CORRELATION

We have learnt how to use graphs, such as scatterplots, to analyze the relationship between two variables. Two variables may be positively or negatively related when different pairs of data show the same pattern. For example, when incomes of individuals rise so does their consumption of goods and services; thus, income and consumption are considered to be positively related. As a person's income rises, the number of bus rides this person takes falls; thus, income and bus riding are negatively related.

Introduction

Figure 13.1 Linear regression and correlation can help you determine if an auto mechanic's salary is related to his work experience.

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability, or your gender or color. The amount you pay a labor is often determined by an initial amount plus an hourly fee.

In this chapter we will begin with correlation, the investigation of relationships among variables that may or may not be founded on a regression model. The variables simply move in the same, or opposite, direction. That is to say, they do not move randomly. Correlation provides a measure of the degree to which this is true. We will be able to formulate models and tests to determine if they are statistically sound. If they are found to be so, then we can use them to make predictions: if as a matter of policy we changed the value of this variable what would happen to this other variable?

The Correlation Coefficient r

As we begin this section we note that the type of data we will be working with has changed. Perhaps unnoticed, all the data we have been using is for a single variable i.e. univariate variable. The type of data described in regression model is **bivariate** data - "bi" for two variables. In reality, statisticians use multivariate data, meaning many variables.

Beginning with a set of data with two independent variables we ask the question: are these related? One way to visually answer this question is to create a scatter plot of the data. We could not do that before when we were doing descriptive statistics because those data were univariate. Now we have bivariate data so we can plot in two dimension.

Correlation, is a statistical measure which quantifies the strength and direction of the relationship between two variables.

The correlation tells us something about the co-movement of two variables, but nothing about why this movement occurred. The correlation coefficient, ρ (pronounced rho), is the mathematical statistic for a population that provides us with a measurement of the strength of a linear relationship between the two variables. For a sample of data, **the statistic, r,** developed by ***Karl Pearson in the early 1900s***, is an estimate of the population correlation and is defined mathematically as:

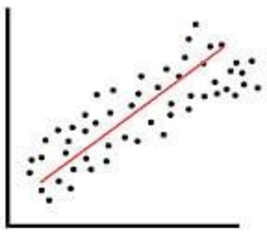
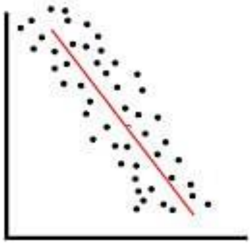
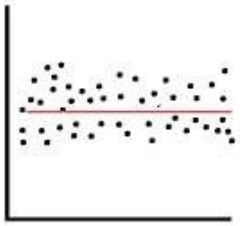
$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\,\sqrt{n\sum y^2 - (\sum y)^2}}$$

ou can apply correlation to a variety of data sets. In some cases, you may be able to predict how things will relate, while in others, the relation will come as a complete surprise. It's important to remember that just because something is correlated doesn't mean it's causal.

Types of correlation

There are three types of correlation

Positive Correlation	**Negative Correlation**	**No Correlation**
A positive correlation means that this linear relationship is positive, and the two variables increase or decrease in the same direction.	A negative correlation is just the opposite. The relationship line has a negative slope, and the variables change in opposite directions, i.e., one variable decreases while the other increases	No correlation simply means that the variables behave very differently and thus, have no linear relationship
$(0 < r \leq 1)$	**$(-1 \leq r < 0)$**	**$(r \approx 0)$**

