Bare ground (%)	30	35	40	35
Concentration	800	1100	1200	1000

Using the data for steeply sloped plots, find the equation of the least squares regression line for predicting y = runoff sediment concentration using x = percentage of bare ground.

b. What would you predict runoff sediment concentration to be for a steeply sloped plot with 18% bare ground?

c. Would you recommend using the least squares regression line from Part (a) to predict runoff sediment concentration for gradually sloped plots? Explain.

SECTION 4.3 Assessing the Fit of a Line

Once the least squares regression line has been obtained, the next step is to examine how effectively the line summarizes the relationship between x and y. Important questions to consider are:

1. Is a line an appropriate way to summarize the relationship between the two variables?
2. Are there any unusual aspects of the data set that you need to consider before using the least squares regression line to make predictions?
3. If you decide that it is reasonable to use the regression line as a basis for prediction, how accurate can you expect the predictions to be?

This section looks at graphical and numerical methods that will allow you to answer these questions. Most of these methods are based on the vertical deviations of the data points from the regression line, which represent the differences between actual y values and the corresponding predicted $\hat{y}$ values from the regression line.

Predicted Values and Residuals

The predicted value corresponding to the first observation in a data set, (x_1, y_1), is obtained by substituting x_1 into the regression equation to obtain $\hat{y}_1$, so

$$\hat{y}_1 = a + bx_1$$

The difference between the actual y value, y_1, and the corresponding predicted value is

$$y_1 - \hat{y}_1$$

This difference, called a *residual*, is the vertical deviation of a point in the scatterplot from the least squares regression line. A point that is above the line results in a positive residual, whereas a point that is below the line results in a negative residual. This is shown in Figure 4.18.

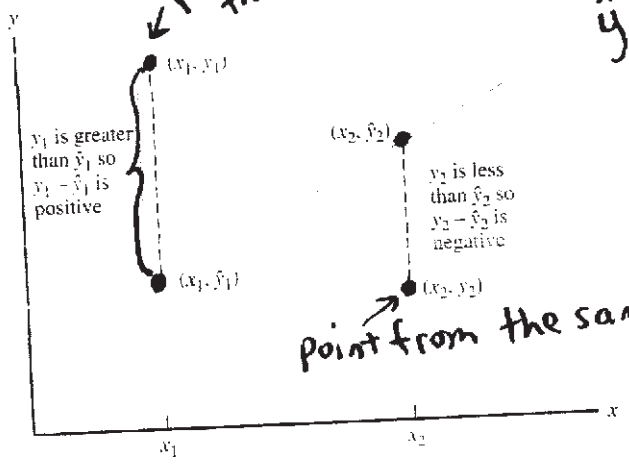
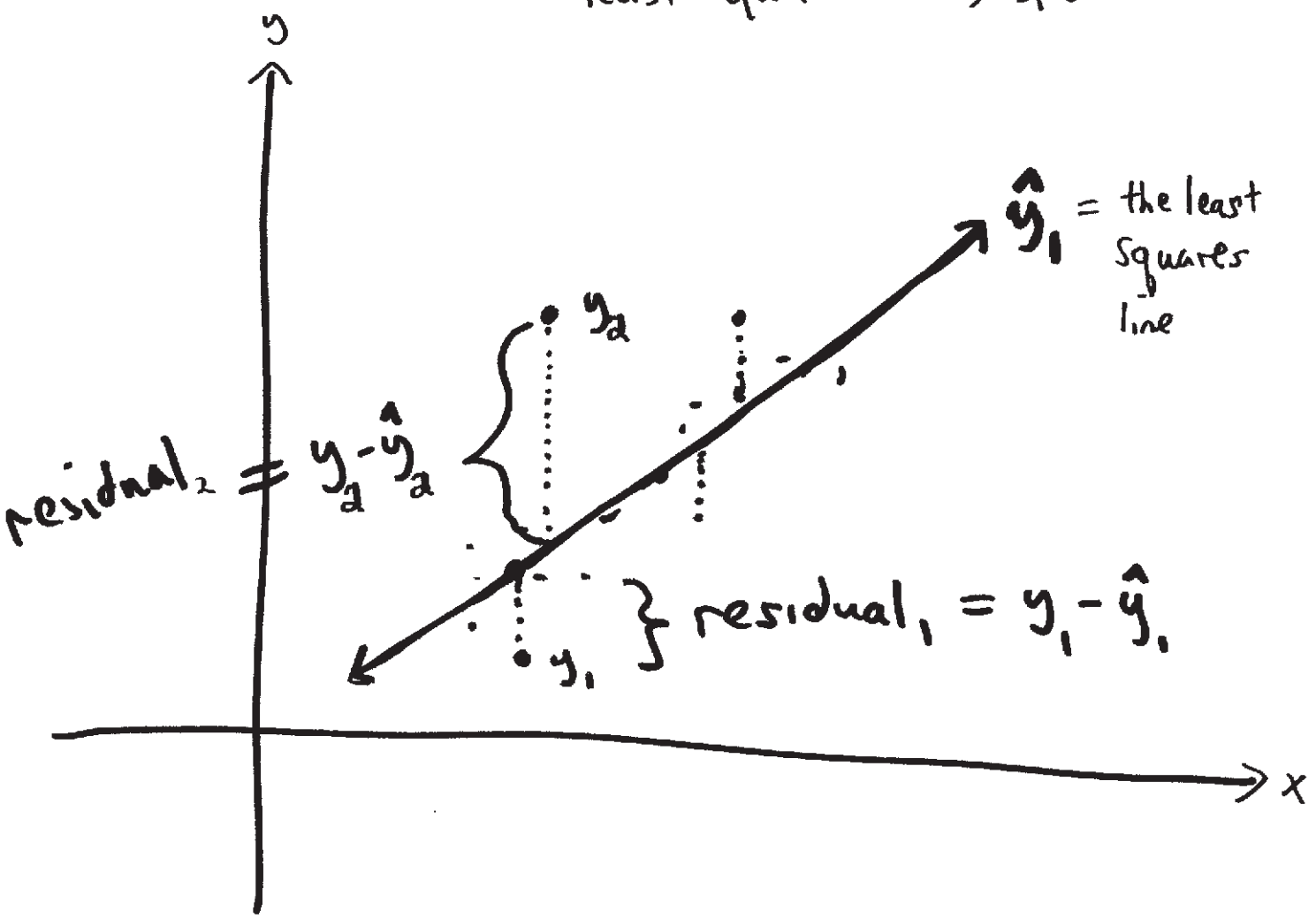


