Correlation

Example of Correlation

Is there an association between:

- Children's IQ and Parents' IQ
- Degree of social trust and number of membership in voluntary association ?
- Urban growth and air quality violations?
- GRA funding and number of publication by Ph.D. students
- Number of police patrol and number of crime
- Grade on exam and time on exam

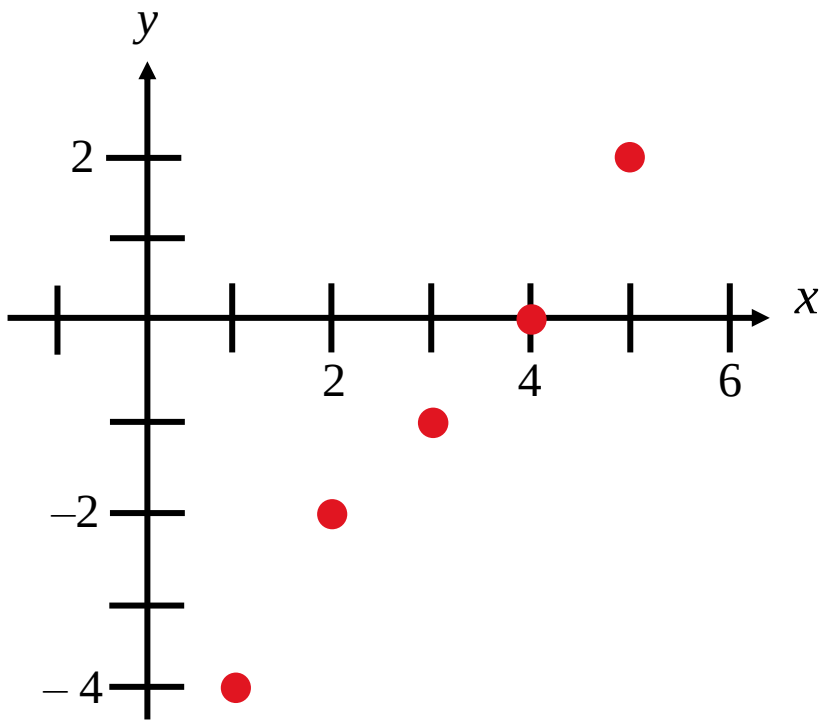
Correlation

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**.

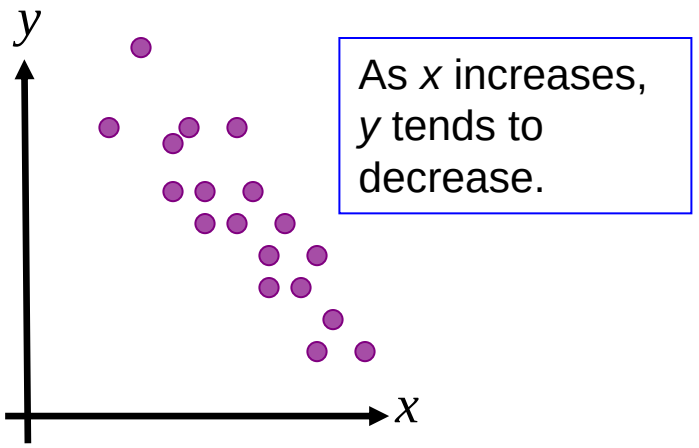
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

