

From the research bench

What is regression to the mean?

Regression to the mean: after an extreme measurement, the next tends to be less extreme. In a group chosen for that extreme, it can look like real change.

Author	Published	Reading time	Words
Sinan Isoglu, MBA (Quantic)	18 October 2026	17 min	3,733
Commercial growth leader, lecturer and doctoral researcher			

Management summary	Regression to the mean is the tendency of a value that was extreme on one measurement to be closer to the average on the next, whenever the two measurements are less than perfectly correlated and equally spread. This concept guide defines it (Barnett, van der Pols and Dobson; Kahneman and Tversky), explains why it happens and why it grows when units are selected for an extreme first value, draws Galton's 1886 table of parents' and children's heights, reports studies in which people predicted as if there were no regression, and sets out how to separate it from a real change: a randomly allocated comparison group, more than one baseline measurement and analysis of covariance. The sources are one tutorial paper with clinical examples, one 1886 study of family height records and one set of prediction studies, mostly with students.
--------------------	--

Contents	§ 1	What is regression to the mean?	2
	§ 2	Why is the second measurement less extreme than the first?	3
	§ 3	What did Galton find in the heights of parents and children?	4
	Figure 1	Galton's mid-parents and the median height of their adult children	5
	§ 4	Why do intuitive predictions ignore regression?	6
	§ 5	How can you tell regression from a real change?	7
	§ 6	How do Barnett and colleagues take it out of an estimate?	8
	§ 7	What would this look like for a sales team flagged after a bad quarter?	9
	Figure 2	A regression check for a flagged group	10
	§ 8	What do a tutorial, an 1886 table and a set of student experiments leave open?	11
	References	12	

What is regression to the mean?

§ 1 of 8

It is what happens between two measurements of the same thing when the first one was unusual. Barnett et al. (2005)¹ open their tutorial by describing it as a statistical phenomenon that can make natural variation in repeated data look like real change, which happens when unusually large or small measurements tend to be followed by measurements closer to the mean (p. 215). Their first section places it: “RTM is a statistical phenomenon that occurs when repeated” measurements are made on the same subject or unit of observation. Kahneman and Tversky (1973)³ stress how common it is: “Regression effects are all about us” (p. 249).

¹ Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215–220.

³ Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.

In this guide, regression to the mean, also written regression toward the mean, means the tendency of a value that was extreme on one measurement to lie closer to the average on the next, on average and even when nothing acts on the unit in between, whenever the two measurements are less than perfectly correlated and equally spread across units: Kahneman and Tversky³ state the link for two variables of equal variance, and Barnett et al.¹’s formula takes one spread for both measurements. It describes a group, not each member: some units that started at the extreme move further out, but the group’s average moves back. It becomes a practical problem when units are chosen because their first value was extreme and the change that follows is read as the effect of whatever was done to them.

Why is the second measurement less extreme than the first?

§ 2 of 8

Because every measurement carries some variation that is not the thing being measured. Barnett et al.¹ put it in one line: “It happens because values are observed with random” error, which they define as non-systematic variation around a true mean, such as random measurement error or random fluctuation in a person. “Systematic error, where the observed values are consistently” biased, does not cause it. Their example is one person with a true mean HDL cholesterol of 50 mg/dl: a reading of 30 is unusually low for that person, and the next reading would more likely lie closer to 50. Since data almost never come without random error, they call it “a common phenomenon” (p. 215).

Selection makes it larger. Among 50 subjects simulated from a hypothetical population, Barnett et al.¹ mark everyone whose first reading falls below 40 mg/dl as a low group for treatment. Of the 21 subjects who could fall into that group on a single reading, only 2 have a true mean below 40 and 19 have a true mean above it; measured again, those 19 will most likely come back closer to their true means, which raises the low group’s average. “Such a change in the group mean may wrongly be attributed to” a real change, and the issue becomes particularly important when an “intervention is applied between the two measurements” (p. 216). Their abstract states the general rule: the effect becomes more noticeable as measurement error increases and when follow-up measurements are examined only in a subgroup selected on a baseline value (p. 215).

They also give the size of the effect for Normally distributed data: the total standard deviation of one reading across the population, which combines the spread between people with the variation within each person, times one minus the correlation between the two measurements, times a factor that grows as the cut-off moves into the tail. So as the correlation “becomes smaller the RTM effect increases”; “the RTM effect is proportional to the population” standard deviation, which in the formula is this total; and it is larger for a cut-off that lies further into the tail of the distribution. With a correlation of 0.64 between the two readings, a total standard deviation of 15 mg/dl and the cut-off at 40 mg/dl, their example yields about 9.6 mg/dl, a “seemingly large increase in the group’s mean HDL cholesterol” with no treatment at all (p. 217).