FIGURE 4.18
Positive and negative residuals

Defn: The standard deviation of the least squares regression line is a measure of how far a typical point will be above or below the least squares line, $\hat{y}_1$.



Defn: The st dev of the least squares regression line is the avg distance, or the distance you would expect a point to be away from the line.

Guideline

A small value of **Standard Deviation** indicates that residuals tend to be small. Because residuals represent $(y-\hat{y}_i)$, the difference (distance) between a predicted y-value and an observed y-value (from the sample), the value of s_e, the standard deviation of the least squares line tells you how much accuracy you can expect when using the least squares regression line to make a prediction.

Formula for stdev

$$s_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}$$

Where $\Sigma(y-\hat{y})^2$ is the sum of each residual distance, squared.

And where n is equal to the number of ordered pairs (points) in the sample.

Outliers

In some data sets, there are values (**observed data points**) called **outliers**. **Outliers are observed data points that are far from the least squares line.** They have large "errors", where the "error" or residual is the vertical distance from the line to the point.

Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called **influential points**. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line is changed significantly.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

Identifying Outliers

We could guess at outliers by looking at a graph of the scatterplot and best fit-line. However, we would like some guideline as to how far away a point needs to be in order to be considered an outlier. **As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best-fit line as an outlier.** The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure will be shown now.

The standard deviation can be used to find outlying data points.

Guideline How to find outliers:

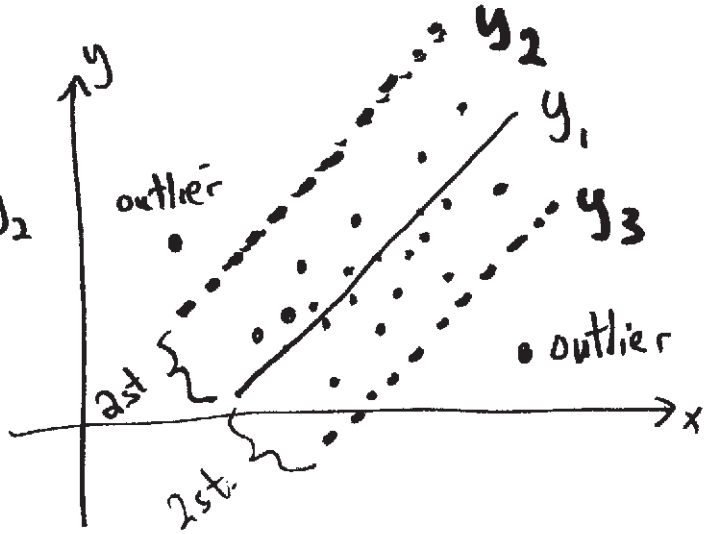
A point is an outlier if it is more than 2 standard deviations above or below the regression line.

To determine if a point is an outlier, do the following:

① Input the following equations into the calculator:

$$\begin{cases} y_1 = a + bx, \text{ the regression line} \\ y_2 = y_1 + 2s_e, \text{ where } s_e \text{ is the st. dev. of the regression line} \\ y_3 = y_1 - 2s_e \end{cases}$$

② If any point is above y_2 or below y_3, then the point is considered to be an outlier.



Exampl~~EXAMP~~

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Are there any outliers?

LinRegTtest

$r = 0.66$ moderate, positive

$\hat{y} = a + bx$

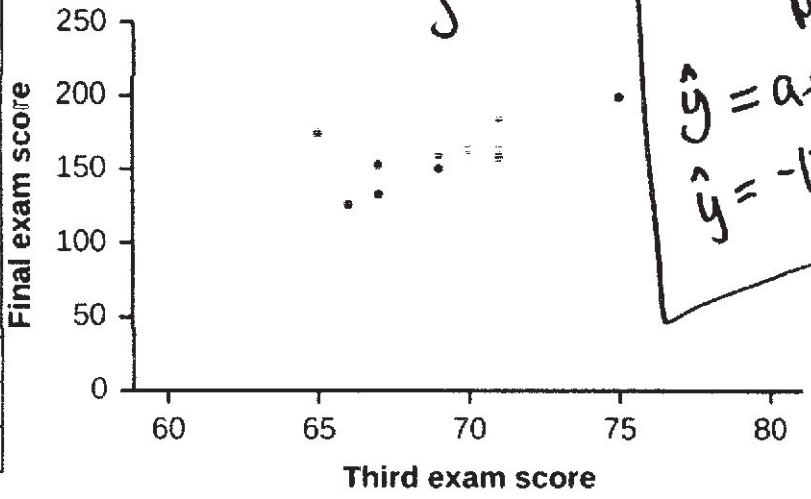
$\hat{y} = -173.5 + 4.8x$

$s_e = 16$

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	139
69	151
69	159

Table 12.3

(a) Table showing the scores on the final exam based on scores from the third exam.



(b) Scatter plot showing the scores on the final exam based on scores from the third exam.

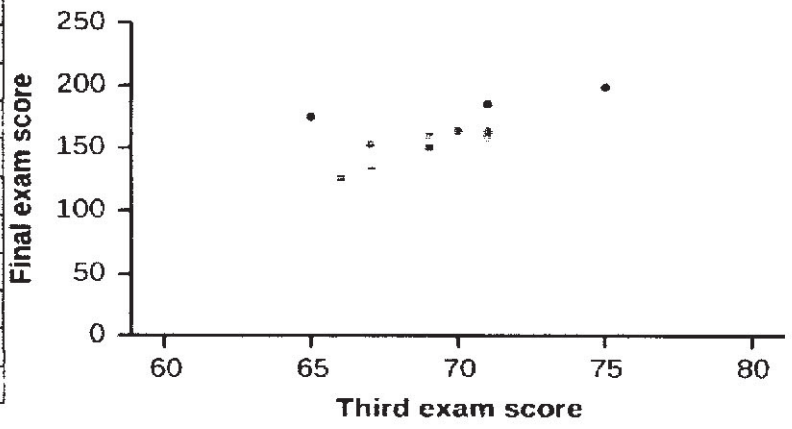
Example~~EXAMP~~

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

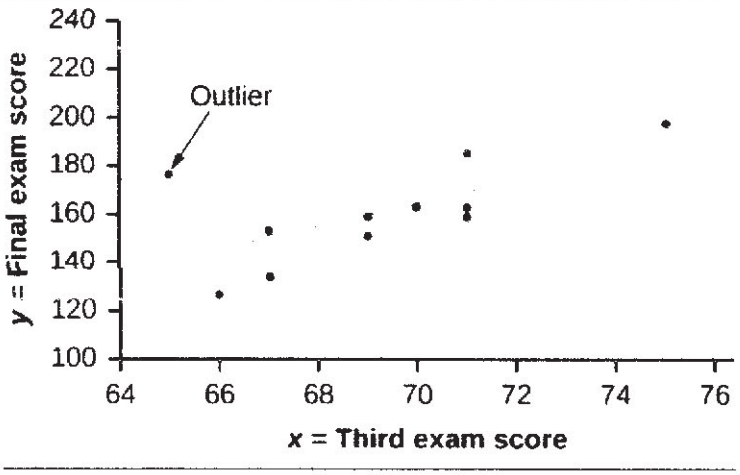
x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Table 12.3

(a) Table showing the scores on the final exam based on scores from the third exam.



(b) Scatter plot showing the scores on the final exam based on scores from the third exam.



