

Relation between two variables

MA 116

June 2025

Items marked with *** are important concepts that may be tested on quizzes or exams.

Response variable vs. Explanatory variable

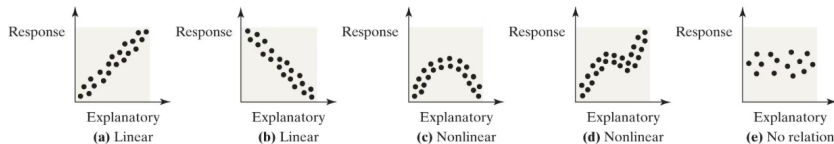
Explanatory variable is used to predict the value of the **response variable**.

Example. Explanatory variable=the speed at which a golf club is swung;
response variable=the distance the golf ball travels.

Scatter diagram

Scatter diagrams show the type of relation that exists between two variables. Our goal in interpreting scatter diagrams is to distinguish scatter diagrams that imply a linear relation, a nonlinear relation, or no relation. Figure 2 displays various scatter diagrams and the type of relation implied.

Figure 2

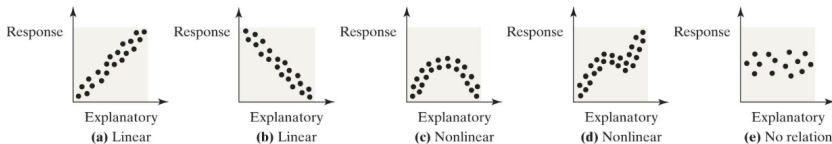


Explanatory variable on the horizontal axis, response variable on the vertical axis.

It is not always clear which variable should be considered the response variable and which the explanatory variable. For example, does high school GPA predict a student's SAT score or can the SAT score predict GPA? The researcher must determine the role of each variable based on the question they want to answer.

Scatter diagrams show the type of relation that exists between two variables. Our goal in interpreting scatter diagrams is to distinguish scatter diagrams that imply a linear relation, a nonlinear relation, or no relation. Figure 2 displays various scatter diagrams and the type of relation implied.

Figure 2

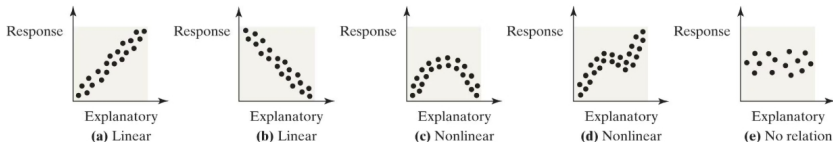


Two kinds of linear relations.

(a) shows a linear relation such that the two variables are linearly related and positively associated: When the explanatory variable assumes a high value, the response variable assumes a high value as well. When the explanatory variable assumes a low value, the response variable assumes a low value as well.

Scatter diagrams show the type of relation that exists between two variables. Our goal in interpreting scatter diagrams is to distinguish scatter diagrams that imply a linear relation, a nonlinear relation, or no relation. Figure 2 displays various scatter diagrams and the type of relation implied.

Figure 2



Two kinds of linear relations.

(b) shows a linear relation such that the two variables are linearly related and negatively associated: When the explanatory variable assumes a high value, the response variable assumes a low value. When the explanatory variable assumes a low value, the response variable assumes a high value.

Are we again dealing with matched pair data?

In a scatter diagram, we are plotting **paired data** with a data point denoted by (x_i, y_i) . The term matched pair data is (in this course) reserved for inference about two means, in which $d_i = x_i - y_i$ makes sense and is what we care about.

Sample linear correlation coefficient

The **sample linear correlation coefficient** is a measure of the strength and direction of the linear relation between two quantitative variables.

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

where

x_i is the i th observation of the explanatory variable

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

s_x is the sample standard deviation of the explanatory variable

y_i is the i th observation of the response 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



(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

Properties of the sample linear correlation coefficient r .

• $-1 \leq r \leq 1$

4. The closer r is to $+1$, the stronger is the evidence of positive association between the two variables. See Figures 4(b) and 4(c).
5. The closer r is to -1 , the stronger is the evidence of negative association between the two variables. See Figures 4(e) and 4(f).
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.** See Figures 4(g) and 4(h).
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.

