Theory & Psychology

Article

Theory & Psychology
21(6) 762–784

DOI: 10.1177/0959354310384910
tap.sagepub.com
SAGE

The mythologization of regression towards the mean

Michael D. Maraun, Stephanie Gabriel, and Jack Martin
Simon Fraser University

Abstract

In the quantitative methodology literature, there now exists what can be considered a received account of the enigmatic phenomenon known as regression towards the mean (RTM), the origins of which can be traced to the work of Sir Francis Galton circa 1885. On the received account, RTM is, variably, portrayed as a ubiquitous, unobservable property of individual-level difference and change phenomena, a force that impacts upon the characteristics of individual entities, an explanation for difference and change phenomena, and a profound threat to the drawing of correct conclusions in experiments. In the current paper, we describe the most essential components of the received account, and offer arguments to the effect that the received account is a mythologization of RTM. In particular, we: (a) describe the scientific and statistical setting in which a consideration of RTM is embedded; (b) translate Galton's discussion of RTM into modern statistical terms; (c) excavate a definition of the concept *regression towards the mean* from Galton's discussion of RTM; and (d) employ the excavated definition to dismantle certain of the most essential components of the received account.

Keywords

bivariate distribution, conditional mean function, Galton, mythology, regression towards the mean

> *At any given moment there is an orthodoxy, a body of ideas, which it is assumed that all right thinking people will accept without question.* (Orwell, 1949, p. 6)

The origins of the widely discussed yet enigmatic phenomenon known as regression towards the mean (RTM) can be traced to the eminent geneticist Sir Francis Galton, circa 1885. Based upon his studies of the distribution of artistic talent and intelligence within the family bloodlines of eminent men, "Galton noted that there was a marked tendency for a steady decrease in eminence the further down or up the family tree one went from

Corresponding author:
Michael D. Maraun, Simon Fraser University, Department of Psychology, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada.
Email: maraun@sfu.ca

the great man" (Stigler, 1997, p. 107), a tendency that Galton (1885) originally labeled *reversion* or *regression towards mediocrity*. Galton eventually sought the assistance of mathematician J.H. Dickson to attempt to produce a complete quantitative formulation of the phenomenon he had observed and described (Stigler, 1997).

Although it remains far from homogeneous in its details, there has come to exist in the literature on quantitative methodology the outline of what can be described as a received account of RTM. On this account, captured perhaps most purely in Stephen Stigler's many papers on the topic, Galton grappled with what has been referred to as the "conundrum" of "why it was that talent or quality once it occurred tended to dissipate rather than grow" (Stigler, 1997, p. 107), and eventually reached a conclusion that has been heralded in modern times as "one of the grand triumphs of the history of science" (Stigler, 1997, p. 107).

The received account of RTM is constituted of claims of the following sorts:

1. "[U]nless there is perfect correlation between X and Y there must be regression towards the average" (Stigler, 1997, p. 105).
2. Regression towards the mean "affect[s] scores on retesting so that they are closer to the population mean" (Streiner, 2001, p. 72).
3. "Suppose the first score is exceptionally high—near the top of the class. How well do we expect the individual to do on the second test? The answer, regression teaches us, is 'less well,' relative to the class's performance" (Stigler, 1997, p. 104).
4. "[T]he regression fallacy is the most common fallacy in the statistical analysis of . . . data" (Friedman, 1992, p. 2131).

On the received account, RTM is a potent but unobservable force that can play havoc with the scientist's attempts to study change and to draw correct conclusions about the efficacy of treatments and interventions. While the received account does not yet contain a definitive explanation of the empirical origins and nature of RTM, proto-explanations that rest on the properties of true and error scores and, hence, that are rooted in classical test theory appear to be ascendant (cf. Karylowski, 1985; Nesselroade, Stigler, & Baltes, 1980).

The case that we will herein present departs considerably from this received depiction, for we will argue that, in his attempts to mathematize the phenomena that were originally the foci of his investigations, Galton lost sight of these phenomena, and unintentionally founded a now long-standing practice of mythologizing RTM, the chief product of which is the received account. The existence of the mythology that is the received account has not only militated against a sound historical account of what Galton did and did not accomplish, but has also interfered with the modern-day researcher's capacity to conceptualize and coherently investigate change phenomena.

In order to support this verdict, it will be necessary to take stock of the key pieces of the puzzle that is the modern-day discussion of RTM. It will be necessary to clarify the Galtonian origins of this puzzle and, in so doing, distinguish RTM from the individual-level phenomena out of which it is constituted. This will, in turn, necessitate a clarification of the essential distinction between phenomena and their statistical representations. The cornerstone of this analysis, and the key to dismantling the mythology that is the received account, will be an excavation of Galton's definition of the concept *regression* *towards the mean* from the talk of bivariate normality and linearity of regression under which it has long been buried. The organization of the paper is as follows: (a) the scientific and statistical setting in which a consideration of RTM is embedded will be described in detail; (b) Galton's discussion of RTM will be given a modern statistical translation; (c) a definition of the concept *regression towards the mean* will be excavated from Galton's discussion of RTM; and (d) with the excavated definition as the primary tool, the key components of the mythology that is the received account will be dismantled.

The embedding context of RTM

For the sake of economy of exposition throughout the remainder of the paper, it will be useful at this point to introduce a set of statistical notation and to review a small number of scientific concepts. Following this, Galton's 1885 work will be reviewed, and his claims cast in this language and notation.

Statistical notation and language

Let P_T denote a population of N individual entities (such as the set of all humans, or all Canadian women under 40, or all gold nuggets weighing more than one ounce). There can be defined, for the purposes of scientific investigation, countless distinct populations P_T. Let (X,Y) denote a pair of variates defined on some population P_T, the scores on each of X and Y being measurements of the standings of individual entities with respect to two types of phenomena under study. This simply means that each $i \in P_T$ yields a score-pair (x_i,y_i). A particular choice of P_T and (X,Y) thus constitutes a *population/phenomenon pairing* (henceforth, P_T/phenomenon pairing).

We will say that X and Y have a *bivariate empirical distribution* in P_T and symbolize this distribution as $F_{x,y}$. We call this an *empirical* distribution function because it is a literal reporting of the proportions of joint occurrences $P(X \le t,\ Y \le s)$ of the values assumed by X and Y *within* P_T. The *empirical moments* of $F_{x,y}$ are functions of the joint proportions $P(X \le t,\ Y \le s)$ and include the population means μ_x and μ_y variances σ_x^2 and σ_y^2, Pearson product– moment correlation ρ_{xy}, all higher-order association parameters, and so on. The empirical moments of $F_{x,y}$ are simply aggregate properties of the set of score pairs (x_i,y_i) yielded by the N entities contained in P_T. Hence, they are aggregate properties produced under a particular P_T/phenomenon pairing. Finally, we will denote the class of all $F_{x,y}$—that is, bivariate distributions produced under P_T/phenomenon pairings—as F.

In contrast to the class F of empirical distribution functions, let T stand for the class of *theoretical* or *mathematical bivariate distributions*. Unlike the elements of F, the elements $T_{x,y}$ of T are specified by mathematical functions of the form $z = f(x,y)$. Some familiar elements of T include the bivariate normal, bivariate gamma, and bivariate uniform distributions. Because theoretical distributions are simply mathematical functions, they have no necessary link to empirical reality. If a scientist wishes to use some $T_{x,y} \in T$ to represent a particular $F_{x,y} \in F$, the onus is on the scientist to choose an element of T that is a satisfactory representation of $F_{x,y}$. A failure to make a satisfactory choice can very easily result in the scientist making erroneous empirical claims.

For any observed score-pair (x_i, y_i) yielded by an $i \in P_T$, difference and change phenomena are captured quantitatively as functions defined on (x_i, y_i), examples being the raw difference $y_i - x_i$, the absolute difference $|y_i - x_i|$, and the ratio $\frac{y_i}{x_i}$. Thus, a quantitative study of difference or change phenomena *at the level of the individual entity* is carried out through a study of functions defined on score pairs (x_i, y_i). If it is observed that, for a particular function of (x_i, y_i), all elements of P_T are such-and-such, then nomothetic knowledge (Lamiell, 1998) has been derived, and it is the job of the scientist to explain why the entities in P_T are as they are. Traditionally, nomothetic explanations are expressed as *individual-level laws*, which are typically represented in the following form:

$$\forall i \in P_T,\ y_i = f\left(x_i, \underline{\omega}\right) \tag{1}$$

This symbolic statement can be interpreted as the claim that "for all elements i contained in population P_T, y_i is causally determined by x_i in conjunction with a vector of ancillary factors $\underline{\omega}$." A law of this sort is typically accompanied by supporting sentences that explain the law's workings and the meanings of the symbols that appear in its representation.

