The Islamic University of Gaza
Faculty of Engineering
Civil Engineering Department

Numerical Analysis
ECIV 3306

Chapter 17

Least Square Regression

Part 5 - 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:**

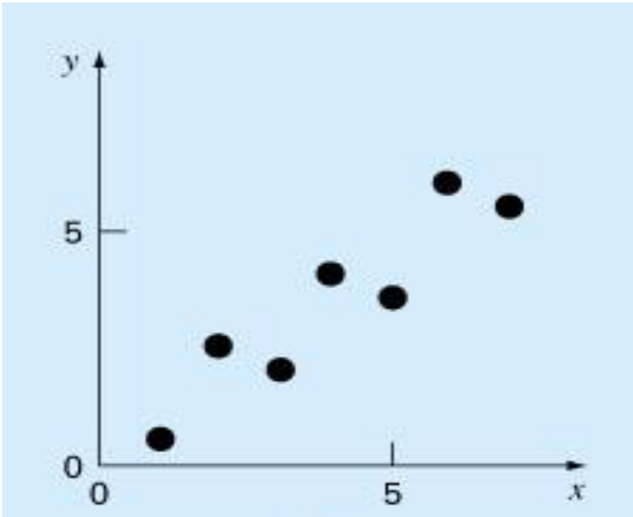
  Data is very precise. The strategy is to pass a curve or a series of curves through each of the points.

Introduction

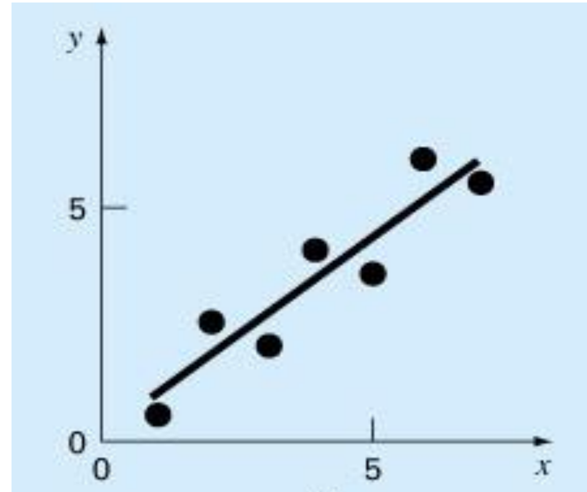
In engineering, two types of applications are encountered:

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

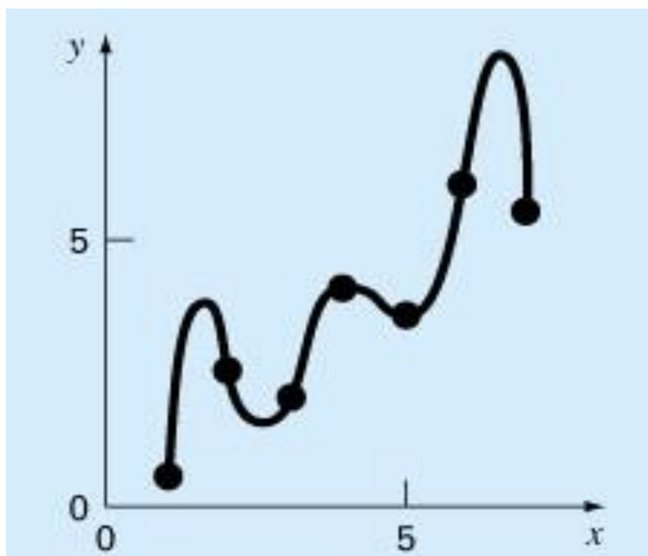
- **Hypothesis testing**. Comparing existing mathematical model with measured data.



(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.