Correlation \neq Causation

If data used in a study are observational, we can not conclude that the two correlated variables have a casual relationship. A linear correlation coefficient that implies a strong positive or negative association (i.e. r close to ± 1) **does not** imply causation if it was computed using observational data.

Example. As air-conditioning bills increases, so does the crime rate. Can I argue that folks should turn off their air conditioners so that crime rates decrease? No. In fact, it is the rising temperature that **causes** both the air-conditioning bills and crime rate to increase.

4.2 Least-Squares Regression

If two variables have a linear relation, we'd like to find a linear equation that describes this relation.



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



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



Least-Square regression line

$$\hat{y} = b_1x + b_0$$

where $b_1 = r \cdot \frac{s_y}{s_x}$ is the slope of the least square regression line, and $b_0 = \bar{y} - b_1\bar{x}$ is the y -intercept of the least squares line. The notation \hat{y} is used in the least square regression line to remind us that it is a predicted value of y for a given value of x .

Properties of this line.

- 1 $\hat{y} = b_1x + b_0$ is a mathematical formula that assigns to **any number** x a number \hat{y} . It extends infinitely in both directions. In practice, the explanatory variable often could assume values within a smaller range. For example, if I have a linear relation with explanatory variable number of cups of coffee a student drink per day, then x can only assume nonnegative integers, despite the fact that \hat{y} has a value $-0.5b_1 + b_0$ at $x = -0.5$.
- 2 The line $\hat{y} = b_1x + b_0$ always passes through the point (\bar{x}, \bar{y}) .

Example.

*** Consider the data given in the table. Compute the linear correlation coefficient r and then the least square regression line.

x	y
1	18
3	13
3	9
6	6
7	4

First compute \bar{x} , s_x , \bar{y} , s_y , since the formula for the linear correlation coefficient r is

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

*** Consider the data given in the table. Compute the linear correlation

x	y
1	18
3	13
3	9
6	6
7	4

coefficient r and then the least square regression line.

find that $\bar{x} = 4$, $s_x = 2.4$, $\bar{y} = 10$, $s_y = 5.6$. Then, note that

We

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

From this we get $r \approx -0.9$.

*** Consider the data given in the table. Compute the linear correlation

x	y
1	18
3	13
3	9
6	6
7	4

coefficient r and then the least square regression line.

find that $\bar{x} = 4$, $s_x = 2.45$, $\bar{y} = 10$, $s_y = 5.61$, $r = -0.946$. Then

We

$$b_1 = r \cdot \frac{s_y}{s_x} = -2.1676$$

and

$$b_0 = \bar{y} - b_1\bar{x} = 18.67.$$

So our least square regression line is

$$\hat{y} = -2.1676x + 18.67$$

Interpret the slope

Given that the least square regression line for a sample data set with explanatory variable=the speed at which a golf club is swung; response variable=the distance the golf ball travels is

$$\hat{y} = 3.166x + 55.797.$$

How do we interpret this slope 3.166?

There are two interpretations.

- 1 If club-head speed increases by 1 mile per hour, the **expected** distance the golf ball will travel increases by 3.166 yards.
- 2 If club-head speed increases by 1 mile per hour, the distance the golf ball travel increases by 3.166 yards, **on average**.

Residual

*** Suppose I have a least square regression line $\hat{y} = b_1x + b_0$ for a sample. For each data point (x_i, y_i) in my sample, its residual is $y_i - \hat{y}$, where the \hat{y} is its value at x_i (i.e. $b_1x_i + b_0$).

Residual is defined for other line

Given another line $f(x) = a_0x + a_1$ we can also consider the residual of a data point (x_i, y_i) with respect to this line, that is, $y_i - f(x_i)$.

Result

Fix a sample with data points like (x_i, y_i) . For any line we can consider the sum of squared residual of this sample $\sum_i (y_i - f(x_i))^2$. The least square regression line of a sample, by design, always have the least sum of squared residual of this sample, among all possible lines of the form $f(x) = a_0x + a_1$.

Goodness of slope interpretation?

Given that the least square regression line for a sample data set with explanatory variable=the speed at which a golf club is swung; response variable=the distance the golf ball travels is

$$\hat{y} = 3.166x + 55.797.$$

Suppose a club-head speed is 103 mph. The least square regression line predicts that the distance of the shot to be
 $\hat{y} = 3.166 \cdot 103 - 55.78 = 270.3$ yards.

