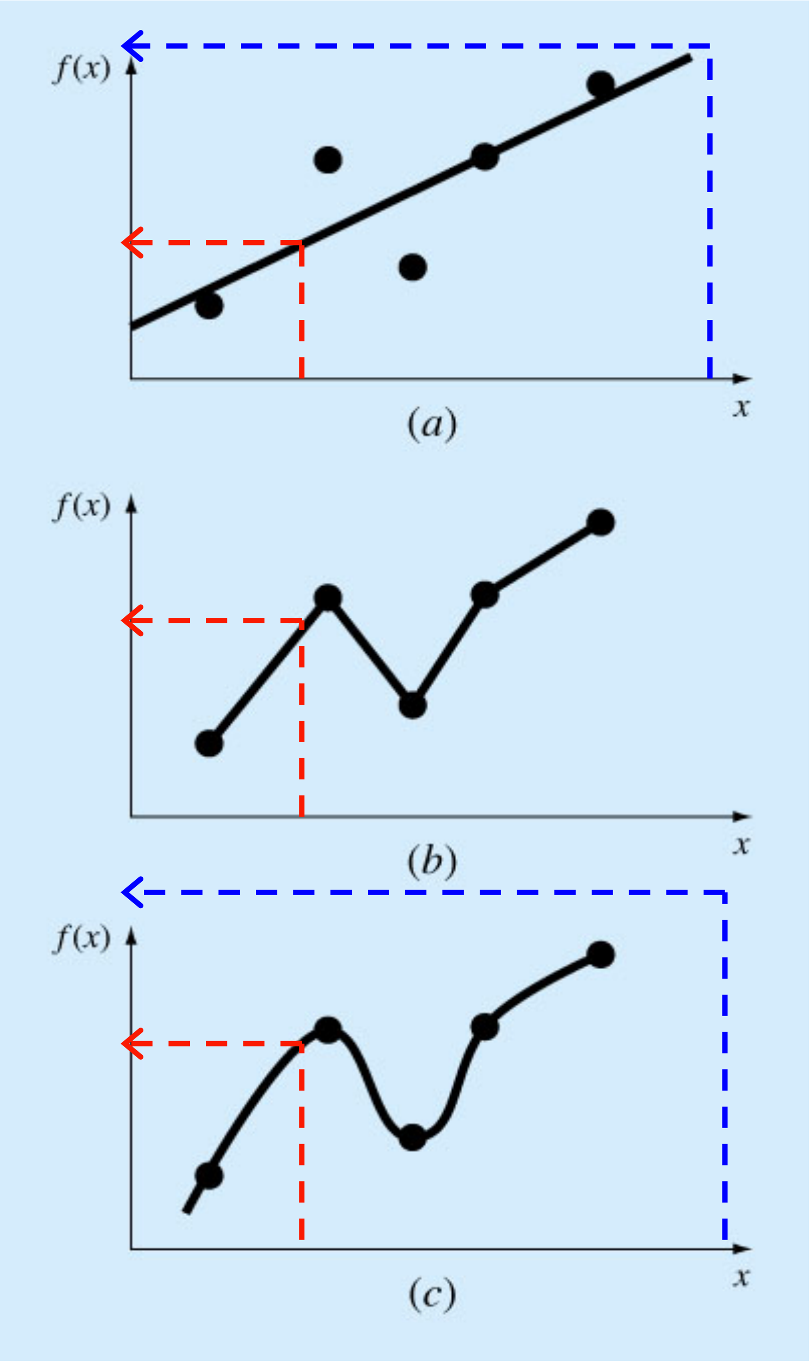
Curve-Fitting

Regression

Some Applications of Curve Fitting

- To fit curves to a collection of discrete points in order to obtain intermediate estimates or to provide trend analysis