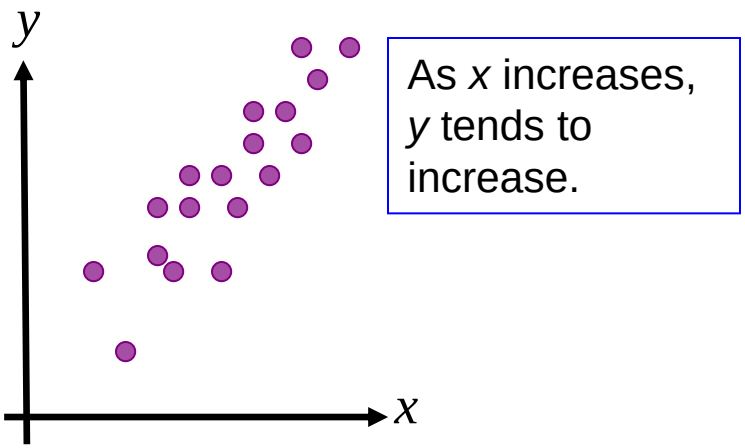
x	1	2	3	4	5
y	-4	-2	-1	0	2



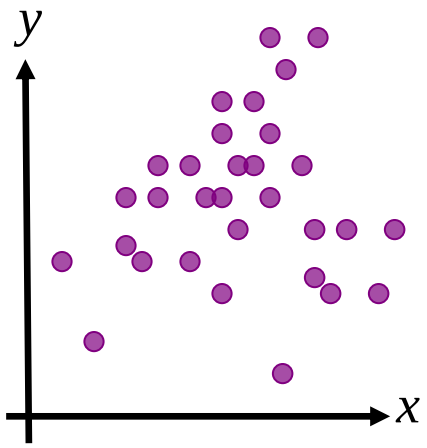
Linear Correlation



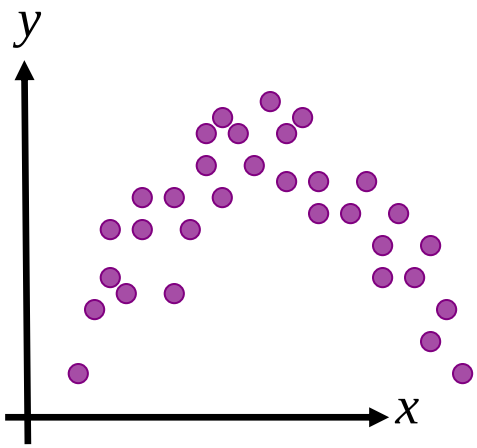
Negative Linear Correlation



Positive Linear Correlation



No Correlation



Nonlinear Correlation

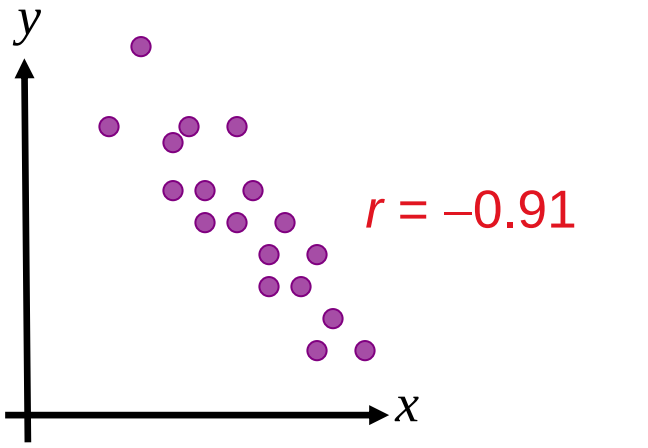
Correlation Coefficient

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

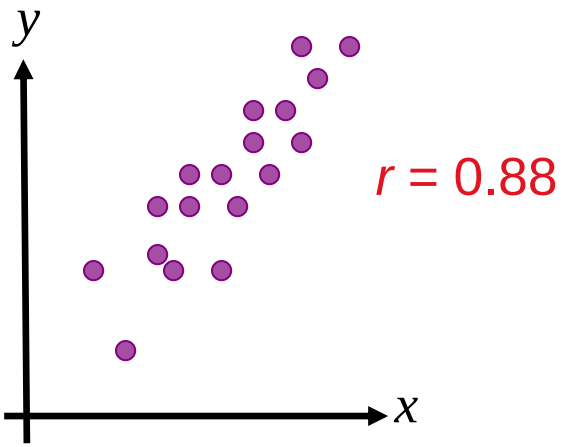
$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}.$$

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.

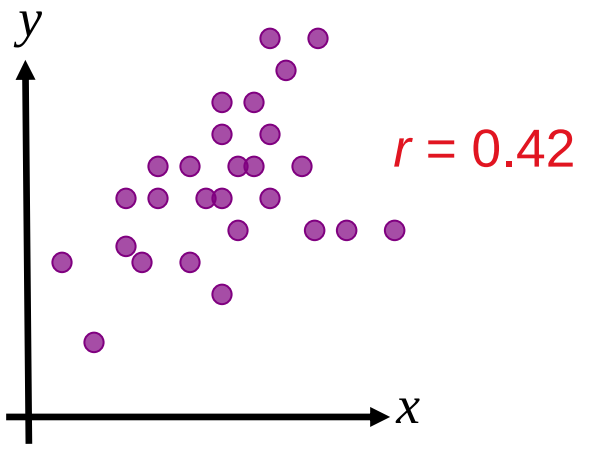
Linear Correlation



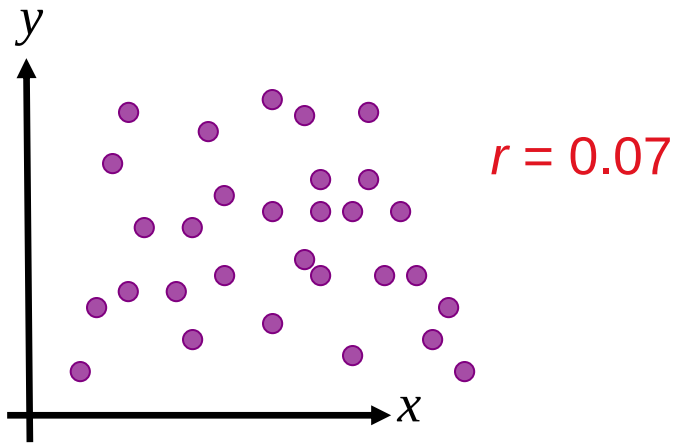
Strong negative correlation



Strong positive correlation



Weak positive correlation



Nonlinear Correlation

Correlation

The ***correlation coefficient*** is a quantitative measure of the strength of the linear relationship between two variables. The correlation ranges from + 1.0 to - 1.0. A correlation of $\pm$ 1.0 indicates a perfect linear relationship, whereas a correlation of 0 indicates no linear relationship.

Correlation

SAMPLE CORRELATION COEFFICIENT

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

where:

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

Correlation

SAMPLE CORRELATION COEFFICIENT

or the algebraic equivalent:

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

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

Sales	Years			
y	x	yx	y^2	x^2
487	3	1,461	237,169	9
445	5	2,225	198,025	25
272	2	544	73,984	4
641	8	5,128	410,881	64
187	2	374	34,969	4
440	6	2,640	193,600	36
346	7	2,422	119,716	49
238	1	238	56,644	1
312	4	1,248	97,344	16
269	2	538	72,361	4
655	9	5,895	429,025	81
563	6	3,378	316,969	36

$\sum y = 4{,}855$ $\sum x = 55$ $\sum xy = 26{,}091$ $\sum y^2 = 2{,}240{,}687$ $\sum x^2 = 4{,}855$

Correlation

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

$$r = \frac{12(26{,}091) - 55(4{,}855)}{\sqrt{[12(329) - (55)^2][12(2{,}240{,}687) - (4{,}855)^2]}}$$

$$= 0.8325$$

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.

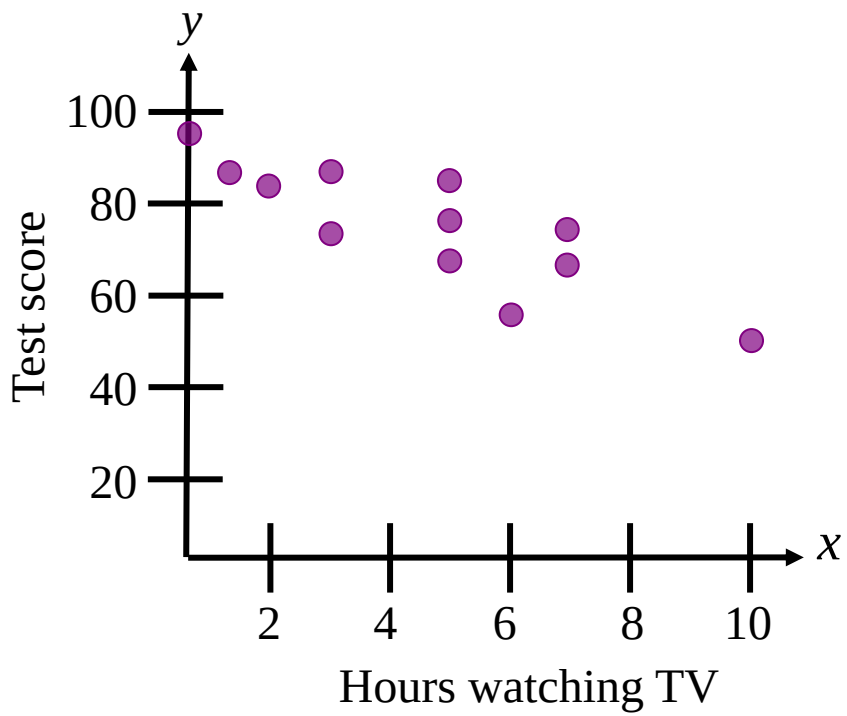
b.) Calculate the correlation coefficient *r*.

