

Chapter 10: Correlation and Regression

Section 10.2 Correlation

Regression and correlation deal with the relationship between paired data.

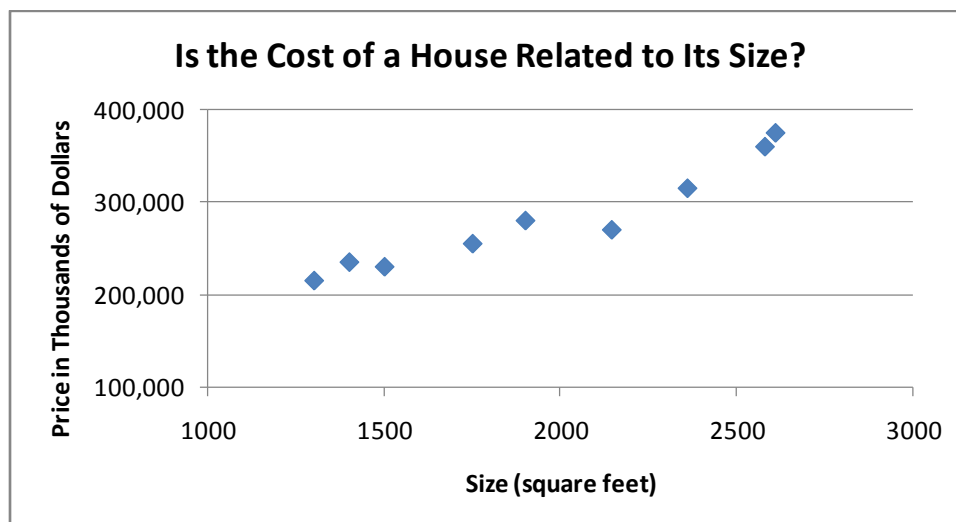
Examples:

- How does lead content of soil vary with distance from a major highway?
- How is a child's vocabulary size related to age?
- How is per pupil spending related to average SAT scores?

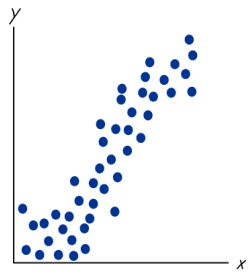
Two questions to ask:

1. Is there a relationship between the variables?
2. If so, can that relationship be described by an equation?

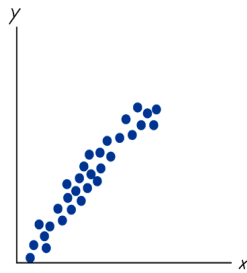
A scatterplot is a graph in which data pairs (x, y) are plotted as individual points on a grid with horizontal axis x and vertical axis y . x is the explanatory variable and y is the response variable.



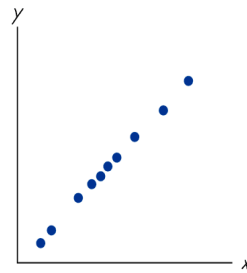
Correlation measures the degree to which two variables are related. In this class, we'll focus on linear correlation. In other words, we'll ask whether the relationship between two variables can be described by a straight line.



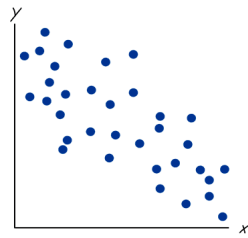
(a) Positive correlation between x and y



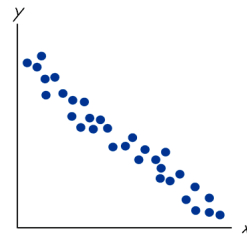
(b) Strong positive correlation between x and y



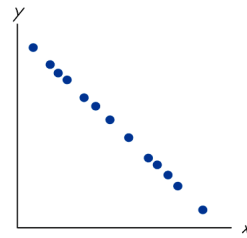
(c) Perfect positive correlation between x and y



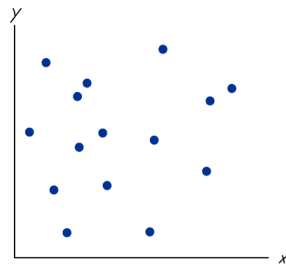
(d) Negative correlation between x and y



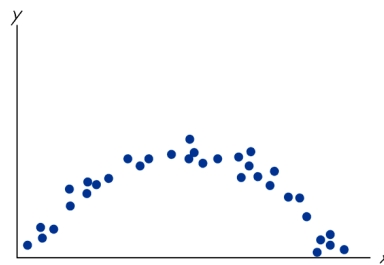
(e) Strong negative correlation between x and y



(f) Perfect negative correlation between x and y



(g) No correlation between x and y



(h) Nonlinear relationship between x and y

The linear correlation coefficient (r) measures the strength of the **linear** relationship between x and y in a sample. r is between -1 and $+1$.

- If $r = -1$, all of the points on the scatterplot of the data lie exactly on a straight line that slopes downward.
- If $r = 1$, all of the points on the scatterplot of the data lie exactly on a straight line that slopes upward.
- If $r = 0$, there is no linear relationship (however, there could be another relationship).
- The closer r is to -1 or 1 , the stronger the linear relationship.

