Correlation and Regression

G. Ramakrishna

Correlation Coefficient

Correlation is a measure of association between variables

A statistic that quantifies a relation between two variables

Can be either positive or negative

Falls between -1.00 and 1.00

The value of the number (not the sign) indicates the strength of the relation

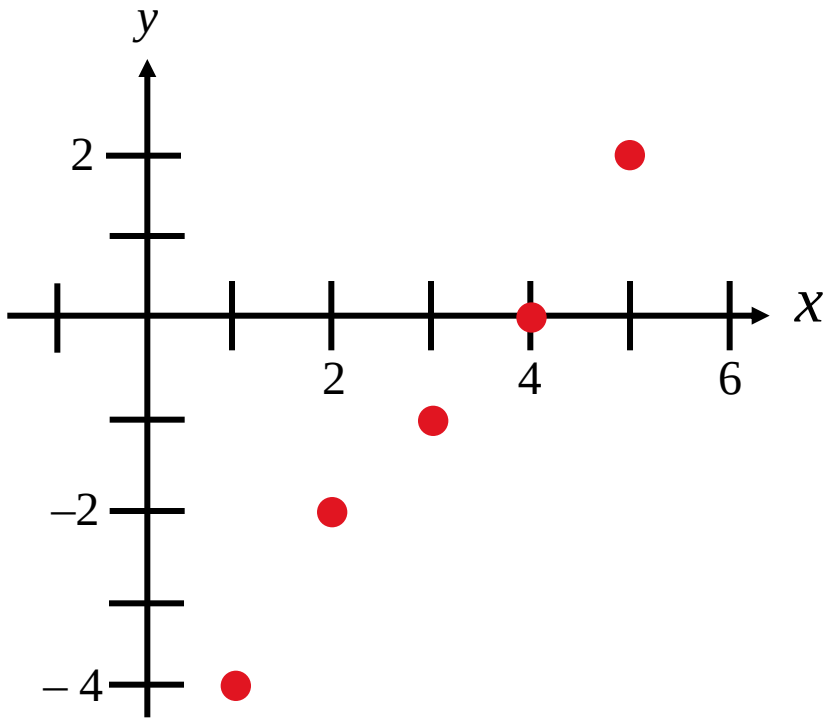
CORRELATION

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**.

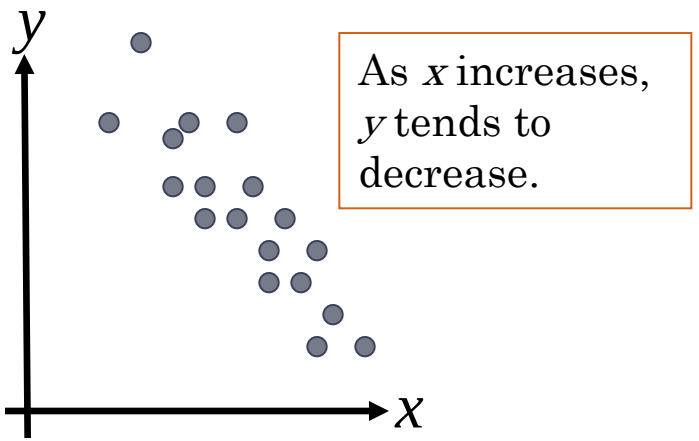
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

