

Linear Models and Scatter Plots

Grapher	Calculator	Return
Help	Scatter Plot	

Contents: This page corresponds to § 2.6 (p. 228) of the text.

Suggested Problems from Text:

p. 234 #5, 6, 11, 12, 13, 14, 15

[Scatter Plots and Correlation](#)

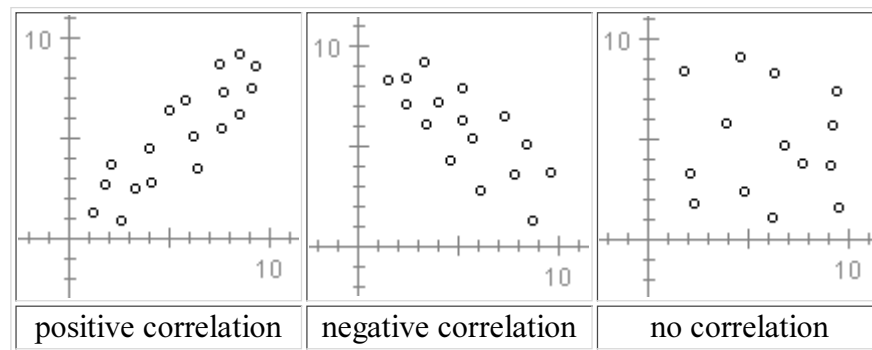
[Fitting a Line to Data Points](#)

Scatter Plots and Correlation

We introduced scatter plots and the Java Scatter Plot program in the section on the [Cartesian Plane](#). In this section we take a closer look at analyzing graphical data.

If a collection of data points has the property that y tends to increase as x increases, then the collection is said to have a **positive correlation**. If y tends to decrease as x increases, then the collection is said to have a **negative correlation**.

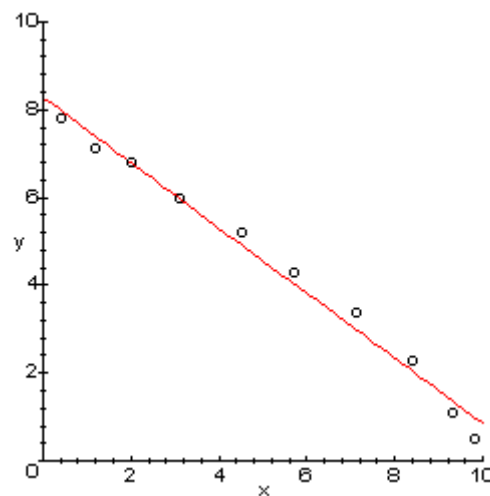
Example 1.



Example 2.

x	0.4	1.2	2.0	3.1	4.5	5.7	7.1	8.4	9.3	9.8
y	7.8	7.1	6.8	6.0	5.2	4.3	3.4	2.3	1.1	0.5

The scatter plot is shown below.



The data set in Example 2 has a negative correlation, and the points are "close" to the line drawn in the scatter plot.

If you were given the scatter plot above and asked to draw the line that was the "best fit" to the data, then you would probably draw a line close to the one we have drawn. The point is that even without defining exactly what we mean by "best fit," you probably have a pretty good idea of what it means.

[Return to Contents](#)

Fitting a Line to Data Points

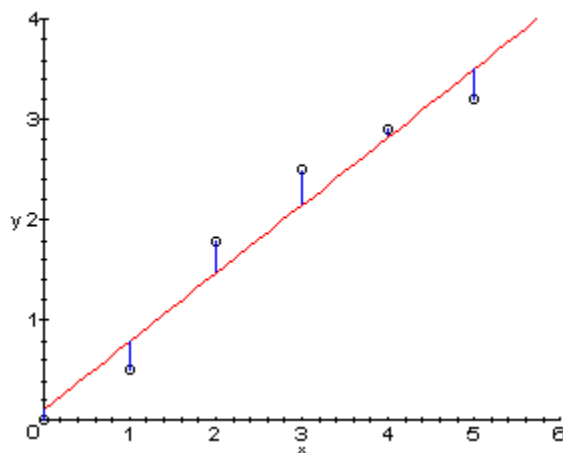
A line fit to data points is called a **regression line**. One way to measure how well a regression line fits the data points is to measure the vertical distance from each data point to the line.

Consider the following data set with six points.

Example 3.

x	0	1	2	3	4	5
y	0	0.5	1.8	2.5	2.9	3.2

This data set has positive correlation, as you can see from the graph.



$$\text{Regression Line: } R = 0.68x + 0.11$$

The regression line is a mathematical model of the relationship between the x and y coordinates. This model can be used in a variety of ways.

For example, suppose that we need to predict the y value corresponding to $x = 1.5$. We do not have a data point with x coordinate 1.5, but since the regression line appears to fit the data reasonably well we could take the value of R when $x = 1.5$ as an approximation. $R(1.5) = 1.13$.

We can measure how well the model fits the data by comparing the actual y values with the R values predicted by the model.

x	0	1	2	3	4	5
y	0	0.5	1.8	2.5	2.9	3.2
R	0.11	0.79	1.47	2.15	2.83	3.51
y - R	-0.11	-0.29	0.33	0.35	0.07	-0.31
$(y - R)^2$	0.0121	0.0841	0.1089	0.1225	0.0049	0.0961

The values in the row $y - R$ are the directed distances from the data points to the line. By directed distance, we mean that it is positive if the data point lies above the line, and negative otherwise. If we add together the directed distances $y - R$, then the positive and negative values tend to cancel and the sum is small. For this reason, the **sum of the squared differences** is used to measure how well the line fits the data.

The regression line for which the sum of the squared differences is smallest is called the **least squares regression line**.

Obviously we cannot test all possible regression lines to find the one with the smallest sum of squared differences! There are formulas for the slope and y-intercept of the least squares regression line, and you will find them on p. 231 of your text. We are not including the formulas on this page because they are somewhat complicated, and these calculations are rarely done by hand.

Exercise 1:

Use either the Java Scatter Plot program or another graphing utility to plot the data points listed in Example 3 and find the least squares regression line. (The regression line R that appears above was obtained from the least squares regression line by rounding the slope and y-intercept to two decimal places.)

To use the Scatter Plot program, recall that you enter the data points in the text box in the upper left hand corner. For instance, to enter the first point $(0,0)$, you type 0,0 in the box and press enter. Then you are ready to enter the next point. Once all the data points are entered, click the **Plot** button to display the scatter plot. Finally, click the **Linear** button to compute the least squares regression line. The formula appears in the text box below the **Linear** button.

When you use a graphing utility to compute the least squares regression line, a number called the **correlation coefficient** is returned along with the formula for the line. In the Scatter Plot program the correlation coefficient is the number in the box beside the **Linear** button. The value of the correlation coefficient from Exercise 1 is approximately 0.97475.

The correlation coefficient ranges in size from -1 up to 1. It is negative when the data set has negative correlation, and positive when there is positive correlation. The size of the absolute value of the correlation coefficient is determined by how well the line fits the data. When the correlation coefficient is close to 1 or -1, then the data points are close to the regression line. If all the data points do lie on a line, then the correlation is exactly -1 or 1.

Exercise 2:

For each of the following three data sets, find the least squares regression line and the correlation coefficient.

(a)

x	1	2	3	4	5	6
y	7	8	5	3	2	2

[Answer](#)

(b)

x	1	2	3	4	5	6
y	0	4	2	4	6	8

[Answer](#)

(c)

x	1	2	3	4	5	6
y	8	1	5	9	0	4

[Answer](#)

[Return to Contents](#)

Grapher	Calculator	Return
Help	Scatter Plot	