One way to quantify the *degree of exceptionality* or *extremity* of an entity's score within the distribution of a variate is to compute the number of standard deviations of the score from the mean of the distribution. Thus, the extremity of a score x_i with respect to the distribution of X is quantifiable as $\left|z_{x_i}\right| = \left|\frac{(x_i - \mu_x)}{\sigma_x}\right|$ and the extremity of a score y_i with respect to the distribution of Y, as $\left|z_{y_i}\right| = \left|\frac{(y_i - \mu_y)}{\sigma_y}\right|$. These quantities may be referred to as the standardized scores on X and Y, respectively. A $\left(\left|z_{x_i}\right|, \left|z_{y_i}\right|\right)$ pair equal to (3,1.5), for example, indicates that the score of entity i on X is three standard deviations from the mean of X, while its score on Y is 1.5 standard deviations from the mean of Y. Alternatively, we might say that the X-score in this example is twice as extreme as the Y-score.

Galton's 1885 work

In his 1885 article "Regression Towards Mediocrity in Hereditary Stature," Galton describes the analysis of his human adult parent/offspring height data set—which he referred to as "the Records"—as follows:

> [I]t 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. (p. 247)

His analyses of "family likeness" in height were based on data that "consisted of the heights of 930 adult children and of their respective parentages, 205 in number" (p. 247).

In particular, the individual entities $i \in P_T$ that were the targets of his investigation were adult, British, human parent/offspring pairs. As a preliminary step in his analysis, Galton equated male and female measurements by multiplying the latter by a factor of 1.08, and calculated the average (arithmetic mean) height of the parents of each adult offspring.

Galton (1885) claimed that "the stature of the children depends closely on the average stature of the two parents" (p. 249). This claim suggests that his aim was to discover a law that would describe, for any $i \in P_T$, the causal dependency of offspring height (z_{yi}) on parental height (z_{xi}).[1] Galton seems to have believed that his "discovery" of RTM represented the fulfillment of this aim, for he concluded that "[t]his law tells heavily against the full hereditary transmission of any gift, as only a few of many children would resemble their mid-parentage" (p. 253), a suggestion that agrees with Galton's frequent references to the *law* of RTM. While his written work suggests that the essence of RTM, to his mind, was that "the structure of adult offspring must on the whole be more mediocre than the structure of their parents," he translated (for reasons that we will later discuss) the notion of "on the whole" into the mathematical expression "on average." Thus, with respect to height, Galton expressed his RTM claim as follows: "[W]e can define the law of regression very briefly. It is that the height-deviate of the offspring is, *on the average* [emphasis added], two-thirds of the height deviate of its mid-parentage" (p. 252). This claim can be represented symbolically as

$$E\left(Z_y \mid Z_x = \delta\right) = \frac{2}{3}\delta, \tag{2}$$

in which $E\left(Z_y \mid Z_x = \delta\right)$ is the mean of Z_y given that Z_x is equal to δ.

Because we will, shortly, provide both a general definition of the concept of RTM, and argue that the phenomenon of RTM has been mythologized, it is essential to distinguish the empirical components of Galton's claim 2 from what follows necessarily from the claim itself. An examination of equation 2 itself and a consideration of Galton's claims ("the experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it"; 1885, p. 246) make it clear that the most fundamental empirical claim embodied in equation 2 is that the conditional mean $E\left(Z_y \mid Z_x = \delta\right)$ of Z_y on Z_x (equivalently, Y on X) is *linear*, or, in other words, has the functional form

$$E\left(Z_y \mid Z_x = \delta\right) = \beta_{y.x}\delta. \tag{3}$$

The coefficient $\beta_{y.x}$ in equation 3 is the slope of the linear regression function, and is equal to the number of standard deviations change in the average of Y associated with a one standard deviation increase in X. Let us point out two important facts related to the claim that $E\left(Z_y \mid Z_x = \delta\right)$ is linear:

1. For any $F_{x,y} \in F$, $E\left(Z_y \mid Z_x = \delta\right)$ is determined by $F_{x,y}$, and, hence, its functional form is a property of the set of N score-pairs (x_i, y_i) yielded by the entities $i \in P_T$. In particular, $E\left(Z_y \mid Z_x = \delta\right)$ can be flatline (when Z_x and Z_y are unrelated), linear, or one of an infinity of nonlinear forms. Thus, for a particular P_T/

phenomenon pairing, the claim of linearity of $E\left(Z_y \mid Z_x = \delta\right)$ need not be true. It is a testable, empirical assertion about $F_{x,y}$, or, equivalently, about the N score pairs yielded by the individual entities under a particular P_T/phenomenon pairing. Whether or not Galton was correct in his claim about the form of $E\left(Z_y \mid Z_x = \delta\right)$ remains open for debate: a recent reanalysis by Wachsmuth, Wilkinson, and Dallal (2003), for example, raises the possibility that Galton misanalyzed his data, and that his conditional mean functions were, in reality, nonlinear.

2. If equation 3 were true for a particular P_T/phenomenon pairing, then the linearity of $E\left(Z_y \mid Z_x = \delta\right)$ would be a *discovered property* of the $F_{x,y}$ generated under the particular P_T/phenomenon pairing (equivalently, a discovered property of the population of score-pairs yielded by the entities $i \in P_T$). That is to say, when it arises in nature, the linearity of $E\left(Z_y \mid Z_x = \delta\right)$ is an *aggregate*, or *population-level property* of a set of score-pairs. Discovered aggregate properties must be explained, and it is the job of the scientist to formulate plausible explanations for the discovered properties of things. However, for a particular P_T/phenomenon pairing, the state of nature described by equation 3 is not equivalent to the discovery of a law that explains phenomena arising at the level of the individual entity. On the contrary, when it arises in nature, the linearity of $E\left(Z_y \mid Z_x = \delta\right)$ has no implications for the characteristics of any particular individual entity $i \in P_T$.

Let us now consider what follows *necessarily* when equation 3 obtains. *If it were true, for a particular P_T/phenomenon pairing, that $E\left(Z_y \mid Z_x = \delta\right)$ was a linear function,* then, because Z_x and Z_y are standardized variates, the slope of $E\left(Z_y \mid Z_x = \delta\right)$ would necessarily be equal to the Pearson product–moment correlation coefficient (PPMC): that is, it would be the case that

$$\beta_{y.x} = \rho_{xy}. \tag{4}$$

Because ρ_{xy} is restricted by its very construction to lie in the interval [-1,1], it would then follow from equation 4 that

$$-1 \leq \beta_{yx} \leq 1. \tag{5}$$

Thus, *if,* for a particular P_T/phenomenon pairing, $E\left(Z_y \mid Z_x = \delta\right)$ happened to be a linear function, it would then follow that $E\left(Z_y \mid Z_x = \delta\right)$ would have to lie within the shaded, bow-tie shaped, region of Figure 1. In other words, with respect to Galton's height data, *linearity of $E\left(Z_y \mid Z_x = \delta\right)$ implies that the average extremity of offspring height is no greater than the extremity of parent height.*

Because $\left|\rho_{xy}\right|$ will assume the value of unity only under manufactured conditions (as when calculated on a data set containing but two observations), practically speaking, if, for a particular P_T/phenomenon pairing, $E\left(Z_y \mid Z_x = \delta\right)$ happens to be linear, it would then follow that the average extremity of offspring height would have to be less than the extremity of parent height. Finally, because the *magnitude of the difference between parent and average offspring extremity* is, under linearity of $E\left(Z_y \mid Z_x = \delta\right)$, equal to

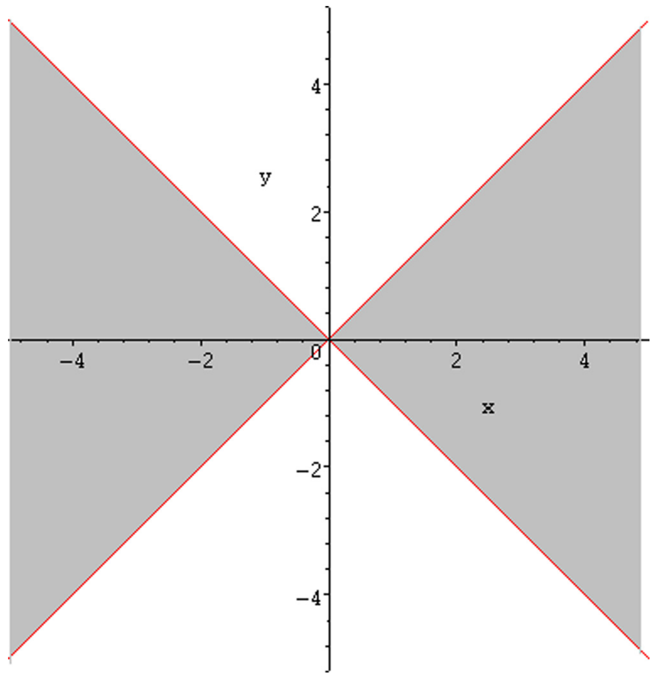


Figure 1. A graphical definition of RTM-E.1

$$\delta - E\left(\left|Z_y\right| \mid \left|Z_x\right| = \delta\right) = \delta - \left|\beta_{y.x}\right|\delta \\ = \delta\left(1 - \left|\rho_{xy}\right|\right), \tag{6}$$

if it did happen to be the case for a particular P_T/phenomenon pairing that $E\left(Z_y \mid Z_x = \delta\right)$ was linear, then the magnitude given in equation 6 would necessarily be a linear, increasing function of $|\delta|$, the extremity of parent height.