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

Chapter 17

Least Squares Regression

Linear Regression

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

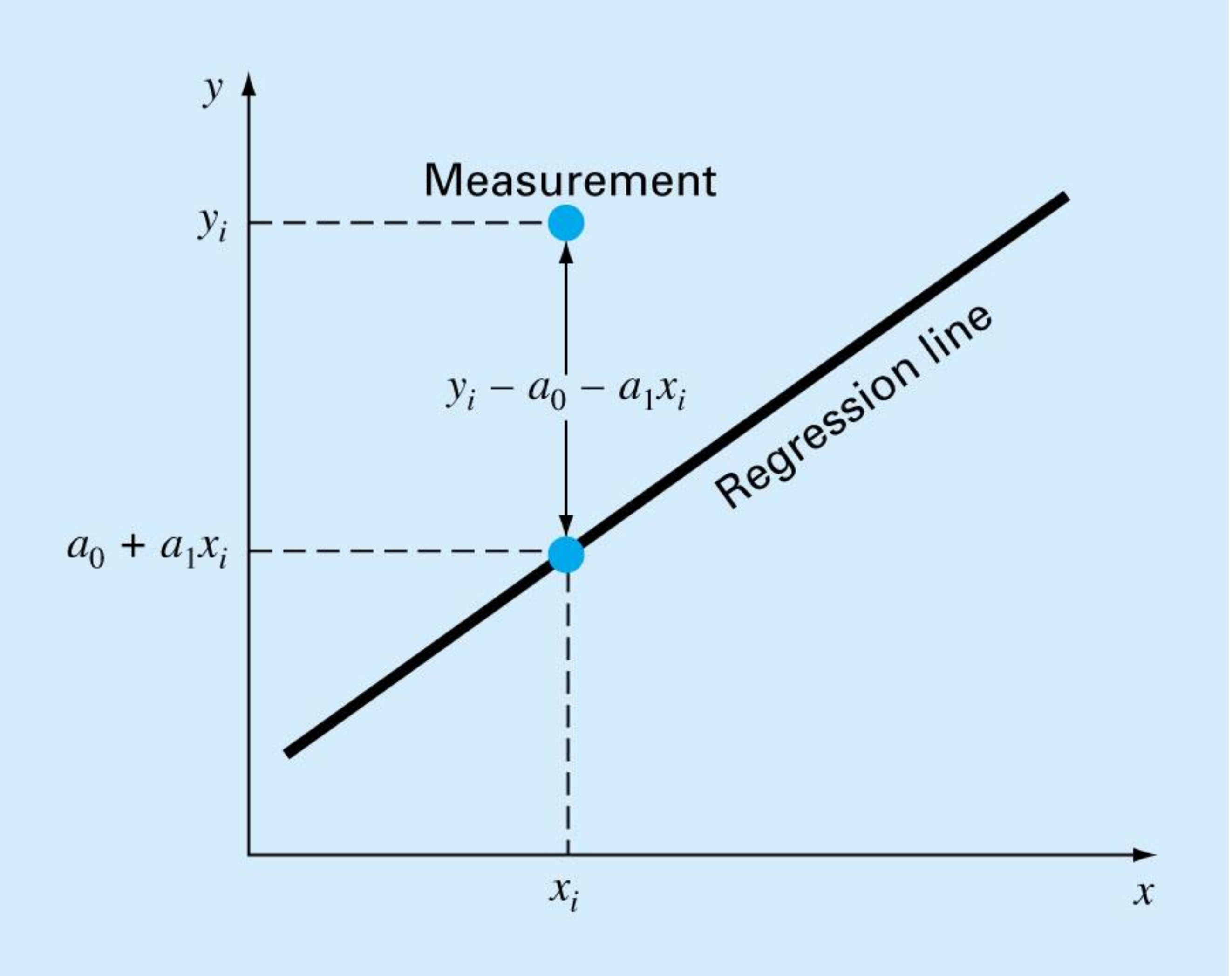
$$y = a_0 + a_1 x + e$$

a_1 - slope

a_0 - intercept

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

Linear Regression: Residual

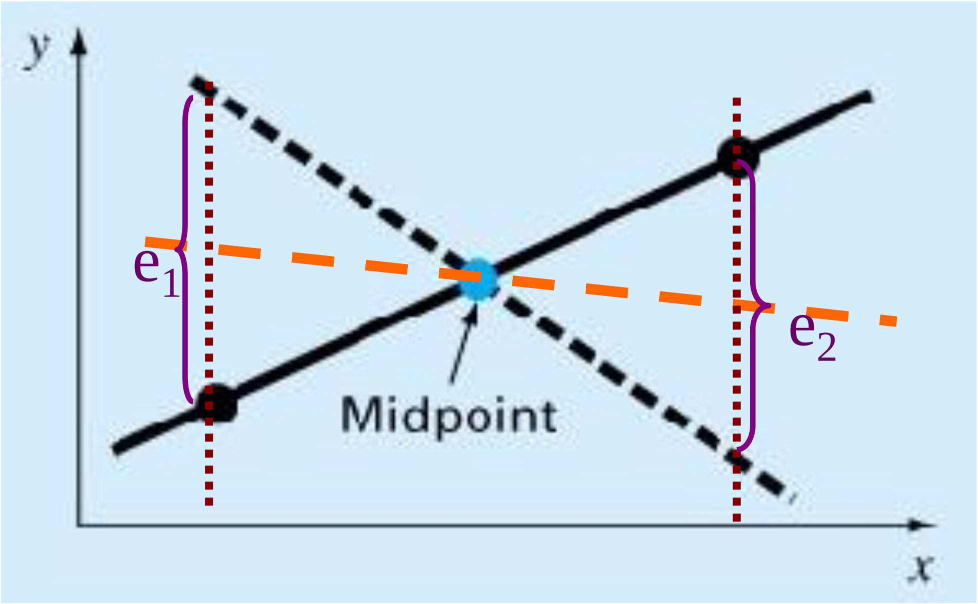


Linear Regression: Question

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

Linear Regression: Criteria for a "Best" Fit

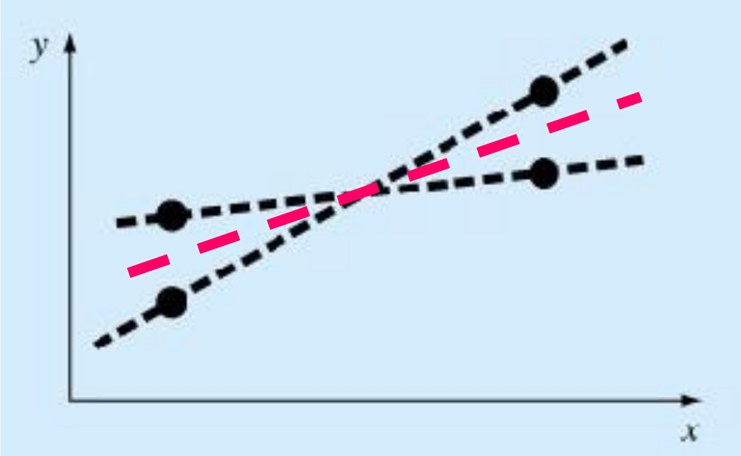
$$\min \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)$$



$e_1 = -e_2$

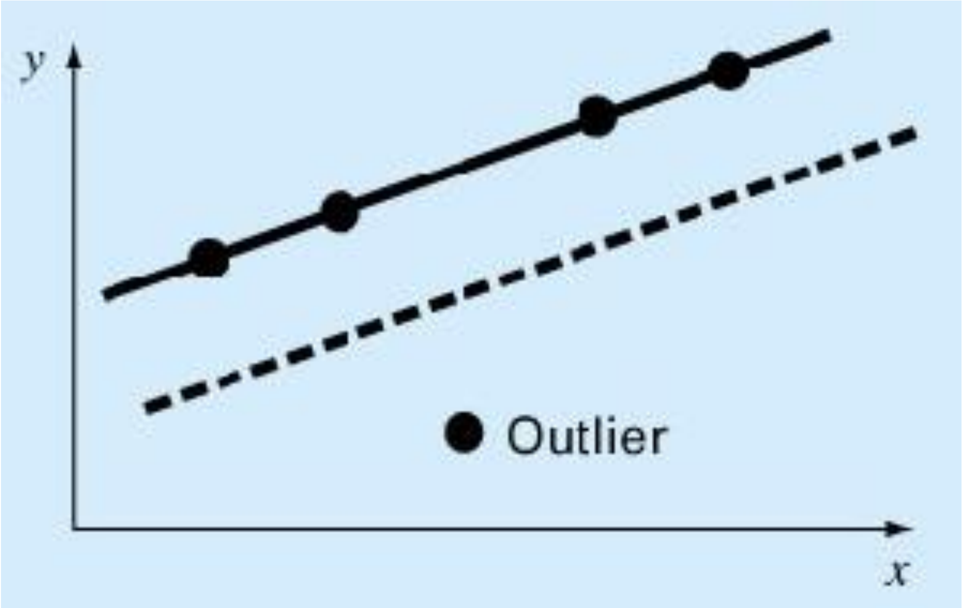
Linear Regression: Criteria for a "Best" Fit