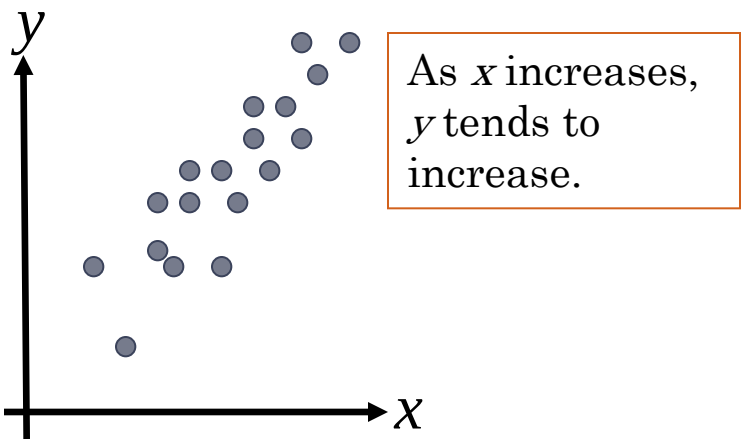
x	1	2	3	4	5
y	-4	-2	-1	0	2



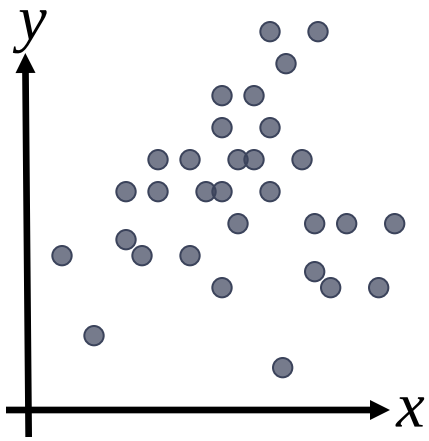
Linear Correlation



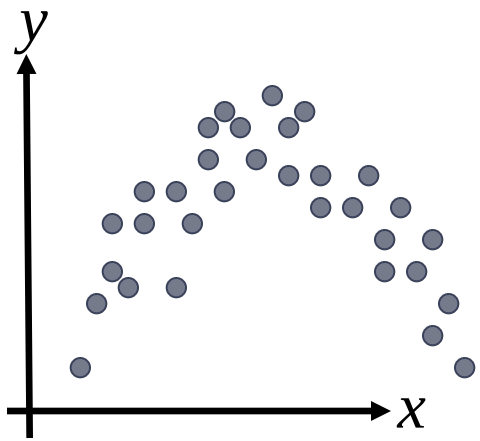
Negative Linear Correlation



Positive Linear Correlation



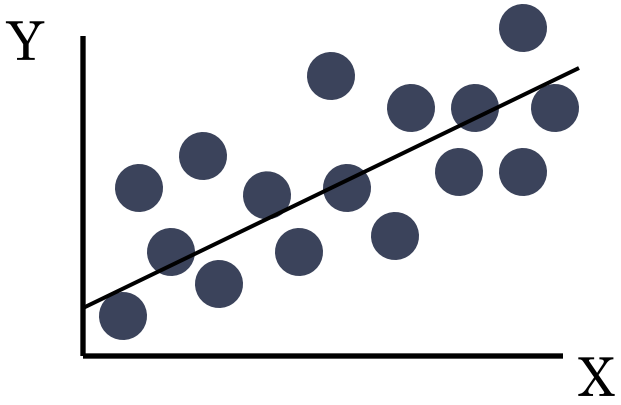
No Correlation

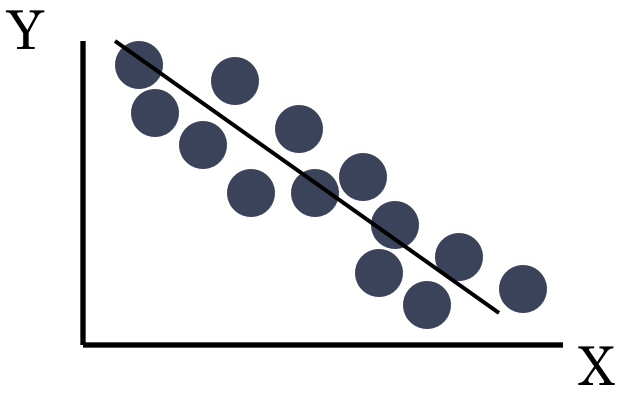


Nonlinear Correlation

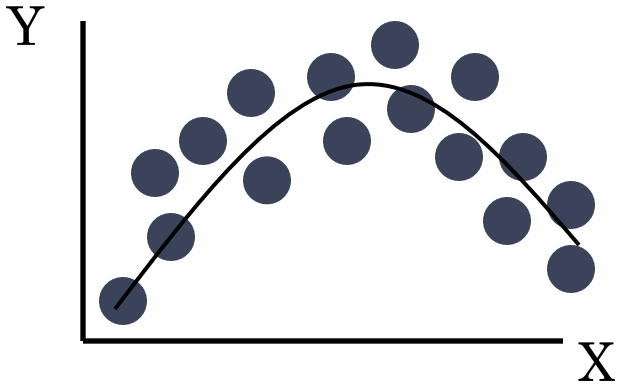
Linear Correlation

