Correlation Regression

Correlation

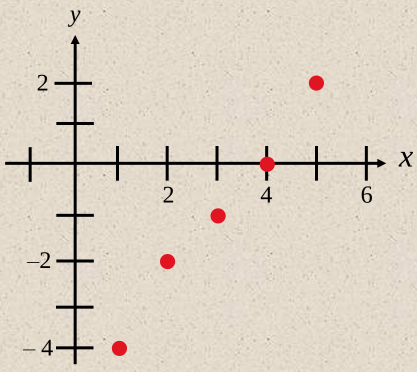
Correlation

A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y) where x is the **independent** (or **explanatory**) **variable**, and y is the **dependent** (or **response**) **variable**.

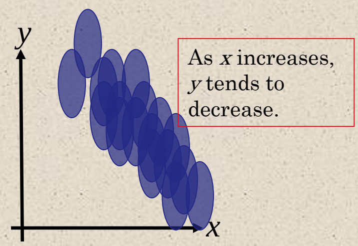
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

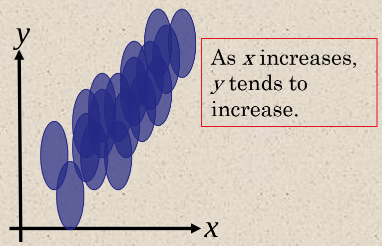
x	1	2	3	4	5
y	– 4	– 2	– 1	0	2



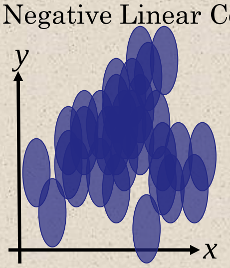
Linear Correlation



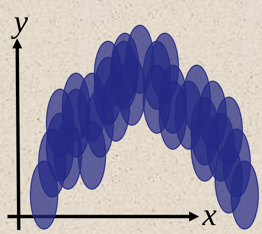
Negative Linear Correlation



Positive Linear Correlation



No Correlation



Nonlinear Correlation

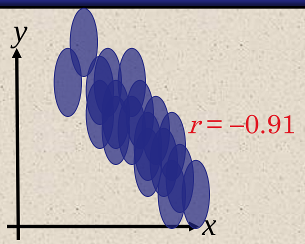
Correlation Coefficient

The **correlation coefficient** is a measure of the strength and the direction of a linear relationship between two variables. The symbol r represents the sample correlation coefficient. The formula for r is

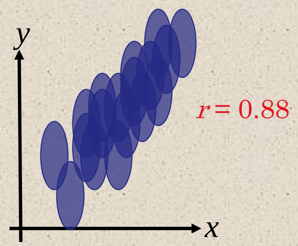
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

The range of the correlation coefficient is –1 to 1. If x and y have a strong positive linear correlation, r is close to 1. If x and y have a strong negative linear correlation, r is close to –1. If there is no linear correlation or a weak linear correlation, r is close to 0.

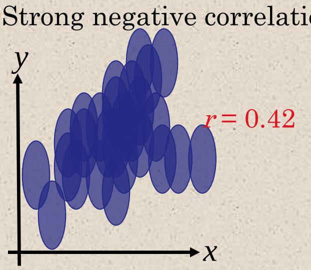
Linear Correlation



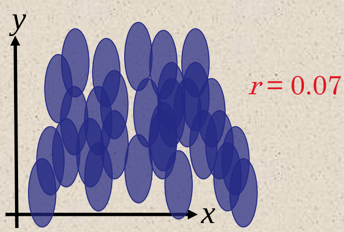
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

Calculating a Correlation Coefficient

Calculating a Correlation Coefficient

In Words	*In Symbols*
1. Find the sum of the *x*-values.	$\sum x$
2. Find the sum of the *y*-values.	$\sum y$
3. Multiply each *x*-value by its corresponding *y*-value and find the sum.	$\sum xy$
4. Square each *x*-value and find the sum.	$\sum x^2$
5. Square each *y*-value and find the sum.	$\sum y^2$
6. Use these five sums to calculate the correlation coefficient.	$r = \dfrac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$

Correlation Coefficient

Example:

Calculate the correlation coefficient r for the following data.

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{5(9) - (15)(-1)}{\sqrt{5(55) - 15^2}\sqrt{5(15) - (-1)^2}}$$

$$= \frac{60}{\sqrt{50}\sqrt{74}} \approx 0.986$$

There is a strong positive linear correlation between x and y.

Correlation Coefficient

Example:

The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Display the scatter plot.