Some Applications of Curve Fitting

- Function approximation
  - e.g.: In the applications of numerical integration

$$f(x) \approx p_n(x) \Rightarrow \int_a^b f(x) \approx \int_a^b p_n(x)$$

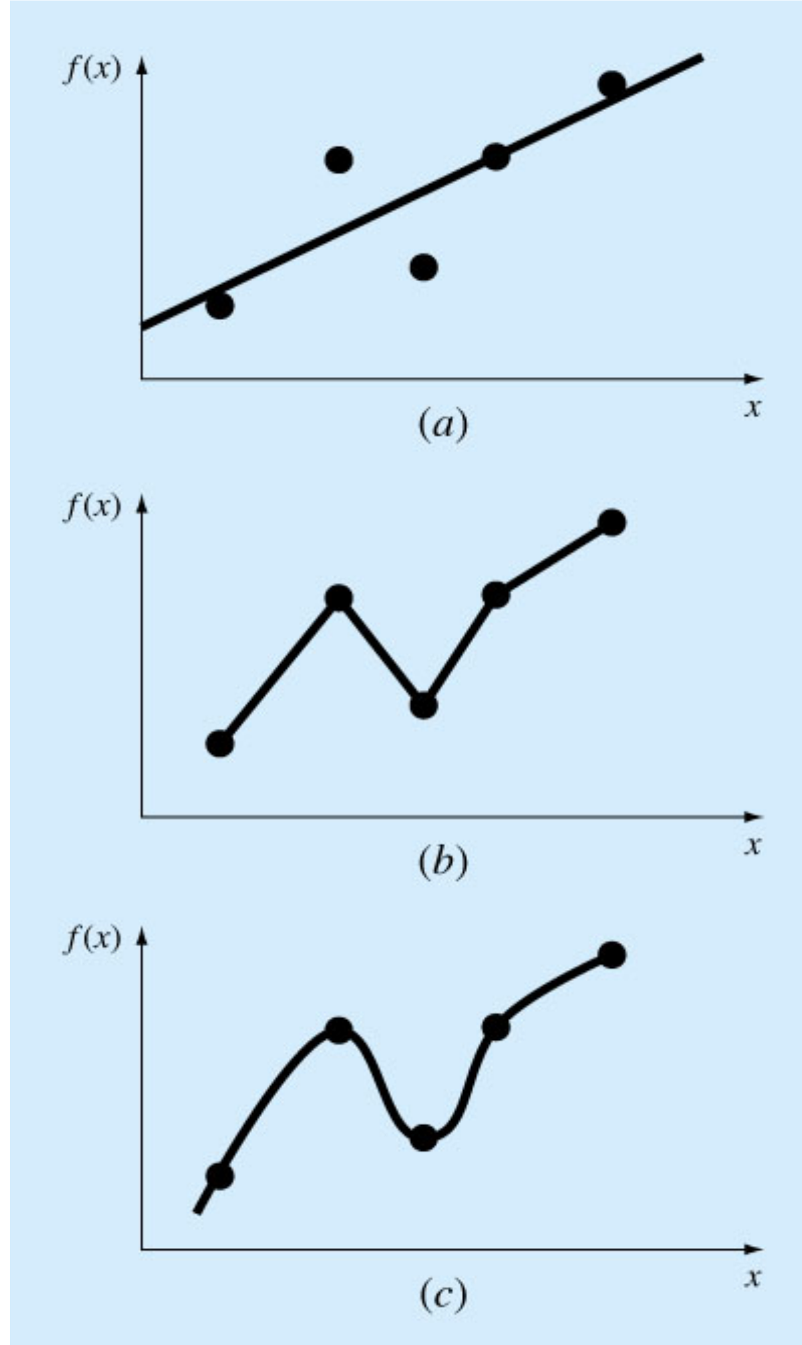
where $p_n(x)$ is an nth order polynomial

- Hypothesis testing
  - Compare theoretical data model to empirical data collected through experiments to test if they agree with each other.

