Regression to the Mean in Flight Tests

Reid Dorsey-Palmateer and Gary Smith

Department of Economics

Pomona College

Claremont, California 91711

contact: Gary Smith, Department of Economics, Pomona College, Claremont, California 91711; 909.607.3135; fax 909.621.8576; gsmith@pomona.edu

* We are very grateful to Cor van der Wal and Lieutenant Corey Johnston for their invaluable help in obtaining and explaining the flight training data used in this paper.

Regression to the Mean in Flight Tests

Abstract

Kahnemann and Tversky report that flight trainees who do well on a maneuver typically do not do as well on the next maneuver. Because they did not understand regression to the mean, the flight instructors attributed this regression to the praise the pilots received—leading to the perverse conclusion that pilots who do well should be criticized. Kahnemann and Tversky do not report any actual data to support this memorable anecdote. This paper uses U. S. Navy flight training data to demonstrate that there is substantial regression to the mean in pilot performances. We also show how these flight scores can be used to assess changes in a pilot's ability as the training proceeds, taking into account the anticipated regression to the mean.

key words: regression to the mean, testing

Regression to the Mean in Flight Tests

Sir Francis Galton (1886) observed regression toward the mean in his seminal study of the heights of parents and their adult children. Observed heights are influenced by genes—tall parents tend to have tall children—but are not determined completely by genes—siblings are not all the same height. Thus persons who are, say, 78 inches tall generally have “tall genes” but may have been pulled above or below their genetically predicted height by environmental factors. The former is more likely because there are many more people with genetically predicted heights below 78 inches than with genetic heights above 78 inches. Thus the observed heights of tall persons usually overstate the genetic heights that they inherit from their parents and pass on to their children. A similar argument applies to relatively short parents.

The educational testing literature provides a theoretical framework for explaining regression to the mean (Kelley 1947; Lord and Novick 1968). A person’s true score is the statistical expected value of their score on a test; the difference between a person’s observed score and true score is called the error score. Those who score the highest on a test are likely to have positive error scores since it would be unusual for someone to score below their true score and still have the highest score on a test. Since a score that is high relative to the group is also likely to be high relative to this person’s true score, their score on another test—either before or after—is likely to regress toward the mean.

Similarly, athletic performances are an imperfect measure of skills and consequently regress. Schall and Smith (2000) looked at major league baseball players who had at least 50 times at bat or 25 innings pitched in two consecutive seasons. Of 4026 players who had batting averages of .300 or higher in any season, 80% did worse the following season. Of 3849 players who had earned run averages of 3.00 or lower in any season, 80% did worse the following season.

In a classic paper, Kahnemann and Tversky (1973) wrote that “regression effects are all about us. In our experience, most outstanding fathers have somewhat disappointing sons, brilliant wives have duller husbands, the ill-adjusted tend to adjust and the fortunate are eventually stricken by ill luck. In spite of these encounters, people do not acquire a proper notion of regression. First, they do not expect regression in many situations where it is bound to occur. Second, as any teacher of statistics will attest, a proper notion of regression is extremely difficult to acquire. Third, when people observe regression, they typically invent spurious dynamic explanations for it."

A specific example Kahnemann and Tversky cite is a problem they gave graduate students in psychology, which described the actual experience of one of the authors in advising the Israeli air force:

> A problem of training. The instructors in a flight school adopted a policy of consistent positive reinforcement recommended by psychologists. They verbally reinforced each successful execution of a flight maneuver. After some experience with this training approach, the instructors claimed that contrary to psychological doctrine, high praise for good execution of complex maneuvers typically results in a decrement of performance on the next try. What should the psychologist say in response?

The flight instructors were unaware of the regression argument and none of the graduate students suggested regression as a possible explanation for the observed data: "The respondents had undoubtedly been exposed to a thorough treatment of statistical regression. Nevertheless, they failed to recognize an instance of regression when it was not couched in the familiar terms of the heights of fathers and sons." Kahnemann and Tversky conclude, "This true story illustrates a saddening aspect of the human condition. We normally reinforce others when their behavior is good and punish them when their behavior is bad. By regression alone, therefore, they are most likely to improve after being punished and most likely to deteriorate after being rewarded.