Kahneman and Tversky³ state the link to correlation without the formula. For two variables of equal variance, saying that one is

regressive with respect to the other is logically equivalent to saying that the correlation between them is less than one, so explaining regression amounts to explaining “why a correlation is less than unity” (p. 250). Barnett et al.¹ add that regression to the mean occurs in any measurement “that is observed with error”, not only in Normal or continuous data: in binary data it can make units change category with no change in what underlies them (p. 219). The measurement error guide separates an indicator from the target it stands for and explains classical error, under which, in a simple regression, noise in the predictor can flatten the estimated slope; regression to the mean is another consequence of the same kind of random error, and Barnett et al. draw the same link (pp. 216 to 217).

What did Galton find in the heights of parents and children?

§ 3 of 8

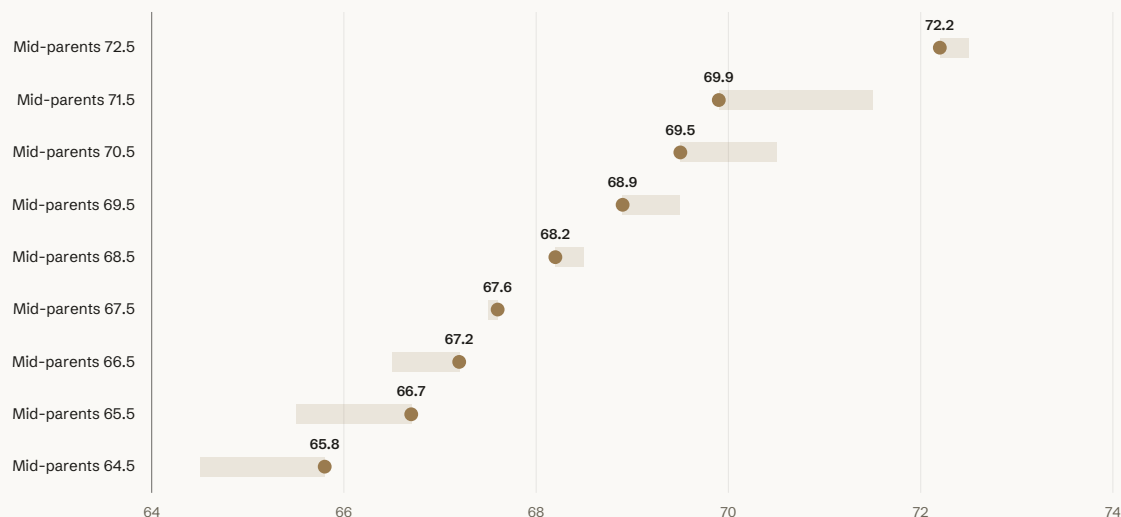
The name comes from him. Barnett et al.¹ date the first discussion to Galton in 1877 and credit him with the phrase (pp. 215, 217); his own 1886 title speaks of regression towards mediocrity. In that paper Galton (1886)² recalls experiments on “seeds of different size but of the same species”, from which the offspring came out less extreme than their parent seeds, smaller than large parents and larger than very small ones, with the average regression proportional to the parent’s deviation from mediocrity, a point well below the average size of the seeds he had bought (p. 246).

² Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.

He then gathered family records of height. His data covered 930 adult children and 205 sets of parents; he multiplied every female height by 1.08 to put it on the male scale and averaged the two parents into one mid-parental height (pp. 247, 250). In that population the level of mediocrity was 68 1/4 inches, and he stated his law: the offspring’s deviation from that level is, on average, two-thirds of the deviation of its mid-parentage (p. 252). He reports finding it “with unexpected coherence and precision” (p. 247), and his Plate IX states that the deviates of the children are to those of “their Mid-Parents as 2 to 3”. Figure 1 draws the medians from his Table I.

Figure 1 Galton's mid-parents and the median height of their adult children

In all nine classes the median adult child stood nearer to 68¼ inches than the mid-parents did: taller than them below that level, shorter above it.