Two Approaches

- *Regression* – Find the "best" curve to fit the points. The curve does not have to pass through the points. (Fig (a))

- *Interpolation* – Fit a curve or series of curves that pass through every point. (Figs (b) & (c))



Curve Fitting

Regression

- Linear Regression
- Polynomial Regression
- Multiple Linear Regression
- Non-linear Regression

Interpolation

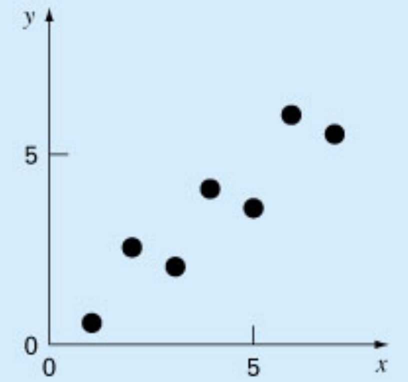
- Newton's Divided-Difference Interpolation
- Lagrange Interpolating Polynomials
- Spline Interpolation

Linear Regression – Introduction

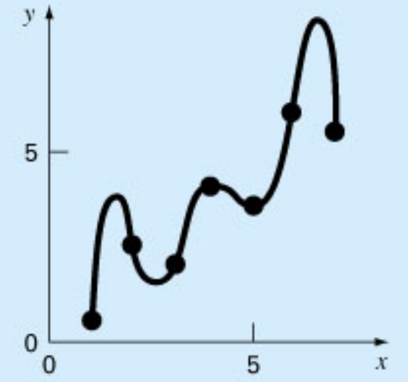
- Some data exhibit a linear relationship but have noises

- A curve that interpolates all points (that contain errors) would make a poor representation of the behavior of the data set.

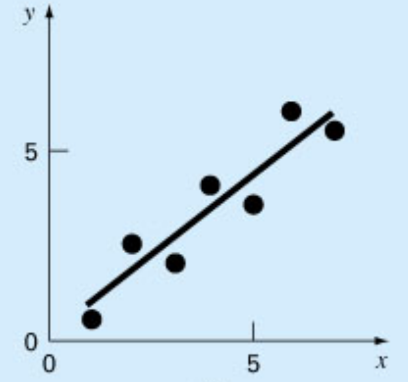
- A straight line captures the linear relationship better.



(a)



(b)

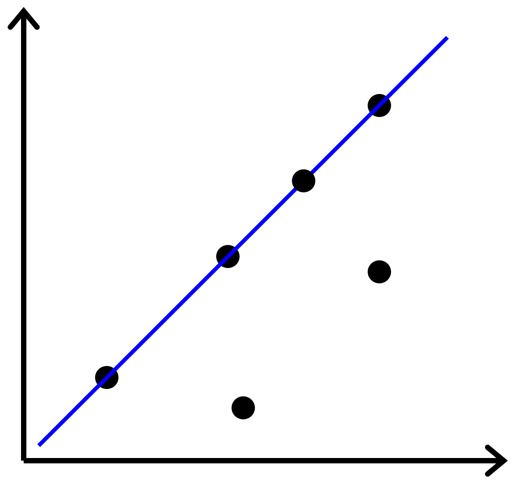


(c)

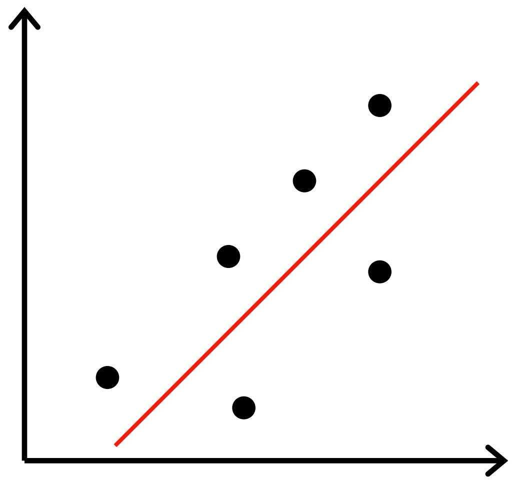
Linear Regression

Objective: Want to fit the "best" line to the data points (that exhibit linear relation).

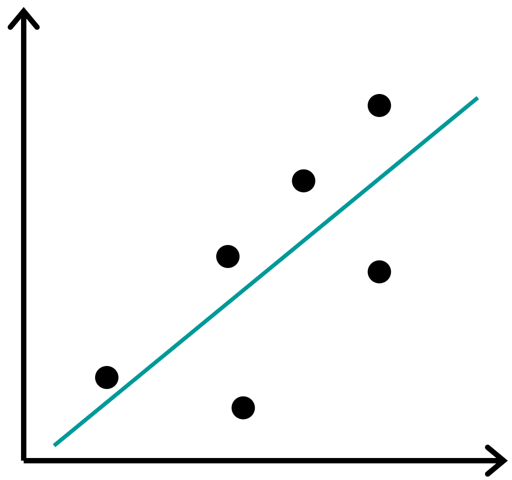
- How do we define "best"?