For the general case of two continuous variates X and Y defined under a particular P_T/phenomenon pairing, the analytic results above can be summarized as two conclusions:

> Conclusion 1: If it happens to be the case that X and Y have a linear relationship (and there is no necessity that they do), then it follows that for any particular degree of extremity $|\delta|$ of X, the average extremity of Y will be less than $|\delta|$.
>
> Conclusion 2: If it happens to be the case that X and Y have a linear relationship (and there is no necessity that they do), then it follows that the magnitude of the difference between $|\delta|$ and the average extremity of Y will be a linearly increasing function of $|\delta|$.

Excavating Galton's intended definition of RTM

Notably absent from the literature on RTM, and, in particular, the received account of RTM, is a definition of the concept *regression towards the mean*. In a largely unspecified fashion, conclusions (1) and (2) are seen as having something to do with RTM, and this has meant that the phenomenon of RTM has been conceptualized as inextricably tied to the linearity of conditional mean functions (cf. Chesher, 1997).[2] As we will shortly discuss, this unwarranted tying of RTM to the linearity of conditional mean functions likely arose because Galton chose to employ the bivariate normal distribution as his chief representational tool, and in the case of the bivariate normal, both conditional mean functions are linear. However, the assignment to linearity and bivariate normality of a fundamental role in the study of RTM represents a fundamental distortion of what Galton had originally meant by *regression towards the mean*. Buried under the misguided preoccupation with bivariate normality and linearity lies Galton's true (linearity/bivariate normality-free) proto-concept *regression towards the mean*. We will now turn to an excavation of this concept.

As is clear from his many writings on the topic, Galton's proto-concept *regression towards the mean* rested on two distinct components: (a) for any particular degree of extremity $|\delta|$ of X, the average extremity of Y is less than $|\delta|$ and (b) the magnitude of the difference between $|\delta|$ and the average extremity of Y is an increasing function of $|\delta|$. These components can be expressed symbolically as

RTM-E.1: $$-|\delta| \leq E\left(Z_y \mid Z_x = \delta\right) \leq |\delta|$$

Equivalently, $$\left|E\left(Z_y \mid Z_x = \delta\right)\right| \leq |\delta|$$

RTM-E.2: $|\delta|$ - $\left|E\left(Z_y \mid Z_x = \delta\right)\right|$ is an increasing function of $|\delta|$.

Thus, Galton's excavated definition can be stated as follows:

> Definition: *regression towards the mean*. The state of affairs in which the conditional mean function $E\left(Z_y \mid Z_x = \delta\right)$ of one standardized variate Z_y, given a second standardized variate Z_x, satisfies both RTM-E.1 and RTM-E.2.

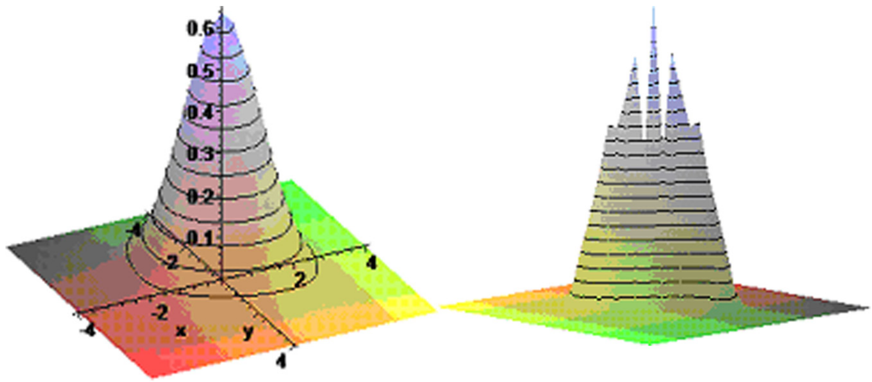
Geometrically, RTM-E.1 is the condition that $E\left(Z_y \mid Z_x = \delta\right)$ is contained within the shaded, bow-tie shaped, region of Figure 1, while RTM-E.2 is the condition that the first derivative of $E\left(Z_y \mid Z_x = \delta\right)$ is less than 1 in quadrants 1 and 3, and greater than -1 in quadrants 2 and 4 of $\mathbb{R}^2$. We will continue to employ the notation RTM when speaking loosely about the phenomenon of regression towards the mean, but will, henceforth, employ the notation *RTM-E* to stand for Galton's, now excavated, concept *regression towards the mean*. Specifically: (a) a conditional mean function $E\left(Z_y \mid Z_x = \delta\right)$ will be said to be *RTM-E consistent* only if it satisfies both RTM-E.1 and RTM-E.2 and (b) a bivariate distribution will be said to be RTM-E consistent only if at least one of its conditional mean functions, $E\left(Z_y \mid Z_x = \delta\right)$ and $E\left(Z_x \mid Z_y = \delta\right)$, is RTM-E consistent. Let T stand for the class of bivariate theoretical distribution functions and F for the class of bivariate empirical distribution functions. Partition T into the subclasses of RTM-E-consistent and RTM-E-inconsistent distribution functions, denoted T_{RTM} and T_{RTMI}, respectively. Analogously, partition F into the subclasses of RTM-E-consistent and RTM-E-inconsistent distribution functions, denoted F_{RTM} and F_{RTMI}. We then have the following:

1. Because the elements of T are specified mathematically, the determination of whether a particular element $T_{x,y} \in T$ belongs to T_{RTM} or T_{RTMI} is made by mathematical analysis. Specifically, on the basis of mathematical analysis, a decision must be made as to whether at least one of the conditional mean functions, $E\left(Z_y \mid Z_x = \delta\right)$ and $E\left(Z_x \mid Z_y = \delta\right)$, of $T_{x,y}$ satisfies both RTM-E.1 and RTM-E.2 (i.e., is RTM-E consistent).
2. Because the elements of F, one for each P_T/phenomenon pairing, are not specified mathematically, the determination of whether a particular element $F_{x,y} \in F$ belongs to F_{RTM} or F_{RTMI} is a matter not for mathematical analysis, but, rather, for empirical investigation. The researcher must draw a sample of score-pairs from $F_{x,y}$ and make an inferential decision about whether at least one of the conditional mean functions, $E\left(Z_y \mid Z_x = \delta\right)$ and $E\left(Z_x \mid Z_y = \delta\right)$, of $F_{x,y}$ satisfies both RTM-E.1 and RTM-E.2 (i.e., is RTM-E consistent).
3. For a $T_{x,y} \in T$ ($F_{x,y} \in F$) for which at least one of $E\left(Z_y \mid Z_x = \delta\right)$ and $E\left(Z_x \mid Z_y = \delta\right)$ is linear, Conclusions 1 and 2 establish that $T_{x,y} \in T_{RTM}$ ($F_{x,y} \in F_{RTM}$). We denote as T_{LRTM} (F_{LRTM}) the subclass of T_{RTM} (F_{RTM}) containing $T_{x,y}$ ($F_{x,y}$) that are RTM-E consistent by virtue of the fact that at least one of $E\left(Z_y \mid Z_x = \delta\right)$ and $E\left(Z_x \mid Z_y = \delta\right)$ is linear and note that $T_{LRTM} \subset T_{RTM} \subset T$ ($F_{LRTM} \subset F_{RTM} \subset F$).
4. Let T_{BVN} be the subclass of T containing the bivariate normal distributions. Then, because for a $T_{x,y} \in T_{BVN}$, *both* $E\left(Z_y \mid Z_x = \delta\right)$ and $E\left(Z_x \mid Z_y = \delta\right)$ are linear, it follows that $T_{x,y} \in T_{RTM}$. We conclude that $T_{BVN} \subset T_{LRTM} \subset T_{RTM} \subset T$.