Each row is one class of mid-parental height. The band runs from the class height to the median height of the adult children in the class, marked by the dot and its value. In every class the median lies nearer to 68 1/4 inches, the level Galton calls mediocrity (p. 252), than the class height does. In the top class, the furthest from 68 1/4 inches, the median lies only 0.3 inch below the class height; that class rests on 6 mid-parents, and the bottom class on 5. Three of the nine gaps are 0.3 inch or less, and Galton heads the children's columns 62.2, 63.2 and so on because reported heights cluster on whole inches (note to Table I). Galton's two-thirds is his summary of these data, drawn as a line in his Plate IX, and is not computed here.

Source Galton (1886), p. 248, Table I: number of adult children of various statures born of 205 mid-parents, heights in inches, all female heights multiplied by 1.08; the row headings and the column of medians. The classes Above and Below print no median and are left out. From the 72.5 class down, the classes hold 6, 11, 22, 41, 49, 33, 20, 12 and 5 mid-parents. Read from the page image of the scanned copy.

Galton's² own explanation was about heredity, not about measurement: "The child inherits partly from his parents, partly from his ancestry", and a remote ancestry averages out to the population: "in other words, it will be mediocre" (p. 252). He also found the relation running the other way. Read by columns instead of rows, the same table shows that the most probable mid-parentage of a man deviates only one-third as much as the man does; and because people near mediocrity are so numerous, an exceptional man is more often the somewhat exceptional son of fairly ordinary parents than the average son of exceptional ones (pp. 253 to 254).

On this guide's reading, that second finding separates regression from a causal pull. Nothing that passes from parents to children can make tall children's parents shorter, yet the parents of tall children lie nearer the middle than their children do, as the children of tall parents lie nearer the middle than their parents. The two directions are not mirror images: Galton stresses that the converse is far from the numerical opposite of his law. That opposite would have the mid-parentage deviate one and a half times as much as the son, where the table gives one-third (pp. 253 to 254). Both rest on an imperfect correlation, though not with the equal variances Kahneman and Tversky³ assume: Galton expects the variability of mid-parental and adult heights to differ by a factor of the square root of 2 (p. 251), so the two fractions need not be equal. Galton's² numbers stand. The later sources treat the heights of parents and children as a familiar case of regression and explain regression in general by imperfect correlation and random variation; neither discusses the remote ancestry Galton reached for.

Why do intuitive predictions ignore regression?

§ 4 of 8

Because people predict by resemblance, and a prediction made that way is a non-regressive prediction: as extreme as the evidence it rests on. Kahneman and Tversky (1973)³ state their thesis plainly: "people predict by representativeness", that is, they choose the outcome that appears most representative of the evidence. As a result, they argue, intuitive predictions are insensitive to how reliable the evidence is and to the prior probability of the outcome, in violation of the logic of statistical prediction, and they trace mistaken intuitions about statistical regression to the same representativeness heuristic (p. 237).

The statistical rule is that predictions should vary only as much as the evidence warrants. When accuracy is perfect, predictions vary as much as the outcomes; when it is zero, the same value is predicted every time; in between, predictions are regressive, and the greater the uncertainty, the smaller their spread (p. 243). People did not follow it. In two studies with descriptions of college freshmen, one group rated each description and another predicted each student's grade point average; the two "groups produced equally extreme" judgements, with average standard deviations of 25.7 and 24.0 in one study and 22.2 and 21.4 in the other, none of the differences significant (p. 245).

A third study gave three groups the same ten percentile scores and asked for a grade point average. Group 1 was told the scores were percentiles of grade point average itself; group 2, that they came from a test of mental concentration described as unreliable from day to day; group 3, that they came from a test of sense of humour. The groups

predicting from the more remote inputs should have been more regressive. The spread of predictions was 0.91 in group 1 and 0.87 in group 2, not a significant difference, and 0.69 in group 3, significantly less than in group 2 (Table 3, p. 247). In the authors' words, "people fail to regress when predicting" achievement from a measure of ability, however unreliable; the predictions from sense of humour were regressive, though by the authors' account not enough.

The same pattern held outside the classroom. Officers on an Israeli army selection board, who knew that their "predictive validity was only moderate" (on the order of 0.20 to 0.40), predicted the officer-school grades of over 200 candidates in a distribution virtually identical to that of the actual grades, except that they predicted fewer failures than occurred: "their predictions were nonregressive" (p. 247). Since intuitive predictions are often not regressive enough, the authors add, the gap between predictions and outcomes is largest at the extremes (p. 249).

How can you tell regression from a real change?