Pass through as many points as possible



Minimize the maximum residual of each point



Each point carries the same weight

Linear Regression

Objective

- Given a set of points

$$(x_1, y_1) , (x_2, y_2), \ldots, (x_n, y_n)$$

- Want to find a straight line

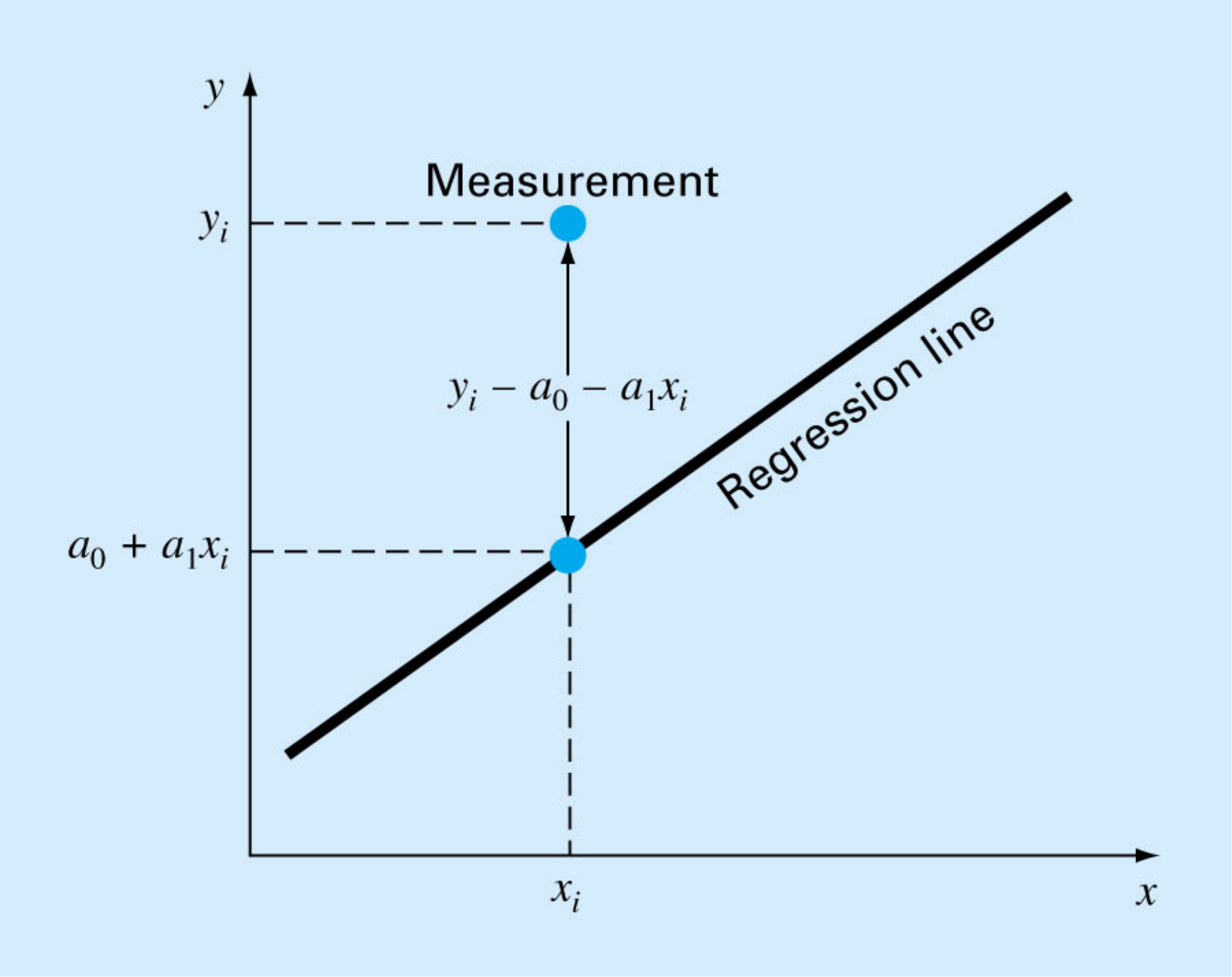
$$y = a_0 + a_1 x$$

that best fits the points.