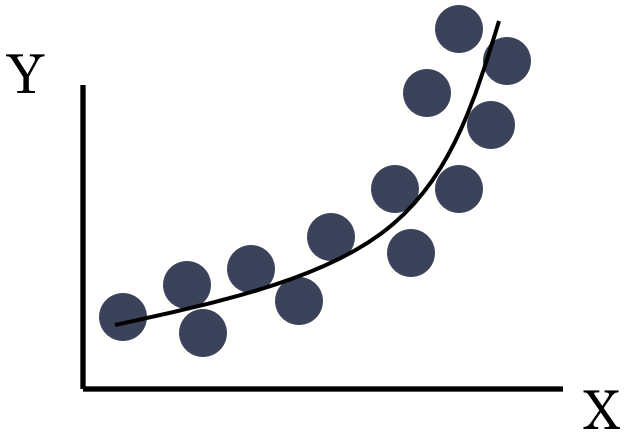
Linear relationships





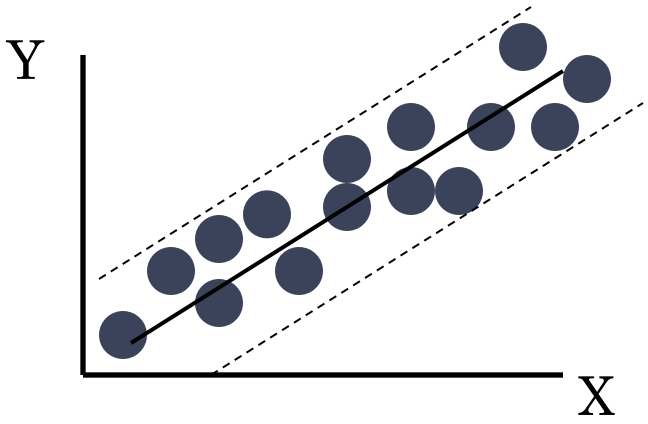
Curvilinear relationships

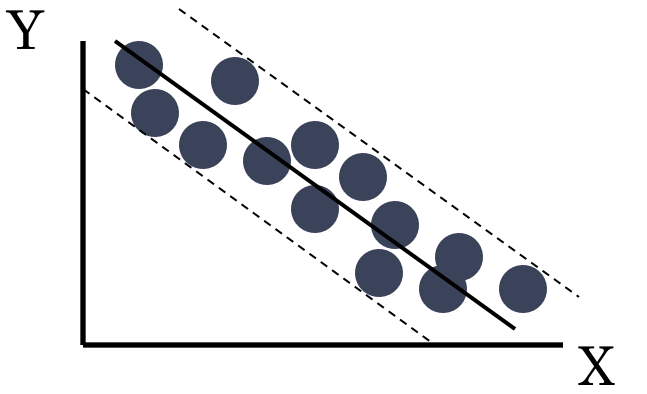




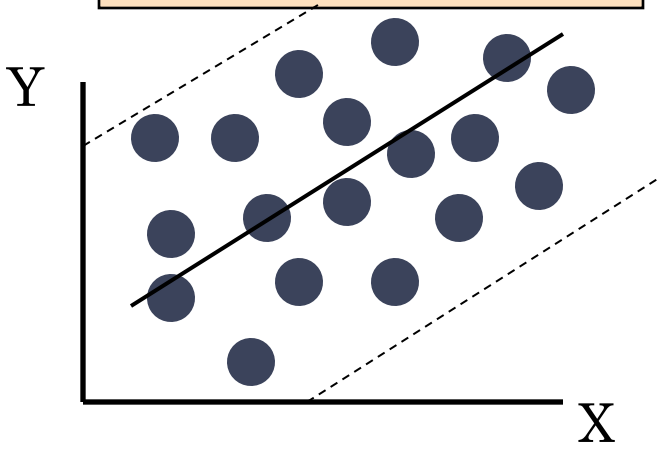
Linear Correlation

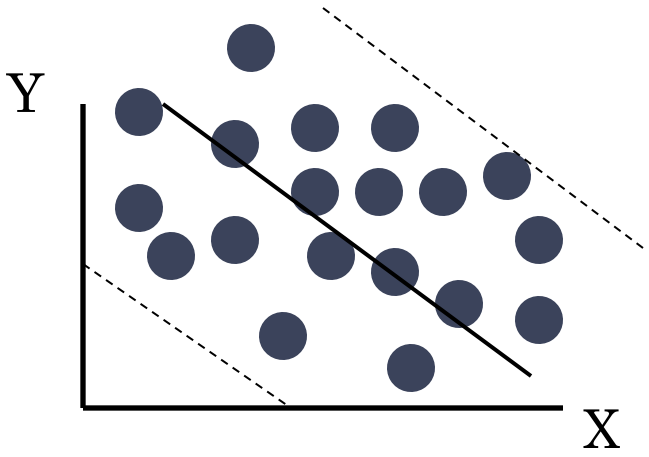
Strong relationships



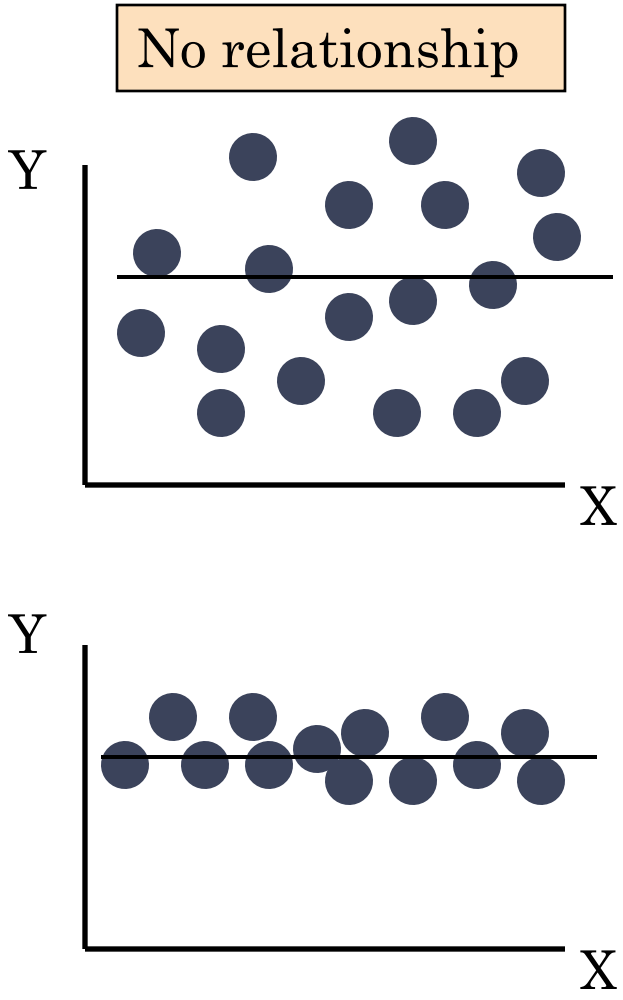


Weak relationships





Linear Correlation



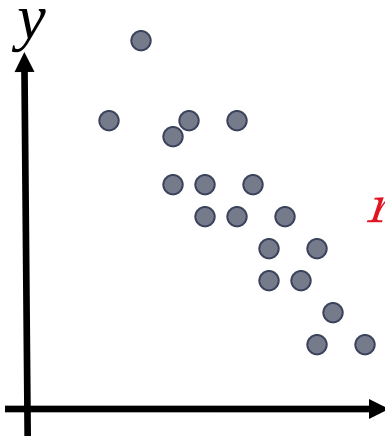
Correlation Coefficient