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



Linear Regression: Criteria for a "Best" Fit

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



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

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

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

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

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

$$\sum a_0 = n a_0$$

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

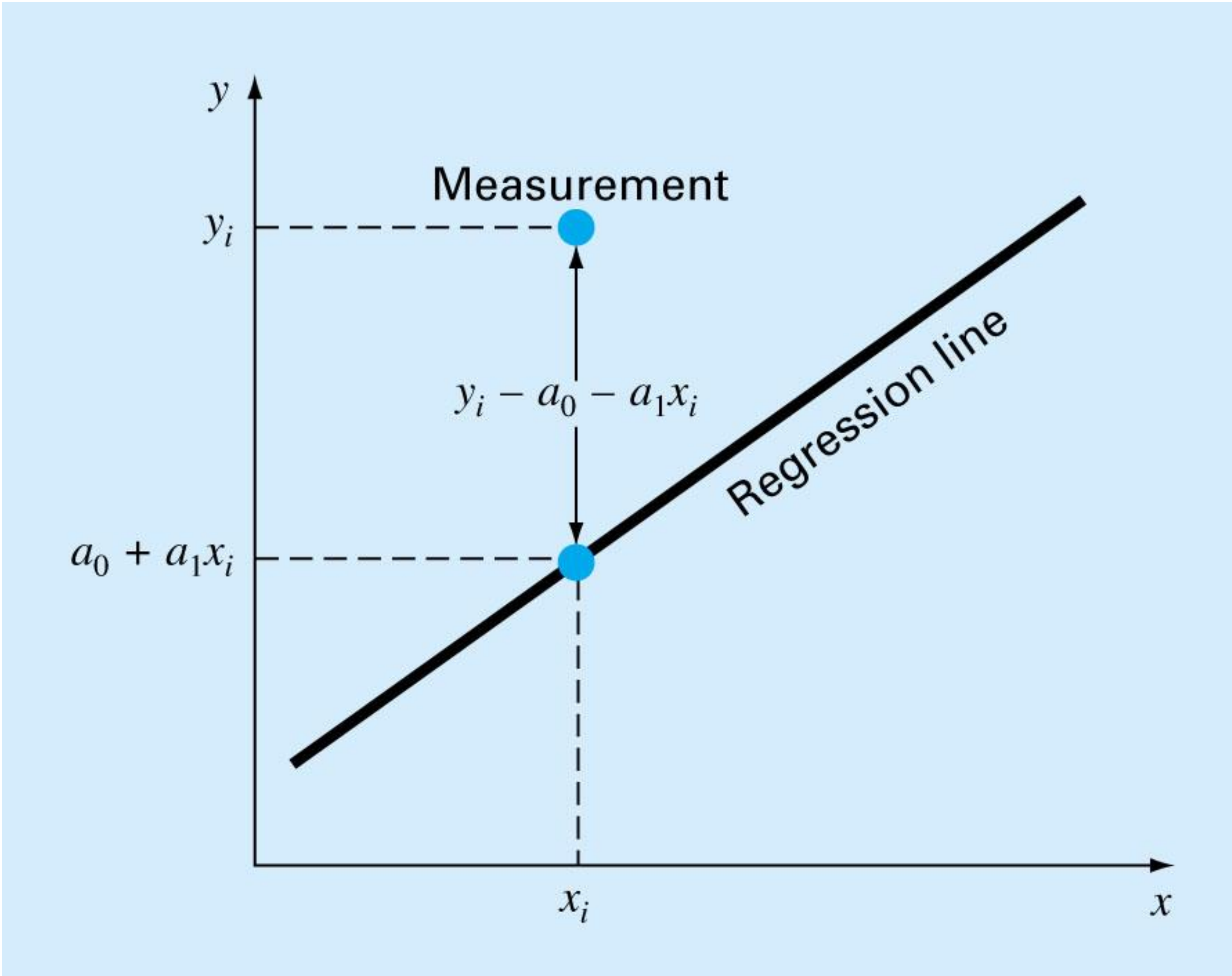
$$\sum y_i x_i = \sum a_0 x_i + \sum a_1 x_i^2$$

} 2 equations with 2 unknowns, can be solved simultaneously

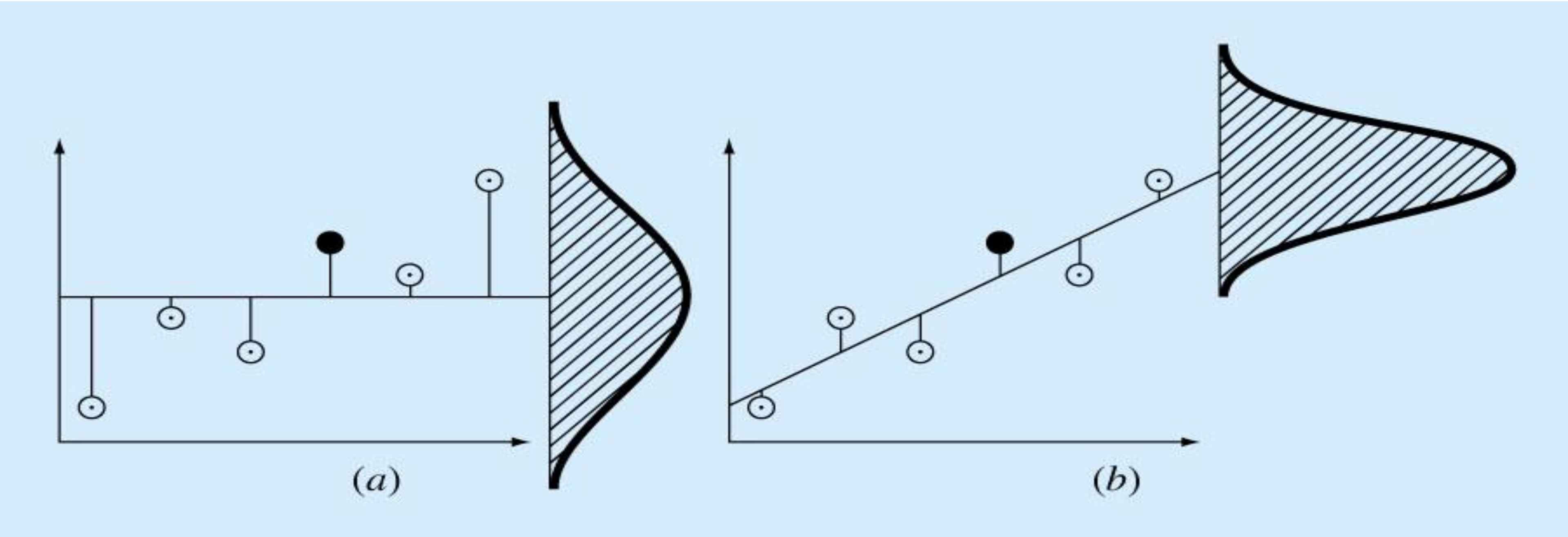
Linear Regression: Determination of ao and a1

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

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



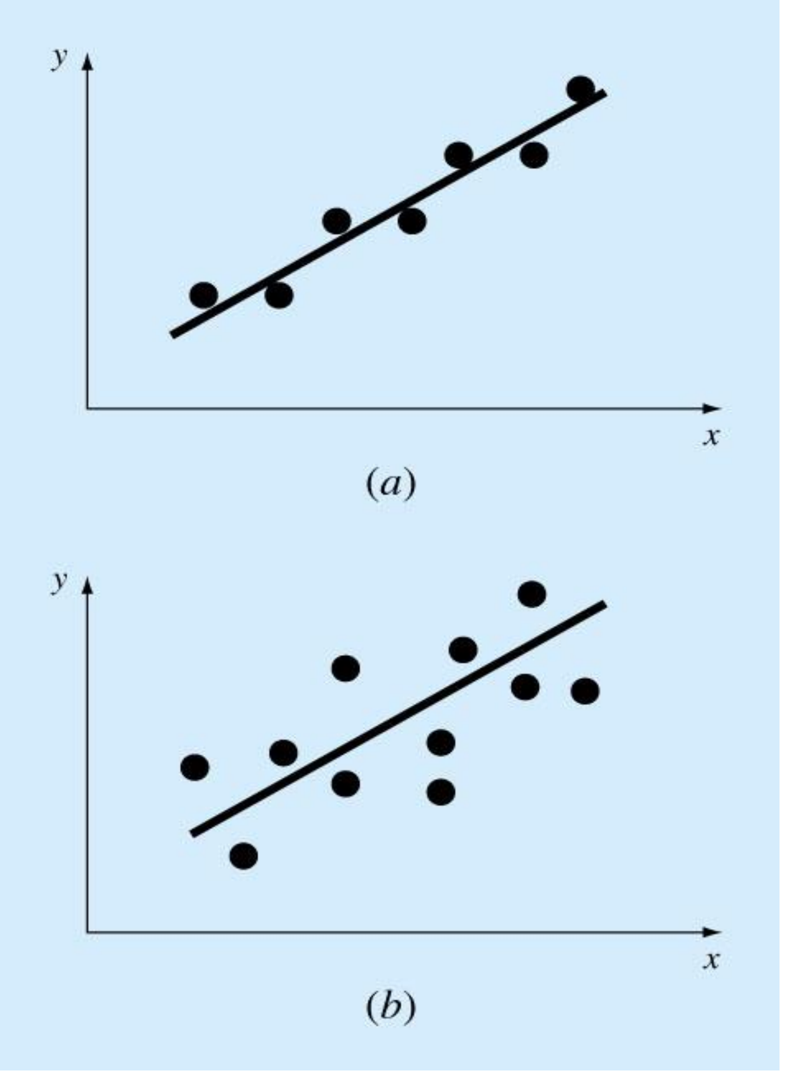
y
Measurement
y_i
$y_i - a_0 - a_1x_i$
Regression line
$a_0 + a_1x_i$
x_i
x



Data spread around Mean

Data spread around best-fit line

Examples of linear regression with (a) small and (b) large residual errors



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)

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

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

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

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

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**

Y = 0.07142857 + 0.8392857 x

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

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

$$S_r = \sum e_i^2 = 2.9911$$

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

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

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

Algorithm for linear regression

```
SUB Regress(x, y, n, a1, a0, syx, r2)

  sumx = 0: sumxy = 0: st = 0
  sumy = 0: sumx2 = 0: sr = 0
  DO i = 1, n
    sumx = sumx + x_i
    sumy = sumy + y_i
    sumxy = sumxy + x_i*y_i
    sumx2 = sumx2 + x_i*x_i
  END DO
  xm = sumx/n
  ym = sumy/n
  a1 = (n*sumxy − sumx*sumy)/(n*sumx2 − sumx*sumx)
  a0 = ym − a1*xm
  DO i = 1, n
    st = st + (y_i − ym)^2
    sr = sr + (y_i − a1*x_i − a0)^2
  END DO
  syx = (sr/(n − 2))^0.5
  r2 = (st − sr)/st

END Regress
```

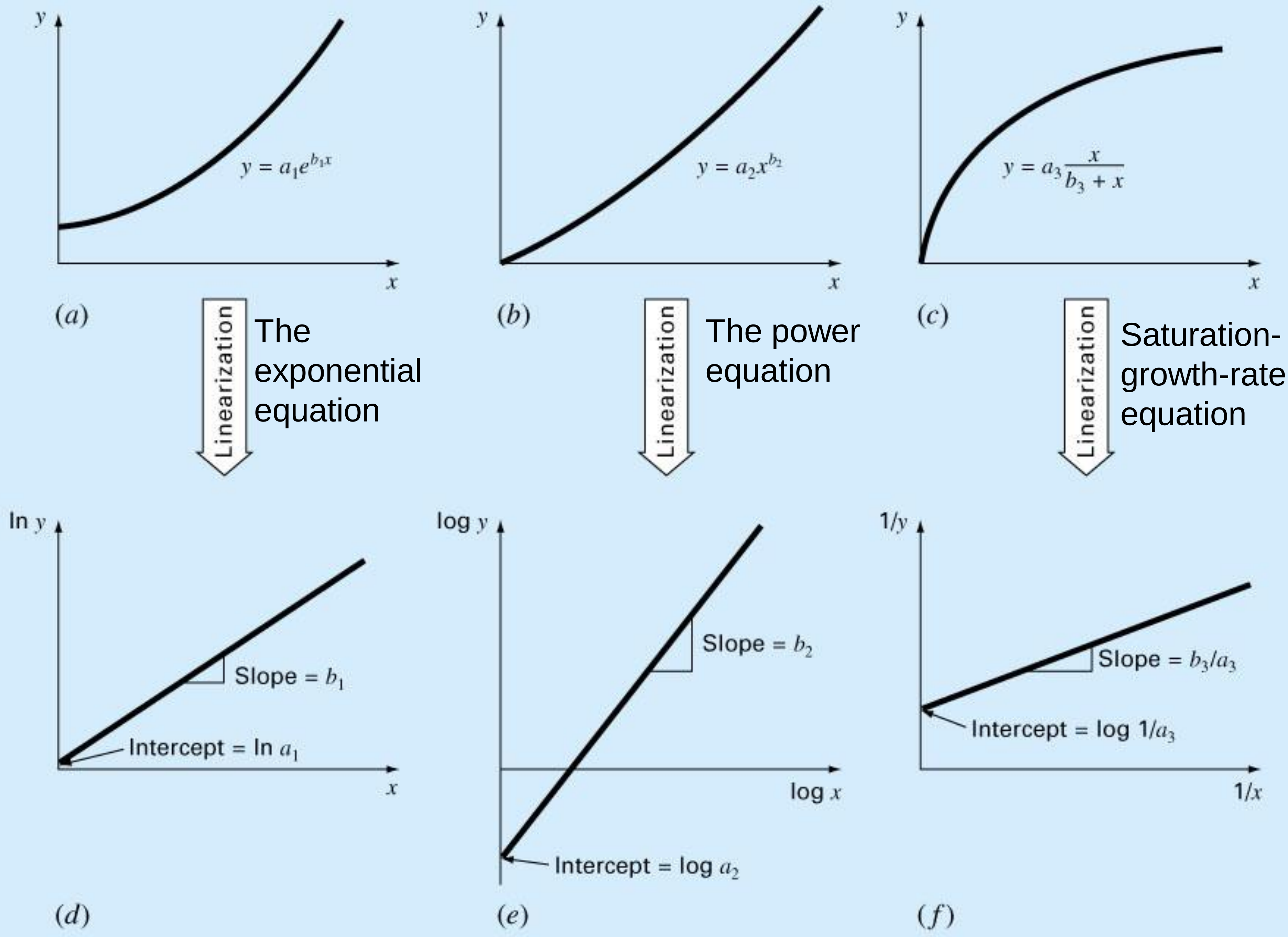
Linearization of Nonlinear Relationships

- The relationship between the dependent and independent variables is linear.
- However, few types of nonlinear functions can be transformed into linear regression problems.

➢ The exponential equation.

➢ The power equation.

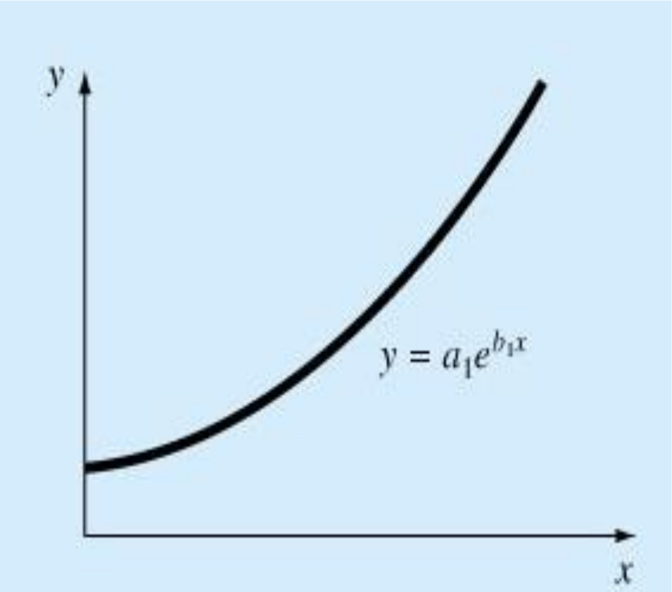
➢ The saturation-growth-rate equation.

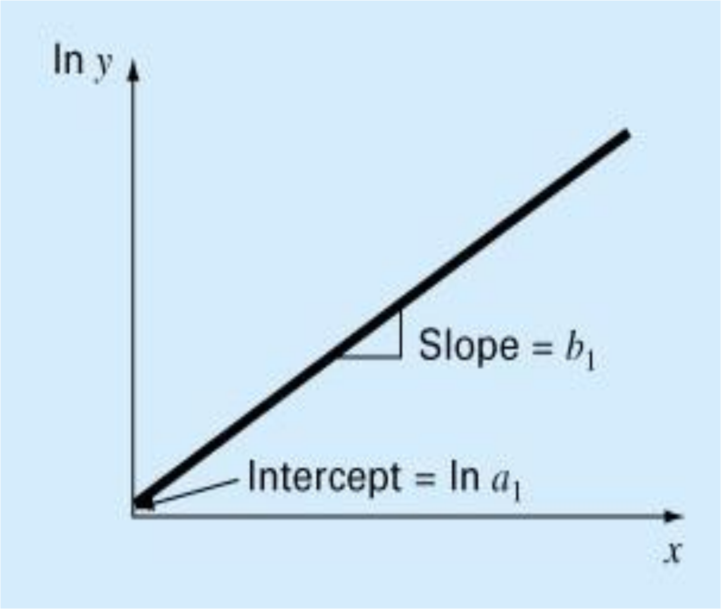


y
$y = a_1 e^{b_1 x}$
x
(a)
Linearization
The exponential equation
y
$y = a_2 x^{b_2}$
x
(b)
Linearization
The power equation
y
$y = a_3 \frac{x}{b_3 + x}$
x
(c)
Linearization
Saturation-growth-rate equation
ln y
Slope = b_1
Intercept = ln a_1
x
(d)
log y
Slope = b_2
log x
Intercept = log a_2
(e)
1/y
Slope = b_3/a_3
Intercept = log $1/a_3$
1/x
(f)