The formula for the linear correlation coefficient:

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

****Note: We will use the calculator instead of this formula.**

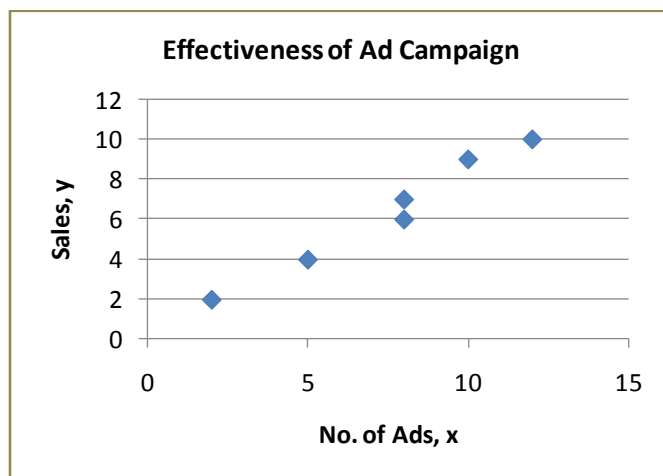
On the calculator, first enter the data into L_1 and L_2 .

Then, STAT → TESTS → LinRegTTest → Make sure $L_1, L_2, \rho \neq 0$ are selected → Calculate

Example: A manager wishes to find out whether there is a relationship between the number of radio ads aired per week and the amount of sales (in thousands of dollars) of a product. The data for the sample are shown. Create a scatterplot and calculate the linear correlation coefficient.

No. of ads (x)	2	5	8	8	10	12
Sales, (y)	2	4	7	6	9	10

Solution:



A scatterplot seems to indicate a linear relationship. Using the calculator, the linear correlation coefficient $r = 0.988$, indicating a strong, positive relationship.

Testing the Significance of r

r represents the linear correlation coefficient for a sample.

ρ represents the linear correlation coefficient for a population.

If r is close to 1, there is a strong positive linear correlation in the sample. We can conclude that there is probably a positive linear correlation in the population.

If r is close to -1, there is a strong negative linear correlation in the sample. We can conclude that there is probably a negative linear correlation in the population.

How close does r have to be to 1 or -1 to conclude that there is a linear relationship in the population?

To answer this question, conduct a hypothesis test:

Method 1: The Critical Value Approach

Step 1: State the null and alternative hypotheses.

$H_0: \rho = 0$ (there is no linear relationship between x and y in the population)

$H_1: \rho \neq 0$ (there is a linear relationship between x and y in the population)

Step 2: Compute r , the sample correlation coefficient using the calculator.

Step 3: Find the critical value (Table A-5 on Page 616) and make a decision.

If $|r| > \text{critical value}$, reject H_0 . Conclude that r is significant and that there is a linear relationship between x and y in the population.

If $|r| < \text{critical value}$, fail to reject H_0 . Conclude that r is not significant and that there is not a linear relationship between x and y in the population.

Method 2: The P-value Approach

Step 1: State the null and alternative hypotheses.

$H_0: \rho = 0$ (there is no linear relationship between x and y in the population)

$H_1: \rho \neq 0$ (there is a linear relationship between x and y in the population)

Step 2: Compute r , the sample correlation coefficient and the p -value using the calculator.

Step 3: Compare the p -value to α .

If the p -value is less than α , reject H_0 . Conclude that r is significant and there is a linear relationship between x and y in the population.

If the p -value is less than α , fail to reject H_0 . Conclude that r is not significant and that there is not a linear relationship between x and y in the population.

Example: A manager wishes to find out whether there is a relationship between the number of radio ads aired per week and the amount of sales (in thousands of dollars) of a product. The data for the sample are shown. Calculate the linear correlation coefficient and determine if it is significant at $\alpha = 0.05$.

No. of ads (x)	2	5	8	8	10	12
Sales, (y)	2	4	7	6	9	10

Previously, we found that $r = 0.988$.

Step 1: State the null and alternative hypotheses.

$H_0: \rho = 0$ (there is no linear relationship between x and y in the population)

$H_1: \rho \neq 0$ (there is a linear relationship between x and y in the population)

Step 2: Compute r, the sample correlation coefficient using the calculator. $r = 0.988$

Step 3: Find the critical value (Table A-5 on Page 616) and make a decision.

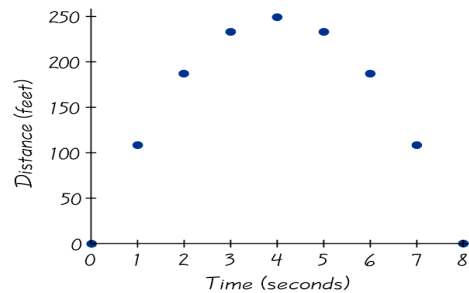
critical value = 0.811

Since $|0.988| > 0.811$, reject H_0 . Conclude that r is significant and that there is a linear relationship between x and y in the population.