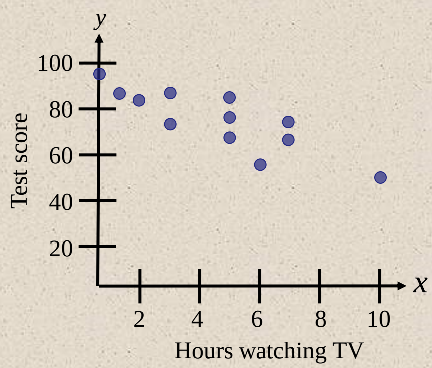
b.) Calculate the correlation coefficient *r*.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50



Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$\sum x = 54$ $\quad$ $\sum y = 908$ $\quad$ $\sum xy = 3724$ $\quad$ $\sum x^2 = 332$ $\quad$ $\sum y^2 = 70836$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{12(3724) - (54)(908)}{\sqrt{12(332) - 54^2}\sqrt{12(70836) - (908)^2}} \approx -0.831$$

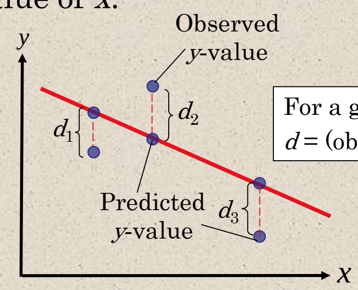
There is a strong negative linear correlation.

As the number of hours spent watching TV increases, the test scores tend to decrease.

Linear Regression

Residuals

After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that can be used to predict the value of y for a given value of x.



For a given x-value,
d = (observed y-value) – (predicted y-value)

Each data point d_i represents the difference between the observed y-value and the predicted y-value for a given x-value on the line. These differences are called **residuals**.

Regression Line

A **regression line**, also called a **line of best fit**, is the line for which the sum of the squares of the residuals is a minimum.

The Equation of a Regression Line

The equation of a regression line for an independent variable x and a dependent variable y is

$$\hat{y} = mx + b$$

where $\hat{y}$ is the predicted y-value for a given x-value. The slope m and y-intercept b are given by

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \text{ and } b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

where $\bar{y}$ is the mean of the y-values and $\bar{x}$ is the mean of the x-values. The regression line always passes through $(\bar{x}, \bar{y})$.

Regression Line

Example:

Find the equation of the regression line.

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{5(9) - (15)(-1)}{5(55) - (15)^2} = \frac{60}{50} = 1.2$$

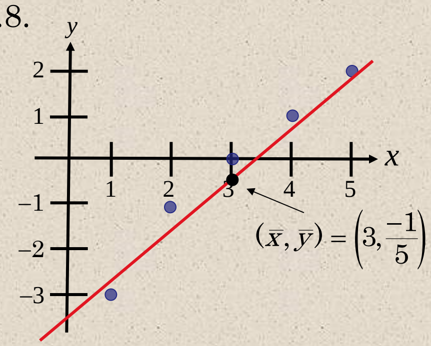
Regression Line

Example continued:

$$b = \bar{y} - m\bar{x} = \frac{-1}{5} - (1.2)\frac{15}{5} = -3.8$$

The equation of the regression line is

$$\hat{y} = 1.2x - 3.8.$$



Regression Line

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Find the equation of the regression line.

b.) Use the equation to find the expected test score for a student who watches 9 hours of TV.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy	0	85	164	222	285	340	380	420	348	455	525	500
x^2	0	1	4	9	9	25	25	25	36	49	49	100
y^2	9216	7225	6724	5476	9025	4624	5776	7056	3364	4225	5625	2500

$\sum x = 54$ $\sum y = 908$ $\sum xy = 3724$ $\sum x^2 = 332$ $\sum y^2 = 70836$

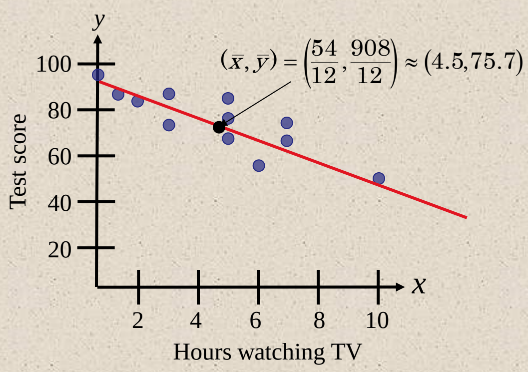
Regression Line

Example continued:

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{12(3724) - (54)(908)}{12(332) - (54)^2} \approx -4.067$$

$$b = \bar{y} - m\bar{x}$$
$$= \frac{908}{12} - (-4.067)\frac{54}{12}$$
$$\approx 93.97$$

$$\hat{y} = -4.07x + 93.97$$



Regression Line

Example continued:

Using the equation $\hat{y} = -4.07x + 93.97$, we can predict the test score for a student who watches 9 hours of TV.

$$\begin{aligned}\hat{y} &= -4.07x + 93.97\\ &= -4.07(9) + 93.97\\ &= 57.34\end{aligned}$$

A student who watches 9 hours of TV over the weekend can expect to receive about a 57.34 on Monday's test.

