Linear Regression and Correlation

MAT 152 – Statistical Methods I

Lecture 2

Instructor: Dustin Roten

Fall 2020

Beyond the Scatter Plot

- Inspecting trends and patterns in scatter plots will reveal insights into correlations of bivariate data; however, more rigor and scrutiny may be needed.

- The **correlation coefficient**, *r*, is a numerical measurement of the strength of a linear association between the independent and dependent variables.

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

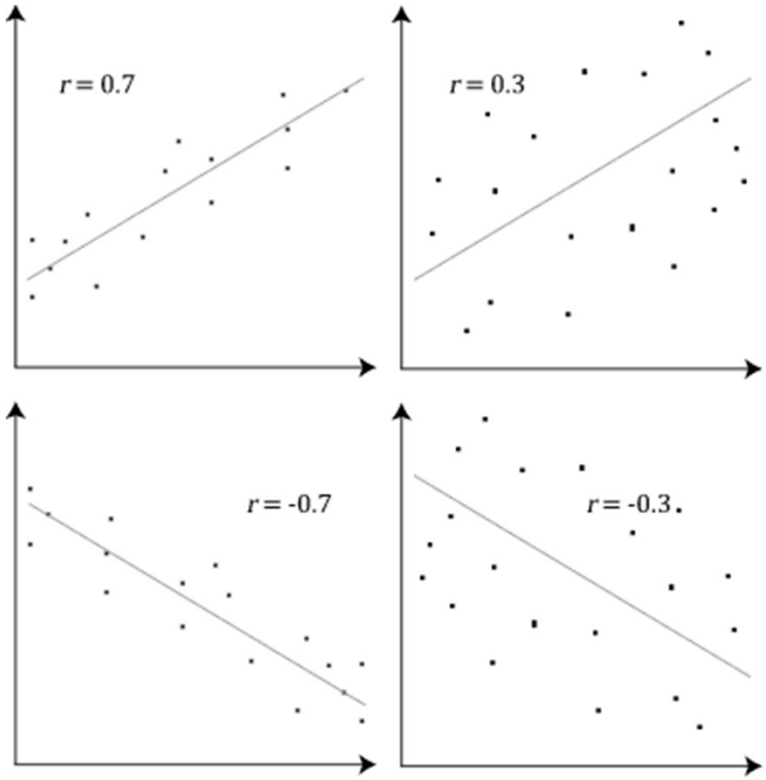
n is the number of data points

Correlation Coefficient

- The correlation coefficient is between -1 and 1.

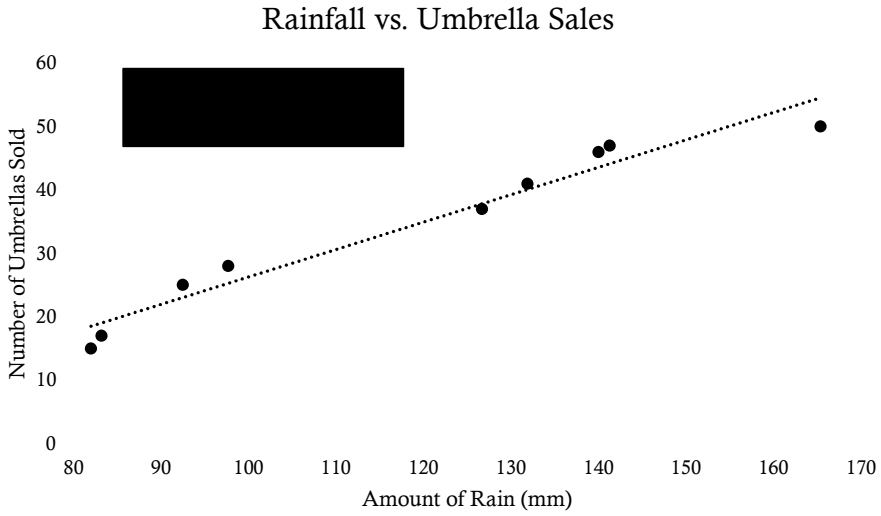
$$-1 \leq r \leq 1$$

- Values close to -1 or 1 indicate strong correlations.
- If *r=0* then there is no correlation.
- If $r = -1$ or 1, there is perfect correlation. (All of the data fall perfectly on the line of best fit.)
- The sign of *r* is the same as the slope, *b*.



Coefficient of Determination

- r^2 is the **coefficient of determination**. It is the square of the correlation coefficient.
- When expressed as a percent, r^2 represents the amount of variation in the dependent variable that can be explained by variation in the independent variable using the regression line.
- Here, we see that 95.42% of the variation in umbrella sales can be explained by the amount of rainfall using the regression line.



$$r = \sqrt{r^2} = \sqrt{0.9542} \approx 0.9768$$

Correlation in Populations

- Can the linear relationship from the sample bivariate data be applied to the population?
- A hypothesis test of **the significance of the correlation coefficient** must be performed.
- The sample correlation coefficient, r, is the estimate of the unknown population coefficient, ρ.
- The hypothesis test asks the question: is ρ close to zero or significantly different from zero?
  - If ρ IS significantly different from zero, the correlation is significant.
  - If ρ is NOT significantly different from zero, the correlation is NOT significant.