Consequently, we are exposed to a lifetime schedule in which we are most often rewarded for punishing others, and punished for rewarding.”

Kahnemann and Tversky do not report any actual data to demonstrate the regression that underlies this example of how experienced observers and bright students are unaware of regression to the mean. We will use U. S. Navy flight training data to show that there is substantial regression to the mean in pilot performances, thereby confirming the memorable anecdote told by Kahnemann and Tversky. We also show how flight scores can be used to assess changes in a pilot’s ability as the training proceeds, taking into account the anticipated regression to the mean—a point that was not made by Kahnemann and Tversky and is no doubt not sufficiently appreciated by most instructors.

A Model

A pilot’s ability □ is the statistical expected value of his or her flight score. The distribution of abilities across pilots has mean □ and standard deviation $\square_{\square}^{2}$. We assume that a pilot’s score X on any particular test differs from ability by an independent and identically distributed error term (or “error score”) □:

$$X = \square + \square \tag{1}$$

If the error terms are independent of abilities, then the variance of scores across pilots is equal to the variance of abilities across pilots plus the variance of the error term and is therefore larger than the variance of abilities:

$$\square_{X}^{2} = \square_{\square}^{2} + \square_{\square}^{2}$$

Because there is more variation in scores than in abilities, observed differences in scores typically overstate the differences in abilities.

A test’s reliability is gauged by the squared correlation $\square^{2}$ between scores and abilities, which equals the ratio of the variance of abilities to the variance of scores:

$$□^2 = \frac{□_{□}^2}{□_{X}^2} \tag{2}$$

The correlation between scores and abilities would be 1 if the standard deviation of the error score were 0, and approaches 0 as the standard deviation of the error score becomes infinitely large.

If we knew each pilot’s ability, Equation 1 could be used to make unbiased predictions of each pilot’s score. However, we are interested in the reverse question—using scores to estimate abilities—which can be answered by Bayes’ Rule. Recognizing that the error term represents the cumulative effects of a great many omitted variables and appealing to the central limit theorem, we assume that the error term in Equation 1 is normally distributed. A convenient conjugate prior for a pilot’s ability is therefore a normal distribution with mean $□_0$ and standard deviation $□_{□}^2$. Our estimate □ of the pilot’s ability is the mean of the posterior distribution for □, which is a weighted average of the prior mean and the pilot’s test score:

$$□ = (1 □ □^2)□_0 + □^2 X \tag{3}$$

If we have no information about a pilot’s ability prior to the test, a natural prior mean is the average ability of the pilots, $□_0 = □$. If, in addition, the mean test score $\overline{X}$ is close to the mean ability □, we can also write the estimated deviation of a pilot’s ability □ from the mean ability □ as a fraction of the deviation of this pilot’s score from the mean score $\overline{X}$:

$$□ □ □ = □^2\left(X □ \overline{X}\right)$$

In this equation, the squared correlation between scores and ability is used to shrink each pilot’s estimated ability toward the mean. For example, if a test’s reliability is $□^2 = 0.5$, those pilots who score 1 point above the mean are estimated to have an ability 0.5 points above the mean. On a comparable test, their scores can consequently be expected to regress to the mean by averaging only 0.5 points above the mean.

Again using the mean score $\overline{X}$ as our prior mean μ_0, we can also write Equation 3 as

$$\mu = \left(1 - \rho^2\right)\overline{X} + \rho^2 X \qquad (4)$$

A pilot's estimated ability is a weighted average of the pilot's score and the mean score, using the squared correlation coefficient as the weight. Truman Kelley derived Equation 4 without using Bayes' Rule and, in classical test theory, this is known as Kelley's equation (Kelley, 1947).

