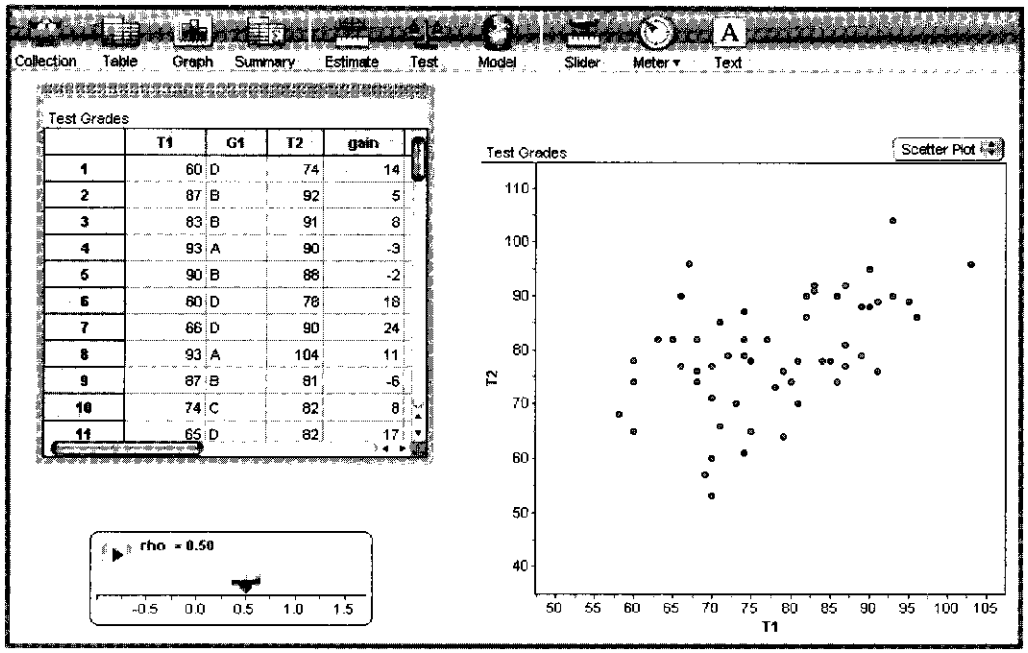
Demo 10: Regression Toward the Mean

Regression toward the mean • The meaning—and asymmetry—of the least-squares line

Let's use a context of test scores, as we did in Demo 7, "Standard Scores." Suppose you do really well on the first test of the semester. Then the second test comes around, and you don't score as high. What happened? Were you just "due" for a bad score? Yes and no. You should rebel against that idea immediately because it sounds like the "gambler's fallacy"—the one whereby you expect tails after a long run of heads. However, it is true that in imperfectly correlated data, if you're extreme on one measurement, you're likely to be less so on another. This is called *regression toward the mean.*



What To Do

▷ Open **Regression Towards the Mean.ftm**. It resembles the illustration.

The case table shows the first 11 of 60 students from a fictitious class record book. **T1** and **T2** are scores on the first two tests; **G1** is the letter grade on that first test. The attribute **gain** is simply **T2 – T1**, the amount that student improved (or declined) in test score. The slider **rho** is the correlation between the two tests. The scatter plot shows the two test scores graphed against each other.

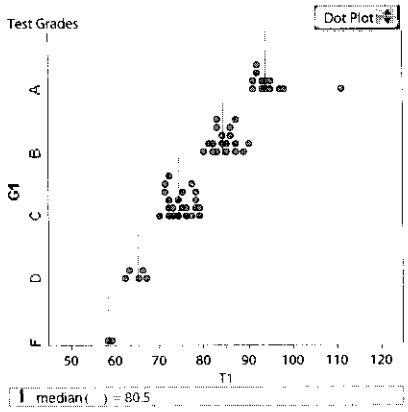
▷ Play with **rho** to see how it works; then reset it at about 0.5.

▷ Let's see what the grading scale is. Drag the attribute **G1** (grade on the first test) to the vertical axis of the graph, replacing **T2**.

▷ Let's also get some summary information on the graph. With the graph selected (it has a border), choose **Plot Value** from the **Graph** menu. The formula editor appears.

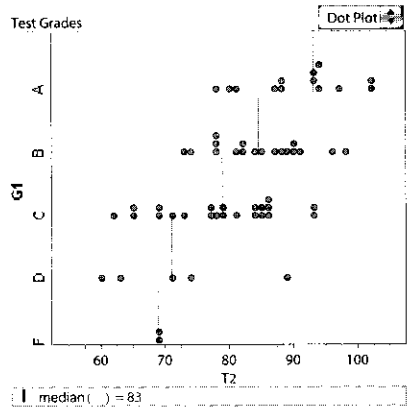
▷ Enter **median( )** and press **OK** to close the editor. Now lines appear showing the median score for each grade. Your graph should resemble the one in the illustration that follows.

This graph shows the simple grading scheme: in the 90s for an A, and so forth.



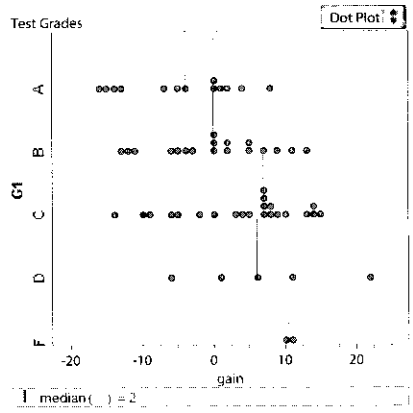
Note: Throughout this demo, feel free to rerandomize the scores. It's possible, given random variation, that some of your graphs will not show what they're supposed to the first time. Choose **Rerandomize** from the **Collection** menu.