The error or residual at each given point can be expressed as

$$e_i = y_i - a_0 - a_1 x_i$$

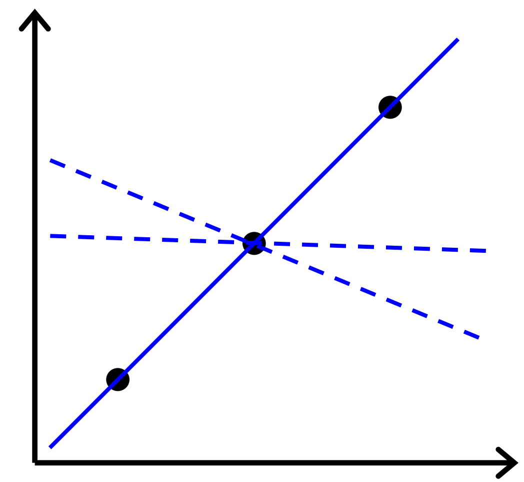
Residual (Error) Measurement



Criteria for a "Best" Fit

- Minimize the sum of residuals

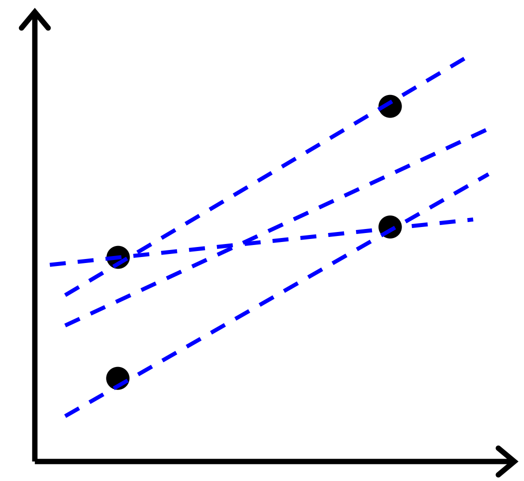
$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)$$



  - Inadequate
  - e.g.: Any line passing through mid-points would satisfy the criteria.

- Minimize the sum of absolute values of residuals (L_1-norm)

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



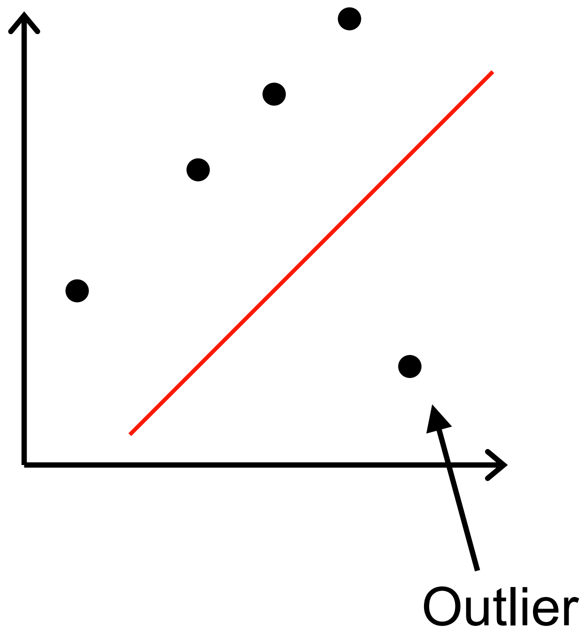
  - "Best" line may not be unique
  - e.g.: Any line within the upper and lower points would satisfy the criteria.

Criteria for a "Best" Fit

- Minimax method: Minimize the largest residuals of all the point (L_∞-Norm)

$$\min_{0 \le i \le n} \max e_i = \min_{0 \le i \le n} \max \left| y_i - a_0 - a_1 x_i \right|$$

  - Not easy to compute
  - Bias toward outlier
  - e.g.: Data set with an outlier. The line is affected strongly by the outlier.