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

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

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

There is a strong negative linear correlation.

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

CURVE FITTING

Describes techniques to fit curves (*curve fitting*) to discrete data to obtain intermediate estimates.

There are two general approaches for curve fitting:

- **Least Squares regression:**

  Data exhibit a significant degree of scatter. The strategy is to derive a single curve that represents the general trend of the data.

- **Interpolation:**

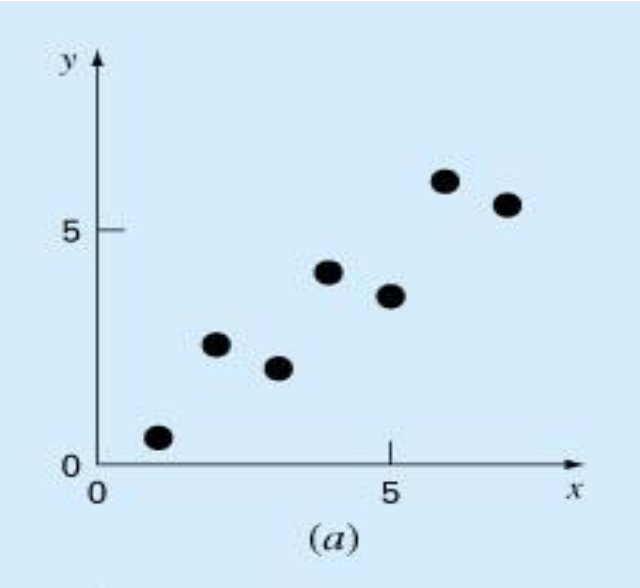
  Data is very precise. The strategy is to pass a curve or a series of curves through each of the points.

Introduction

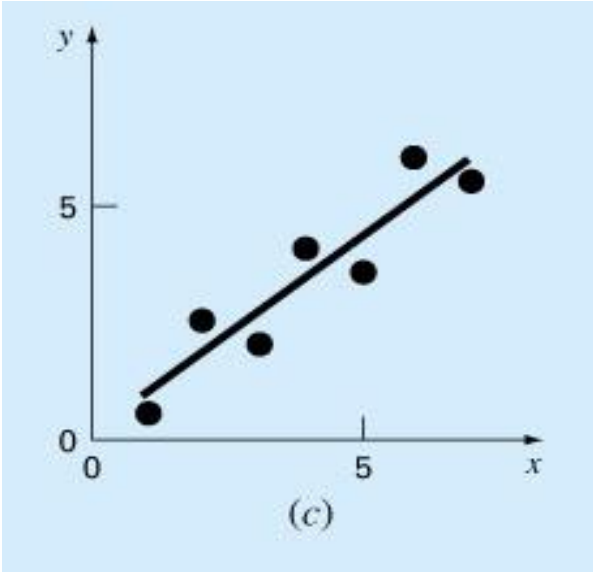
In engineering, two types of applications are encountered:

- **Trend analysis**. Predicting values of dependent variable, may include extrapolation beyond data points or interpolation between data points.

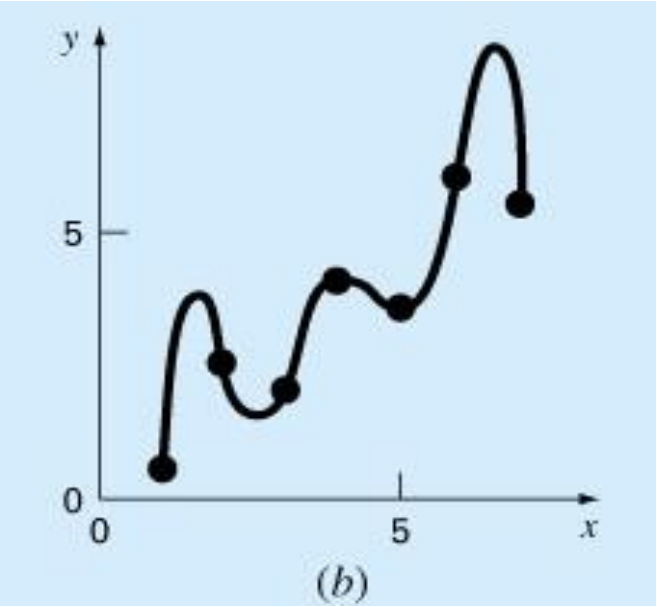
- **Hypothesis testing**. Comparing existing mathematical model with measured data.



(a)



(c)



(b)

Mathematical Background

- ***Arithmetic mean***. The sum of the individual data points (yi) divided by the number of points (n).

$$\bar{y} = \frac{\sum y_i}{n}, i = 1, \ldots, n$$

- ***Standard deviation***. The most common measure of a spread for a sample.

$$S_y = \sqrt{\frac{S_t}{n-1}}, \quad S_t = \sum (y_i - \bar{y})^2$$

Mathematical Background (cont'd)

- ***Variance***. Representation of spread by the square of the standard deviation.

$$S_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

$$S_y^2 = \frac{\sum y_i^2 - \left(\sum y_i\right)^2 / n}{n-1}$$

- ***Coefficient of variation***. Has the utility to quantify the spread of data.

$$c.v. = \frac{S_y}{\bar{y}} 100\%$$

Least Squares Regression

Linear Regression

Fitting a straight line to a set of paired observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