Using the p-value method, from the calculator, the p-value = 0.0002. Since the p-value is less than α , reject H_0 .

Errors involving Correlation

1: There may be some relationship between x and y even when there is no significant linear correlation.



2: It is wrong to conclude that correlation implies causality. There may be a *lurking variable*. The correlation could be due to a variable that is unknown to the researcher or not accounted for in the study.

Example: Can you identify the lurking variables?

1. Children with bigger feet spell better.

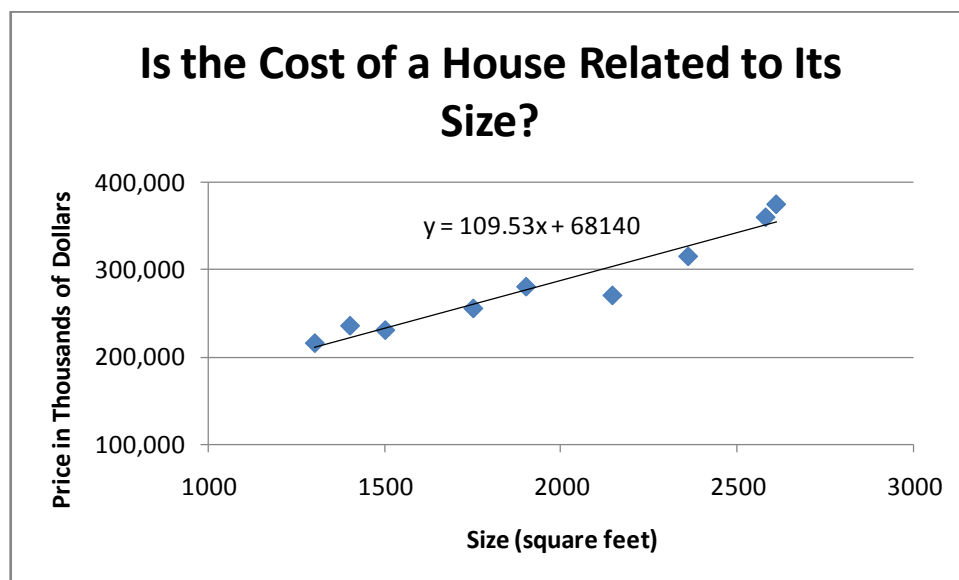
Lurking variable: Age—as children get older, their feet grow and they spell better.

2. A survey of the world's nations in 2004 shows a strong positive correlation between percentage of the country using cell phones and life expectancy in years.

Lurking variable: Wealth—Wealthier countries tend to have longer life expectancy and higher cell phone usage.

Section 10.3: Regression

We'd like to describe the linear relationship (if there is one!) between two variables x and y by finding the equation of the line that best represents this relationship. The graph of the regression equation is called the regression line or the least squares line.



Important! When the correlation is not significant, the data cannot be described by a line. Do not calculate the regression equation if the correlation is not significant!

Finding the Equation of the Regression Line $\hat{y} = a + bx$

$$\text{Slope: } b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$\text{Intercept: } a = \bar{y} - b\bar{x}$$

Note: We will use the calculator to find the slope and intercept of the regression line.

On the calculator, first enter the data into L_1 and L_2 ,

Then,

STAT \rightarrow TESTS \rightarrow LinRegTTest \rightarrow Make sure $L_1, L_2, \rho \neq 0$ are selected \rightarrow Calculate

a = y-intercept

b = slope

$y = a + bx$ is the equation of the best-fit line through the data. You can plot the line and use it to make predictions as long as r is significant.

If r is not significant, the best predicted y-value is \bar{y} .

Example: A manager wishes to find out whether there is a relationship between the number of radio ads aired per week and the amount of sales (in thousands of dollars) of a product. The data for the sample are shown. Create a scatter diagram and calculate the linear correlation coefficient. If the linear correlation is significant, calculate the equation of the regression line. Use the equation of the regression line to estimate the amount of sales when seven radio ads are aired.

No. of ads (x)	2	5	8	8	10	12
Sales, (y)	2	4	7	6	9	10

For this data set, we found that $r = 0.988$ and that r is significant. Therefore, the relationship between x and y can be modeled by a line of the form $y = a + bx$, where a = y-intercept and b = slope.

From the calculator, $a = 0.0738$ and $b = 0.83465$. The equation of the regression line is given by $y = 0.0738 + 0.8346x$.

To predict the amount of sales when seven radio ads are aired, let $x = 7$.

$y = 0.0738 + 0.8346(7) = 5.916$ thousand dollars (or \$5,916).

You should make predictions only for values of x that are between observed x values (interpolation) and not beyond observed x values (extrapolation).

Coefficient of Determination (r^2)

The value r^2 is the ratio of explained variation over total variation. The value of r^2 is the proportion of the variation of y that is explained by the linear relationship between x and y .

Example: Find and interpret the coefficient of determination for the previous problem.

$$r^2 = 0.988^2 = 0.976$$

97.6% of the variation in sales can be explained by the linear relationship between the number of ads and sales.