Note: Minimax method is sometimes well suited for fitting a simple function to a complicated function. (Why?)

Least-Square Fit

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

- Minimize the sum of squares of the residuals (L_2-Norm)
- Unique solution
- Easy to compute
- Closely related to statistics

How to find a_0 and a_1 that minimize $\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$

Least-Squares Fit of a Straight Line

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

To minimize $S_r(a_0, a_1)$, we can find a_0, a_1 that satisfy

$$\frac{\partial S_r}{\partial a_0} = 0$$

$$\Rightarrow -2\sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) = 0$$

$$\frac{\partial S_r}{\partial a_1} = 0$$

$$\Rightarrow -2\sum_{i=1}^{n} x_i (y_i - a_0 - a_1 x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (x_i y_i - a_0 x_i - a_1 x_i^2) = 0$$

Least-Squares Fit of a Straight Line

$$\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_0 - \sum_{i=1}^{n} a_1 x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n a_0 - a_1 \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow n a_0 + \left(\sum_{i=1}^{n} x_i\right) a_1 = \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n}(x_i y_i - a_0 x_i - a_1 x_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} a_0 x_i - \sum_{i=1}^{n} a_1 x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - a_0 \sum_{i=1}^{n} x_i - a_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \left(\sum_{i=1}^{n} x_i\right) a_0 + \left(\sum_{i=1}^{n} x_i^2\right) a_1 = \sum_{i=1}^{n} x_i y_i$$

These are called the *normal equations*.

How do you find a_0 and a_1?

Least-Squares Fit of a Straight Line

$$\begin{aligned} n a_0 &+ \left(\sum_{i=1}^{n} x_i\right) a_1 &= \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a_0 &+ \left(\sum_{i=1}^{n} x_i^2\right) a_1 &= \sum_{i=1}^{n} x_i y_i \end{aligned} \Rightarrow \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

Solving the system of equations yields

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad a_0 = \frac{\sum y_i - a_1 \sum x_i}{n} = \bar{y} - a_1 \bar{x}$$

Statistics Review

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (Mean)}$$

$$S_t = \sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ (Sum of squares of the residuals)}$$

$$S_y = \sqrt{\frac{S_t}{n-1}} \text{ (Standard deviation)}$$

- Mean – The "best point" that minimizes the sum of squares of residuals.
- Standard deviation – Measure how the sample (data) spread about the mean.
  - The smaller the standard deviation the better the mean describes the sample.

Quantification of Error of Linear Regression

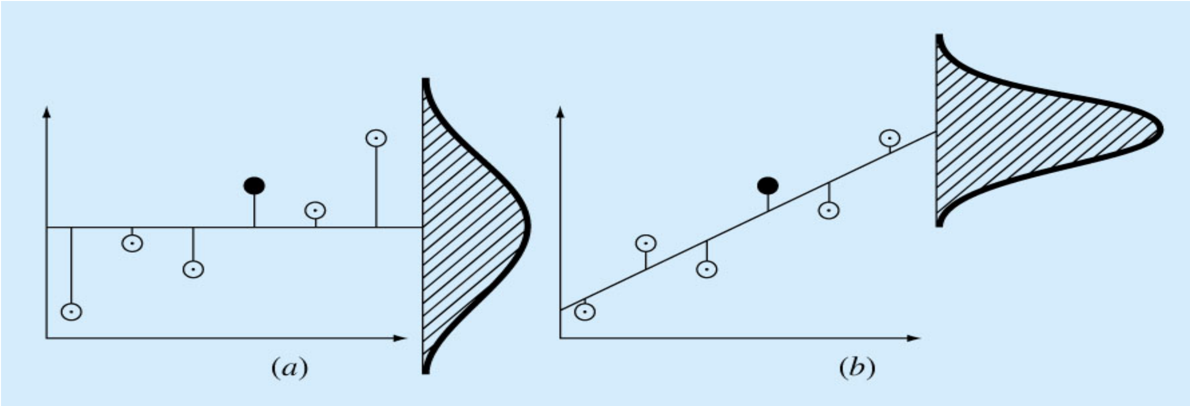
$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

$$S_{y/x} = \sqrt{\frac{S_r}{n-2}}$$

$S_{y/x}$ is called the *standard error of the estimate*.

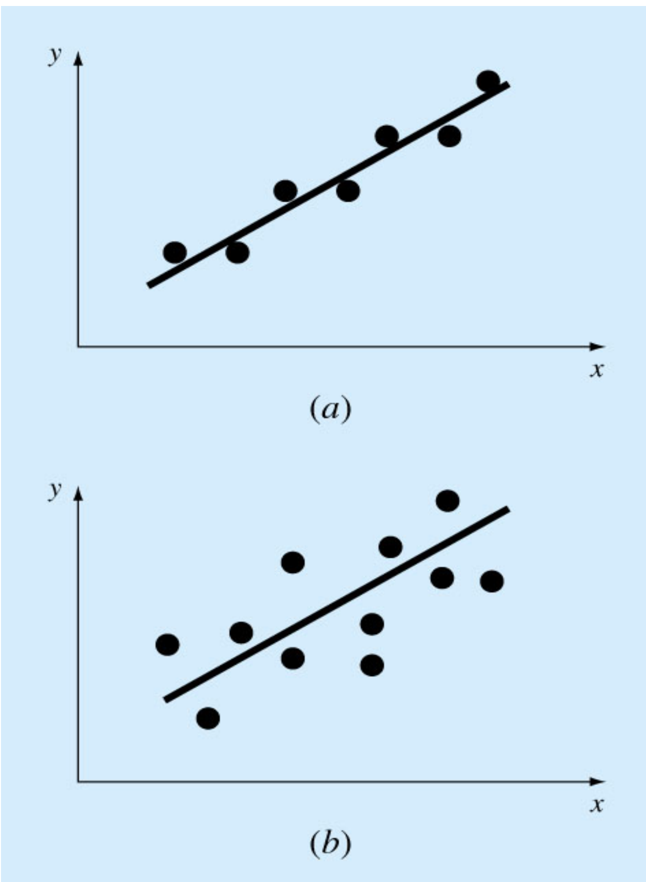
Similar to "standard deviation", $S_{y/x}$ quantifies the spread of the data points around the regression line.

The notation "y/x" designates that the error is for predicted value of y corresponding to a particular value of x.



(a) Spread of the data around the mean of the dependent variable.

(b) Spread of the data around the best-fit line.



Linear regression with (a) small and (b) large residual errors.

"Goodness" of our fit

- Let S_t be the sum of the squares around the mean for the dependent variable, y

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Let S_r be the sum of the squares of residuals around the regression line
- 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.

"Goodness" of our fit

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

r^2 : coefficient of determination

r : correlation coefficient

- 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.

- e.g.: $r^2=0.868$ means 86.8% of the original uncertainty has been "explained" by the linear model.

Polynomial Regression

Objective

- Given n points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- Want to find a polynomial of degree m

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$$

that best fits the points.

The error or residual at each given point can be expressed as

$$e_i = y_i - a_0 - a_1 x - a_2 x^2 - \ldots - a_m x^m$$

Least-Squares Fit of a Polynomial

The procedures for finding $a_0, a_1, \ldots, a_m$ that minimize the sum of squares of the residuals is the same as those used in the linear least-square regression.

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_o - a_1 x_i - a_2 x_i^2 - \ldots - a_m x_i^m)^2$$

$$\text{Setting } \frac{\partial S_r}{\partial a_j} = 0 \quad \text{for } j = 0,1,\ldots,m \text{ yields}$$

$$\sum_{i=1}^{n} x_i^j (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_m x_i^m) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i^j (a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_m x_i^m) = x_i^j y_i$$

Least-Squares Fit of a Polynomial

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{m+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \cdots & \sum x_i^{m+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \sum x_i^{m+2} & \cdots & \sum x_i^{2m} \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \\ \vdots \\ \sum x_i^m y_i \end{bmatrix}$$

To find a_0, a_1, …, a_n that minimize S_r, we can solve this system of linear equations.