Interpretation Let the study finds r=0.5 This value indicates a positive a moderate strength of correlation.(means that while there is a noticeable relationship between study time and exam scores, it is not the only factor influencing exam performance. Students who study more tend to score higher, but there are enough exceptions that other factors should also be considered.	**Interpretation** If r= −0.85, there is a strong negative correlation (The study finds that as the number of hours of daily physical activity increases, the risk of developing chronic diseases decreases. This relationship is strong because regular physical activity is known to significantly reduce the risk of various chronic diseases such as diabetes, heart disease, and hypertension.	**Interpretation**: If is r is close to 0, there is no linear relationship between two variables . For example, r = 0.02 would indicate virtually no correlation, (means that changes in shoe size have no predictable impact on intelligence)
 Positive Correlation	 Negative Correlation	 No Correlation
Ice Cream Sales vs. Temperature: As temperatures rise (one variable), ice cream sales (the other variable) tend to increase as well. This is a positive correlation because both variables move in the same direction (upward in this case). ・ **Study Hours vs. Exam Scores:** Generally, students who dedicate more hours to studying (one variable) tend to score higher on exams (the other variable). This is another example of a positive correlation, with both variables increasing together. ・ **Height vs. Weight:** Taller people (one variable) tend to weigh more (the other variable) on average. This is a positive correlation, though not necessarily a perfect one, as body types can vary.	**Exercise vs body fat:** As the amount of exercise someone does increases ,their body fat percentage tends to decrease. This is a negative correlation because the variables move in opposite directions. More exercise is generally associated with lower body fat percentage ・ **Driver Experience and Auto Insurance:** When it comes to auto insurance, there's a strong negative correlation between driver experience and the amount you pay in premiums. The more experience you have on the road, the less you'll likely pay for insurance. ・ **Price of a Good vs. Demand:** As the price of a good increases (one variable), the demand for that good (the other variable) generally decreases. This is a negative correlation, with higher prices leading to lower demand.	**Shoe Size vs. IQ:** There's no logical reason why foot size (one variable) should relate to intelligence (the other variable). In this case, the variables would likely show a scattered pattern with no particular trend, indicating no correlation. ・ **Coffee Consumption vs Musical Preference:** There's no scientific evidence suggesting that coffee consumption (one variable) directly affects musical taste (independent) These are independent aspects of a person's life. ・ **Playing Video Games vs. Liking Pizza:** Whether someone enjoys video games (one variable) has no inherent connection to their preference for pizza (the other variable). There would likely be no trend in the data, indicating no correlation.

What the VALUE of r tells u

- The value of r is always between –1 and +1
- The size of the correlation r indicates the strength of the linear relationship between two variables. Values of r close to –1 or to +1 indicate a stronger linear relationship.
- If r=0 there is absolutely no linear relationship between two variables (no linear correlation).
- If r=1, there is perfect positive correlation. If r=–1, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line:

What the SIGN of r tells us

- A positive value of r means that when first variable increases second variable tends to increase and when second variable decreases the first tends to decrease.

- A negative value of r means that when first variable increases second tends to decrease and when second variable decreases first variable tends to increase (negative correlation).

Strength of Coefficient of Correlation

Coefficient of correlation	Degree of Association
± 0.8 to ± 1	Strong
± 0.5 to ± 0.8	Moderate
± 0.2 to ± 0.5	Weak
0 to ± 0.2	negligible