How does the outlier affect the best fit line?

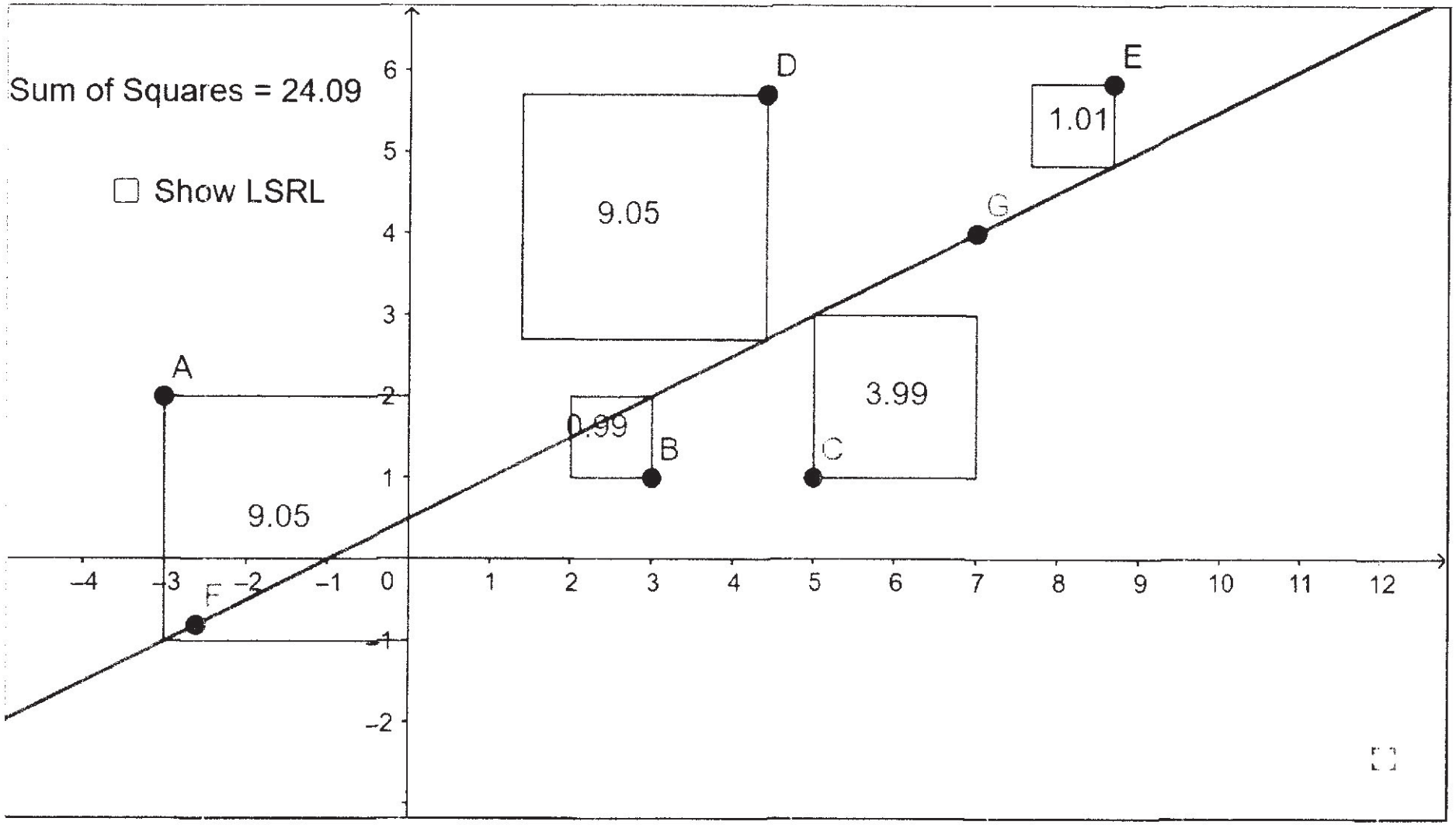
~~Numerically and~~ graphically, we have identified the point (65, 175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error, we should fix the error if possible, or delete the data. If the data is correct, we would leave it in the data set. For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on and delete the outlier, so that we can explore how it affects the results, as a learning experience.

Compute a new best-fit line and correlation coefficient using the ten remaining points:

On the TI-83, TI-83+, TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, the new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$

The new line with $r = 0.9121$ is a stronger correlation than the original ($r = 0.6631$) because $r = 0.9121$ is closer to one. This means that the new line is a better fit to the ten remaining data values. The line can better predict the final exam score given the third exam score.


Sum of Squares = 24.09
Show LSRL
A
B
C
D
E
F
G
9.05
0.99
3.99
1.01
9.05