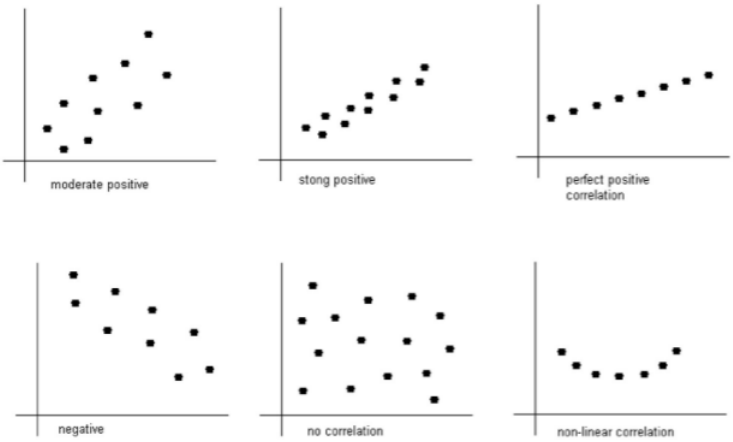
10.2: Correlation

A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

When you see a pattern in the data you say there is a correlation in the data. Though this book is only dealing with linear patterns, patterns can be exponential, logarithmic, or periodic. To see this pattern, you can draw a scatter plot of the data.

Remember to read graphs from left to right, the same as you read words. If the graph goes up the correlation is positive and if the graph goes down the correlation is negative.

The words " weak", "moderate", and "strong" are used to describe the strength of the relationship between the two variables.



Figures

The **linear correlation coefficient** is a number that describes the strength of the linear relationship between the two variables. It is also called the Pearson correlation coefficient after Karl Pearson who developed it. The symbol for the sample linear correlation coefficient is *r*. The symbol for the population correlation coefficient is ρ (Greek letter rho).

The formula for *r* is

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Where

$SS_x = \sum(x - \overline{x})^2$
$SS_y = \sum(y - \overline{y})^2$
$SS_{xy} = \sum(x - \overline{x})(y - \overline{y})$

Assumptions of linear correlation are the same as the assumptions for the regression line:

a. The set (*x*, *y*) of ordered pairs is a random sample from the population of all such possible (*x*, *y*) pairs.
b. For each fixed value of *x*, the *y* -values have a normal distribution. All of the *y* -distributions have the same variance, and for a given *x*-value, the distribution of *y*-values has a mean that lies on the least squares line. You also assume that for a fixed *y*, each *x* has its own normal distribution. This is difficult to figure out, so you can use the following to determine if you have a normal distribution.
    i. Look to see if the scatter plot has a linear pattern.
    ii. Examine the residuals to see if there is randomness in the residuals. If there is a pattern to the residuals, then there is an issue in the data.

Note

Interpretation of the correlation coefficient

r is always between -1 and 1. r = -1 means there is a perfect negative linear correlation and r = 1 means there is a perfect positive correlation. The closer r is to 1 or -1, the stronger the correlation. The closer r is to 0, the weaker the correlation.

CAREFUL: r = 0 does not mean there is no correlation. It just means there is **no linear correlation**. There might be a very strong curved pattern.

r

How strong is the positive relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in *Table* 10.2.1. Find the correlation coefficient and interpret that value.

Table 10.2.1: Alcohol and Calorie Content in Beer without Outlier

Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution

State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

Assumptions check:

From Example 10.2.2, the assumptions have been met.

To compute the correlation coefficient using the TI-83/84 calculator, use the LinRegTTest in the STAT menu. The setup is in *Figure* 10.2.2. The reason that >0 was chosen is because the question was asked if there was a positive correlation. If you are asked if there is a negative correlation, then pick <0. If you are just asked if there is a correlation, then pick $\neq 0$. Right now the choice will not make a different, but it will be important later.

