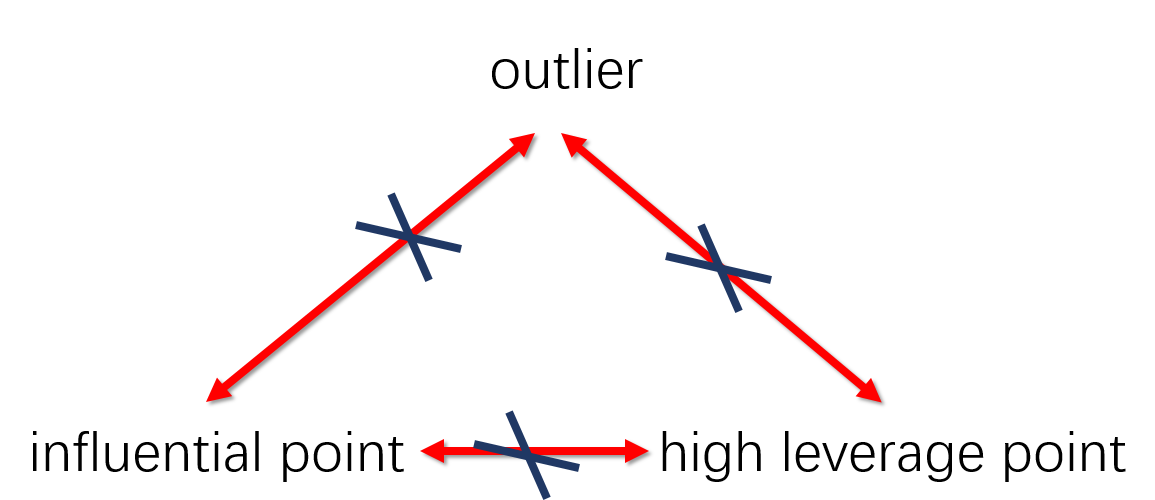
**5.** outlier, high leverage point and influential point

