12.2.5: Outliers

A scatter plot should be checked for outliers. An outlier is a point that seems out of place when compared with the other points. Some of these points can affect the equation of the regression line.

Should linear regression be used with this data set?

x 1 3 8 2 1 3 2 2 3 1 y 2 3 8 2 3 1 3 1 2 1

Solution

A regression analysis for the data set was run on Excel.

Regression Statistics	
Multiple R	0.844
R Square	0.712
Adjusted R Square	0.676
Standard Error	1.176
Observations	10

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.406	0.618	0.658	0.529
x	0.844	0.19	4.446	0.002

If we test for a significant correlation:

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

The correlation is $r = 0.844$ and the p-value is 0.002, which is less than α = 0.05, so we would reject H_0 and conclude there is a significant relationship between x and y.

However, if look at the scatterplot in Figure 12-18, with the regression equation we can clearly see that the point $(8, 8)$ is an outlier. The outlier is pulling the slope up towards the point $(8, 8)$.

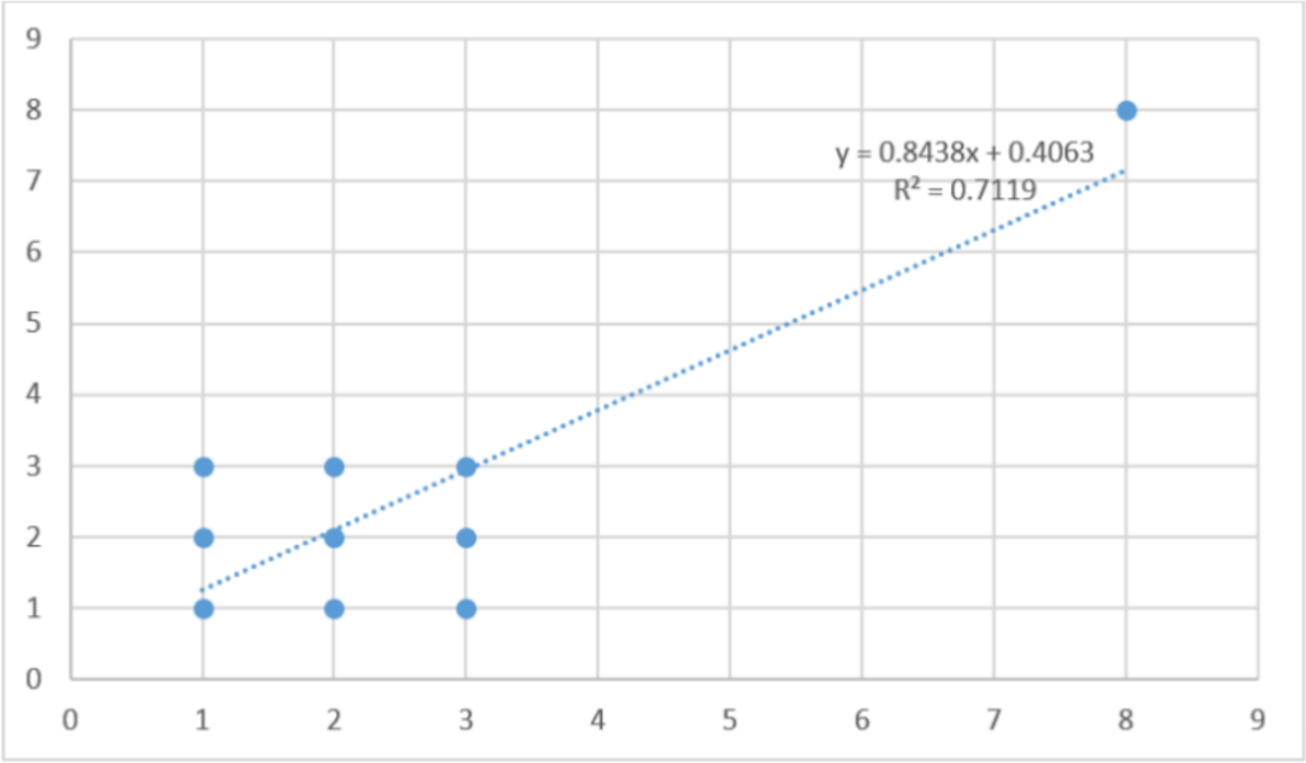

Figure 12-18: Scatterplot of data with an outlier.

If we were to take out the outlier point $(8, 8)$ and run the regression analysis again on the modified data set we get the following Excel output.

Regression Statistics	
Multiple R	0
R Square	0
Adjusted R Square	-0.142857
Standard Error	0.92582
Observations	9

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2	0.816	2.449	0.044
x	0	0.378	0	1

See Figure 12-19: note the correlation is now 0 and the p-value is 1, so there is no relationship at all between x and y.