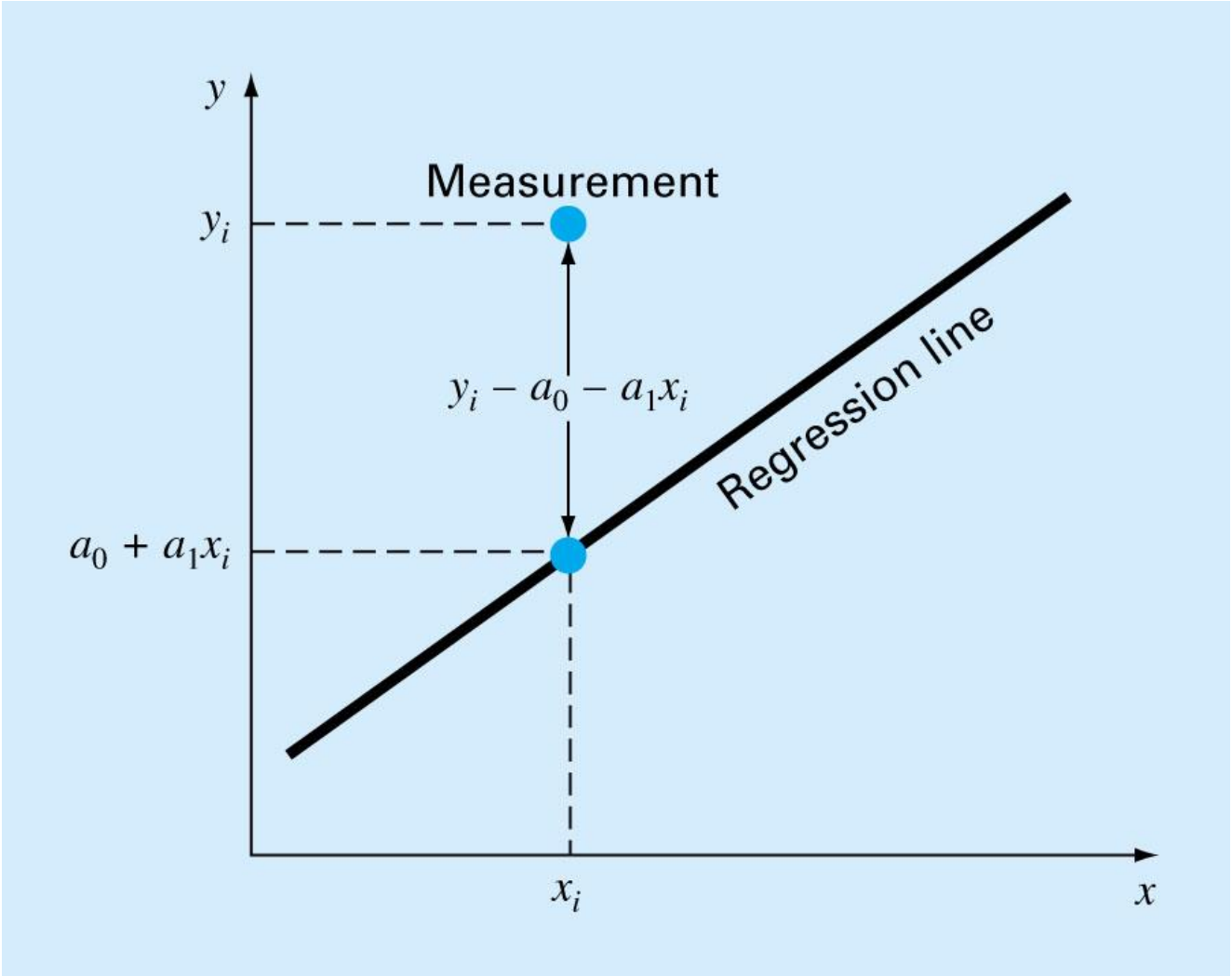
$$y = a_0 + a_1 x + e$$

a_1 - slope

a_0 - intercept

e - error, or residual, between the model and the observations

Linear Regression: Residual

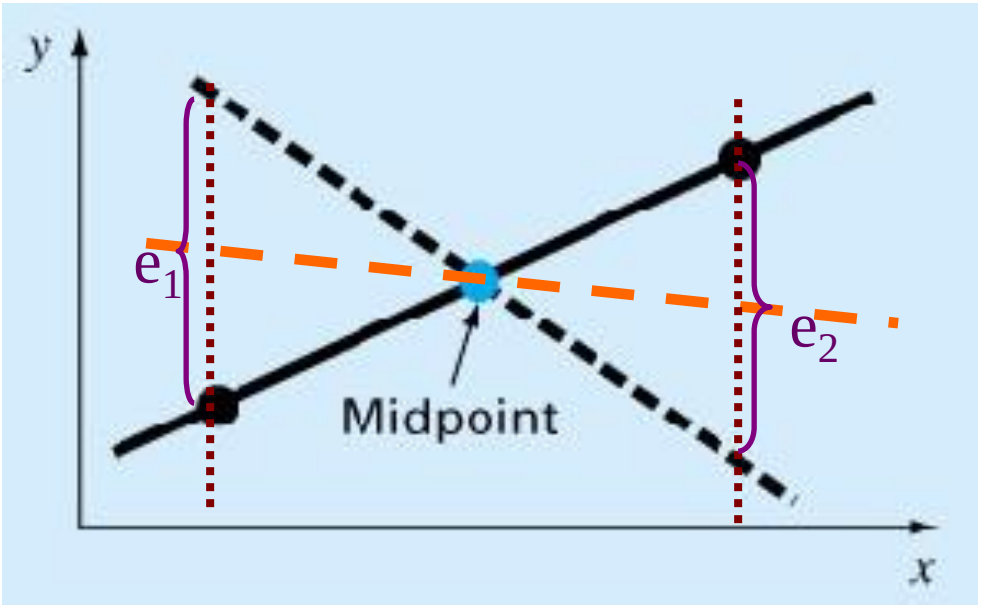


Linear Regression: Question

How to find a_0 and a_1 so that the error would be minimum?