§ 5 of 8

Start by assuming it is there. Barnett et al.¹ write that "One should assume that RTM has taken place unless the data" show otherwise, and their discussion adds that it "needs to be ruled out as a cause of an observed" change before any other explanation is sought. Their first check is a plot, "a scatterplot of change" (follow-up minus baseline) against the baseline. Their key messages put the question to ask of it: "is there more change at the tails of the baseline measurements?" They also give a public-health case they think was missed: part of the 90 per cent fall in meningitis C cases in the UK after an immunisation programme could be due to a very bad year being likely to be followed by a better one (pp. 217, 219, 220).

Kahneman and Tversky³ explain why it is so easy to miss. People do not expect regression in many situations where "it is bound to occur", find the concept hard to learn, and, when they do see regression, invent explanations for it in terms of some dynamic cause. Asked for a 95 per cent interval around the true IQ of a person who scored 140 on a test described as a true score plus random error, 73 of 108 graduate students in psychology gave intervals symmetric around 140; 24 gave regressive intervals and 11 counterregressive ones. Most predicted as if "the value of 140 was a true score" (p. 250). Among the spurious explanations offered in the social sciences, they list dynamic principles invoked to explain why businesses that did exceptionally well at one point tend to do worse later, and they refer readers to earlier discussions of regression fallacies in research (p. 250).

Their last example is a story from one of the authors' own experience. Flight instructors who praised every good manoeuvre reported that the next one was usually worse, and concluded that praise hurt. Performance is not perfectly reliable and improves slowly from one manoeuvre to the next, so pilots who did exceptionally well on one trial were likely to do worse on the next whatever the instructor said. The instructors "discovered the regression but attributed it" to the harm done by praise. Since people tend to reward good behaviour and punish bad, the authors conclude that, by regression alone, others are most likely to improve after punishment and to "deteriorate after being rewarded". None of the graduate students given the problem suggested regression as the cause (pp. 250 to 251).

Put together, on this guide's reading, the warning signs are four: the units were chosen because a first value was extreme; that value is a single measurement; nothing selected the same way was left untreated; and the change is largest where the first values were most extreme. When all four hold and a story about cause is ready to hand, the story has not yet been tested against regression, and accepting it untested is the regression fallacy.

How do Barnett and colleagues take it out of an estimate?

§ 6 of 8

At two stages: when the study is designed and when the data are analysed. The first design is random allocation. "If subjects are randomly allocated to comparison groups the" responses from all groups should be equally affected by RTM, so the mean change in an untreated group, such as a group given a placebo, "provides an estimate of the change caused by" regression to the mean, plus any placebo effect, and the difference between the groups' mean changes estimates the treatment effect after adjusting for it (p. 217). The [A/B test guide](#) covers randomised experiments of this kind; where a comparison group cannot be randomised, a difference-in-differences comparison has to argue that both groups would have moved in parallel without treatment, which the [parallel trends guide](#) sets out. On this guide's reading, a group picked for an extreme first value is a hard case for that argument, because it is the group expected to regress.

The second design selects on more than one measurement. "To reduce the variability we can select subjects using two or more baseline measurements", applying the cut-off to their mean or to a later measurement. The reduction is largest from the first to the second baseline and smaller after that. In a more realistic model the authors cite, in which two measurements from the same person are

more alike than measurements from two people, more baselines never remove the effect entirely, and the key messages say the design is unlikely to eliminate the problem completely (pp. 218, 220).

At the analysis stage they give two methods. If the population mean and standard deviation and the within-subject standard deviation are known or can be estimated, the expected regression from their formula can be subtracted from the observed change. The approach “often more practical” is analysis of covariance (ANCOVA), which “adjusts each subject’s follow-up measurement according to their” baseline measurement: the follow-up is regressed on the centred baseline as a covariate and on group membership, and the group coefficient is the treatment effect adjusted for regression to the mean. They recommend ANCOVA wherever a t-test could be used (pp. 218, 219).

Their worked example uses log-transformed blood levels of betacarotene from a subset of a randomised trial, 96 pairs of measurements in all, 52 with a supplement and 44 with a placebo. Restricted to the 49 pairs with a baseline below 0.5 $\mu\text{M/l}$, the placebo group’s mean showed a possible rise of 0.24 (95 per cent CI -0.04 to 0.53); the formula put the expected regression at 0.29, so the authors conclude “that there was no real” change in the placebo group. The treatment effect in that subset was 0.84 (0.27 to 1.41) from t-tests on the changes and 0.86 (0.37 to 1.36) from ANCOVA, whose interval was narrower because the baseline explained part of the variance in the outcome (Tables 1 and 2, p. 219). Because allocation in the trial was random, the study was protected against regression to the mean at the design stage.