In the absence of a definition of the concept *regression towards the mean*, neither a determination of whether a $T_{x,y} \in T$ is a member of T_{RTM} or T_{RTMI}, nor a determination of whether a $F_{x,y} \in F$ is a member of F_{RTM} or F_{RTMI}, can be made. Given the, until now, glaring absence of a definition of this concept, it is hardly surprising that little is known about the contents and relative sizes of T_{RTM} and T_{RTMI}, on the one hand, and the contents and relative sizes of F_{RTM} and F_{RTMI}, on the other. As we have observed, it is known that T_{LRTM}, hence, T_{BVN}, are subclasses of T_{RTM}. We are unaware of any other analytic work on this topic. In our brief consideration of T_{RTMI}, we had little difficulty in finding example members of this subclass, the two distributions depicted in Figure 2 being examples.[3] We are unaware of any empirical studies in which legitimate assignments to F_{RTM} and F_{RTMI} were made of the $F_{x,y}$ produced under particular P_T/phenomenon pairings.

The mythologization of regression towards the mean

The received account of RTM is a paradigm case of a phenomenon having been mythologized. On the received account, RTM is, variably, portrayed as a ubiquitous,



(a) (b)

Figure 2. Examples of bivariate distributions belonging to T_{RTMI}

unobservable property of individual-level difference and change phenomena, a force that impacts upon the characteristics of individual entities, an explanation for difference and change phenomena, and a profound threat to the drawing of correct conclusions in experiments. Shortly, we will turn to the task of dismantling this mythology, but before doing so, will describe the manner in which Galton inadvertently sowed the seeds of its growth.

Galton's initial regression towards mythologization

While Galton was indisputably a heavy-weight of science, we do not believe that the conclusions he drew on the topic of "reversion towards mediocrity" represent, as Stigler has claimed, one of the grand triumphs of the history of science. Quite to the contrary, Galton's conclusions were in good part mythology. There are two primary routes by which he built mythological elements into his study of eminence. First, his correlational approach to the investigation of phenomena that arose at the level of the individual entity generated a levels of analysis dilemma that ultimately blurred the boundary between individual- and population-level phenomena. The blurring of this boundary subsequently led him to misinterpret his results. Secondly, in employing the bivariate normal distribution to represent his findings, he lost sight of the boundary between mathematical representation and empirical reality, and this ultimately led him to project mathematical necessities onto reality. We will consider each of these routes in turn.

Recall that, as Stigler (1997) notes, Galton collected data which suggested "that there was a marked tendency for a steady decrease in eminence the further down or up the family tree one went from the great man" (p. 107). With respect to the phenomenon of height, Galton believed that he had established that offspring were more mediocre (less extreme) than their parents. Thus, family lines, or, in the case of the height data, parent/offspring pairs, were the individual entities to which Galton's observations referred. Because Galton's observations were made at the level of the individual entity, scientific explanations of these observations would have to have taken the form of individual-level laws—that is, laws that held for all $i \in P_T$. If Galton had succeeded in discovering even one such law, he would have made a contribution to nomothetic knowledge: that is, knowledge of properties or relations that hold for all members in a class of objects under study. Such properties or relations can be considered, in a certain sense, to be universal.

While Galton portrayed RTM as a "law" that explained his observations, this designation was, in fact, a misnomer. Neither RTM-E, nor the sub-species of it, LRTM, that is brought about by the linearity of a conditional mean function is an individual-level law that holds over the class of entities on which Galton's observations were made. They are neither characteristics that hold for the individual entities $i \in P_T$ under study, nor do they explain anything about hereditary transmission down family lines, nor do they explain why it is that an offspring is more mediocre (or less mediocre or equally mediocre) than its parents when it is so. To our knowledge, science has thus far failed to formulate a law that describes the causal story that underlies the determination of the extremity of a human's height. If, however, such a law had been formulated, this law would undoubtedly have involved biological factors such as diet, the presence or absence of various environmental stressors, and parental genes. Because a parent's primary contribution to the extremity of the height of its offspring is made via genetics, the discovery of a law that explained hereditary transmission would have spelled an end to the *relevance* of the variate "observed extremity of parent height" that featured so very prominently in Galton's research: the parent's *genes* could simply have been observed.

Even allowing that the science of Galton's time was not sufficiently advanced to afford him the opportunity to discover relevant hereditary laws, his commitment to the correlational approach to science, of which he and Pearson were originators (Danziger, 1990), itself militated against his proper handling of the empirical questions that he had set out to answer. The correlational approach involves the calculation of quantities that are defined only when the quantity known as the variance—that is, $E\,(x-\mu)^2$ —is non-zero. The variance is *always* equal to zero when calculated on the basis of a single score yielded by an individual entity. Thus, in order for the correlational approach to be operational, the individual entity had to give way to the *population* as the target of investigation. While Galton set out to make observations on, and provide explanations with respect to, individual entities (such as the family line or the parent/offspring pair), his approach to the problem necessitated that his analyses had as their targets an entirely different type of entity: the population. Thus, for example, the notion of "on the whole," which should properly have played the role of a rider to an individual-level law, was instead interpreted by Galton as "on average," thus invoking an aggregate property of a population of scores. All told, Galton sowed the seeds of a levels of analysis dilemma that has grown into one of the key components of the mythology that is the modern-day received account of RTM.

Giving birth to this levels of analysis problem was not Galton's only foray into the domain of myth generation. With the aid of mathematician J.H. Dickson, he formulated a representation of his findings in terms of the bivariate normal distribution, and, in so doing, seems to have lost track of the boundary between mathematics (in particular, mathematical necessities arising from the structure of the bivariate normal distribution) and empirical reality. As Wachsmuth et al. (2003) put it, “Galton derived his theory [of natural inheritance] by looking at data, but the lens he used profoundly shaped what he saw” (p. 190). The unquestioned employment of this very same lens by the generation of methodologists who have followed in Galton’s footsteps has, we suggest, played a significant role in their distorting the phenomenon of RTM and in their creating, out of these distortions, a full-blown mythology (cf. Freedman, 1985).

According to Galton (1889), “However paradoxical it may appear at first sight, it is theoretically a necessary fact, and one that is clearly confirmed by observation, that the structure of adult offspring must on the whole be more *mediocre* than the structure of their parents” (p. 95). In fact, there is nothing at all paradoxical about large decrements in extremity within a family line. Nature is the way it is, and while it never ceases to amaze, it is not paradoxical. When the issue is “hereditary transmission,” large decrements are neither more shocking nor more important than small decrements, or increments, or instances of no change at all. Regardless of the particulars of the change or difference phenomena observed, what is required is a scientific explanation of these phenomena, and it is the job of the scientist to formulate scientific explanations. Moreover, when it is observed to be the case, it is not by theoretical necessity that “adult offspring are on the whole more mediocre than their parents.” Theoretical necessities arise within axiomatized logical systems such as mathematics and statistics, and are established by proof. Thus, it is a “theoretical necessity” that LRTM obtains when $E\left(Z_y \mid Z_x = \delta\right)$ is linear and the absolute value of the Pearson–product moment correlation is less than unity. It is a theoretical necessity that bivariate normal distributions have linear conditional expectations and, hence, are both LRTM- and RTM-E consistent. It is, however, neither a theoretical necessity that a particular element of F has linear conditional mean functions, nor that it is satisfactorily represented by the bivariate normal distribution. Thus, contra Galton’s assertion, it is not a necessity that “the structure of adult offspring must on the whole be more *mediocre* than the structure of their parents.” Determinations of whether or not potential properties hold for elements of F must be made on a case-by-case basis, and take the form of data-based inferential decisions. Galton supplanted the making of case-by-case determinations on the basis of empirical evidence with the making of mathematical deductions on the basis of the bivariate normal distribution, a mathematical function whose adequacy as a representer of bivariate phenomena is always open to question, and this can hardly be seen as scientific progress.

What is not what in RTM

From the seeds sown by Galton’s initial confusions has grown a complex, ramifying mythology of RTM that we call the received account. We will turn now to an elucidation of a number of its most salient characteristics. The style of exposition is to state what RTM is not, the argumentation resting chiefly on simple comparisons to the excavated definition RTM-E. [4]

RTM is not a necessity. Contrary to the message of many discussions of RTM, there is no *tendency* for, and certainly no necessity that, bivariate distribution functions, be they either elements of T or F, be RTM-E consistent. A determination must be made for each distribution function. As we indicated earlier on in the paper, elements of T are assigned to either T_{RTM} or T_{RTMI} on the basis of mathematical analysis and elements of F are assigned to either F_{RTM} or F_{RTMI} on the basis of empirical investigation. And while the concept necessity can, at least, be applied coherently in the case of T (e.g., as we earlier summarized, if, for $T_{x,y} \in T$, at least one of $E\left(Z_y \mid Z_x = \delta\right)$ and $E\left(Z_x \mid Z_y = \delta\right)$ is linear, then $T_{x,y} \in T_{LRTM} \subset T_{RTM}$; i.e., it can be shown in mathematics that membership of $T_{x,y}$ in T_{RTM} is a necessary condition of $T_{x,y}$ having at least one linear conditional mean function), it cannot be applied coherently in the case of F. This is because the members of F, they being empirical distribution functions, are not specified by mathematical functions. Hence, properties of the members of F cannot be deduced, nor can propositions pertaining to them be proven, on the basis of mathematics.

There exists no evidence that RTM is a ubiquitous phenomenon. Regardless of whether the focus is on members of T or F, the determination of whether distribution functions are RTM-E consistent must be made on a case-by-case basis. Each distribution function of interest must be assessed for conformity to RTM-E.1 and RTM-E.2. For class T, the determination of whether or not a given member is an element of T_{RTM} or T_{RTMI} is made via mathematical analysis. It follows, then, that a determination of the *relative sizes* of the subclasses T_{RTM} and T_{RTMI} must be made via mathematical analysis. For class F, on the other hand, the determination of whether or not a given member is an element of F_{RTM} or F_{RTMI} is made via empirical investigation. Accordingly, a determination of the relative sizes of the subclasses F_{RTM} and F_{RTMI} must be made via empirical investigation (i.e., by inferential investigations of the $F_{x,y}$ produced under many different P_T/phenomenon pairings).

As we indicated earlier in the paper, to our knowledge, there have not been carried out *any* studies in which determinations have been made about whether particular bivariate distributions are RTM-E consistent. We cannot see how there could have been, given that, until now, Galton's definition of the concept *regression towards the mean* has been buried, and that without a definition it is not possible to judge whether a distribution has the RTM property. Consequently, virtually nothing is known about the relative sizes of the subclasses T_{RTM} and T_{RTMI}, on the one hand, and the relative sizes of the subclasses F_{RTM} and F_{RTMI}, on the other (hence, whether or not it is true that RTM-E-consistent distributions are ubiquitous). In the absence of evidence relevant to a resolution of these issues, it is simply unscientific to presume, as many seem to do, that RTM is a ubiquitous phenomenon.[5] The investigative work, either mathematical (in the case of class T) or empirical (in the case of class F) in nature, has yet to be carried out. A possible explanation for the existence of the unfounded view that RTM is a ubiquitous phenomenon is that, for some time now, researchers have been unquestioningly (and wrongly) taking bivariate normality and linearity of conditional mean functions, both of which imply RTM-E consistency, to be ubiquitous states of nature.

RTM is not a consequence of selecting on a variate. It has become commonplace for experts in methodology to claim that RTM is "a problem of selection" (see, e.g., Copas, 1997; Lin & Hughes, 1997). Consider any bivariate distribution: that is, any element of T or F.

The concept of *selecting on X* is a synonym for *conditioning on X* (Kotz, Balakrishnan, & Johnson, 2000). The distribution of Y for a sub-population selected on X (e.g., the sub-population for which $X \leq 200$, or $30 \leq X \leq 70$, or $X = -2$, etc.) is simply the conditional distribution of Y given that X satisfies the conditions ω that define the sub-population. It is uncontroversial that, unless X and Y are statistically independent, the conditional distribution of Y given that $X \in \omega$ will not be equivalent to the unconditional distribution of Y. Thus, a study of Y carried out on a sub-population selected on X (i.e., on a conditional distribution) will likely yield very different quantitative results than an analogous study carried out on the full population (i.e., on the unconditional distribution of Y).

There have been documented many examples in which a researcher's failure to acknowledge the level of conditioning on which his or her results were based led to him or her making erroneous claims about the phenomena under study. A classical concern of statisticians has been the derivation of mathematical results linking conditional and unconditional features of distributions under particular sets of assumptions, and, in particular, the problem of estimating parameters of an unconditional distribution (say, an unconditional treatment effect) on the basis of knowledge of parameters of a conditional distribution (e.g., the treatment effect observed on a selected sub-population; see, e.g., Chuang-Stein & Tong, 1997; Mee & Chua, 1991; Ostermann, Willich, & Ludtke, 2008). While the possibility of differences in results generated under a selection on X, on the one hand, and on the full population, on the other, is a serious scientific and statistical problem, it has nothing to do with RTM. In particular, selecting on a variate is neither equivalent to, nor a cause of, RTM.

For a given element of either T or F, consider the conditional distributions of *Y(X)* defined under the *finest* level of selection on *X(Y)* possible. These distributions are simply the conditional distributions of *Y(X)* given that *X(Y)* assumes a particular value *x(y)*, and are determined, as with any conditional distribution, by the bivariate distribution of X and Y. Each conditional distribution $Y|X$=x $[X|Y=y]$ has an associated conditional mean function $\mathrm{E}(Z_y|Z_x = \delta)$ $[\mathrm{E}(Z_x|Z_y = \delta)]$. If either one or both of $\mathrm{E}(Z_y|Z_x = \delta)$ and $\mathrm{E}(Z_x|Z_y = \delta)$ satisfies both RTM-E.1 and RTM-E.2, then the bivariate distribution in question is RTM-E consistent. Otherwise, it is not.

Selecting on *X(Y)*—conditioning on *X(Y)*—could not possibly *give rise to* the property of RTM, because RTM is the condition that at least one of $\mathrm{E}(Z_y|Z_x = \delta)$ and $\mathrm{E}(Z_x|Z_y = \delta)$ satisfies both RTM-E.1 and RTM-E.2, and selecting on *X(Y)* has no impact on the conditional mean functions $\mathrm{E}(Z_y|Z_x = \delta)$ and $\mathrm{E}(Z_x|Z_y = \delta)$. The conditional mean functions $\mathrm{E}(Z_y|Z_x = \delta)$ and $\mathrm{E}(Z_x|Z_y = \delta)$ are determined by the bivariate distribution of Z_y and Z_x. Hence, whether or not at least one of the conditional mean functions $\mathrm{E}(Z_y|Z_x = \delta)$ and $\mathrm{E}(Z_x|Z_y = \delta)$ satisfies both RTM-E.1 and RTM-E.2, thereby having the RTM property, is *determined* by the bivariate distribution of Z_y and Z_x itself. Some bivariate distributions have the RTM property and others do not. Selecting on *X(Y)*, on the other hand, is an *operation* that can be carried out on any bivariate distribution, regardless of whether or not the distribution has the RTM property (one simply specifies the condition ω on X that defines the sub-population on which conditioning will occur).

Individual-level change and difference phenomena are not instances of RTM. As suggested earlier in the paper, Galton introduced into research a levels of analysis dilemma. This dilemma has now grown into an endemic misportrayal of difference and change phenomena—phenomena defined on individual score-pairs (x_i, y_i)—as *instances* of RTM. For example, Campbell and Kenny (1999) claim that "[i]t is indeed possible for the same score to regress both up and down, depending on how the sample is defined" (pp. 29–30) and that "Galton eventually realized that regression was not a biological force but an inherent feature of change" (pp. 2–3). According to Reichardt, "Galton even demonstrated convincingly that *individual* [italics added] height regresses to the mean across generations" (as cited in Campbell & Kenny, 1999, p. ix).

Consider the $F_{x,y}$ produced under a particular P_T/phenomenon pairing. For any entity $i \in P_T$, difference and change phenomena are captured by taking functions *f(.)* on the score-pair (x_i,y_i). The score $v_i = f(x_i, y_i)$ that particular entity i yields on an appropriately chosen function is an instance of individual-level change or difference phenomena. Entity i's yielding of the score v_i on function *f(.)* is the outcome of the joint action of a set of currently unknown causes, and it is the job of the scientist to reveal the identities of these causes. However, individual-level difference and change phenomena—that is, phenomena captured by an appropriately chosen function defined on the score-pairs (x_i,y_i)—*are not instances of RTM*. They are simply individual-level difference and change phenomena whose causal stories are in need of elucidation. Consider the notorious special case in which *f(.)* is taken to be the difference in extremity λ_i=|z_{xi}|-|z_{yi}|. For individual entity i, λ_i will be either positive, negative, or zero. A positive value of λ_i would indicate that entity i had been less extreme on Y than on X. It would not, however, be an instance of RTM, but, rather, of individual-level difference or change phenomena that is in need of an explanation.

RTM is a potential property of a bivariate distribution. Specifically, it signifies the special case in which both RTM-E1 and RTM-E2 are satisfied for at least one of E(Z_y|Z_x = δ) and E(Z_x|Z_y = δ). Because RTM is a potential property of a bivariate distribution, it clearly cannot be a potential property of an individual-level change or difference phenomenon: bivariate distributions are populations (aggregations) *constituted of* such individual-level phenomena.

RTM is not a basis for making predictions about temporally sequenced observations. Stigler (1997) makes the following claim: "Suppose the first score is exceptionally high—near the top of the class. How well do we expect the individual to do on the second test? The answer, regression teaches us, is 'less well,' relative to the class's performance" (p. 104). According to Dallal (2000), "The regression effect causes an individual's expected post-test measurement to fall somewhere between her pre-test measurement and the mean pre-test measurement" (p. 1). Finally, the description provided by Wikipedia runs as follows: "[R]egression toward the mean refers to the phenomenon that a variable that is extreme on its first measurement will tend to be closer to the centre of the distribution on a later measurement" (Regression Toward the Mean n.d.). The idea captured in these quotes is the following: at time 1, individual entity $i \in P_T$ yields score x_i and this score reflects its position with respect to a particular phenomenon; it is known that RTM, while unobservable, will impact upon entity i's expression of the phenomenon in such a manner as to make it less extreme; a subsequent measurement y_i of entity i's position with respect to the phenomenon will therefore be closer to the population mean than it otherwise would have been; knowing that this is the manner in which RTM operates, it is only reasonable to predict that an initial measurement x_i will be more extreme than a subsequent one y_i.

There are many features of this portrayal of RTM that are confused. In the first place, it rests on the tacit assertion that bivariate distribution functions are standardly RTM-E consistent. In fact, the proportion of distribution functions that have the RTM property is currently unknown. More fundamentally, a predictive case is set with respect to a *particular* bivariate distribution, and each particular bivariate distribution either is or is not RTM-E consistent. Assessments of RTM-E consistency must be made on a case-by-case basis. Thus, generalities of the type offered up in the above quotes amount to methodological dogma.

In the second place, RTM is a property of a bivariate state of affairs *that has already happened*. A distribution is RTM-E consistent only if at least one of $E(Z_y|Z_x = \delta)$ and $E(Z_x|Z_y = \delta)$ satisfies both RTM-E.1 and RTM-E.2, and these conditional mean functions are not even defined unless there is already in existence a population of N score-pairs (x_i,y_i). Thus, it is nonsensical to envision RTM as existing contemporaneously with X, but not Y, and as exerting causal influence so as to produce y_i that are less extreme than were the corresponding x_i. Thirdly, even if it were the case that RTM could exist contemporaneously with X, but not Y, RTM is a static potential property of a conditional mean function. As such, it possesses no causal force, and so cannot bring about change, predictable or otherwise.

If it *had* been established (rather than presumed) that an $F_{x,y} \in F$ was RTM-E consistent, then certain kinds of non-temporal prediction-like claims could be justifiably made. For example, if it were known that the score on Z_x for a particular entity $i \in P_T$ was equal to δ^*, the minimum *mean square error* prediction of entity i's score on Z_y would be equal to $E(Z_y|Z_x = \delta^*)$. Because, from RTM-E.1, $|E(Z_y|Z_x = \delta^*)| \le |\delta^*|$, an assertion to the effect that "entity $i \in P_T$ is less extreme on Z_y than it is on Z_x" would be in keeping with the minimum mean square error prediction. However: (a) there would be no *necessity* that entity $i \in P_T$ was less extreme on Z_y than it was on Z_x and (b) it is not clear why this brand of prediction would be of any importance, for to establish that $F_{x,y}$ is RTM-E consistent would require that one *knew* the values on Z_x and Z_y yielded by all entities $i \in P_T$. Thus, for entity $i \in P_T$, with $Z_x = \delta^*$, one would not need to predict anything, but, rather, would *know* entity i's score on Z_y, hence could *determine* whether or not entity i was less extreme on Z_y than it was on Z_x.

RTM is not an explanation of anything. Equal in popularity to the misportrayal of individual-level change and difference phenomena as instances of RTM is the misportrayal of RTM as an *explanation* of such phenomena. Campbell and Kenny (1999) list many examples of individual-level change and difference phenomena that, according to them, can be explained by RTM, including the phenomenon of movie sequels being less outstanding than the originals that spawned them and that of high school students who have performed poorly on an aptitude test improving after taking an SAT course. A sporting example of theirs reads as follows:

> Baseball pundits have given several explanations of the sophomore jinx [the phenomenon by which professional baseball players awarded Rookie honours in their first year fail to perform similarly well in subsequent seasons]. One is that in the second year the pressure has been increased by winning the award and that creates performance anxiety. A second explanation is that the motivation to play well declines in the second year. . . . However, the strongest and most plausible account of the jinx is regression toward the mean. (pp. 42–44)

On the received account,

> Regression toward the mean is only one of several plausible rival hypotheses of change over time. History, maturation, instrumentation, and testing are potentially plausible explanations of "change." However, regression toward the mean is perhaps the most pernicious plausible rival hypothesis because it is universal. (Campbell & Kenny, 1999, p. 51)

Along similar lines, RTM is said to be a compelling candidate explanation for a person's movement from an extreme position in a group to a less extreme position (Fitzmaurice, 2000).

If individual $i \in P_T$ happened to have had a rookie batting average of .314, and to have dropped to an average of .214 in his sophomore season, then his performance would have declined by $(x_i\text{-}y_i) = .100$ raw units. If his performance happened to have declined in relative terms as well, then $z_{xi}\text{-}z_{yi}$ would also be positive. Any time that change has been observed to occur, it has occurred because a constellation of (at least initially) unknown causal factors brought it about. An intensive scientific effort might eventually result in the formulation of an individual-level explanatory law that describes the joint action of these causes in bringing about the change. As with human performance phenomena in general, the causal story underlying a decline in batting average is undoubtedly a complicated one. However, contra Campbell and Kenny (1999), it is precisely factors such as performance anxiety, motivational elements, and physiological characteristics, and certainly *not* RTM, that are candidates for inclusion in a law that explains the decline in the batting averages of those baseball players who experience such a decline. It is confused to offer up RTM as "the most plausible account of the jinx" because RTM is that property of an element of T or F that obtains when at least one of $E(Z_y|Z_x = \delta)$ and $E(Z_x|Z_y = \delta)$ satisfies both RTM-E.1 and RTM-E.2, and conditional mean functions do not impact upon the performances of baseball players any more than do modes, medians, and partial correlation coefficients.

The fact that commentators have misportrayed RTM as an explanatory device does not mean that explanations have no place in the science of RTM. If RTM has been established to be a property of an element of F, then it is the job of the scientist to explain why this state of affairs came to be. It must be noted, however, that to formulate an explanation as to *why* RTM is a property of the distribution function in question will be an even greater challenge than to formulate an individual-level law. This is because RTM is a population-level phenomenon and, as such, its explanation requires (a) knowledge of the individual level law $z_{y_i} = f\left(z_{x_i}, \underline{\omega}\right)$ relating z_{y_i} to z_{x_i} and (b) knowledge of a law that explains the particular distribution of z_{x_i} that, through the individual-level law, yields the RTM-E-consistent proportions $P\left(Z_{x_i} \le a, Z_{y_i} \le b\right) = P\left(Z_{x_i} \le a, f\left(Z_{x_i}, \underline{\omega}\right) \le b\right)$ that constitute the RTM-E-consistent $F_{x,y}$. In other words, an explanation of the existence (or absence) of RTM-E in P_T would be roughly equivalent to an explanation of why the joint empirical distribution $F_{x,y}$ arises under the individual-level law $z_{y_i} = f\left(z_{x_i}, \underline{\omega}\right)$.