Linear Regression: Criteria for a "Best" Fit

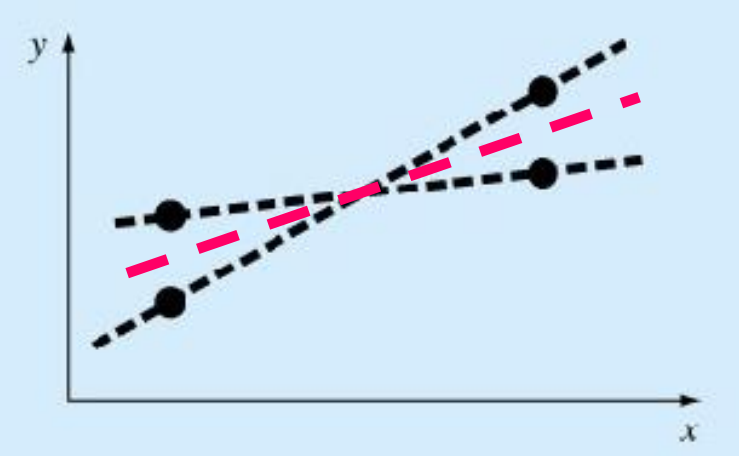
$$\min \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)$$



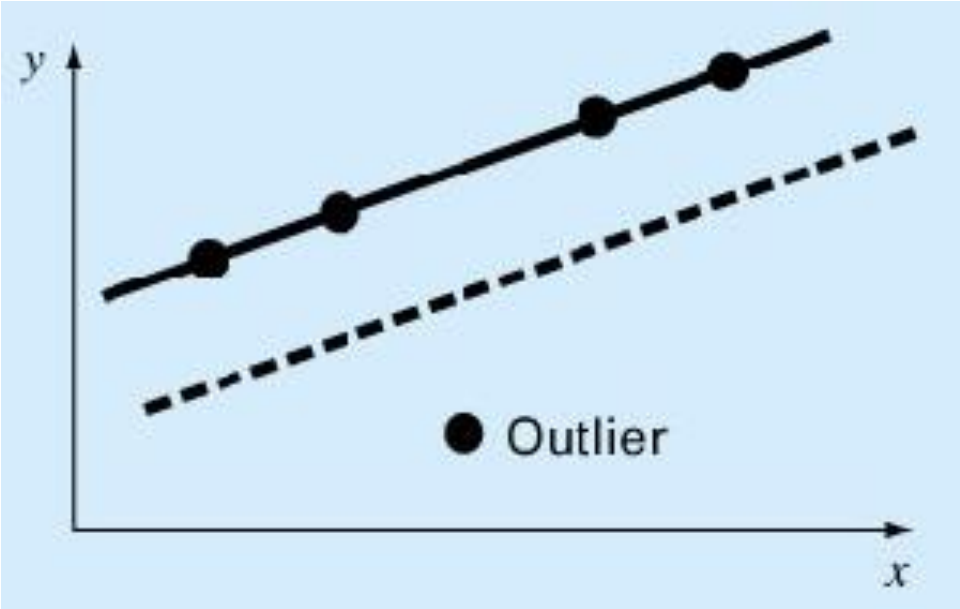
$e_1 = -e_2$

Linear Regression: Criteria for a “Best” Fit

$$\min \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - a_0 - a_1 x_i|$$



Linear Regression: Criteria for a “Best” Fit



Linear Regression: Least Squares Fit

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i, \text{measured} - y_i, \text{model})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

$$\min S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

Yields a unique line for a given set of data.

Linear Regression: Least Squares Fit

$$\min\ S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

The coefficients a_0 and a_1 that minimize S_r must satisfy the following conditions:

$$\begin{cases} \dfrac{\partial S_r}{\partial a_0} = 0 \\ \dfrac{\partial S_r}{\partial a_1} = 0 \end{cases}$$

Linear Regression: Determination of a_o and a_1

$$\frac{\partial S_r}{\partial a_o} = -2\sum (y_i - a_o - a_1 x_i) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum \left[(y_i - a_o - a_1 x_i) x_i \right] = 0$$

$$0 = \sum y_i - \sum a_0 - \sum a_1 x_i$$

$$0 = \sum y_i x_i - \sum a_0 x_i - \sum a_1 x_i^2$$

$$\sum a_0 = n a_0$$

$$n a_0 + \left(\sum x_i \right) a_1 = \sum y_i$$

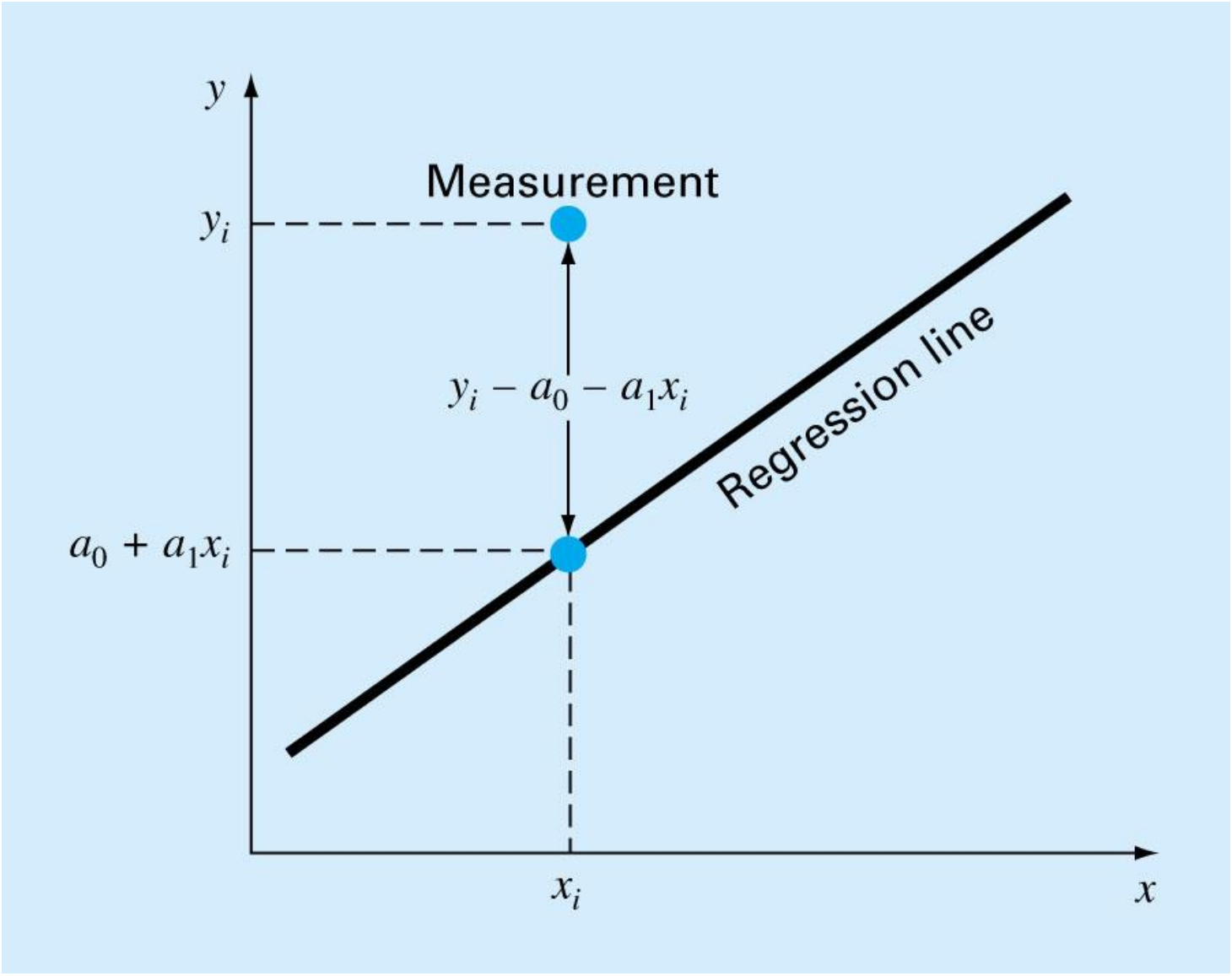
$$\sum y_i x_i = \sum a_0 x_i + \sum a_1 x_i^2$$

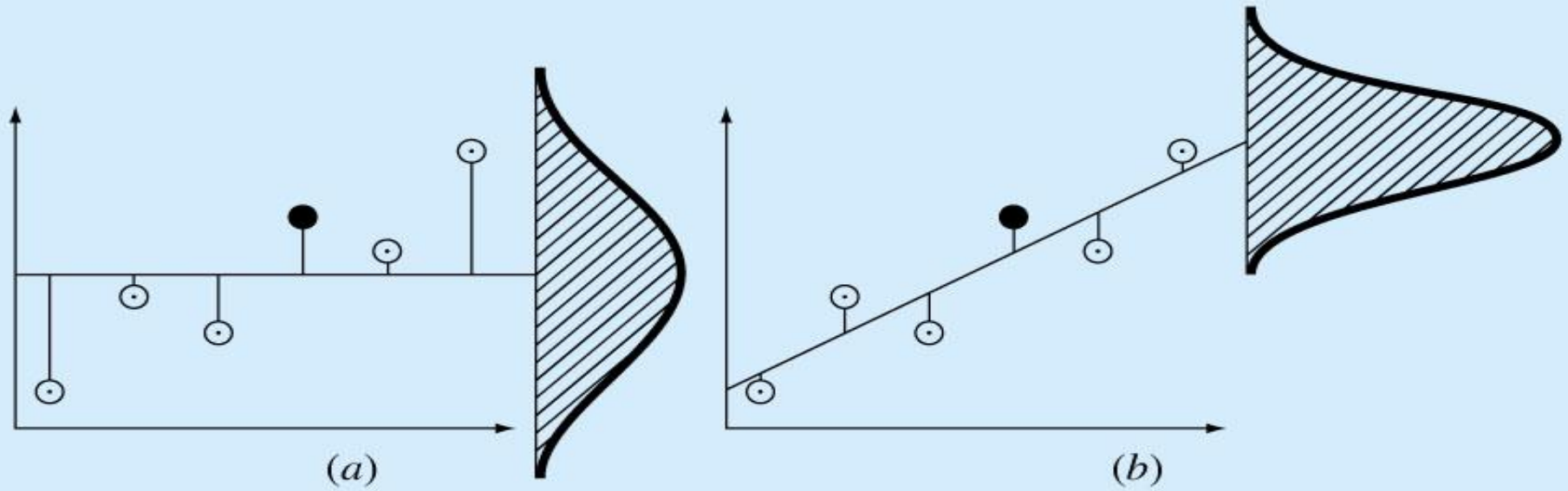
2 equations with 2 unknowns, can be solved simultaneously

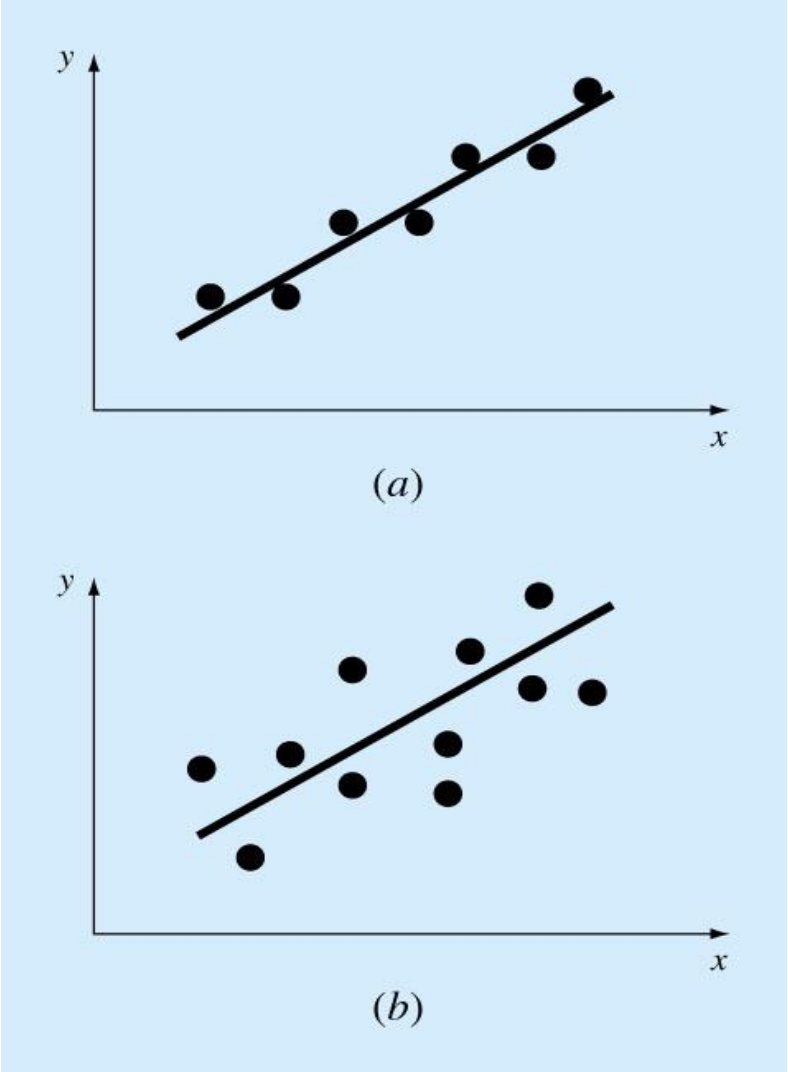
Linear Regression: Determination of ao and a1