What would this look like for a sales team flagged after a bad quarter?

§ 7 of 8

A worksheet comes first. It puts questions drawn from Barnett et al.¹s remedies to any change measured in units chosen for an extreme value: what they were selected on, how many measurements stood behind the selection, who was selected the same way and left untreated, how much change was therefore expected with no action, what changed, and what remains once the expected change and the baseline are accounted for.

Figure 2 A regression check for a flagged group

Write down the selection, the baselines, the comparison, the expected change, the observed change and the adjusted effect before crediting the action.

Selection	Baselines	Comparison group	Expected	Observed	Adjusted
Which first value, which cut-off?	How many measurements before?	Selected the same way, untreated?	Change with no action, and why?	Change after the action?	Effect net of comparison and baseline?

If the comparison group moved as much as the flagged group, the action has not been shown to work.

One row per claim that an action changed a flagged group. A row with one baseline and no comparison cannot separate the action from regression to the mean.

Source Author's decision worksheet. The selection and baselines fields follow Barnett et al. (2005), pp. 216 to 218; the comparison and expected fields follow Barnett et al. (2005), p. 217; the adjusted field follows Barnett et al. (2005), pp. 218 and 219. Applying it to sales or account data is this guide's reasoning. The fields are reader inputs; no company data is supplied.

An illustration, with no company data. A sales director ranks forty account executives by third-quarter bookings, puts the bottom ten on a coaching plan and gives the top ten an award. In the fourth quarter the bottom ten's average rises and the top ten's falls, and the obvious reading is that coaching works and awards breed complacency, the flight instructors' reading of praise. But a quarter's bookings can carry variation that is not skill: a deal can slip past the quarter's end, a large order can land early. If one quarter's bookings correlate with the next well below one and spread about as widely, both groups were expected to move toward the team's average with no coaching and no award, and how far depends on that correlation, which the director has not measured. In Figure 2 the row reads: selected on one quarter, bottom ten; one baseline; no comparison; so no estimate of the expected change, and no adjusted effect.

The same budget allows a cleaner design. Of the ten who qualify, coach a randomly chosen five this quarter and the other five the next, and compare the change in the two halves over the same quarter: with five in each half the comparison is noisy, but regression to the mean is expected to affect both halves alike. Select on the average of two or more quarters rather than one. And analyse next-quarter bookings with prior bookings as a covariate rather than as a simple before-and-after difference. The same row can be filled for an account flagged after a sharp fall in orders and handed to a retention team. These are the author's examples and have not been tested; none of the three sources studied sales or account data.

What do a tutorial, an 1886 table and a set of student experiments leave open?

§ 8 of 8

How large regression to the mean is in any commercial measure. Barnett et al. (2005)¹ is a tutorial: its formula assumes Normally distributed data, its examples are clinical, and its worked example is a random subset of one trial. Galton (1886)² worked from family records collected by offering prizes, with heights he judged liable to an error that as often as not exceeded two-thirds of an inch, partly because it was unclear whether shoes had been worn (p. 252); his Table I counts 928 children, as does p. 254, where his description of the data on p. 247 says 930, and the later sources account for regression in general by random error and imperfect correlation, not by ancestry. Kahneman and Tversky (1973)³ studied judgements, mostly by students, about hypothetical people, with one real selection board; their flight-school case is a single anecdote.

On this guide's reading, two things carry over with some confidence: any change measured in units chosen for an extreme first value contains some regression to the mean unless the design removes it, and people, analysts included, tend to explain that regression with a cause. How large the regression is in quarterly bookings, account revenue or customer scores depends in each case on how strongly one period correlates with the next and whether the two periods spread alike, and the sources do not say.

 End of essay

References

- 1 Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215–220. <https://doi.org/10.1093/ije/dyh299>
- 2 Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263. <https://doi.org/10.2307/2841583>
- 3 Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. <https://doi.org/10.1037/h0034747>

How to cite this essay

Isoglu, S. (2026, October 18). What is regression to the mean? Sinan Isoglu.
<https://isoglu.com/insights/journal/what-is-regression-to-the-mean/>
Online version: <https://isoglu.com/insights/journal/what-is-regression-to-the-mean/>

Sinan Isoglu, MBA (Quantic)

Commercial growth leader, lecturer and doctoral researcher

© 2026 Sinan Isoglu · isoglu.com · All rights reserved