In practice, we don't observe abilities and consequently cannot use data for scores and abilities to estimate ρ^2 from Equation 1. Instead, we can use the observed scores X and Y on two comparable tests to estimate ρ^2 because it can be shown that the population correlation between the two scores ρ_{XY} is proportional to the squared correlation between scores and abilities:

$$\rho_{XY} = \frac{\sigma_X}{\sigma_Y}\rho^2$$

Thus the test's reliability can be estimated from

$$\rho^2 = \frac{\sigma_Y}{\sigma_X}\rho_{XY} \qquad (5)$$

If the two tests have equal standard deviations, then $\rho^2 = \rho_{XY}$. Thus we can use the scores on comparable tests to estimate the reliability and then use Equation 4 to estimate each pilot's ability.

Alternatively, we can use least squares estimates of the following equation to predict each pilot's score on a comparable test:

$$Y = \alpha + \beta X + \varepsilon$$

With a large sample, the estimated slope approaches

$$b = \frac{\sigma_Y}{\sigma_X}\rho_{XY} \qquad (6)$$

Because the least squares line goes through the mean values of the variables, the predicted deviation of a pilot's score $\hat{Y}$ from the mean score on one test is equal to the slope times the deviation of this pilot's score from the mean score on the other test:

$$\hat{Y} - \overline{Y} = b\left(X - \overline{X}\right) \qquad (7)$$

If the two tests have the same mean, Equations 4 and 7 are equivalent.

Data

Our data come from the final stage of a six-phase aviation training track that qualifies Naval Aviators for their first operational flying tour on a fleet squadron. In this carrier qualifying phase, the pilots practice solo landings of their particular fleet aircraft (in this case, the E-2C Hawkeye, or the C-2A Greyhound) on a Naval aircraft carrier under the supervision of a landing signal officer stationed on the carrier who grades each flight and debriefs each pilot thoroughly after each flight.

In a carrier landing the pilot aims for the middle one of three arresting wires on the deck of an aircraft carrier. Here is Lieutenant Corey Johnston's description of a carrier landing:

> The day time landing pattern begins overhead the ship at anywhere from 2000' – 5000'. When aircraft are able to land, they circle and descend to establish themselves at 800' and three miles behind the aircraft carrier at a minimum of 250 knots. As the aircraft passes over the carrier, they will go into a level 2G break turn to the left to bleed energy and get to the airspeed required to drop flaps and landing gear. The aircraft will drop landing gear, flaps, and arresting hook and descend to 600' on a down wind track approximately one mile abeam the ship. When the aircraft is abeam the aft portion of the ship, the start a descending 180 degree

> left hand turn. The turn is conducted at 20-22 degrees angle of bank, starting at 200-300 foot-per-minute rate of descent. After 90 degree of turn, the aircraft will be at 450’ and increase the rate of descent to 500-600’ feet-per-minute. The pilot will maneuver the aircraft to be at 325’-375’ crossing the ships wake and begin to visually assess where he is on glideslope. This will take him to the start of the pass. The pass begins when the aircraft rolls out on centerline of the carrier....
>
> [I]t is the nose attitude of the aircraft in the landing configuration which determines the appropriate landing attitude of the carrier aircraft. In civilian and air force flying, the aircraft is flared to slow down and land. In carrier flying, the aircraft is kept at a constant attitude all the way around the turn to landing. If the aircraft nose is pitched down, we call the aircraft fast. If the nose is pitched up, we call the aircraft slow. A fast aircraft will raise the arresting hook up possibly causing the aircraft to miss all of the wires. We call this a bolter. If the aircraft does engage the wires, there is a possibility of overstressing the arresting gear motors or breaking the arresting hook off of the aircraft. If the aircraft gets slow, there is a possibility of stalling the aircraft (running out of lift over the wings) and crashing, or what we call an in-flight engagement of the arresting gear which usually results in blown tires or a broken airplane.

The possible scores are shown in Table 1. In our particular data set, there were a total of 1828 graded flights for ten different pilots tested on 14 days during the period March 29, 2004 to April 22, 2004. Seven of the pilots flew E-2C Hawkeyes and three flew C-2A Greyhounds. In our data set, the only scores given were 1, 2, 3, and 4; there were no scores of 0 (unsafe), 2.5 (bolter), or 5 (perfect).