<u>EXAMPLE</u>

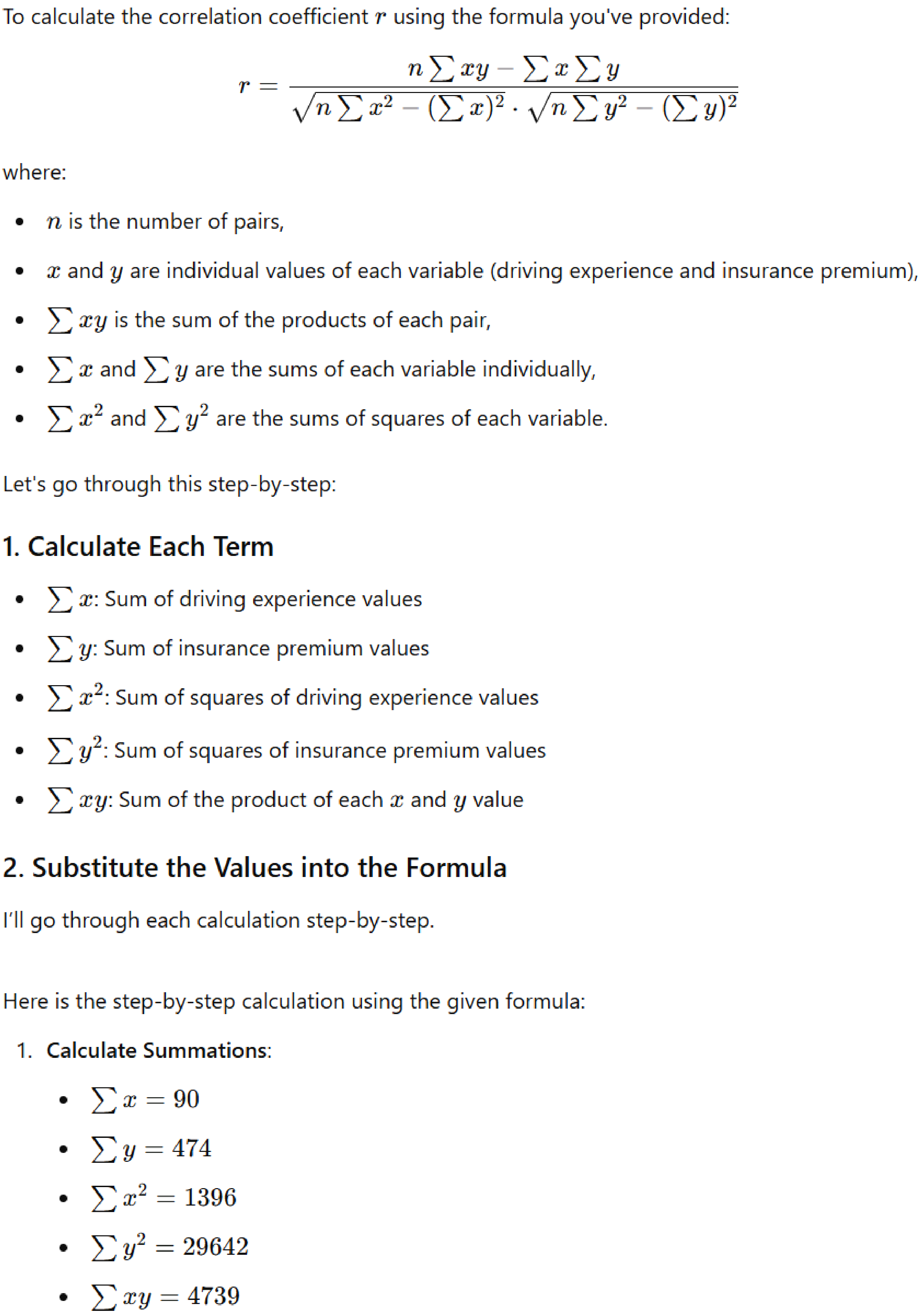
Lets consider the example

A random sample of eight drivers insured with a company and having similar auto insurance policies was selected. The following table lists their driving experiences (in years) and monthly auto insurance premiums.

a. Calculate coefficient of correlation r , interpret the value of r.
b. Interpret r.

A complete example of regression analysis.

PhotoDisc, Inc./Getty Images



To calculate the correlation coefficient r using the formula you've provided:

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

where:

- n is the number of pairs,
- x and y are individual values of each variable (driving experience and insurance premium),
- $\sum xy$ is the sum of the products of each pair,
- $\sum x$ and $\sum y$ are the sums of each variable individually,
- $\sum x^2$ and $\sum y^2$ are the sums of squares of each variable.

Let's go through this step-by-step:

1. Calculate Each Term

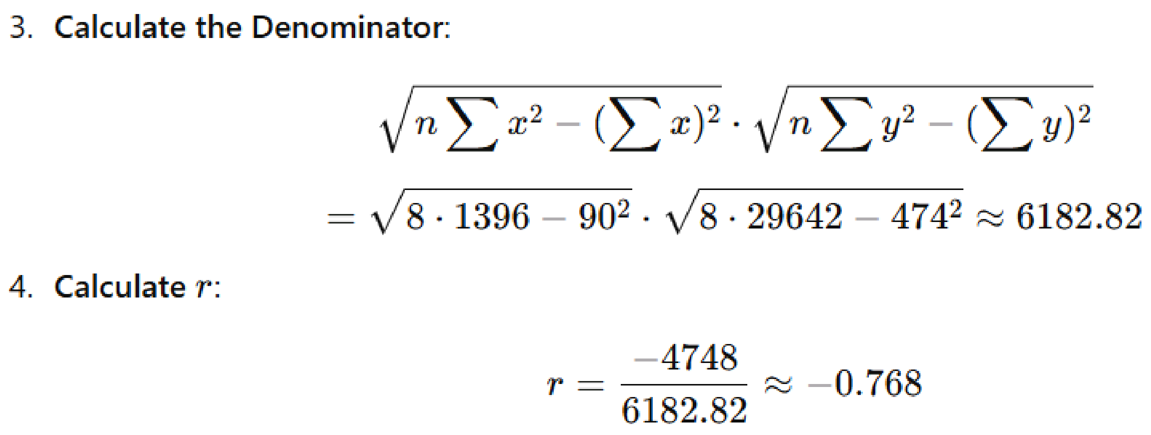
- $\sum x$: Sum of driving experience values
- $\sum y$: Sum of insurance premium values
- $\sum x^2$: Sum of squares of driving experience values
- $\sum y^2$: Sum of squares of insurance premium values
- $\sum xy$: Sum of the product of each x and y value

2. Substitute the Values into the Formula

I'll go through each calculation step-by-step.