How good is this prediction? In other words, we want to figure out how good the least square regression line describes how changes in the explanatory variable affect the value of the response variable.

To answer this question, let's define a new quantity

Coefficient of determination R^2

Consider the response variable y and its variation s_y^2 . Due to the existence of a linear relation between x and y , it makes sense to say that part of s_y^2 is explained by the least square regression line. It can be shown that (although beyond the scope of this course)

$$s_y^2 = \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y})^2 + \sum_i (\hat{y} - \bar{y})^2,$$

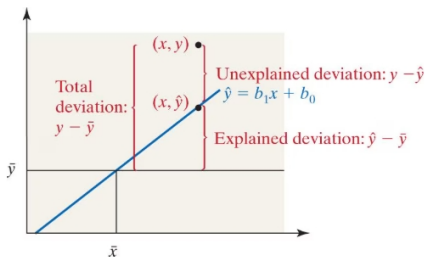
where $s_y^2 = \sum_i (y_i - \bar{y})^2$ is the total deviation, $\sum_i (y_i - \hat{y})^2$ is called the unexplained deviation, and $\sum_i (\hat{y} - \bar{y})^2$ is called the explained deviation (note in this context $\hat{y} = b_1 x_i + b_0$ is the \hat{y} value at each x_i). Define

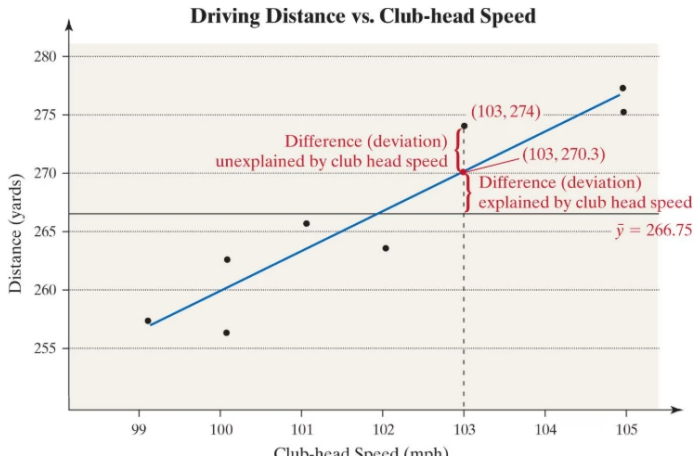
$$R^2 = \frac{\sum_i (\hat{y} - \bar{y})^2}{s_y^2} = 1 - \frac{\sum_i (y_i - \hat{y})^2}{s_y^2}$$

Coefficient of determination R^2

R^2 is $\frac{\text{explained variation}}{\text{total variation}}$. It is the proportion of total variation in the response variable that is explained by the least-squares regression line.


- 1 We always have that $0 \leq R^2 \leq 1$.
- 2 The closer R^2 is to 1, the better the least-squares regression line describes the relation between the explanatory and response variables.





Note the unexpected variation $\sum_i (y_i - \hat{y})^2$ is just the sum of squared residue. The smaller the sum of squared residue of a sample $\{(x_i, y_i)\}$ and its least-squares regression line is, the larger R^2 will be (i.e. closer to 1).

Example.

17.  **An Unhealthy Commute** (Refer to [Problem 27](#), Section 4.1.) The following data represent commute times (in minutes) and score on a well-being survey.


Commute Time (minutes), x	Gallup-Healthways Well-Being Index Composite Score, y
5	69.2
15	68.3
25	67.5
35	67.1
50	66.4
72	66.1
105	63.9



Source: The Gallup Organization.

- 1 Compute the least-squares regression line.
- 2 Predict the well-being index of a person whose commute is 30 minutes.

Example.

17.  **An Unhealthy Commute** (Refer to [Problem 27](#), Section 4.1.) The following data represent commute times (in minutes) and score on a well-being survey.

Commute Time (minutes), x	Gallup-Healthways Well-Being Index Composite Score, y
5	69.2
15	68.3
25	67.5
35	67.1
50	66.4
72	66.1
105	63.9



Source: The Gallup Organization.

- 1 Suppose Barbara has a 20-minute commute and scores 67.333 on the survey. Is Barbara more “well-off” than the typical individual who has a 20-minute commute?
- 2 Compute R^2 .