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$


y
Measurement
y_i
$y_i - a_0 - a_1x_i$
Regression line
$a_0 + a_1x_i$
x_i
x


(a)
(b)


y
x
(a)
y
x
(b)

Error Quantification of Linear Regression

- Total sum of the squares around the mean for the dependent variable, y, is S_t

$$S_t = \sum (y_i - \bar{y})^2$$

- Sum of the squares of residuals around the regression line is S_r

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_o - a_1 x_i)^2$$

Error Quantification of Linear Regression

- S_t-S_r quantifies the improvement or error reduction due to describing data in terms of a straight line rather than as an average value.

$$r^2 = \frac{S_t - S_r}{S_t}$$

r^2: coefficient of determination

r : correlation coefficient

Error Quantification of Linear Regression

For a perfect fit:

- $S_r = 0$ and $r = r^2 = 1$, signifying that the line explains 100 percent of the variability of the data.
- For $r = r^2 = 0$, $S_r = S_t$, the fit represents no improvement.

Least Squares Fit of a Straight Line: Example

Fit a straight line to the x and y values in the following Table:

x_i	y_i	$x_i y_i$	x_i^2
1	0.5	0.5	1
2	2.5	5	4
3	2	6	9
4	4	16	16
5	3.5	17.5	25
6	6	36	36
7	5.5	38.5	49
28	**24**	**119.5**	**140**

$$\sum x_i = 28 \quad \sum y_i = 24.0$$

$$\sum x_i^2 = 140 \quad \sum x_i y_i = 119.5$$

$$\bar{x} = \frac{28}{7} = 4$$

$$\bar{y} = \frac{24}{7} = 3.428571$$

Least Squares Fit of a Straight Line: Example (cont'd)

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= \frac{7 \times 119.5 - 28 \times 24}{7 \times 140 - 28^2} = 0.8392857$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$= 3.428571 - 0.8392857 \times 4 = 0.07142857$$

Y = 0.07142857 + 0.8392857 x

Least Squares Fit of a Straight Line: Example (Error Analysis)

x_i	y_i	$(y_i - \bar{y})^2$	e_i^2
1	0.5	8.5765	0.1687
2	2.5	0.8622	0.5625
3	2.0	2.0408	0.3473
4	4.0	0.3265	0.3265
5	3.5	0.0051	0.5896
6	6.0	6.6122	0.7972
7	5.5	4.2908	0.1993
28	**24.0**	**22.7143**	**2.9911**

$$S_t = \sum (y_i - \bar{y})^2 = 22.7143$$

$$S_r = \sum e_i^{\,2} = 2.9911$$

$$r^2 = \frac{S_t - S_r}{S_t} = 0.868$$

$$r = \sqrt{r^2} = \sqrt{0.868} = 0.932$$

Least Squares Fit of a Straight Line: Example (Error Analysis)

•The standard deviation (quantifies the spread around the mean):

$$s_y = \sqrt{\frac{S_t}{n-1}} = \sqrt{\frac{22.7143}{7-1}} = 1.9457$$

•The standard error of estimate (quantifies the spread around the regression line)

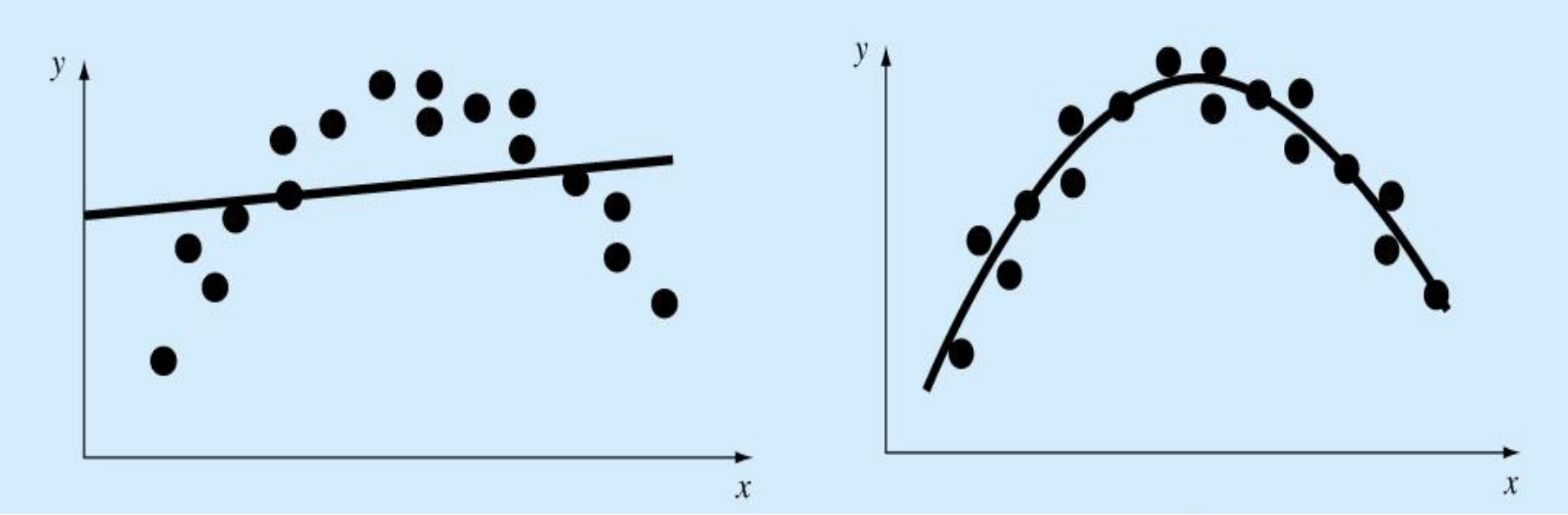
$$s_{y/x} = \sqrt{\frac{S_r}{n-2}} = \sqrt{\frac{2.9911}{7-2}} = 0.7735$$

Because $S_{y/x} < S_y$, the linear regression model has good fitness

Polynomial Regression

- Some engineering data is poorly represented by a straight line.
- For these cases a curve is better suited to fit the data.
- The least squares method can readily be extended to fit the data to higher order polynomials.