Here is the step-by-step calculation using the given formula:

1. **Calculate Summations**:
   - $\sum x = 90$
   - $\sum y = 474$
   - $\sum x^2 = 1396$
   - $\sum y^2 = 29642$
   - $\sum xy = 4739$



3. **Calculate the Denominator:**

$$\sqrt{n\sum x^2-(\sum x)^2}\cdot\sqrt{n\sum y^2-(\sum y)^2}$$

$$=\sqrt{8\cdot 1396-90^2}\cdot\sqrt{8\cdot 29642-474^2}\approx 6182.82$$

4. **Calculate r:**

$$r=\frac{-4748}{6182.82}\approx -0.768$$

INTERPRETATION

The correlation coefficient r is approximately −0.768, confirming a strong negative correlation between driving experience and insurance premium

Linear Regression and Correlation Homework

Solve these problems, then check your answers against the given solutions.

Exercise 1

For each situation below, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than all other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

Exercise 2

Over a period of one year a greengrocer sells tomato at six different prices (x pence per kilogram). He calculates the average number of kilograms y, sold per day at each of the six different prices. From these data the followings are calculated

$$\sum x = 200 \qquad \sum y = 436 \qquad \sum xy = 12\,515$$

$$\sum x^2 = 7250 \qquad \sum y^2 = 39\,234 \qquad n = 6$$

Calculate the value of coefficient of correlation r , interpret its value.

Exercise 3

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	1050	428	362	529	790	1454
Stories	57	28	26	40	60	110

a. Using "stories" as the independent variable and "height" as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Interpret your scatter plot
c. Calculate the least squares regression line. Put the equation in the form of: $\hat{y}=a+bx$
d. What is the slope of the least squares (best-fit) line? Interpret the slope.
e. Find the correlation coefficient. Interpret the degree of association?
f. Find the estimated heights for 32 stories and for 94 stories.

Exercise 4

We are interested in exploring the relationship between the weight of a vehicle and its fuel effciency (gasoline mileage). The data in the table show the weights, in pounds, and fuel efficiency, measured in miles per gallon, for a sample of 12 vehicles.

Weight (x)	2715	2570	2610	2750	3000	3410	3640	3700	3880	3900	4060	4710
Fuel Efficiency (y)	24	28	29	38	25	22	20	26	21	18	18	15

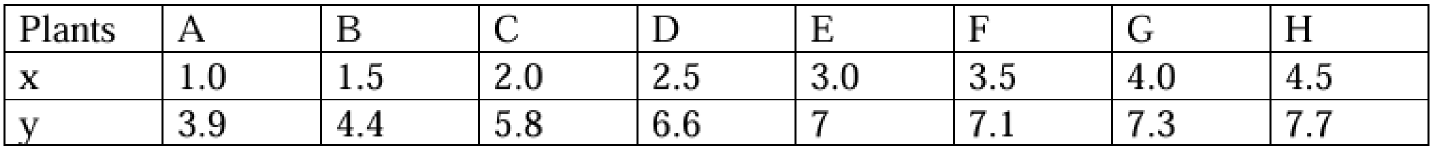
a. Graph a scatterplot of the data. Interpret the pattern.
b. Find the correlation coefficient r. interpret 'r'
c. Find the equation of the best fit line.
d. Write the sentence that interprets the meaning of the slope of the line in the context of the data.
e. For the vehicle that weight 5000 pounds, find the predicted value of response variable$\hat{y}$

Steps to be followed

- Plot the scatter diagram, observe the pattern
- Interpret the pattern of scatter plot
- Find out the sample correlation coefficient "r"
- Interpret r
- Find the values of intercept and slope
- Find the least square line of equation
- Interpret the slope and intercept
- Find the estimated value of fuel efficiency when weight of vehicle is 5000 pounds

Exercise 5

(a) Define Regression and correlation
(b) A biologist assume that there is a linear relationship between the amount of fertilizers supplies to tomato plants and the subsequent yield of tomato obtained. Eight tomato plants are selected randomly for the study and treated weekly with a solution with x grams of fertilizer was dissolve in a fixed quantity of water. The yield y kilograms are recorded.



Plants	A	B	C	D	E	F	G	H
x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	3.9	4.4	5.8	6.6	7	7.1	7.3	7.7

1) Plot a scatter diagram of yield y against amount of fertilizers x
2) Calculate the equation of least square regression line of y on x
3) Estimate the yield of plants treated weekly with 3.2 grams of fertilizer
4) Give interpretation to your analysis. 7.7
5) Indicate why it may not be appropriate to use your equation to predict the yield of plants treated weekly with 20 grams of fertilizer