For comparing successive flights, we separated the data by day, in that we only compared each flight score with the previous flight that day and did not compare the first flight of the day with the last flight the previous day. We assume that, while abilities vary across pilots, each pilot's ability does not change substantially between successive landings on a given day. A pilot's ability may improve with practice, but the improvement should be small in comparing two successive landings. In our data set, the average difference between successive scores is 0.01.

We also calculated each pilot's average score for the day and compared these averages across successive days. One of our research questions is whether scores, whether looking at successive flights or days, regress to the mean. Another is how we might use the daily average scores to assess changes in abilities from one day to the next, taking into account the anticipated regression to the mean.

Analysis

Overall, there were 6.2% scores of 4, 57.7% scores of 3, 30.6% scores of 2, and 5.6% scores of 1. The average score was 2.71 with a standard deviation of 0.64. Table 2 shows the means and standard deviations of the scores for each of the 10 pilots, and also the correlations between scores on successive flights and between average scores on successive days.

Table 3 shows the movements between flight scores. It was not literally true that every high score was followed by a lower score and every low score was followed by a higher score. In fact, 48% of the time, the scores on consecutive landings were identical—no doubt a reflection of the fact that although performances are undoubtedly on a continuum, only four scores were given and 88% of these scores were 2s and 3s. Still, there is clearly regression to the mean even in the limited scores in our data.

For the scores at the extreme, there is nowhere to go but toward the middle and most did: 83% of the 1s improved and 77% of the 4s worsened. Of the 2s, 7% did worse and 58% did better on the next landing. Of the 3s, 32% did worse and 7% did better on the next landing.

A least squares regression of the a flight score X on the previous flight score X(-1) gives this fitted equation, X = 2.09 + 0.21X(-1), with a 0.21 correlation between scores on successive flights. Table 2 shows that the correlation between successive flight scores for each of the pilots ranged from 0.04 to 0.27. A least squares regression of the day's average score A on the average score the previous day A(-1) gives this equation, A = 1.49 + 0.48A(-1) with a 0.51 correlation between successive scores. Because the observed standard deviations of A and A(-1) are not exactly equal, the 0.48 slope differs slightly from the 0.51 correlation.

These equations show regression to the mean for individual flights and for the daily average flight scores. Equation 7 shows that there is regression to the mean if the estimated slope of the relationship between scores is less than 1. Here, the 0.21 slope of the regression equation for individual flights means that a pilot whose flight score is 1.00 above (or below) the mean is predicted to average only 0.21 above (or below) the mean on the next flight. The 0.48 slope of the regression equation for daily average scores means that a pilot whose average score is 1.00 above (or below) the mean one day is predicted to average only 0.48 above (or below) the mean the next day. These results are consistent with the observation by Kahnemann and Tversky that good performances tend to be followed by a "decrement of performance" (and, also, that below-average performances tend to be followed by improved performances).

The regression phenomenon does not depend on the sequence of tests, but rather on the argument that persons with relatively high or low scores on *any* test are usually closer to the mean in abilities than their scores would indicate. Thus, if we were to predict a pilot's previous flight score from the current flight score, or to predict a pilot's previous day's average score from today's average score, we would also find regression to the mean. For our data, the least squares line for successive flights is X(-1) = 2.07 + 0.21X, and the least squares line for successive days is A(-1) = 1.22 + 0.54A. A pilot whose flight score is 1.00 above (or below) the mean is predicted to average only 0.21 above (or below) the mean on the previous flight. A pilot whose average score is 1.00 above (or below) the mean one day is predicted to average only 0.54 above (or below) the mean the previous day.

These data can also be used to demonstrate how to take the anticipated regression into account in assessing changes in a pilot's ability since the regression argument can be used to predict the magnitude of the drop off in above-average scores and the improvement in below-average scores. The appropriate question is whether the observed regression is larger or smaller than that predicted by purely statistical arguments. Specifically, suppose that the test's reliability for these flight scores is $\square^2$ and that a pilot's first score is X when the average first score is $\overline{X}$. We then estimate the ability of this pilot by Equation 4 using $\overline{X}$ as the prior mean for the pilot's ability:

$$\square = \square 1 \square \square^{2} \square \overline{X} + \square^{2} X$$

This pilot would demonstrate improved ability if his or her next score is above $\square$.