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

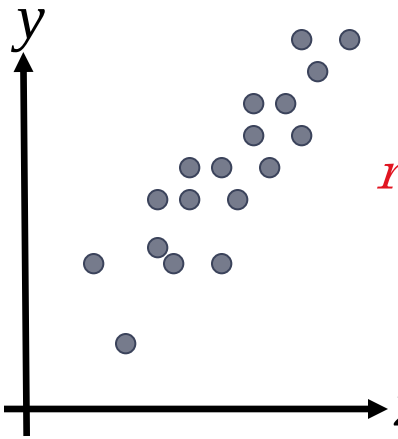
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}.$$

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.

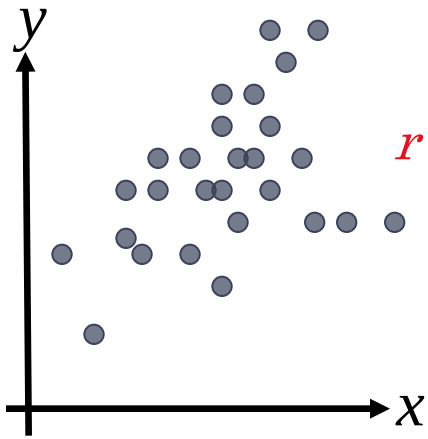
Linear Correlation



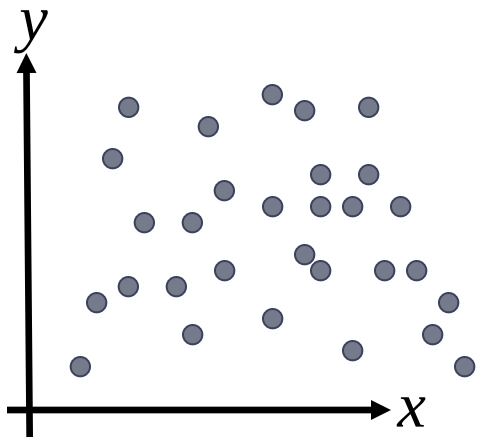
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

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*.