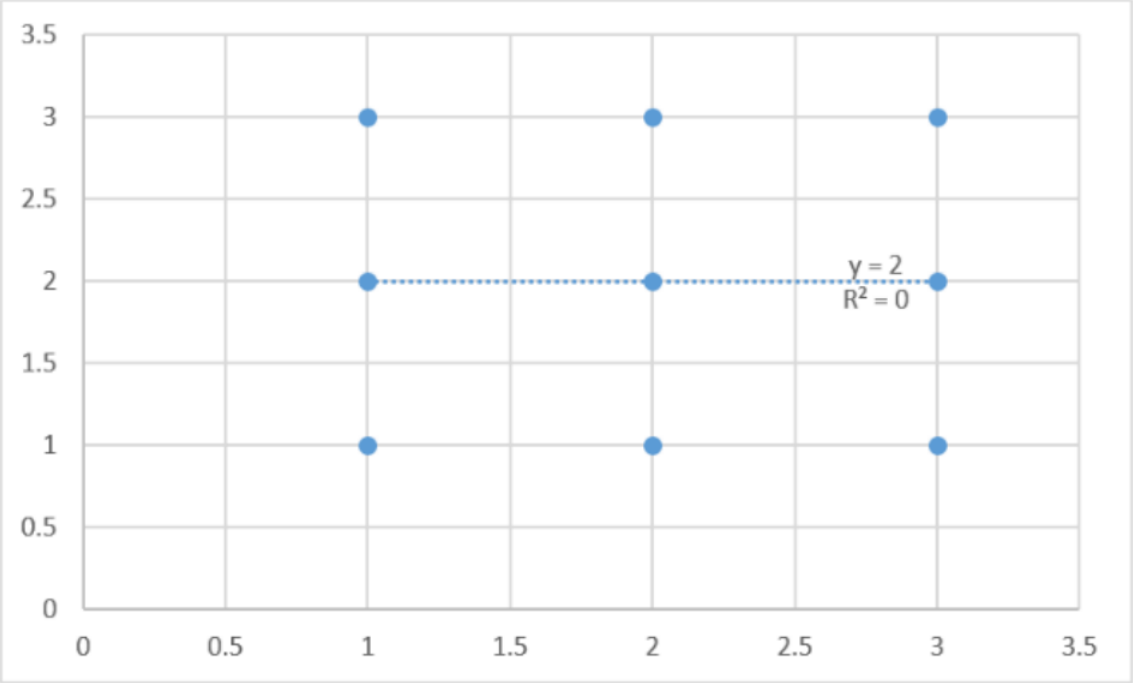

Figure 12-19: Scatterplot of the same data as above with outlier removed.

This type of outlier is called a **leverage point**. Leverage points are positioned far away from the main cluster of data points on the x-axis.

There is another type of outlier called an **influential point**. Influential points are positioned far away from the main cluster of data points on the y-axis. There is an option in most software packages to get the "standardized" residuals. Standardized residuals are z-scores of the residuals. Any standardized residual that is not between -2 and 2 may be an outlier. If it is not between -3 and 3 then the point is an outlier. When this happens, the points are called influential points or influential observations.

Use technology to compute the standardized residuals. Should linear regression be used with this data set?

x 1 3 2 2 4 5 7 9 6 8 y 1 3 10 2 4 5 7 9 6 8

Solution

A regression analysis for the given data set was run on Excel, producing the following results:

Observation	Predicted y	Residuals	Standard Residuals
1	2.974	-1.974	-0.831
2	4.339	-1.339	-0.564
3	3.656	6.344	2.671
4	3.656	-1.656	-0.698
5	5.022	-1.022	-0.430
6	5.705	-0.705	-0.297
7	7.070	-0.070	-0.030
8	8.436	0.564	0.237
9	6.388	-0.388	-0.163
10	7.753	0.247	0.104

The point $(2, 10)$ shown in Figure 12-20 is pulling the left side of the line up and away from the points that form a line. This influential point changes the y-intercept and slope.

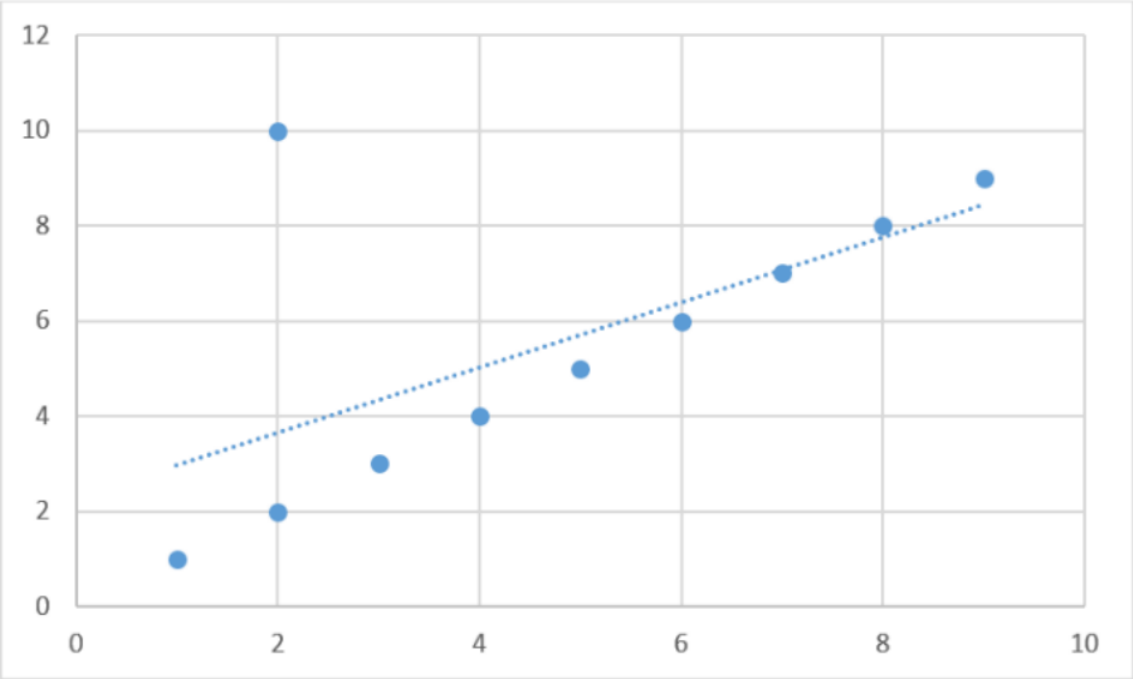

Figure 12-20: