Lecture Two

Section 2.1 – Scatter Diagrams and Correlation

Draw and Interpret Scatter Diagrams

Definition

The ***response variable*** is the variable whose value can be explained by the value of the ***explanatory*** or ***predictor variable***.

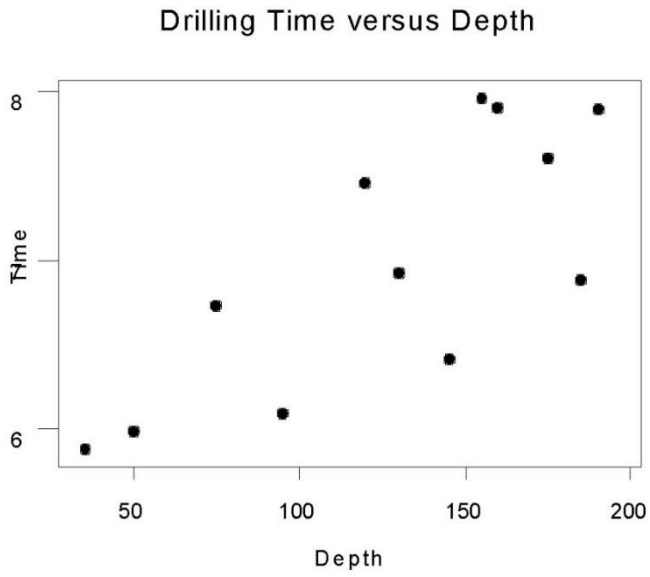
A ***scatter diagram*** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.

Example

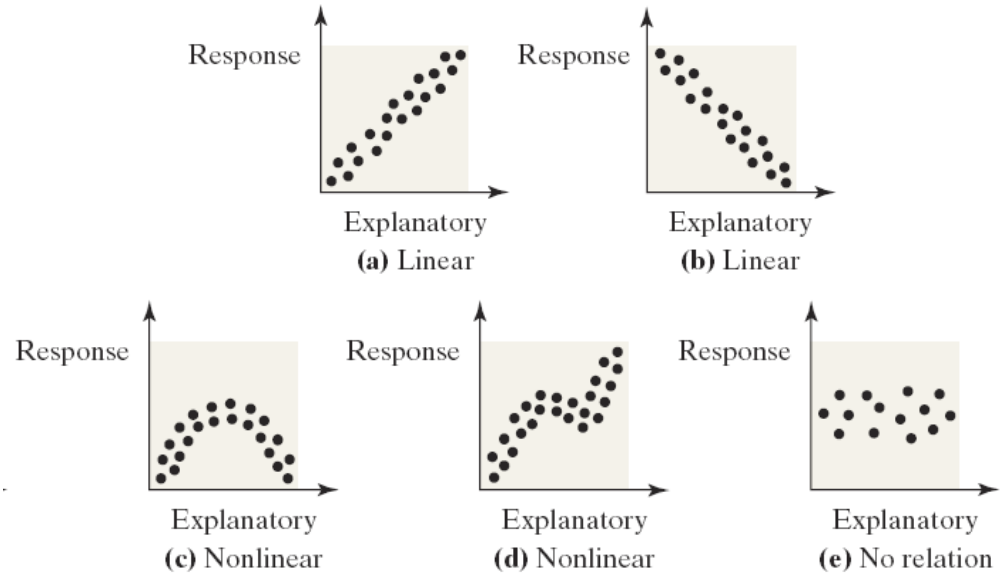
The data shown to the right are based on a study for drilling rock. The researchers wanted to determine whether the time it takes to dry drill a distance of 5 feet in rock increases with the depth at which the drilling begins. So, depth at which drilling begins is the explanatory variable, x, and time (in minutes) to drill five feet is the response variable, y. Draw a scatter diagram of the data.

Depth at Which Drilling Begins ***x*** (in feet)	**Time to Drill 5 feet** ***y*** (in feet)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

Solution



Various Types of Relations in a Scatter Diagram



- Two variables that are linearly related are ***positively associated*** when above-average values of one variable are associated with above-average values of the other variable and below-average values of one variable are associated with below-average values of the other variable. That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.
- Two variables that are linearly related are ***negatively associated*** when above-average values of one variable are associated with below-average values of the other variable. That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

Definitions

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

The ***linear correlation coefficient*** **(r)** or ***Pearson product moment correlation coefficient*** is a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient. We present only the formula for the sample correlation coefficient

Requirements

1. The sample of paired (x, y) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating r with and without the outliers included.

Notation for the Linear Correlation Coefficient

n number of pairs of sample data

$\sum$ denotes the addition of the items indicated.

$\sum x$ denotes the sum of all x-values.

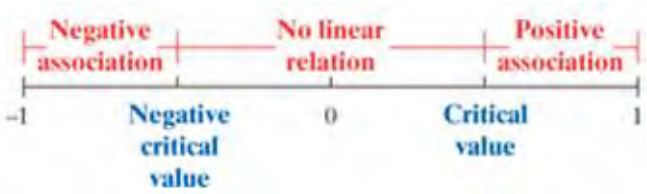
$\sum x^2$ indicates that each x-value should be squared and then those squares added.

$\left(\sum x\right)^2$ indicates that the x-values should be added and then the total squared

$\sum xy$ indicates that each x-value should be first multiplied by its corresponding y-value. After obtaining all such products, find their sum.

r = linear correlation coefficient for ***sample*** data.

ρ = linear correlation coefficient for **population** data.



Formula

The linear correlation coefficient r measures the strength of a linear relationship between the paired values in a sample.

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

- Computer software or calculators can compute r

Sample Linear Correlation Coefficient

$$r = \frac{\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

where

$\bar{x}$ is the sample mean of the explanatory variable

s_x is the sample standard deviation of the explanatory variable

$\bar{y}$ is the sample mean of the response variable

s_y is the sample standard deviation of the response variable

n is the number of individuals in the sample

✓ *Know that the methods of this section apply to a linear correlation. If you conclude that there does not appear to be linear correlation, know that it is possible that there might be some other association that is not linear.*

Properties of the Linear Correlation Coefficient *r*

1. The value of r is always between –1 and 1, inclusive. That is $-1 \le r \le 1$
2. If $r = +1$, then a perfect positive linear relation exists between the two variables.
3. If $r = -1$, then a perfect negative linear relation exists between the two variables.
4. The closer r is to +1, the stronger is the evidence of positive association between the two variables.
5. The closer r is to –1, the stronger is the evidence of negative association between the two variables.
6. If r is close to 0, then little or no evidence exists of a linear relation between the two variables. So r close to 0 does not imply no relation, just no linear relation.
7. The linear correlation coefficient is a unitless measure of association. So the unit of measure for x and y plays no role in the interpretation of r.
8. The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

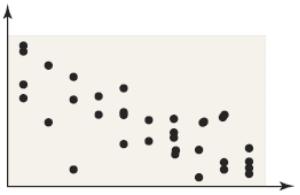
(a) Perfect positive linear relation, $r = 1$

(b) Strong positive linear relation, $r \approx 0.9$

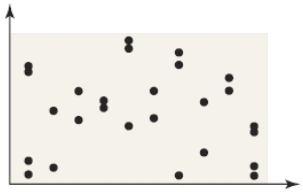
(c) Moderate positive linear relation, $r \approx 0.4$

(d) Perfect negative linear relation, $r = -1$

(e) Strong negative linear relation, $r \approx -0.9$

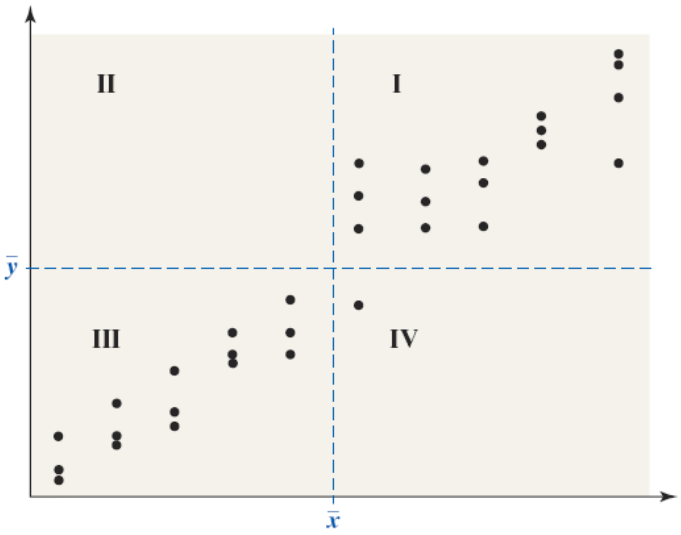


(f) Moderate negative linear relation, $r \approx -0.4$



(g) No linear relation, r close to 0.

(h) No linear relation, r close to 0.



Example

Determine the linear correlation coefficient of the drilling data

Depth at Which Drilling Begins x (in feet)	Time to Drill 5 feet y (in feet)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

Solution

x	y	$\frac{x_i-\bar{x}}{s_x}$	$\frac{y_i-\bar{y}}{s_y}$	
35	5.88	−1.74712	−1.41633	2.474501
50	5.99	−1.45992	−1.27544	1.862051
75	6.74	−0.98126	−0.31486	0.308958
95	6.1	−0.59833	−1.13456	0.678839
120	7.47	−0.11967	0.620111	−0.07421
130	6.93	0.0718	−0.07151	−0.00513
145	6.42	0.358998	−0.72471	−0.26017
155	7.97	0.550463	1.260501	0.693859
160	7.92	0.646196	1.196462	0.77319
175	7.62	0.93394	0.812228	0.758129
185	6.89	1.12486	−0.12274	−0.13807
190	7.9	1.220592	1.170846	1.429126
$\bar{x}=126.25$	$\bar{y}=6.9858$			8.50104
$s_x=52.23$	$s_y=0.781$			

$$r=\frac{\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)}{n-1}$$
$$=\frac{8.501037}{12-1}$$
$$=0.7728$$

OR

Example

The paired pizza/subway fare costs are shown in the table below. Use computer software with these paired sample values to find the value of the linear correlation coefficient r for the paired sample data.

Table *– Cost of a Slice of Pizza, subway Fare, and the CPI*						
Year	**1960**	**1973**	**1986**	**1995**	**2002**	**2003**
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

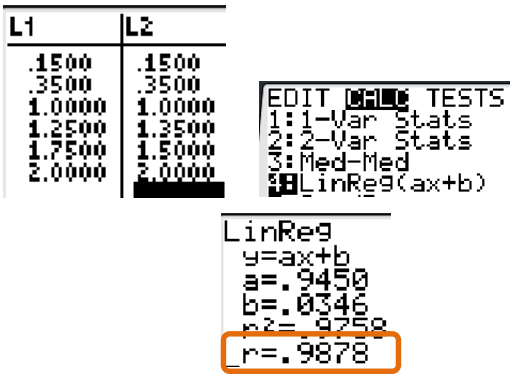
Solution

x (Pizza)	y (Subway)	x^2	y^2	xy
0.15	0.15	0.0225	0.0225	0.0225
0.35	0.35	0.1225	0.1225	0.1225
1.00	1.0	1.0	1.0	1.0
1.25	1.35	1.5625	1.8225	1.6875
1.75	1.50	3.0625	2.250	2.6250
2.0	2.0	4.0000	4.0000	4.0000
$\sum x = 6.50$	$\sum y = 6.35$	$\sum x^2 = 9.77$	$\sum y^2 = 9.2175$	$\sum xy = 9.4575$

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

$$= \frac{6(9.4575) - (6.5)(6.35)}{\sqrt{6(9.77) - (6.5)^2} \cdot \sqrt{6(9.2175) - (6.35)^2}}$$

$$= 0.9878$$



Difference between Correlation and Causation

According to data obtained from the Statistical Abstract of the United States, the correlation between the percentage of the female population with a bachelor's degree and the percentage of births to unmarried mothers since 1990 is 0.940.

Does this mean that a higher percentage of females with bachelor's degrees causes a higher percentage of births to unmarried mothers?

Certainly not! The correlation exists only because both percentages have been increasing since 1990. It is this relation that causes the high correlation. In general, time series data (data collected over time) may have high correlations because each variable is moving in a specific direction over time (both going up or down over time; one increasing, while the other is decreasing over time).

When data are observational, we cannot claim a causal relation exists between two variables. We can only claim causality when the data are collected through a designed experiment.

Another way that two variables can be related even though there is not a causal relation is through a *lurking variable*.

A ***lurking variable*** is related to both the explanatory and response variable.

> For ***example***, ice cream sales and crime rates have a very high correlation. Does this mean that local governments should shut down all ice cream shops? No! The lurking variable is temperature. As air temperatures rise, both ice cream sales and crime rates rise.

Example

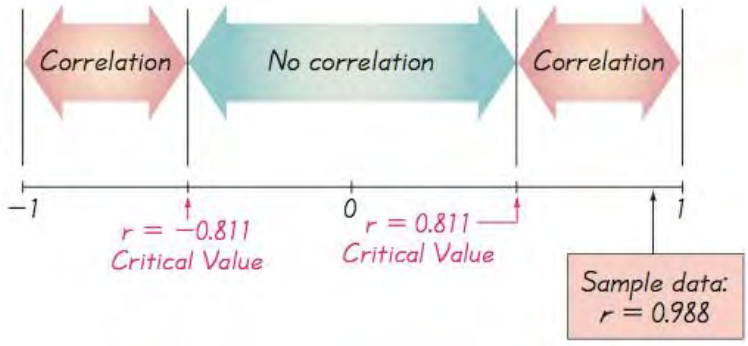
In prospective cohort studies, data are collected on a group of subjects through questionnaires and surveys over time. Therefore, the data are observational. So the researchers cannot claim that increased cola consumption causes a decrease in bone mineral density.
Some lurking variables in the study that could confound the results are:

- body mass index
- height
- smoking
- alcohol consumption
- calcium intake
- physical activity

The authors were careful to say that increased cola consumption is associated with lower bone mineral density because of potential lurking variables. They never stated that increased cola consumption causes lower bone mineral density.

Using Table Critical Values of Spearman's Rank Correlation Coefficient r_s to Interpret r:

If $|r|$ exceeds the value in Critical Values of Spearman's Rank Correlation Coefficient **Table**, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.



Interpreting *r*: Explained Variation

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

Example

Using the pizza subway fare costs in Table below, we have found that the linear correlation coefficient is $r = 0.988$. What proportion of the variation in the subway fare can be explained by the variation in the costs of a slice of pizza?

Table *– Cost of a Slice of Pizza, subway Fare, and the CPI*

Year	**1960**	**1973**	**1986**	**1995**	**2002**	**2003**
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

Solution

With $r = 0.988$, we get $r^2 = 0.976$.

We conclude that 0.976 (or about 98%) of the variation in the cost of a subway fares can be explained by the linear relationship between the costs of pizza and subway fares. This implies that about 2% of the variation in costs of subway fares cannot be explained by the costs of pizza.

Common Errors Involving Correlation

1. ***Causation***: It is wrong to conclude that correlation implies causality.
2. ***Averages***: Averages suppress individual variation and may inflate the correlation coefficient.
3. ***Linearity***: There may be some relationship between x and y even when there is no linear correlation.

TI-83/84 PLUS Enter the paired data in lists L1 and L2, then press **STAT** and select **TESTS.** Using the option of **LinRegTTest** will result in several displayed values, including the value of the linear correlation coefficient *r*. To obtain a scatterplot, press **2nd,** then **Y =** (for STAT PLOT). Press **Enter** twice to turn Plot 1 on, then select the first graph type, which resembles a scatterplot. Set the *X* list and *Y* list labels to L1 and L2 and press the **ZOOM** key, then select **ZoomStat** and press the **Enter** key.

Exercises *Section* 2.1 – Correlation

1. For each of several randomly selected years, the total number of points scored in the Super Bowl football game and the total number of new cars sold in The U.S. are recorded. For this sample of paired data

a) What does r represent?

b) What does ρ represent?

c) With our doing any research or calculations, estimate the value of r.

2. The heights (in inches) of a sample of eight mother/daughter pairs of subjects measured. Using Excel with the paired mother/daughter heights, the linear correlation coefficient is found to be 0.693. Is there sufficient evidence to support the claim that there is a linear correlation between the heights of mothers and the heights of their daughters? Explain.

3. The heights and weights of a sample of 9 supermodels were measured. Using a TI calculator, the linear correlation coefficient is found to be 0.360. Is there sufficient evidence to support the claim that there is a linear correlation between the heights and weights of supermodels? Explain.

4. Given the table below

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

a) Construct a scatterplot

b) Find the value of linear correlation coefficient r and then determine whether there is sufficient evidence to support the claim of a linear correlation between the 2 variables.

c) Identify the feature of the data that would be missed if part (*b*) was completed without constructing the scatterplot.

5. Given the table below

x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

a) Construct a scatterplot

b) Find the value of linear correlation coefficient r and then determine whether there is sufficient evidence to support the claim of a linear correlation between the 2 variables.

c) Identify the feature of the data that would be missed if part (*b*) was completed without constructing the scatterplot.

6. The paired values of the Consumer Price Index (CPI) and the cost of a slice of pizza are shown below

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

a) Construct a scatterplot

b) Find the value of linear correlation coefficient r.

7. Listed below are systolic blood pressure measurements (in mm HG) obtained from the same woman.

Right Arm	102	101	94	79	79
Left Arm	175	169	182	146	144

a) Construct a scatterplot

b) Find the value of linear correlation coefficient *r*.

8. Listed below are costs (in dollars) of air fares for different airlines from NY to San Francisco. The costs are based on tickets purchased 30 days in advance and one day in advance.

30 Days	244	260	264	264	278	318	280
One Day	456	614	567	943	628	1088	536

a) Construct a scatterplot

b) Find the value of linear correlation coefficient *r*.

9. Listed below are repair costs (in dollars) for cars crashed at 6 mi/h in full-front crash tests and the same cars crashed at 6 mi/f in full-rear crash tests.

Front	936	978	2252	1032	3911	4312	3469
Rear	1480	1202	802	3191	1122	739	2767

a) Construct a scatterplot

b) Find the value of linear correlation coefficient *r* .

10. For the following data:

x	2	4	7	7	9
y	1.6	2.1	2.4	2.6	3.2

a) Draw a scatter diagram

b) Compute the correlation coefficient

c) Comment on the type of the relation that appears to exist between *x* and *y*.

11. A pediatrician wants to determine the relation that may exist between a child's height and head circumference. She randomly selects 8 children, measures their height and head cieecumference, and obtains the data shown in the table.

***Height* (*in.*)**	27.25	25.75	26.25	25.75	27.75	26.75	25.75	26.75
***Head Circumference* (*in.*)**	17.6	17	17.2	16.9	17.6	17.3	17.2	17.4

a) Draw a scatter diagram

b) Compute the correlation coefficient

c) If the pediatrician wants to use height to predict head circumference, determine which variable is the explanatory variable and which is the response variable.

d) Does a linear relation exist between height and head circumference?

12. An engineer wanted to determine how the weight of a car affects gas mileage. The accompanying data represent the weight of various domestic cars and their mileages in the city for the 2008 model year. Suppose that we add Car 12 to the original data. Car 12 weights 3,305 *lbs*. and gets 19 miles per gallon.

Car	***Weight (lbs)***	***Miles / Gal***
1	3,775	21
2	3,964	17
3	3,470	21
4	3,175	22
5	2,580	27
6	3,730	18
7	2,605	26
8	3,772	17
9	3,310	20
10	2,991	25
11	2,752	26

a) Draw a scatter diagram

b) Compute the correlation coefficient.

c) Compute the correlation coefficient with Car 12 included

d) Compare the correlation coefficient in part (*b*) & (*c*), and why are the results reasonable.

e) Suppose that we add Car 13 (a hybrid car) to the original data (remove the Car 12). Car 13 weighs 2,890 *lbs*. and gets 60 miles per gallon. Compute the linear coefficient with Car 13 included,