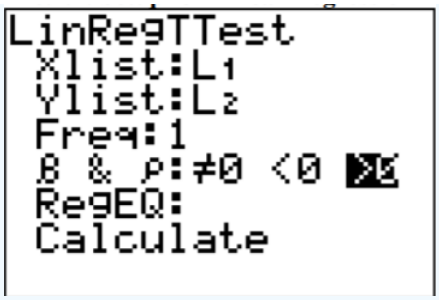


Figure 10.2.2: Setup for Linear Regression Test on TI-83/84

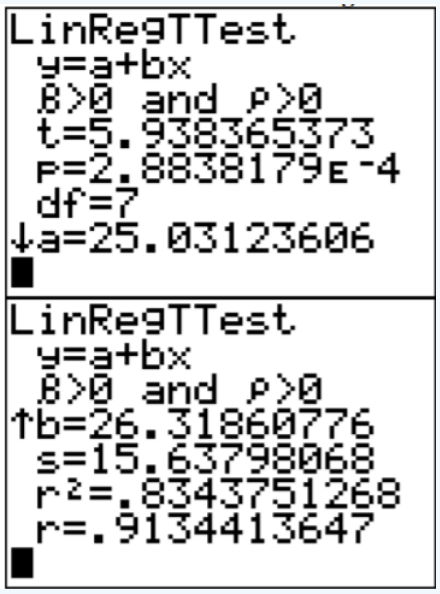


Figure 10.2.3: Results for Linear Regression Test on TI-83/84

To compute the correlation coefficient in R, the command is cor(independent variable, dependent variable). So for this example the command would be cor(alcohol, calories). The output is

[1] 0.9134414

The correlation coefficient is $r = 0.913$. This is close to 1, so it looks like there is a strong, positive correlation.

Causation

One common mistake people make is to assume that because there is a correlation, then one variable causes the other. This is usually not the case. That would be like saying the amount of alcohol in the beer causes it to have a certain number of calories. However, fermentation of sugars is what causes the alcohol content. The more sugars you have, the more alcohol can be made, and the more sugar, the higher the calories. It is actually the amount of sugar that causes both. Do not confuse the idea of correlation with the concept of causation. Just because two variables are correlated does not mean one causes the other to happen.

Example 10.2.2 correlation versus Causation

a. A study showed a strong linear correlation between per capita beer consumption and teacher's salaries. Does giving a teacher a raise cause people to buy more beer? Does buying more beer cause teachers to get a raise?
b. A study shows that there is a correlation between people who have had a root canal and those that have cancer. Does that mean having a root canal causes cancer?

Solution

a. There is probably some other factor causing both of them to increase at the same time. Think about this: In a town where people have little extra money, they won't have money for beer and they won't give teachers raises. In another town where people have more extra money to spend it will be easier for them to buy more beer and they would be more willing to give teachers raises.

b. Just because there is positive correlation doesn't mean that one caused the other. It turns out that there is a positive correlation between eating carrots and cancer, but that doesn't mean that eating carrots causes cancer. In other words, there are lots of relationships you can find between two variables, but that doesn't mean that one caused the other.

Remember a correlation only means a pattern exists. It does not mean that one variable causes the other variable to change.

Explained Variation

As stated before, there is some variability in the dependent variable values, such as calories. Some of the variation in calories is due to alcohol content and some is due to other factors. How much of the variation in the calories is due to alcohol content?

When considering this question, you want to look at how much of the variation in calories is explained by alcohol content and how much is explained by other variables. Realize that some of the changes in calories have to do with other ingredients. You can have two beers at the same alcohol content, but beer one has higher calories because of the other ingredients. Some variability is explained by the model and some variability is not explained. Together, both of these give the total variability. This is

$$\begin{array}{ccccc} \text{(total variation)} & = & \text{(explained variation)} & + & \text{(unexplained variation)} \\ \sum(y-\overline{y})^2 & = & \sum(\hat{y}-\overline{y})^2 & + & \sum(y-\hat{y})^2 \end{array}$$

Note

The proportion of the variation that is explained by the model is

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

This is known as the **coefficient of determination**.