Linearization of Nonlinear Relationships

1. The exponential equation.

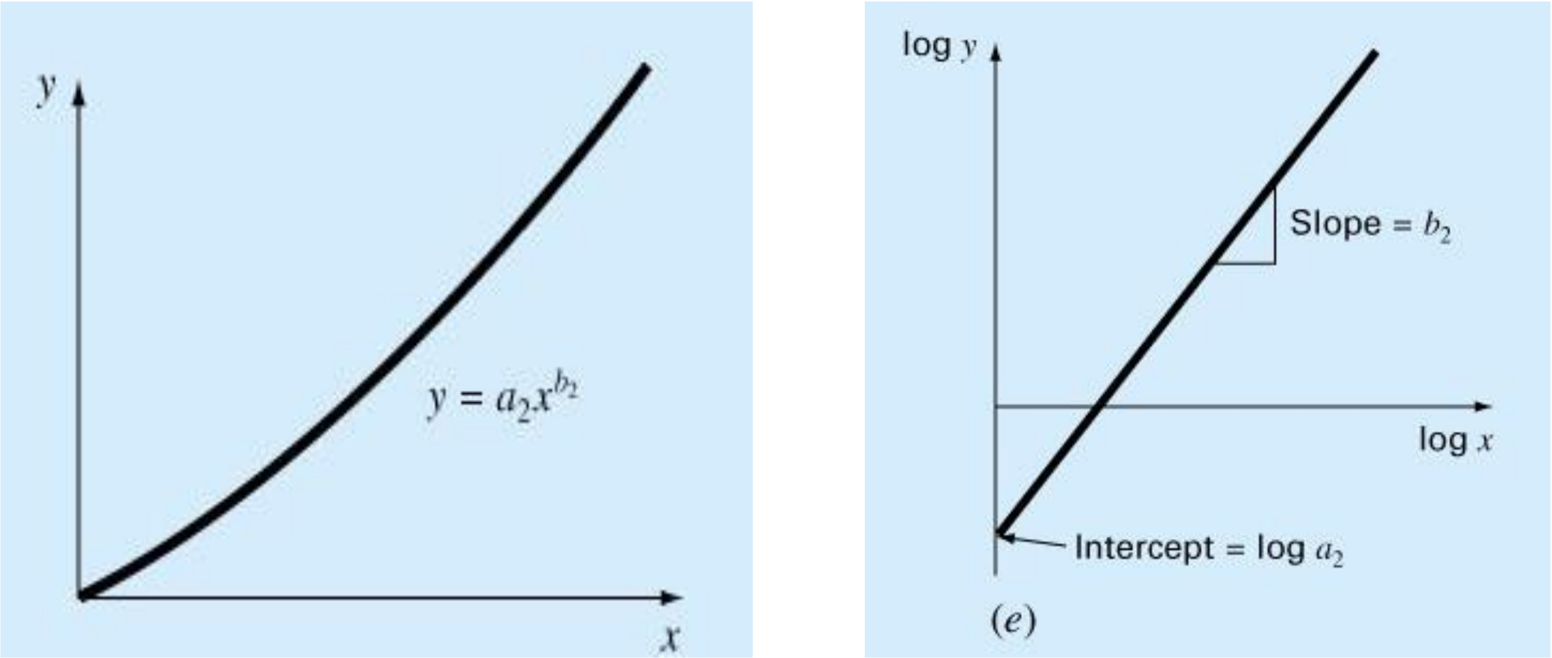




$$\ln y = \ln a_1 + b_1 x$$

Linearization of Nonlinear Relationships

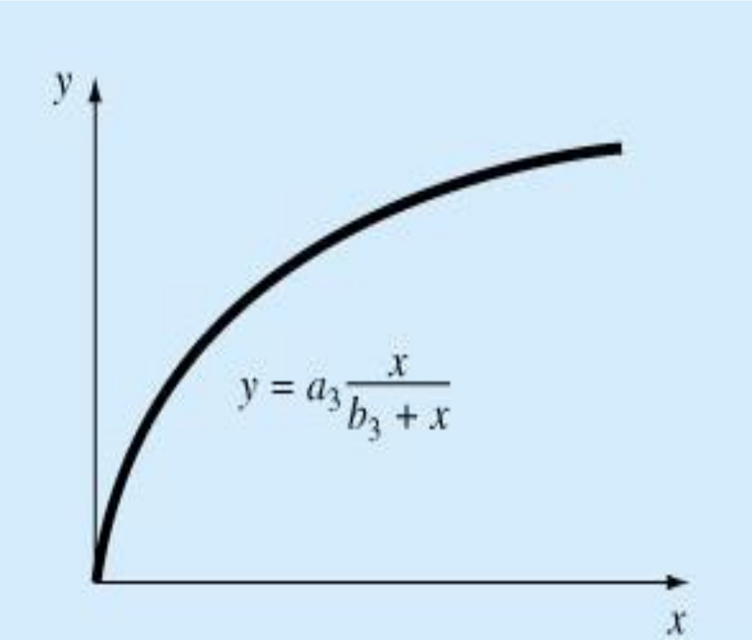
2. The power equation

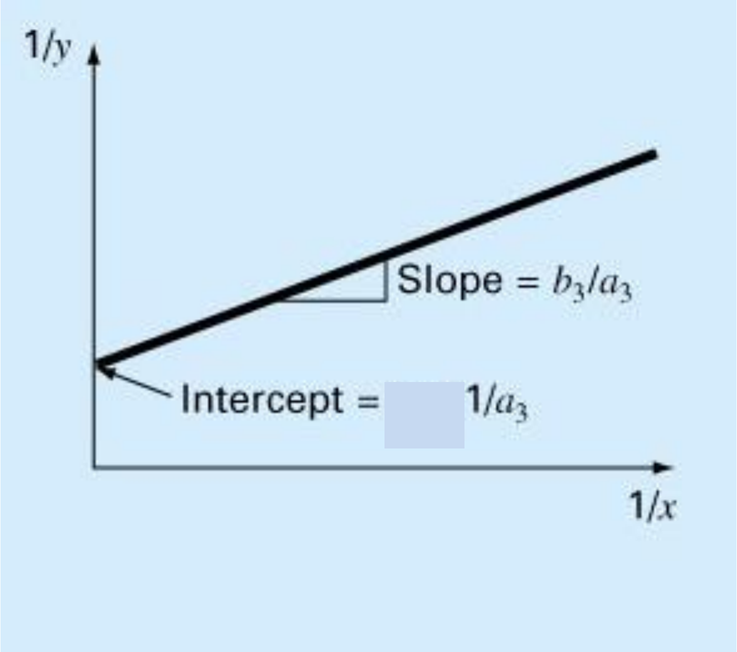


$$\log y = \log a_2 + b_2 \log x$$

Linearization of Nonlinear Relationships

3. The saturation-growth-rate equation





$$\frac{1}{y} = \frac{1}{a_3} + \frac{b_3}{a_3}\left(\frac{1}{x}\right)$$

Example

Fit the following Equation:

$$y = a_2 x^{b_2}$$

to the data in the following table:

x_i	y_i	$X=\log x_i$	$Y=\log y_i$
1	0.5	0	-0.301
2	1.7	0.301	0.226
3	3.4	0.477	0.534
4	5.7	0.602	0.753
5	8.4	0.699	0.922
15	19.7	2.079	2.141

$$\log y = \log(a_2 x^{b_2})$$

$$\log y = \log a_2 + b_2 \log x$$

$$\text{let } Y = \log y,\ X = \log x,$$

$$a_0 = \log a_2,\ a_1 = b_2$$

$$Y = a_0 + a_1 X$$

Example

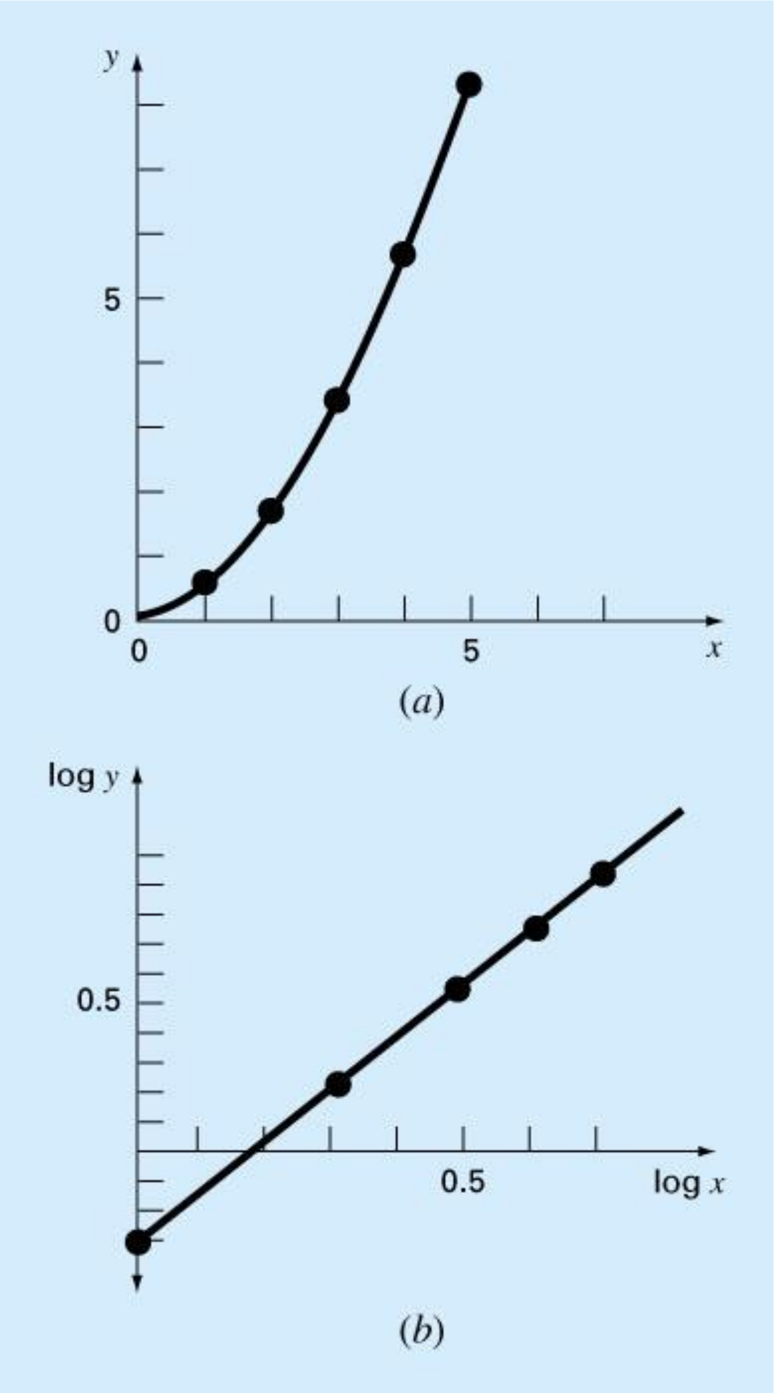
	Xi	Yi	X^*_i=Log(X)	Y^*_i=Log(Y)	X*Y*	X*^2
	1	0.5	0.0000	-0.3010	0.0000	0.0000
	2	1.7	0.3010	0.2304	0.0694	0.0906
	3	3.4	0.4771	0.5315	0.2536	0.2276
	4	5.7	0.6021	0.7559	0.4551	0.3625
	5	8.4	0.6990	0.9243	0.6460	0.4886
Sum	15	19.700	**2.079**	**2.141**	**1.424**	**1.169**

$$\begin{cases} a_1 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \dfrac{5\times 1.424 - 2.079\times 2.141}{5\times 1.169 - 2.079^2} = 1.75 \\ a_0 = \bar{y} - a_1\bar{x} = 0.4282 - 1.75\times 0.41584 = -0.334 \end{cases}$$