Polynomial Regression (cont'd)



A parabola is preferable

Polynomial Regression (cont'd)

- **A 2nd order polynomial (quadratic)** is defined by:

$$y = a_o + a_1 x + a_2 x^2 + e$$

- The residuals between the model and the data:

$$e_i = y_i - a_o - a_1 x_i - a_2 x_i^2$$

- The sum of squares of the residual:

$$S_r = \sum e_i^2 = \sum \left(y_i - a_o - a_1 x_i - a_2 x_i^2 \right)^2$$

Polynomial Regression (cont'd)

$$\frac{\partial S_r}{\partial a_o} = -2\sum (y_i - a_o - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum (y_i - a_o - a_1 x_i - a_2 x_i^2) x_i = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum (y_i - a_o - a_1 x_i - a_2 x_i^2) x_i^2 = 0$$

$$\sum y_i = n \cdot a_o + a_1 \sum x_i + a_2 \sum x_i^2$$

$$\sum x_i y_i = a_o \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3$$

$$\sum x_i^2 y_i = a_o \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4$$

3 linear equations with 3 unknowns (a_o,a_1,a_2), can be solved

Polynomial Regression (cont'd)

- A system of 3x3 equations needs to be solved to determine the coefficients of the polynomial.

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{Bmatrix}$$

- The standard error & the coefficient of determination

$$s_{y/x} = \sqrt{\frac{S_r}{n-3}} \qquad r^2 = \frac{S_t - S_r}{S_t}$$

Polynomial Regression (cont'd)

General:

The mth-order polynomial:

$$y = a_o + a_1 x + a_2 x^2 + \ldots\ldots + a_m x^m + e$$

- A system of (m+1)x(m+1) linear equations must be solved for determining the coefficients of the mth-order polynomial.
- The standard error:

$$s_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}}$$

- The coefficient of determination:

$$r^2 = \frac{S_t - S_r}{S_t}$$

Polynomial Regression- Example

Fit a second order polynomial to data:

x_i	y_i	x_i^2	x_i^3	x_i^4	x_iy_i	$x_i^2y_i$
0	**2.1**	0	0	0	0	0
1	**7.7**	1	1	1	7.7	7.7
2	**13.6**	4	8	16	27.2	54.4
3	**27.2**	9	27	81	81.6	244.8
4	**40.9**	16	64	256	163.6	654.4
5	**61.1**	25	125	625	305.5	1527.5
15	**152.6**	**55**	**225**	**979**	**585.6**	**2489**

$$\sum x_i = 15$$

$$\sum y_i = 152.6$$

$$\sum x_i^2 = 55$$

$$\sum x_i^3 = 225$$

$$\sum x_i^4 = 979$$

$$\sum x_i y_i = 585.6$$

$$\sum x_i^2 y_i = 2488.8$$

$$\bar{x} = \frac{15}{6} = 2.5, \quad \bar{y} = \frac{152.6}{6} = 25.433$$

Polynomial Regression- Example (cont'd)

- The system of simultaneous linear equations:

$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{Bmatrix}$$

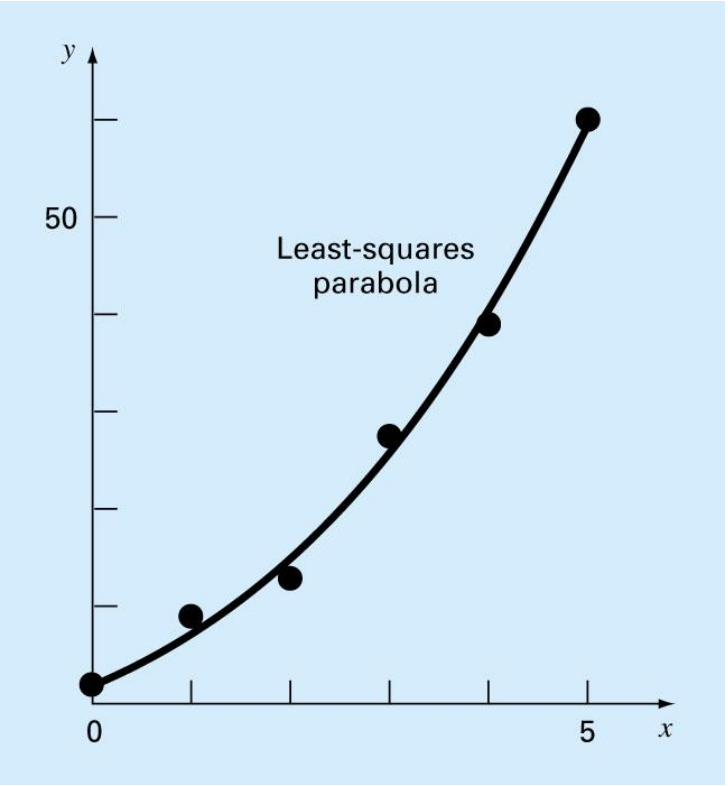
$$a_0 = 2.47857,\ a_1 = 2.35929,\ a_2 = 1.86071$$

$$y = 2.47857 + 2.35929\,x + 1.86071\,x^2$$

$$S_t = \sum (y_i - \bar{y})^2 = 2513.39 \qquad S_r = \sum e_i^{\,2} = 3.74657$$

Polynomial Regression- Example (cont'd)

x_i	y_i	y_{model}	e_i^2	$(y_i-y`)^2$
0	2.1	2.4786	0.14332	544.42889
1	7.7	6.6986	1.00286	314.45929
2	13.6	14.64	1.08158	140.01989
3	27.2	26.303	0.80491	3.12229
4	40.9	41.687	0.61951	239.22809
5	61.1	60.793	0.09439	1272.13489
15	**152.6**		**3.74657**	**2513.39333**



- The standard error of estimate:

$$s_{y/x} = \sqrt{\frac{3.74657}{6-3}} = 1.12$$

- The coefficient of determination:

$$r^2 = \frac{2513.39 - 3.74657}{2513.39} = 0.99851, \quad r = \sqrt{r^2} = 0.99925$$