Hypothesis Testing

- Null Hypothesis
  - $H_0: \rho = 0$
  - The population correlation coefficient is NOT significantly different from zero. There is no significant linear relationship between x and y in the population.
- Alternative Hypothesis
  - $H_a: \rho \neq 0$
  - The population correlation coefficient IS significantly different from zero. There IS a significant linear relationship between x and y in the population.

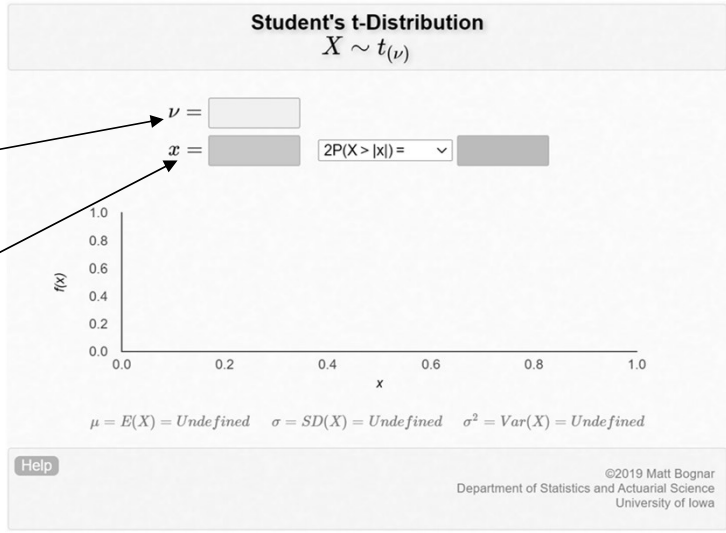
- If the p-value is less than α, reject the null hypothesis.
- If the p-value is more than α, do not reject the null hypothesis.

Hypothesis Testing

1. Perform the linear regression on the bivariate data.
2. Obtain the correlation coefficient.
3. Calculate the degrees of freedom:

$$\nu = df = n - 2$$

1. Calculate the test statistic: $t = \frac{r \cdot \sqrt{\nu}}{\sqrt{1 - r^2}}$
2. Perform a two-tailed test using a t-distribution to obtain a p-value.



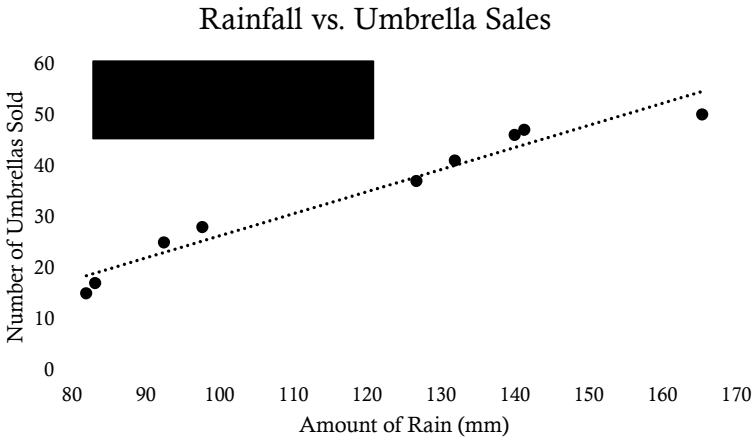
Example

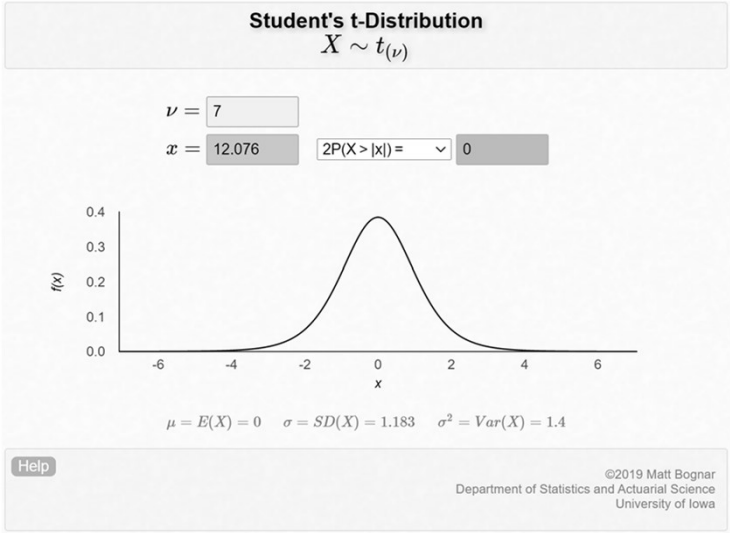
From a previous example, determine if the correlation coefficient is significant.

$$\nu = df = n - 2 = 9 - 2 = 7$$

$$t = \frac{r \cdot \sqrt{\nu}}{\sqrt{1 - r^2}} = \frac{0.9768 \cdot \sqrt{7}}{\sqrt{1 - 0.9542}} \approx 12.076$$

$p < \alpha$ so the correlation is significant.