This argument applies to individual flights and to daily averages. We will focus on the latter. If we assume that the scoring standards are the same for each flight and for each day, then the 0.51 correlation between scores on successive days is an estimate of the test's reliability. To illustrate how a pilot's scores are adjusted for the regression phenomenon, consider a pilot whose score is 2.78 on the first training day, when the average score is 2.35. This pilot's estimated ability is 2.57:

$$\begin{aligned} \square &= \square 1 \square \square^{2} \square \overline{X} + \square^{2} X \\ &= (1 \square 0.51)(2.35) + 0.51(2.78) \\ &= 2.57 \end{aligned}$$

This pilot would demonstrate an improved ability if, on the next day, the pilot's score is above 2.57. Notice particularly that if the pilot's score falls from 2.78 to 2.71, this is actually good news in that the score did not fall as much as predicted by regression to the mean.

If we were now to make a new estimate of ability based on the latest score, the prior mean in Equation 3 is now 2.57 and we regress this new 2.71 score towards the previous 2.57 estimate of ability:

$$\begin{aligned} \square &= (1 \square 0.51)(2.57) + 0.51(2.71) \\ &= 2.64 \end{aligned}$$

Thus the 2.71 score leads us to revise our estimate of the pilot's ability upward from 2.57 to 2.64. In general, the estimate of ability is revised upward or downward depending on whether the most recent score is above or below the previous estimate of ability.

These were in fact the actual first two day's scores for Pilot 10 in Table 2. Figure 1 shows this pilot's observed scores and estimated ability through 10 days of testing. By the 10th day, the pilot's estimated ability had risen from 2.57 to 3.06. Nine of the 10 pilots showed an increase in ability over the course of the flight tests. The exception is Pilot 9, whose ability was initially estimated to be 2.45, revised upward for a while and then revised downward to 2.45, the same as on the first day of training.

Discussion

When tests are an imperfect measure of ability, scores are typically farther from the mean than are abilities. Because those with relatively high or low scores are usually closer to the mean in abilities than their scores would indicate, their scores generally regress to the mean when they are retested. The sequence doesn't matter in that those who score the highest and lowest on the retest generally had scores closer to the mean on the initial test.

Kahnemann and Tversky report that Israeli Air Force instructors observed that good performances were often followed by "a decrement of performance on the next try," which the instructors attributed to the pilots being praised but which could be explained simply by regression to the mean. Our analysis of U. S. Navy flight training data confirms that there is substantial regression to the mean in individual flight scores and in average daily scores.

Estimates of a pilot's ability should take this regression into account. Those who score the highest are probably above-average pilots, but not as far above-average as their scores suggest. Improved estimates of their ability can be obtained by taking the regression effect into account. Similarly, assessments of the effects of training on ability should take regression into account. If those pilots who were above-average slip less and those who were below-average improve more than predicted by regression to the mean, the training is succeeding.

References

Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature, Journal of the Anthropological Institute, 15, 246–263.

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

Kelley, T. L. (1947). Fundamentals of statistics. Cambridge, MA: Harvard University, 409.

Lord, F. M., & Novick, M. R. (1968). Statistical theory of mental test scores. Reading, MA; Addison-Wesley.

Schall, T. & Smith, G. (2000). Baseball players regress toward the mean, The American Statistician, 54, 231-235.

Table 1 Possible Scores

5.0	perfect pass
4.0	reasonable deviations with good correction
3.0	reasonable deviation
2.5	bolter
2.0	below average but safe pass; own waveoff; waveoff pattern, or no-grade bolter
1.0	waveoff
0.0	unsafe, gross deviation inside waveoff point

Table 2 Pilot Data

		scores		correlation	
Pilot	observations	mean	standard deviation	flights	days
1	183	2.67	0.67	0.24	0.60
2	201	2.71	0.56	0.18	0.09
3	203	2.60	0.69	0.20	0.47
4	204	2.56	0.65	0.27	0.56
5	192	2.78	0.62	0.16	0.26
6	150	2.94	0.63	0.20	0.59
7	201	2.69	0.62	0.04	0.62
8	182	2.62	0.66	0.19	-0.13
9	171	2.69	0.67	0.24	0.03
10	141	2.92	0.56	0.09	0.09
total	1828	2.71	0.64	0.21	0.51

note: The correlations are between successive flights on the same day (“flights”) and between successive daily averages (“days”)