To find the coefficient of determination, you square the correlation coefficient. In addition, r^2 is part of the calculator results.

Example 10.2.3 finding the coefficient of determination

Find the coefficient of variation in calories that is explained by the linear relationship between alcohol content and calories and interpret the value.

Solution

From the calculator results,

$r^2 = 0.8344$

Using R, you can do (cor(independent variable, dependent variable))^2. So that would be (cor(alcohol, calories))^2, and the output would be

[1] 0.8343751

Or you can just use a calculator and square the correlation value.

Thus, 83.44% of the variation in calories is explained to the linear relationship between alcohol content and calories. The other 16.56% of the variation is due to other factors. A really good coefficient of determination has a very small, unexplained part.

and r^2

How strong is the relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer," 2011), and the data are in Example 10.2.1. Find the correlation coefficient and the coefficient of determination using the formula.

Solution

From Example 10.2.2, $SS_x = 12.45, SS_y = 10335.5556, SS_{xy} = 327.6667$

Correlation coefficient:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{327.6667}{\sqrt{12.45 * 10335.5556}} \approx 0.913$$

Coefficient of determination:

$r^2 = (r)^2 = (0.913)^2 \approx 0.834$

Now that you have a correlation coefficient, how can you tell if it is significant or not? This will be answered in the next section.

Homework

Exercise 10.2.1

For each problem, state the random variables. Also, look to see if there are any outliers that need to be removed. Do the correlation analysis with and without the suspected outlier points to determine if their removal affects the correlation. The data sets in this section are in section 10.1 and will be used in section 10.3.

1. When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone 1 (in cm) were collected and are in Example 10.2.5("Prediction of height," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
2. Example 10.2.6contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
3. The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in Example 10.2.7. Find the correlation coefficient and coefficient of determination and then interpret both.
4. The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information is available for the year 2011 are in Example 10.2.8. Find the correlation coefficient and coefficient of determination and then interpret both.
5. The height and weight of baseball players are in Example 10.2.9("MLB heightsweights," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
6. Different species have different body weights and brain weights are in Example 10.2.10 ("Brain2bodyweight," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
7. A random sample of beef hotdogs was taken and the amount of sodium (in mg) and calories were measured. ("Data hotdogs," 2013) The data are in Example 10.2.11. Find the correlation coefficient and coefficient of determination and then interpret both.
8. Per capita income in 1960 dollars for European countries and the percent of the labor force that works in agriculture in 1960 are in Example 10.2.12("OECD economic development," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
9. Cigarette smoking and cancer have been linked. The number of deaths per one hundred thousand from bladder cancer and the number of cigarettes sold per capita in 1960 are in Example 10.2.13("Smoking and cancer," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
10. The weight of a car can influence the mileage that the car can obtain. A random sample of cars weights and mileage was collected and are in Example 10.2.14("Passenger car mileage," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.
11. There is a negative correlation between police expenditure and crime rate. Does this mean that spending more money on police causes the crime rate to decrease? Explain your answer.
12. There is a positive correlation between tobacco sales and alcohol sales. Does that mean that using tobacco causes a person to also drink alcohol? Explain your answer.
13. There is a positive correlation between the average temperature in a location and the morality rate from breast cancer. Does that mean that higher temperatures cause more women to die of breast cancer? Explain your answer.
14. There is a positive correlation between the length of time a tableware company polishes a dish and the price of the dish. Does that mean that the time a plate is polished determines the price of the dish? Explain your answer.

Answer

Only the correlation coefficient and coefficient of determination are given. See solutions for the entire answer.

1. $r = 0.9578$, $r^2 = 0.7357$

3. $r = -0.9313$, $r^2 = 0.8674$

5. $r = 0.6605$, $r^2 = 0.4362$

7. $r = 0.8871$, $r^2 = 0.7869$

9. $r = 0.7036$, $r^2 = 0.4951$

11. No, see solutions.

13. No, see solutions.