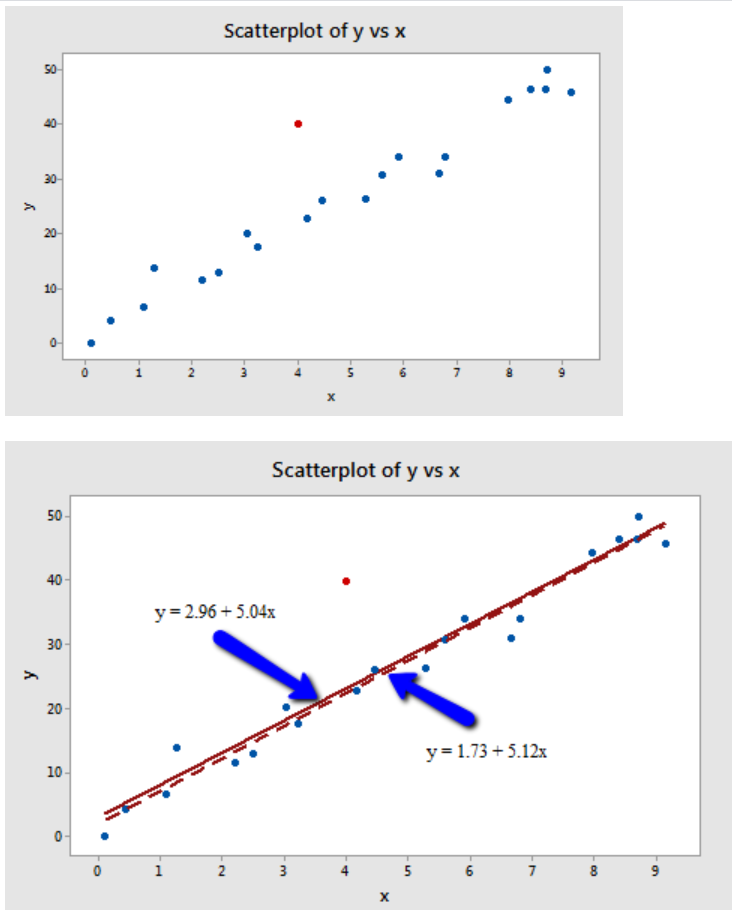
**Def:**

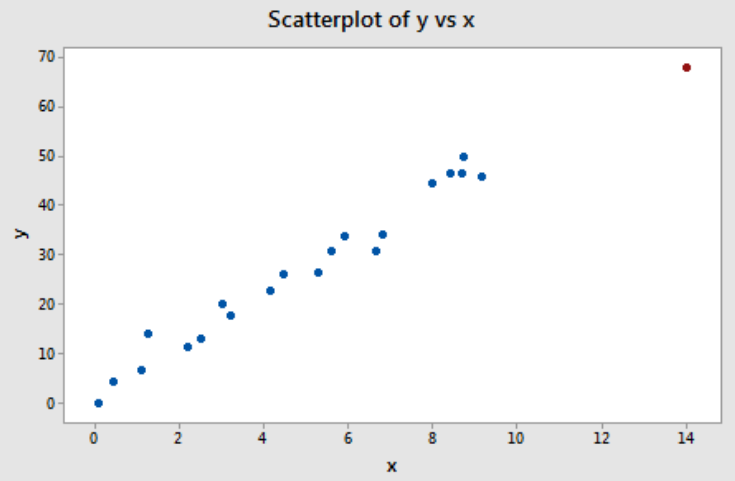
* **outlier** is a data point

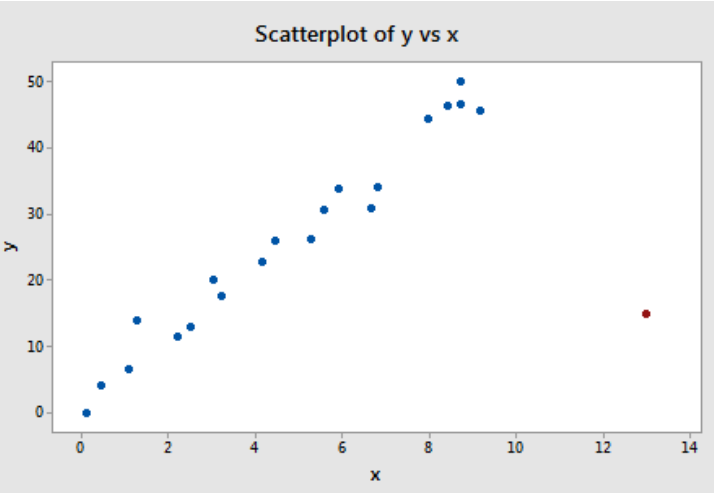
* high **leverage** point:

* **influential** point:

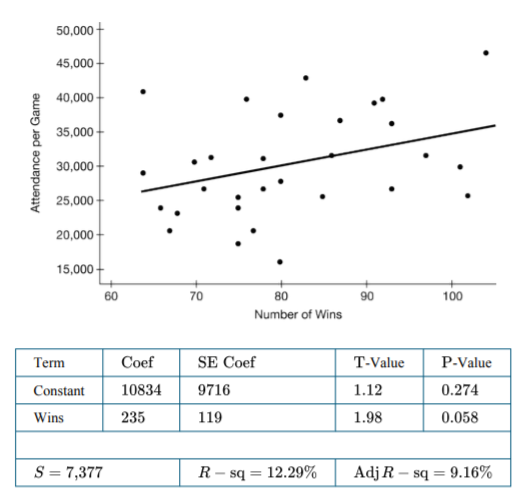








1. There are n pairs of numbers. Suppose x-bar is 3, Sx is 5, y-bar is 5 and Sy is 6. What is the least squares regression line?



1. A response variable appears to be exponentially related to the explanatory variable. The regression line is ln(y) = 1.64 – 0.88x, what is the

predicted value of x=3.1?

1. Which of the following is not appropriate for displaying continuous data?
2. Dotplot
3. Stemplot
4. **Bar chart**
5. Boxplot
6. Histogram
7. Suppose the salaries for a graduation class are as follows:

|  |  |
| --- | --- |
| Number of students | Salary ($) |
| 10 | 15000 |
| 17 | 20000 |
| 38 | 25000 |
| 21 | 30000 |
| 26 | 35000 |

What is the mean salary for the class?

1. 25441.25
2. **26607.14**
3. 30000.31
4. 30533.45
5. 32523.61
6. Given the cumulative frequency table below, what are the mean and median of the distribution?

|  |  |
| --- | --- |
| Value | Cumulative frequency |
| 2 | 15 |
| 3 | 25 |
| 5 | 45 |
| 7 | 95 |
| 10 | 100 |

1. **5.6 , 7**
2. 5.6 , 3
3. 5.4 , 5
4. 5.4 , 7
5. 4.8 , 7
6. The following scatterplot shows the number of wins and the attendance per game for 30 baseball teams in 2017. Also shown are the least-squares regression line and computer output.



1. Interpret the slope of the least-squares regression line in context.

**For each additional win, the predicted attendance per game increases by 235 people.**

1. Explain why it is not reasonable to use the least-squares regression model to predict attendance per game for 0 wins.

**It is not reasonable to predict the value of attendance per game when the number of wins is equal to 0. The number of wins in the data set only includes values from 64 to 104, so we cannot be confident that the linear model is a good predictor of attendance per game if we were to** extrapolate **outside of this interval, including at x=0.**

1. What is the value of the correlation coefficient for the sample?



1. If the point representing wins and attendance of people per game is removed from the set of data and a new regression analysis is conducted, how would the following be impacted? Explain your reasoning.

(i) The slope of the least-squares line:

(ii) The correlation coefficient:



