value for the father and vice versa. If you hit an extreme value for the father, you are likely to hit a closer-to-the-mean value for the son. Look at the first paragraph above. That's what is written there.

Again, let's not forget that conditional independence is a prerequisite for identical PDFs, which in turn allow us to predict regression toward the mean. This might not always be the case. For instance, if height is also affected by nutritional status, and Smith's father grew up malnourished – X_{son} and X_{father} are not independent variables. There is no conditioning on the same value of nutritional status. As a result, Smith's PDF is different from his father's PDF and sampling from his father's PDF is not like sampling again from Smith's PDF. We should not expect the father to be taller than his short son "because of regression toward the mean".

Figure 5.



Consider an opposite example where Smith is "tall enough" and Smith's father grew up malnourished, as before. The father is likely to be shorter than the son, but it is not "regression toward the mean". It is simply a shift of the father's PDF to the left.

The moral of these stories is deeper than it might seem. No observation may be predicted, or explained, by only one theory; there are always competing explanations. A thoughtful mind would say that an observation is "compatible with regression toward the mean", and a critical mind will try to explain why the variables denoting the attribute are indeed conditionally independent (or why they are not).[f]

Acknowledgement: Doron Shahar – for helpful exchanges and comments on a draft manuscript.

References

1. Barnett AG et al. Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology* 2005;34:215-220

2. Senn S. Francis Galton and regression to the mean. *Significance* 2011;8:124-126

[e] To simplify, I provide a hypothetical example where the genes of height originate only in the father. Galton's work used the average height of both parents.

[f] You will encounter all kinds of minds, of course.

3. Shahar DJ. Deciding on a measure of effect under indeterminism. *Open Journal of Epidemiology* 2016;6:198-232

4. Shahar E, Shahar DJ. Marginal structural models: much ado about (almost) nothing. *Journal of Evaluation in Clinical Practice* 2013 Feb;19(1):214-22. Epub 2011 Aug 24

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