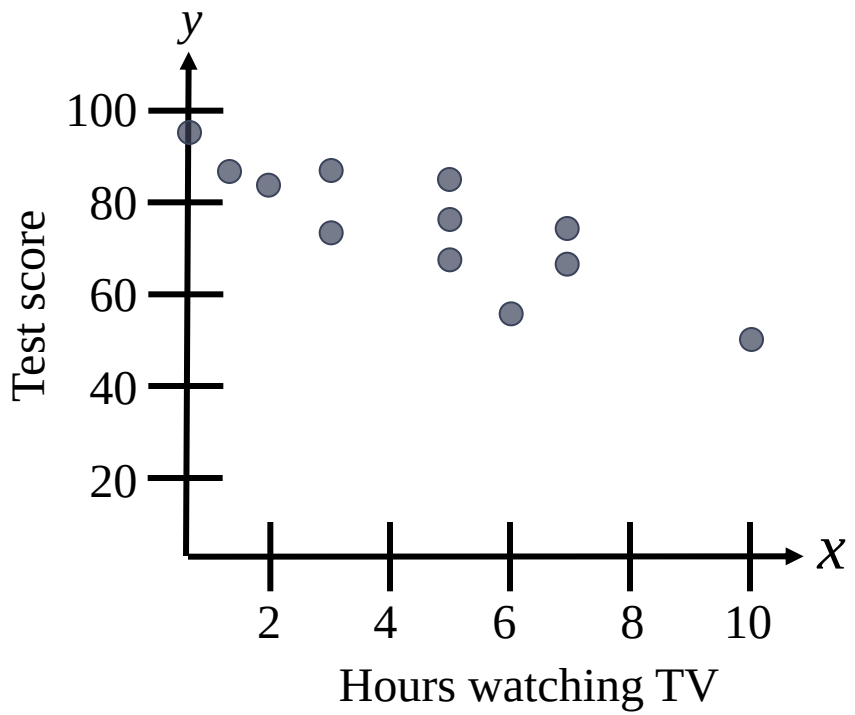
Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Continued.

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



Continued.

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

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

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

There is a strong negative linear correlation.

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

Testing a Population Correlation Coefficient

Once the sample correlation coefficient r has been calculated, we need to determine whether there is enough evidence to decide that the population correlation coefficient ρ is significant at a specified level of significance.

If $|r|$ is greater than the critical value, there is enough evidence to decide that the correlation coefficient ρ is significant.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

For a sample of size $n = 6$, ρ is significant at the 5% significance level, if $|r| > 0.811$.

Hypothesis Testing for *P*

The *t*-Test for the Correlation Coefficient

A ***t*-test** can be used to test whether the correlation between two variables is significant. The **test statistic** is r and the **standardized test statistic**

$$t = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

follows a t-distribution with $n - 2$ degrees of freedom.

Correlation and Causation

The fact that two variables are strongly correlated does not in itself imply a cause-and-effect relationship between the variables.

If there is a significant correlation between two variables, you should consider the following possibilities.

1. **Is there a direct cause-and-effect relationship between the variables?**
   Does x cause y?
2. **Is there a reverse cause-and-effect relationship between the variables?**
   Does y cause x?
3. **Is it possible that the relationship between the variables can be caused by a third variable or by a combination of several other variables?**
4. **Is it possible that the relationship between two variables may be a coincidence?**

Linear Regression

Simple Regression

A statistical model that utilizes <u>one</u> **quantitative** *independent* variable "X" to estimate the **quantitative** *dependent* variable "Y."

- The purpose of regression is to estimate, explain. Predict and evaluate the relation between variables.
- Linear and non- linear regressions relate to how we have entered the coefficients in the model.
- Linear regression estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.

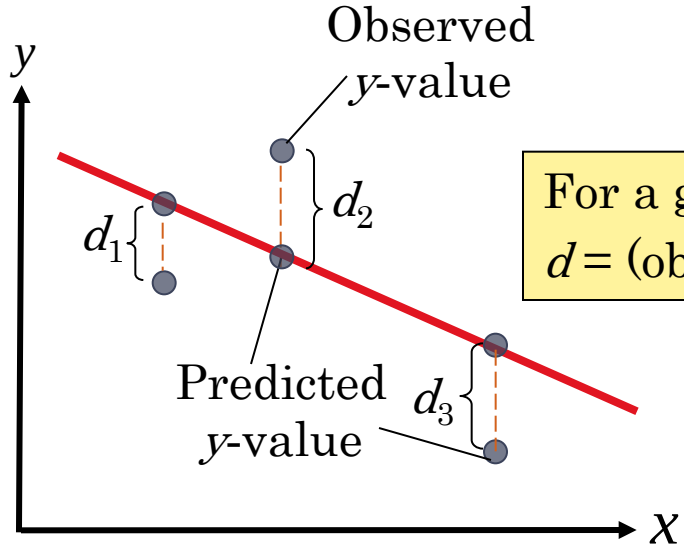
19

Assumptions

- **Linearity** - the Y variable is linearly related to the value of the X variable.
- **Independence of Error** - the error (residual) is independent for each value of X.
- **Homoscedasticity** - the variation around the line of regression be constant for all values of X.
- **Normality** - the values of Y be normally distributed at each value of X.

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

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
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

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

Continued.

