Regression: What a Line!

I mention earlier that a line is the best way to summarize the relationship in the scatterplot in Figure 14-1. It's possible to draw an infinite amount of straight lines through the scatterplot. Which one best summarizes the relationship?

Intuitively, the "best fitting" line ought to be the one that goes through the maximum number of points and isn't too far away from the points it doesn't go through. For statisticians, that line has a special property: If you draw that line through the scatterplot, then draw distances (in the vertical direction) between the points and the line, and then square those distances and add them up, the sum of the squared distances is a minimum.

Statisticians call this line the *regression line,* and indicate it as

$$y' = a + bx$$

Each y' is a point on the line. It represents the best prediction of y for a given value of x.

To figure out exactly where this line is, you calculate its slope and its intercept. For a regression line, the slope and intercept are called *regression coefficients.*

The formulas for the regression coefficients are pretty straightforward. For the slope, the formula is

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

The intercept formula is

$$a = \bar{y} - b\bar{x}$$

Table 14-2 SAT Scores and GPAs for 20 Sahutsket University Students

Student	*SAT*	*GPA*
1	990	2.2
2	1150	3.2
3	1080	2.6
4	1100	3.3
5	1280	3.8
6	990	2.2
7	1110	3.2
8	920	2.0
9	1000	2.2
10	1200	3.6
11	1000	2.1
12	1150	2.8
13	1070	2.2
14	1120	2.1
15	1250	2.4
16	1020	2.2
17	1060	2.3
18	1550	3.9
19	1480	3.8
20	1010	2.0
Mean	1126.5	2.705
Variance	26171.32	0.46
Standard Deviation	161.78	0.82