Linearization of Nonlinear Functions: Example

$$\log y = -0.334 + 1.75 \log x$$

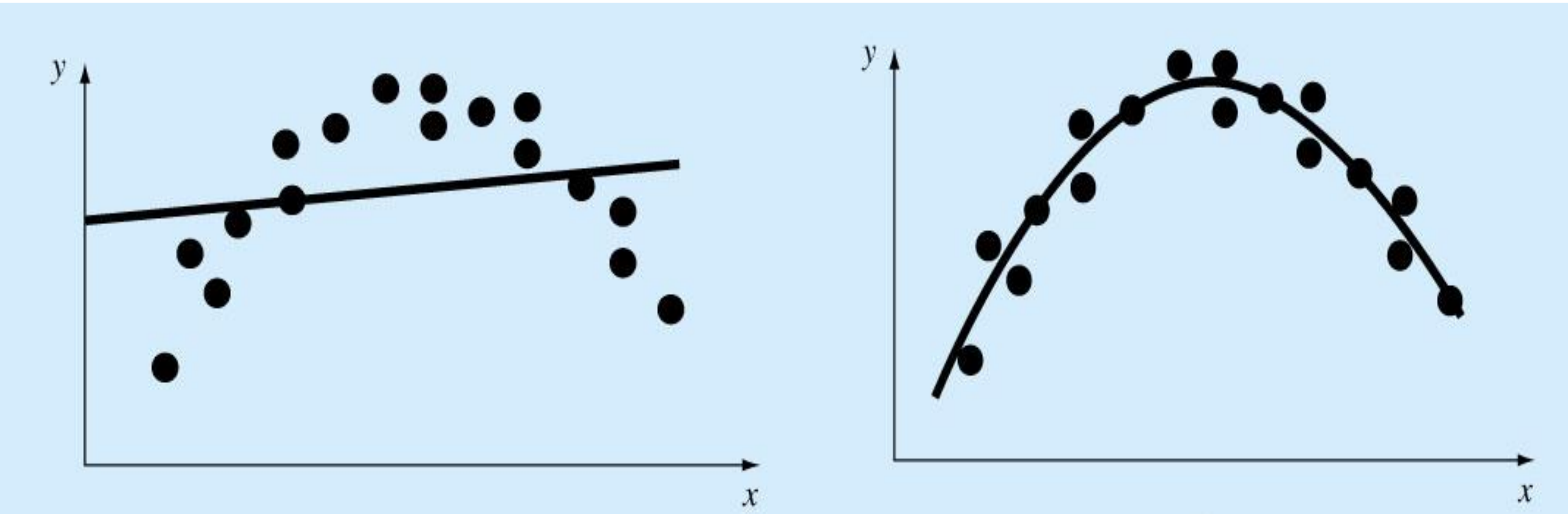
$$y = 0.46x^{1.75}$$



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)

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

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

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

$$\left.\begin{aligned}
\sum y_i &= n \cdot a_o + a_1 \sum x_i + a_2 \sum x_i^2 \\
\sum x_i y_i &= a_o \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 \\
\sum x_i^2 y_i &= a_o \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4
\end{aligned}\right\}$$

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:

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

- 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	$x_i y_i$	$x_i^2 y_i$
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:

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

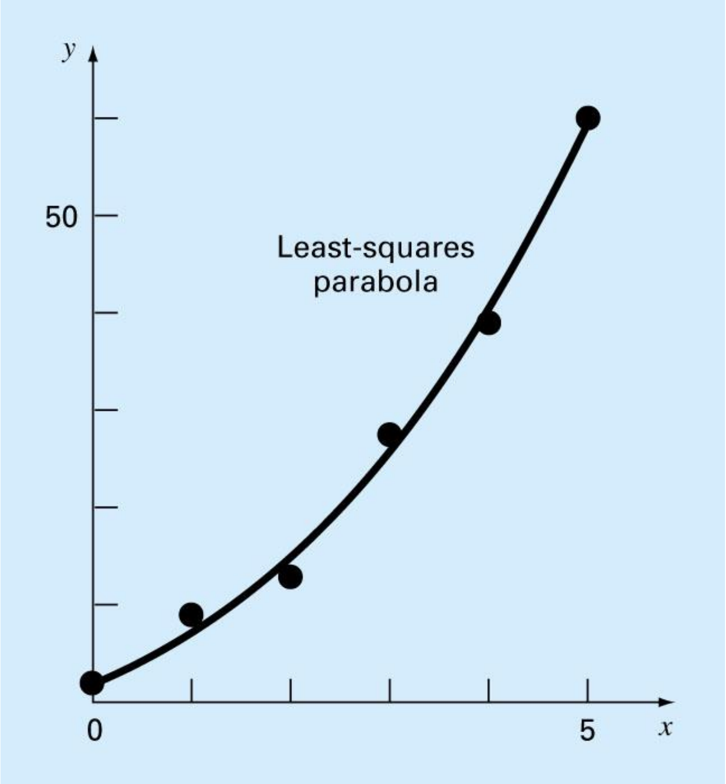
$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{Bmatrix}$$

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

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

Polynomial Regression- Example (cont'd)

x_i	y_i	y_{model}	e_i^2	$(y_i-\bar{y})^2$
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**



$$S_t = \sum (y_i - \bar{y})^2 = 2513.39$$

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

$$S_r = \sum e_i^2 = 3.74657$$

•The standard error of estimate:

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

•The coefficient of determination:

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

Nonlinear Regression

- Consider the previous exponential regression:

$$y = f(x_i) = a_o(1 - e^{-a_1 x})$$

- The sum of the squares of the residuals:

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

- The criterion for least squares regression is:

$$\frac{\partial S_r}{\partial a_o} = 0 \quad \& \quad \frac{\partial S_r}{\partial a_1} = 0$$

Nonlinear Regression

$$y = f(x_i) = a_o(1 - e^{-a_1 x})$$

$$S_r = \sum_{i=1}^{n}\Big(y_i - a_o(1 - e^{-a_1 x_i})\Big)^2 = \sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2$$

$$\frac{\partial S_r}{\partial a_o} = 0 \quad \& \quad \frac{\partial S_r}{\partial a_1} = 0$$

$$\frac{\partial S_r}{\partial a_o} = -2\sum_{i=1}^{n}\big(y_i - f(x_i)\big)\left(\frac{\partial f(x_i)}{\partial a_o}\right) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n}\big(y_i - f(x_i)\big)\left(\frac{\partial f(x_i)}{\partial a_1}\right) = 0$$

Nonlinear Regression

$$\sum_{i=1}^{n}\left(y_i - f(x_i)\right)\left(\frac{\partial f(x_i)}{\partial a_o}\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i - f(x_i)\right)\left(\frac{\partial f(x_i)}{\partial a_1}\right) = 0$$

- The partial derivatives are expressed at every data point (i) in terms of a_o and a_1.
- Thus, the above leads to 2 equations in 2 unknowns which can be solved iteratively for a_o and a_1.