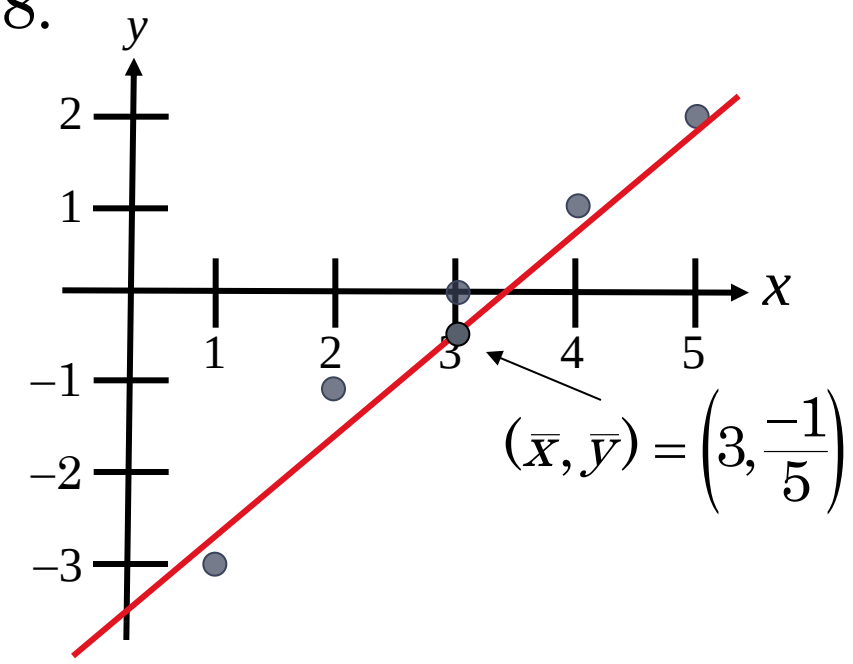
Regression Line

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

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

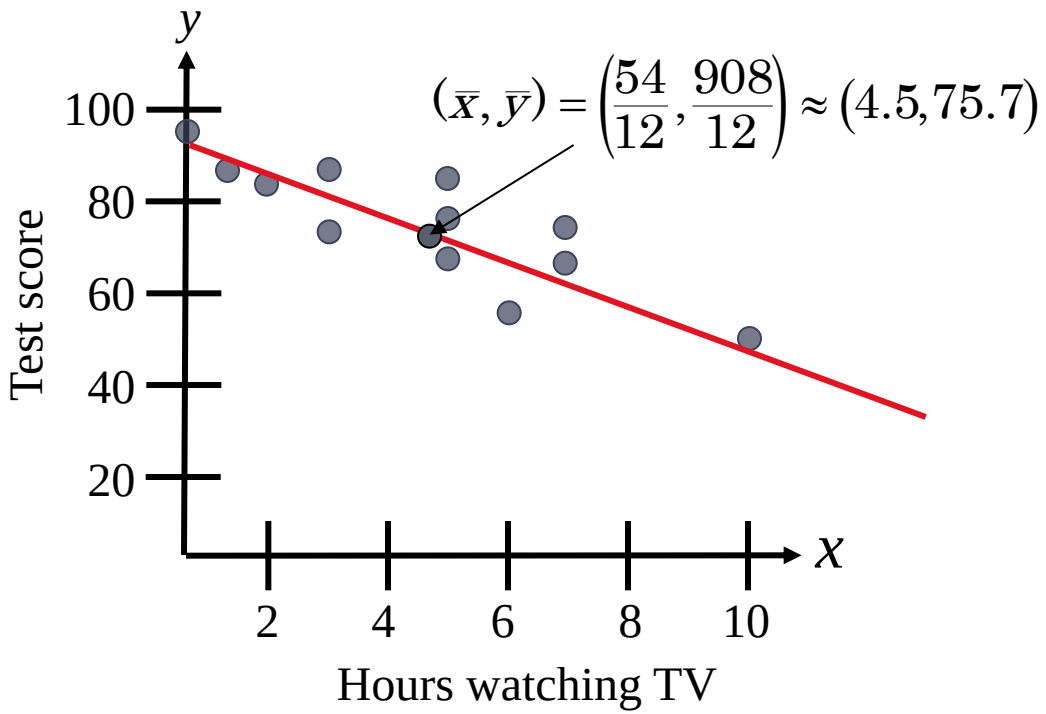
Regression Line

Example continued:

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

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

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



Continued.

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.