RTM is neither a force, nor a causal mechanism. A sub-species of the "RTM as explanation" mythology misportrays RTM as a force or causal agent that affects the scores yielded by individual entities. Consider, for example, the following quotes:

> The primary effect of this is to affect scores on retesting so that they are closer to the population mean. (Streiner, 2001, p. 72)

> By virtue of RTM, we can expect to see a mean reduction from pretreatment to posttreatment, regardless of the efficacy of the treatment. (Fitzmaurice, 2000, p. 81)

> [T]he net effect of regression toward the mean is to shift a selected group's average (either high or low) closer to the mean of the entire population. (Streiner, 2001, p. 75)

> Galton initially viewed RTM as a biological fact. It is true that biology needs to overcome RTM. . . . [The form of inheritance that humans have, whereby a child receives half of her chromosomes from each parent,] prevents species from regressing toward the mean and so promotes biodiversity. (Campbell & Kenny, 1999, pp. 17–18)

The idea manifested in these quotes is that RTM is a constituent of reality, a force, that exists independently of the entities $i \in P_T$ and the score-pairs (x_i, y_i) that they yield, and that inexorably draws the y_is closer to the population mean than were the x_is. We will simply remind the reader that RTM is a static, potential property of a bivariate distribution, and, hence, is no more a force or causal mechanism than is a mode, median, or Pearson product–moment correlation. It does not *affect* anything. It has no effects.

The final quote exemplifies the depths of confusion to which science can sink when under the sway of a mythology. When a particular entity $i \in P_T$ happens to be a living entity, the score-pair (x_i, y_i) it yields, and, hence, the value it yields on the function $|z_{xi}|$-$|z_{yi}|$, is brought about, in part, by particular of its anatomical and physiological characteristics (i.e., its biology). By this it is meant that features of the biology of individual entity i are causally responsible, to some unknown degree, for entity i's having yielded the particular value on the score-pair (x_i, y_i) that it did, in fact, yield. Now, imagine pooling all of the individual-level score-pairs (x_i, y_i), one pair for each entity $i \in P_T$, each score-pair having been determined to some degree by the biology of the entity that yielded it. It is this *collection* of score-pairs that either is or is not RTM-E consistent.

Thus, while biology stands for a set of processes operating at the level of the individual, RTM, when it arises, is a static, aggregate property of a collection of score-pairs. As such, it possesses no causal force. Clearly, then, it is incoherent to envision biology (causally operative at the level of the individual entity) as having to overcome RTM (a potential static property of a population of score-pairs). Biology and RTM are not causal competitors, but, rather, completely different kinds of things.

RTM is not a statistical artifact (it is a property of a conditional mean function). The reader frequently reads comments to the effect that "RTM is an artifact that as easily fools statistical experts as lay people" (Campbell & Kenny, 1999, p. xiii) or that "[a]ll too often the statistical fact of RTM is given a substantive meaning that is unwarranted" (Campbell & Kenny, 1999, p. 19) or that "[r]egression toward the mean is . . . not a true process working through time but a methodological artifact" (Campbell & Kenny, 1999, p. 18). All of this flies on the wings provided by the misportrayal of RTM as a rival hypothesis or competitor for the role of causal explanation of some phenomenon. In fact, there is one set of conditions under which RTM could sensibly be said to be artifactual, and the components of this set are: (a) a particular $T_{x,y} \in T$ is RTM-E consistent; (b) $T_{x,y}$ is employed by a researcher as a formal representation of the $F_{x,y}$ yielded under a particular P_T/phenomenon pairing; and (c) $T_{x,y}$ is, in fact, an inadequate representation of $F_{x,y}$. In such a case, the claim that RTM is a discovered property would follow not from the state of reality, but, rather, from the choice of $T_{x,y}$ as a representation of $F_{x,y}$. Under this set of conditions, the tool of representation can be said to have cast shadows upon reality.

Inappropriate representational practices notwithstanding, there is nothing whatsoever artifactual about RTM. A bivariate distribution, either an element of T or an element of F, is or is not RTM-E consistent. When mathematical proof establishes an element of T, or empirical investigation establishes an element of F, to be RTM-E consistent, then a discovery has been made about the aggregate behavior of a particular population of score-pairs, and, in particular, the forms of the conditional mean functions of this population.

RTM is not a threat to the making of valid causal ascriptions. Of all of the components of the modern mythologization of RTM, perhaps the most commonly expressed is that RTM plays havoc with the making of the causal ascriptions that are the standard products of experimentation. An entire book, Campbell and Kenny's *A Primer on Regression Artifacts* (1999), has been written on the subject of the dangers posed by the ever-present rival hypothesis that is RTM. On the received account, "the regression fallacy is the most common fallacy in the statistical analysis of . . . data" (Friedman, 1992, p. 2131). Streiner (2001) provides an example of this fallacy that, with respect to its tone and message, is standard fare:

> In another study, patients with schizophrenia are selected if their scores on a measure of social functioning are below some criterion. They are entered into a program emphasizing social skills training, work-appropriate behavior, and independent living. At the end of 6 months, most of the patients have significantly higher scores on the scale, and the investigators conclude that this intervention is highly successful with these people…Are the researchers justified in their enthusiasm for these treatments? The answer is a resounding "No," for a multitude of reasons . . . a . . . possible explanation for the results [is] "regression toward the mean." (p. 73)

On the received account, even the employment of a sound experimental design cannot overcome the corruptive influence of RTM. As Campbell and Kenny (1999) explain:

> Even if the sample is randomly selected from the population, there is still regression toward the mean. Random samples have the desirable feature of their means being relatively near the population mean and so there is less regression toward the mean—but there is still some regression. (p. 51)

> Even with a control group, regression toward the mean still creates major interpretive problems. (p. 52)

This component of the received account is confused: RTM can neither interfere with the making of causal ascriptions, nor threaten the correctness of conclusions, causal or otherwise, that are made on the basis of experiments. Entity $i \in P_T$ can be said to have changed just when (x_i-y_i) does not equal zero. When $|z_{xi}|-|z_{yi}|>0$, entity i can be said to have a less extreme position within the distribution of variate Y than it does within the distribution of variate X. As noted previously, any particular instance of change is brought about by a constellation of causes whose joint operation is largely unobservable. It is the job of the scientist to discover the processes, forces, and entities that were jointly responsible for bringing about this change. Because the action of the constellation of causes is not observable, it is always a risky business to nominate particular constituents of reality as belonging to this constellation.

The researcher's ally in this dangerous game of causal ascription is sound experimental design. In fact, the degree of confidence assignable to a particular ascription is largely determined by the particulars of the experimental design under which the score-pairs (x_i,y_i) were produced. Sound experimental designs are those designs that enable the researcher to eliminate many candidate causes, leaving only a few that remain plausible. Weak designs are those that eliminate very few candidates, thus rendering the researcher's ascription of causal status to particular factors arbitrary, if not simply fatuous.

Let us return to Streiner's (2001) example, which can be unpacked as follows: (a) the population under study is selected to be those individuals possessing social functioning scores beneath some criterion; (b) during a 6-month period, these individuals receive an intervention; (c) at the end of the 6-month period, the individuals are scored a second time with respect to their level of social functioning; and (d) for most of the individuals, $(y_i-x_i) > 0$. In the first place, regardless of the fact that the individual entities under study were initially low with respect to the distribution of social functioning, there did not exist any necessity that they improve. They simply did. For each entity i, the scores y_i and x_i were brought about by a constellation of causes, the elements of which are unknown. If, as in Streiner's example, most of the entities yielded $(y_i-x_i)>0$, then it was the joint causal action of the elements of this constellation that brought this condition about, and the scientist's task is to discover the identities of these elements.

As indicated by Streiner, it would, indeed, be a serious mistake to ascribe causal status to the intervention, but not because of anything to do with RTM. The problem with ascribing causal status to the intervention described in Streiner's example is that the research design he describes is a weak one, and does not provide a basis for eliminating rival candidate causes such as mortality and age, among countless others. RTM, however, is not the right *kind* of thing to be a candidate rival hypothesis. To correctly claim that RTM has arisen under a particular P_T/phenomenon pairing is to claim that one or both of $E(Z_y|Z_x = \delta)$ and $E(Z_x|Z_y = \delta)$ are a certain way, namely that at least one of these conditional mean functions satisfies both RTM-E.1 and RTM-E.2. That is to say, it is to claim that a very particular *state of affairs* has come into existence, namely that: (RTM-E.1) for any particular degree of extremity $|\delta|$ of X, the average extremity of Y is less than $|\delta|$; and (RTM-E.2) the magnitude of the difference between $|\delta|$ and the average extremity of Y is an increasing function of $|\delta|$.

States of affairs such as the values of the medians of a bivariate distribution happening to be equal to 11.5 and 2.3, or the value of the Pearson product–moment correlation happening to be equal to .45, or RTM happening to be a property of a particular bivariate distribution, have no causal powers whatsoever, and, hence, can have no impact upon the individual entities $i \in P_T$. The particular state of affairs that is RTM does not *explain* anything about individual-level difference and change phenomena, but is, rather, the kind of thing that, itself, is in need of an explanation (as would be a Pearson product–moment correlation that turned out to be .45).