▷ Now drag **T2** to the *horizontal* axis, replacing **T1**:



This graph shows you how the scores on the second test broke down when grouped by grades on the first test. You'll see that, as we would expect, students who did better on the first test generally did better on the second, though there are exceptions.

▷ But now drag **gain** to that horizontal axis, replacing **T2**. The trend reverses: Those who got an A on the first test had the least improvement. In fact, they had a median **gain** of −12 points.



This is the phenomenon we're trying to demonstrate. Even though the higher-scoring students on the first test will generally score higher than others on the second, the performance of that *group* will generally decline. You see this phenomenon in many different settings. In one famous data set (I first saw it in *Statistics,* by Freedman, Pisani, and Purves), you can see that, while tall fathers generally beget tall sons, the tallest men tend to have sons shorter than themselves, while the shortest men have sons taller than themselves.

Questions

1 What happens to this **gain** graph when you change **rho**?

2 When you flip coins and get six heads in a row, the chance of tails is still one-half if it's a fair coin. Yet here, it looks as if when you get a good score, you are more likely to get a lower score next time. What is it about this situation that makes it different? **Sol**

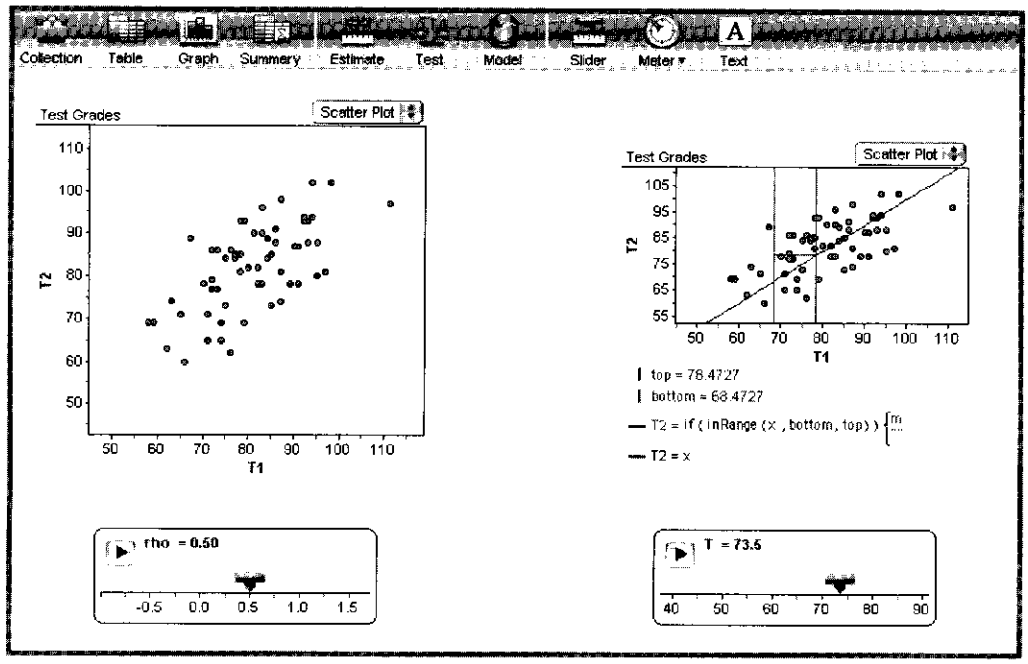
Theory Corner

If you're wondering how we generated that correlated data, you can read about it in "Simulating Correlated Data" and why that works theoretically in "How to Generate Correlated Data" in Appendix B.

Extension

Let's look at this another way, connecting the idea of regression toward the mean with that of the least-squares linear regression line.

▷ Shrink the scatter plot and drag it to the left (covering the table if you wish) to reveal another, fancy scatter plot we have hidden behind the first one. Your document will now look something like this:



This new scatter plot—of **T2** versus **T1**, as we had at the beginning—now has four lines on it. The diagonal line is the line **T2 = T1.** That is, if you're below that line, your score got worse between the two tests. Two vertical lines delimit a region, and the short *horizontal* line between them shows the average score on **T2** for people between the lines. Finally, the slider **T** determines the center of the region; when you slide it, the vertical lines move.

▷ Move the slider. Watch the level of the horizontal line. In general, you should see that it is below the diagonal line for the high scores, and above it for the low scores—which is just what we saw in the other graph.

▷ One more thing: Click on the graph to select it. Then choose **Least-Squares Line** from the **Graph** menu. The least-squares line appears.

▷ Now drag the slider again. The horizontal line will track the least-squares line. That is, *the least-squares line is the line that goes through the mean value for each vertical strip of data.*

Another observation about this last graph: When the points form a characteristic "error ellipse" as they do here, the least-squares line does *not* go along the ellipse's major axis. It's shallower. And if you switch the axes, the same effect occurs. How can that be? It's because the least-squares line is *asymmetrical;* here it's the regression of **T2** on **T1**, not the other way around.

In contrast, the calculation of the correlation coefficient is symmetrical: The correlation of y and x is the same as that of x and y.