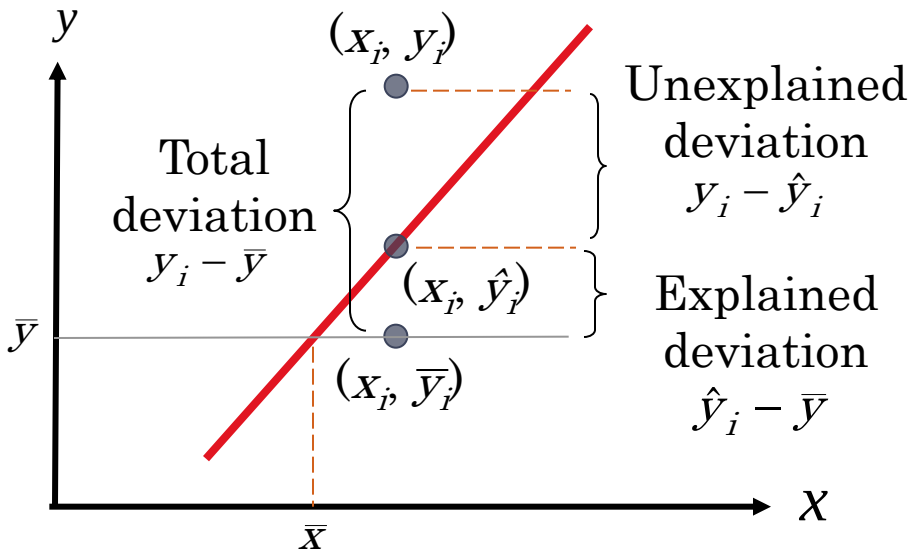
Variation About a Regression Line

To find the total variation, you must first calculate the **total deviation**, the **explained deviation**, and the **unexplained deviation**.

$$\text{Total deviation} = y_i - \bar{y}$$

$$\text{Explained deviation} = \hat{y}_i - \bar{y}$$

$$\text{Unexplained deviation} = y_i - \hat{y}_i$$



Variation About a Regression Line

The **total variation** about a regression line is the sum of the squares of the differences between the *y*-value of each ordered pair and the mean of *y*.

$$\text{Total variation} = \sum (y_i - \bar{y})^2$$

The **explained variation** is the sum of the squares of the differences between each predicted *y*-value and the mean of *y*.

$$\text{Explained variation} = \sum (\hat{y}_i - \bar{y})^2$$

The **unexplained variation** is the sum of the squares of the differences between the *y*-value of each ordered pair and each corresponding predicted *y*-value.

$$\text{Unexplained variation} = \sum (y_i - \hat{y}_i)^2$$

Total variation = Explained variation + Unexplained variation

Coefficient of Determination

The **coefficient of determination r^2** is the ratio of the explained variation to the total variation. That is,

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

Example:
The correlation coefficient for the data that represents the number of hours students watched television and the test scores of each student is $r \approx -0.831$. Find the coefficient of determination.

$$r^2 \approx (-0.831)^2$$
$$\approx 0.691$$

About 69.1% of the variation in the test scores can be explained by the variation in the hours of TV watched. About 30.9% of the variation is unexplained.

The Standard Error of Estimate

When a $\hat{y}$-value is predicted from an x-value, the prediction is a point estimate.

An interval can also be constructed.

The **standard error of estimate s_e** is the standard deviation of the observed y_i-values about the predicted $\hat{y}$-value for a given x_i-value. It is given by

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

where n is the number of ordered pairs in the data set.

The closer the observed y-values are to the predicted y-values, the smaller the standard error of estimate will be.

The Standard Error of Estimate

Example:

The regression equation for the following data is

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

Find the standard error of estimate.

x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	– 3	– 2.6	0.16
2	– 1	– 1.4	0.16
3	0	– 0.2	0.04
4	1	1	0
5	2	2.2	0.04
			$\sum = 0.4$

Unexplained variation

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{0.4}{5-2}} \approx 0.365$$

The standard deviation of the predicted y value for a given x value is about 0.365.

The Standard Error of Estimate

Example:
The regression equation for the data that 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 is

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

Find the standard error of estimate.

Hours, x_i	0	1	2	3	3	5
Test score, y_i	96	85	82	74	95	68
$\hat{y}_i$	93.97	89.9	85.83	81.76	81.76	73.62
$(y_i - \hat{y}_i)^2$	4.12	24.01	14.67	60.22	175.3	31.58

Hours, x_i	5	5	6	7	7	10
Test score, y_i	76	84	58	65	75	50
$\hat{y}_i$	73.62	73.62	69.55	65.48	65.48	53.27
$(y_i - \hat{y}_i)^2$	5.66	107.74	133.4	0.23	90.63	10.69

Continued.