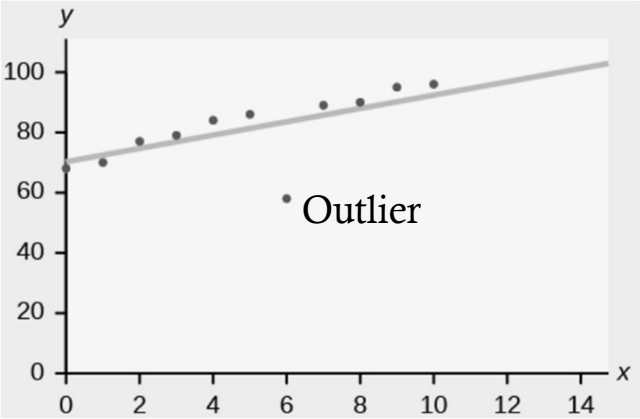


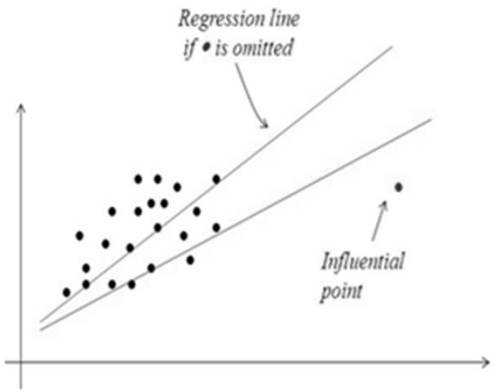
Outliers

- **Outliers** are observed data points that are far from the least squares line. They have large residuals.

- Outliers need to be examined closely. Sometimes, outliers occur because of errors in data collection and should be removed. Other times, outliers contain interesting information and should remain in the data.

- **Influential points** are points that lie far to the left or the right of the rest of the data.





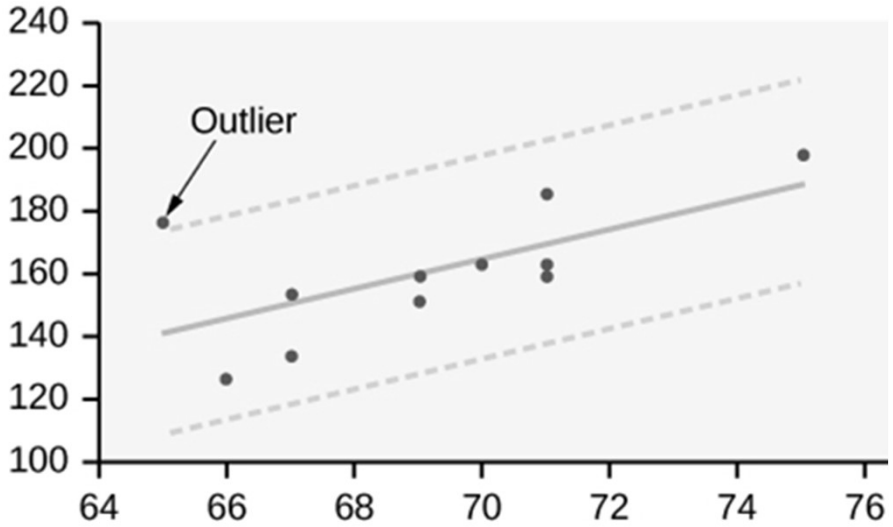
Outliers

- Find the standard deviation of the residuals.

$$s = \sqrt{\frac{SSE}{n-2}}$$

- SSE is the sum of the squared errors. (Standard deviation of the residuals)

- Any values outside 2 standard deviations is considered an outlier.



A Quick Review

- The correlation coefficient, r, measures the strength of the correlation. Its value is between -1 and 1. Strong correlations are near -1 or 1. Weak correlations are closer to 0.

- The coefficient of determination, r^2, determines how much variation in the dependent variable can be explained by the independent variable using the line of best fit.

- The test statistic is calculated by: $t = \frac{r \cdot \sqrt{v}}{\sqrt{1-r^2}}$

- The standard error of the residuals is calculated by: $s = \sqrt{\frac{SSE}{n-2}}$

- Outliers are outside the $\pm 2s$ range.