Nothing necessitates that RTM, the state of affairs in which at least one of $E(Z_y|Z_x = \delta)$ and $E(Z_x|Z_y = \delta)$ satisfies both RTM-E.1 and RTM-E.2, be a property of the bivariate distribution produced under a particular P_T/phenomenon pairing. When RTM happens to be a property of the bivariate distribution produced under a particular P_T/phenomenon pairing, it happens to be so by virtue of the fact that the entities $i \in P_T$ yielded an RTM-consistent population of score-pairs. And the entities $i \in P_T$ yielded an RTM-consistent population of score-pairs as a result of the joint action of a constellation of causal factors. To ask *why* it is that RTM obtains under a particular P_T/phenomenon pairing is to inquire as to the identities of these causal factors. RTM does not make anything happen; it is a state of affairs that happens (or doesn't). Thus, it is deeply confused to view RTM as a rival hypothesis, something that could, unbeknownst to the researcher, have brought about observed change phenomena, the cause of which was then wrongly identified to be an intervention.

Conclusion

Within the domain of quantitative methodologies, mythologies flourish under a particular set of ingredients that include, but are not limited to, conceptual equivocation (or, in the worst cases, the total failure to pin down key concepts with definitions), the misidentification of related concepts, the dogmatic adherence to favored statistical props, the failure to clarify relationships between statistical tools of representation and the components of empirical settings that they were designed to represent, and the projection onto empirical reality of mathematical necessities that follow from the distribution theory in play when distribution functions are employed.

While the mythology that is the received account of regression towards the mean features all of these ingredients, the most damaging of all is the fact that, while Galton provided the necessary ingredients, no definition of the concept *regression towards the mean* was ever laid down. This meant that contributors to the literature on regression towards the mean could each form their own loose feelings about what they meant by the notion. Some took RTM to be the empirical (non-necessary) proposition that the second of two ordered scores will be less extreme than the first, while others took it to be a law to the same effect. Still others took RTM to be a mathematical necessity that followed from linearity, or bivariate normality, or under selection. The ultimate consequence of this conceptual equivocation was that a claim made about the phenomenon of regression towards the mean could not be adjudicated for its correctness. Hence, along with claims that would turn out to be correct, claims that are incorrect or even incoherent have lingered, and, through frequent repetition, have come to attain the status of unquestioned soundbites. Long-term allegiance to the dogma that bivariate normality and linearity are descriptions of nature led to a blurring of the boundary between mathematical necessity and empirical fact, thereby providing additional nutrients for the growth of mythology. And the production and maintenance of a mythology such as the received account of regression towards the mean undermines the doing of sound science.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. When the topic of discussion is the causal relation between parent and offspring height, the parameterization of height measurements is arbitrary. That is to say, a causal law expressed in terms of raw scores x and y and one expressed in terms of standard scores z_x and z_y provide equivalent characterizations of the causal relationship between the two phenomena of interest.
2. For example, virtually all of the papers in the 1997 special edition of *Statistical Methods in Medical Research* dedicated to the issue of regression towards the mean took linearity (or bivariate normality, which implies linearity) as a starting point. And while linearity or bivariate normality are *mathematically* convenient, there is neither any necessity that bivariate empirical phenomena be normally distributed, nor that they have linear conditional mean functions.
3. For the sake of brevity, mathematical details have been omitted. Please contact the corresponding author for an appendix in which a detailed description of these distributions is provided.
4. This section's title is a nod to Louis Guttman's 1977 paper "What Is Not What in Statistics."
5. "RTM is as inevitable as death and taxes" (Reichardt, 1999, p. ix); "Regression is ubiquitous in medical research and can very easily lead the unwary researcher astray" (Fitzmaurice, 2000, p. 81); "Over-time correlations are less than perfect because people change, and these changes imply that regression toward the mean is an omnipresent phenomenon" (Campbell & Kenny, 1999, p. 19).

References

Campell, D., & Kenny, D. (1999). *A primer on regression artifacts*. New York, NY: Guilford.

Chesher, A. (1997). Non-normal variation and regression to the mean. *Statistical Methods in Medical Research, 6,* 147–166.

Chuang-Stein, C., & Tong, D.M. (1997). The impact and implication of regression to the mean on the design and analysis of medical experiments. *Statistical Methods in Medical Research, 6,* 115–128.

Copas, J. (1997). Using regression models for prediction: Shrinkage and regression to the mean. *Statistical Methods in Medical Research, 6,* 167–183.

Dallal, G. (2000). The regression effect/The regression fallacy. Retrieved from http://www.jerrydallal.com/LHSP/regeff.htm

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. New York, NY: Cambridge University Press.

Fitzmaurice, G. (2000). Regression to the mean. *Nutrition*, *16,* 81–82.

Freedman, D.A. (1985). Statistics and the scientific method. In W.M. Mason & S.E. Fienberg (Eds.), *Cohort analysis in social research: Beyond the identification problem* (pp. 343–390). New York, NY: Springer-Verlag.

Friedman, M. (1992). Do old fallacies ever die? *Journal of Economic Literature*, *30*, 2129–2132.

Galton, F. (1885). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, *15*, 246–263.
Galton, F. (1889). *Natural inheritance*. London, UK: Macmillan.
Guttman, L. (1977). What is not what in statistics. *The Statistician, 26*, 81–107.
Karylowski, J. (1985). Regression toward the mean effect: No statistical background required. *Teaching of Psychology*, *12*, 229–230.
Kotz, S., Balakrishnan, N., & Johnson, N. (2000). *Continuous multivariate distributions: Vol. 1. Models and applications* (2nd ed.). Toronto, Canada: Wiley.
Lamiell, J. (1998). "Nomothetic" and "idiographic": Contrasting Windelband's understanding with contemporary usage. *Theory & Psychology*, *8*, 23–38.
Lin, H., & Hughes, M. (1997). Adjusting for regression toward the mean when variables are normally distributed. *Statistical Methods in Medical Research, 6,* 129–146.
Mee, R.W., & Chua, T.C. (1991). Regression toward the mean and the paired sample t-test. *The American Statistician*, *45*, 39–42.
Nesselroade J., Stigler, S., & Baltes, P. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*, *88,* 622–637.
Orwell, G. (1949). *Nineteen eighty-four: A novel.* London, UK: Secker & Warburg.
Ostermann, T., Willich, S., & Ludtke, R. (2008). Regression toward the mean—A detection method for uknown population mean based on Mee & Chua's algorithm. *BMC Medical Research Methodology*, *8,* 1–9.
Regression toward the mean. (n.d.). In *Wikipedia.* Retrieved from http://en.wikipedia.org/wiki/Regression_toward_the_mean
Reichardt, S. (1999). Introduction. In D. Campbell & D. Kenny, *A primer on regression artifacts* (pp. ix-xi). New York, NY: Guilford.
Stigler, S. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research*, *6,* 103–114.
Streiner, D.L. (2001). Regression toward the mean: Its etiology, diagnosis and treatment. *Canadian Journal of Psychiatry*, *46,* 72–76.
Wachsmuth, A., Wilkinson, L., & Dallal, E. (2003). Galton's bend: An undiscovered nonlinearity in Galton's family stature regression data and a likely explanation based on Pearson and Lee's stature data. *The American Statistician*, *57*, 190–192.

Michael D. Maraun is a Professor in the Department of Psychology at Simon Fraser University. His work involves the identification and dissolution of confusions and incoherences that arise in the doing of empirical science. His most important work to date is his unpublished manuscript, Myths and Confusions: Psychometrics and the Latent Variable Model. This manuscript can be found online at http://www.sfu.ca/~maraun/Mikes%20page-%20Myths%20and%20Confusions.html Address: Simon Fraser University, Department of Psychology, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada. [email: maraun@sfu.ca]

Stephanie Gabriel is an M.Sc candidate in the Department of Statistics and Actuarial Science at Simon Fraser University. Her research interests include philosophy of statistics and the history of the Bayes/Frequentist controversy. Address: Simon Fraser University, Statistical and Actuarial Science, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada. [email: sgabriel@sfu.ca]

Jack Martin is Burnaby Mountain Endowed Professor of Psychology at Simon Fraser University. His research interests are in philosophy and history of psychology, social-developmental psychology, and educational psychology, with particular emphasis on the psychology of selfhood and personhood. Address: Simon Fraser University, Department of Psychology, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada. [email: martina@sfu.ca]