The Standard Error of Estimate

Example continued:

$$\sum(y_i - \hat{y}_i)^2 = 658.25$$

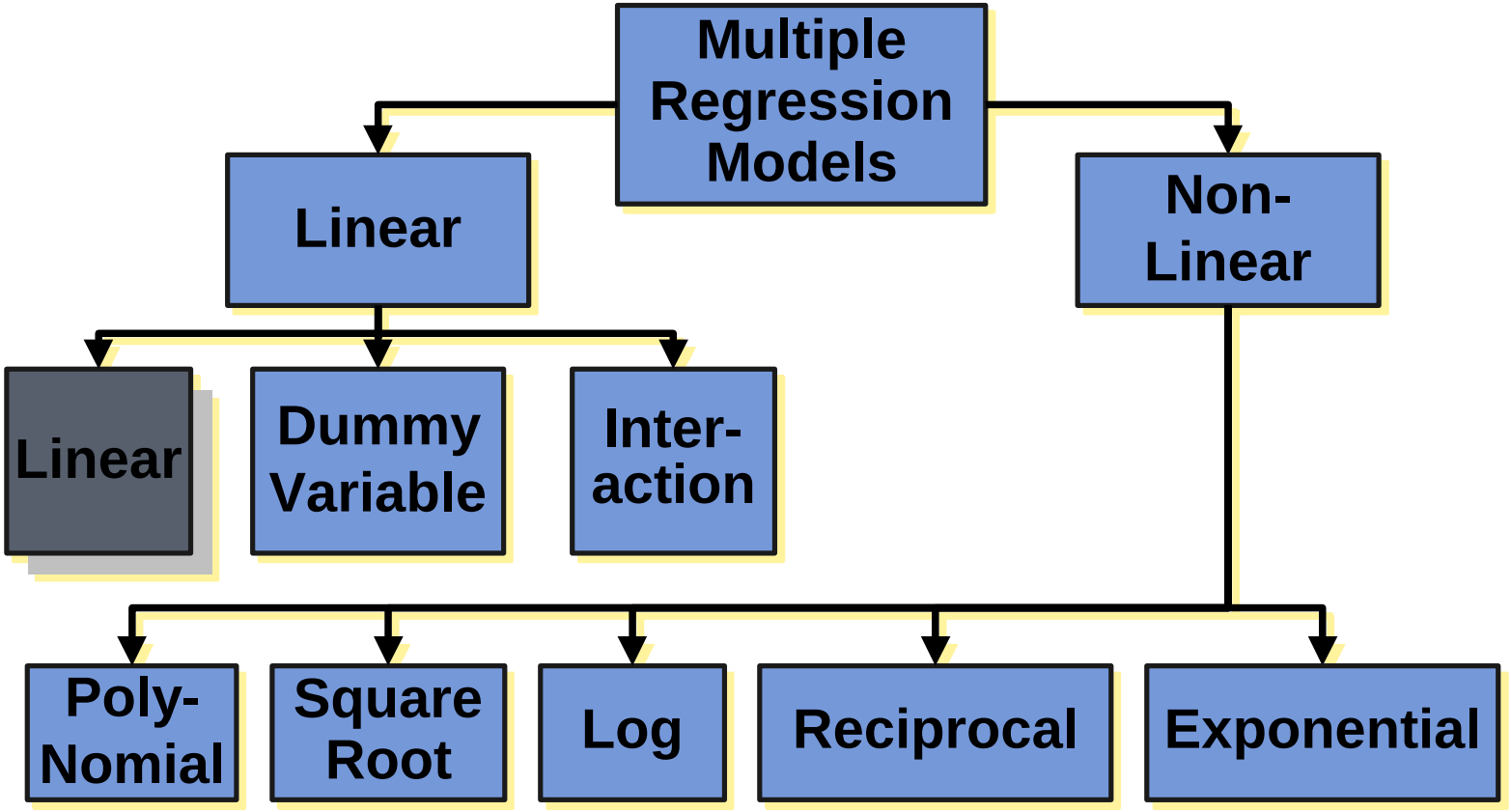
Unexplained variation

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{658.25}{12-2}} \approx 8.11$$

The standard deviation of the student test scores for a specific number of hours of TV watched is about 8.11.

Multiple Regression

Multiple Regression Models



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.

Multiple Regression

- Interpreting a Multiple Regression Equation
- First, let us review how to interpret a bivariate regression equation.
- In the equation
- $y = \alpha + b_1X_1 + e$
- **α = the predicted value of y when $X_1 = 0$**
- **b_1= for every one unit increase in X_1, we predict y to increase by b_1**

Multiple Regression

- Let us say we had the following multiple regression equation:
- $y = \alpha + b_1X_1 + b_2X_2 + e$
- We interpret the equation in the following way:
- α = the predicted value of y when all X's =0
- b_1= for every one unit increase in X_1, we predict y to increase by b_1, holding all other X's equal.
- b_2= for every one unit increase in X_2, we predict y to increase by b_2, holding all other X's equal.

HOW TO TEST HYPOTHESES IN MULTIPLE REGRESSION

- t_1 = $\underline{beta_1}$___ for the first independent variable
- stand err b_1
- And
- t_2 = $\underline{beta_2}$___ for the second independent variable
- stand err b_2
- And if you have three independent variables
- t_3 = $\underline{beta_3}$___ for the third independent variable
- stand err b_3

Thank you