Simple Linear Regression Analysis

The following table shows the sales revenue and advertisement expenditure data of a firm. Get the regression equation of sale revenue on advertisement expenditure. Also, provide an estimate of the sales revenue for an advertisement expenditure of $9 million.

Sales Revenue ($ Million)	*Advertisement Expenditure ($ Million)*
16	4
12	3
20	7
17	2
18	6
25	8

Simple Linear Regression Analysis

Solution

Suppose Y is sales revenue and X is advertisement expenditure.

Table 13.1:Calculation of Estimated Regression Equation

Y	X	$(X-\overline{X})^2$	$(X-\overline{X})\ (Y-\overline{Y})$
16	4	$(4-5)^2 = 1$	$(4-5)(16-18) = 2$
12	3	$(3-5)^2 = 4$	$(3-5)(12-18) = 12$
20	7	$(7-5)^2 = 4$	$(7-5)(20-18) = 4$
17	2	$(2-5)^2 = 9$	$(2-5)(17-18) = 3$
18	6	$(6-5)^2 = 1$	$(6-5)(18-18) = 0$
25	8	$(8-5)^2 = 9$	$(8-5)(25-18) = 21$
$\Sigma Y = 108$	$\Sigma X = 30$	$\sum(X-\overline{X})^2 = 28$	$\sum(X-\overline{X})(Y-\overline{Y}) = 42$

$$\hat{\propto} = \overline{Y} - \hat{\beta}\,\overline{X}$$

$$\hat{\beta} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

Or

$$\hat{\beta} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}$$

$$\overline{Y} = \frac{108}{6} = 18$$

$$\overline{X} = \frac{30}{6} = 5$$

$$\hat{\beta} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} = \frac{42}{28} = 1.5$$

$$\hat{\propto} = \overline{Y} - \hat{\beta}\,\overline{X} = 18 - (1.5)(5) = 10.5$$

The estimated regression equation is:

$$\hat{Y} = 10.5 + 1.5X$$

If the advertisement is \$9 million, then estimated sale would be \$24 million

$$\hat{Y} = 10.5 + 1.5(9) = 24$$

Multiple Regression Equation

In many instances, a better prediction can be found for a dependent (response) variable by using more than one independent (explanatory) variable.

For example, a more accurate prediction of Monday's test grade from the previous section might be made by considering the number of other classes a student is taking as well as the student's previous knowledge of the test material.

A **multiple regression equation** has the form

$$\hat{y} = b + m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_kx_k$$

where x_1, x_2, x_3,…, x_k are independent variables, b is the y-intercept, and y is the dependent variable.

* Because the mathematics associated with this concept is complicated, technology is generally used to calculate the multiple regression equation.

Predicting y-Values

After finding the equation of the multiple regression line, you can use the equation to predict y-values over the range of the data.

Example:
The following multiple regression equation can be used to predict the annual U.S. rice yield (in pounds).

$$\hat{y} = 859 + 5.76x_1 + 3.82x_2$$

where x_1 is the number of acres planted (in thousands), and x_2 is the number of acres harvested (in thousands).
(Source: U.S. National Agricultural Statistics Service)

a.) Predict the annual rice yield when $x_1 = 2758$, and $x_2 = 2714$.

b.) Predict the annual rice yield when $x_1 = 3581$, and $x_2 = 3021$.

Predicting *y*-Values

Example continued:

a.) $\hat{y} = 859 + 5.76x_1 + 3.82x_2$

$= 859 + 5.76(2758) + 3.82(2714)$

$= 27{,}112.56$

The predicted annual rice yield is 27,1125.56 pounds.

b.) $\hat{y} = 859 + 5.76x_1 + 3.82x_2$

$= 859 + 5.76(3581) + 3.82(3021)$

$= 33{,}025.78$

The predicted annual rice yield is 33,025.78 pounds.