Table 3 Movements Between Scores

			Current Score			
Previous Score	1	2	3	4	5	total
1	17	40	44	0	0	101
2	35	187	283	18	0	523
3	42	279	602	65	0	988
4	1	17	58	23	0	99
5	0	0	0	0	0	0
total	95	523	987	106	0	1711

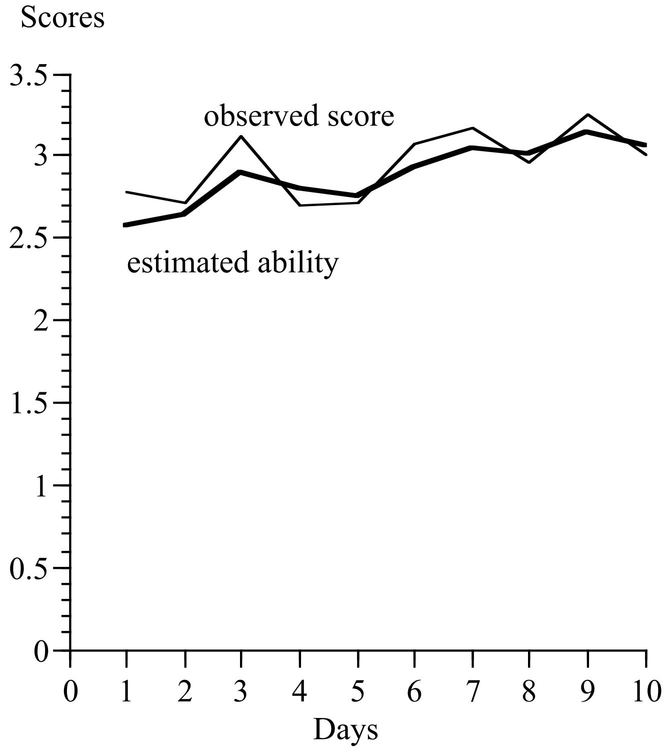

Figure 1 Observed Scores Estimated Ability for Pilot 10

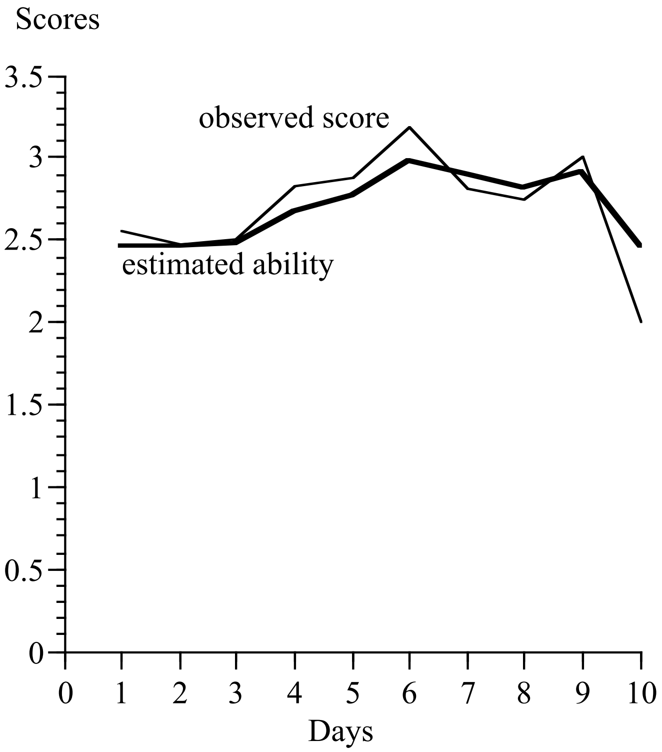

Figure 2 Observed Scores Estimated Ability for Pilot 9