The standard error of the estimate becomes

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

Multiple Linear Regression

- In linear regression, y is a function of one variable.
- In multiple linear regression, y is a linear function of multiple variables.
- Want to find the best fitting linear equation

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m$$

- Same procedure to find $a_0, a_1, a_2, \ldots, a_m$ that minimize the sum of squared residuals
- The standard error of estimate is $S_{y/x} = \sqrt{\dfrac{S_r}{n-(m+1)}}$

General Linear Least Square

- All of simple linear, polynomial, and multiple linear regressions belong to the following general linear least squares model:

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \ldots + a_m z_m + e$$

where

z_i are different functions of x's (can be any kind of functions)

- It is called "linear" because the dependent variable, y, is a linear function of a_i's.

How Other Regressions Fit Into Linear Least Square Model

- Polynomial:

$$y = a_0(1) + a_1(x) + a_2(x^2) + \ldots + a_m(x^m) + e$$

$$\text{i.e., } z_0 = x^0 = 1, z_1 = x, z_2 = x^2, \ldots, z_m = x^m$$

- Multiple linear:

$$y = a_0(1) + a_1(x_1) + a_2(x_2) + \ldots + a_m(x_m) + e$$

$$\text{i.e., } z_0 = 1, z_1 = x_1, z_2 = x_2, \ldots, z_m = x_m$$

- Others:

$$y = a_0(\sin x_1) + a_1(\ln x_1) + a_2(x_2 \cos x_3) + \frac{a_3}{x_1 x_2} + e$$

$$\text{i.e., } z_0 = \sin x_1, z_1 = \ln x_1, z_2 = x_2 \cos x_3, z_3 = (x_1 x_2)^{-1}$$

General Linear Least Square

- Given n points, we have

$$y_j = a_0 z_{0j} + a_1 z_{1j} + a_2 z_{2j} + \ldots + a_m z_{mj} + e_i, \qquad j = 1, \ldots, n$$

where z_{ij} represents the value of function z_i at the j^{th} point.

- We can express the above equations in matrix form as

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{e} \quad or \quad \begin{bmatrix} y_1 \\ y_n \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix}$$

General Linear Least Square

The sum of squares of the residuals can be calculated as

$$S_r = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} a_j z_{ji} \right)^2$$

To minimize S_r, we can set the partial derivatives of S_r to zeroes and solve the resulting normal equations.

The normal equations can be expressed concisely as

$$\mathbf{Z}^T \mathbf{Z} \mathbf{a} = \mathbf{Z}^T \mathbf{y}$$

How should we solve this system?

Example

X	3	5	6
Y	4	1	4

- Find the straight line that best fit the data in least-square sense.
- A straight line can be expressed in the form $y = a_0 + a_1 x$. That is, with $z_0 = 1$, $z_1 = x$.
- Thus we can construct $\mathbf{Z}$ as

$$\mathbf{Z} = \begin{bmatrix} 1 & 3 \\ 1 & 5 \\ 1 & 6 \end{bmatrix}$$

Example

Our objective is to solve $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \mathbf{Z}^T\mathbf{y}$, or

$$\begin{bmatrix} 1 & 1 & 1 \\ 3 & 5 & 6 \end{bmatrix}\begin{bmatrix} 1 & 3 \\ 1 & 5 \\ 1 & 6 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 5 & 6 \end{bmatrix}\begin{bmatrix} 4 \\ 1 \\ 4 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 3 & 14 \\ 14 & 70 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 9 \\ 41 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 4 \\ -3/14 \end{bmatrix} \text{ or } \begin{bmatrix} 4 \\ -0.2143 \end{bmatrix}$$

The best line is $\mathrm{y} = 4 - 0.2143x$

Solving $\mathbf{Z}^T\mathbf{Z}\mathbf{a} = \mathbf{Z}^T\mathbf{y}$

Note: $\mathbf{Z}$ is an n by $(m+1)$ matrix.

- Gaussian or LU decomposition
  - Less efficient

- Cholesky decomposition
  - Decompose $\mathbf{Z}^T\mathbf{Z}$ into $\mathbf{R}^T\mathbf{R}$ where $\mathbf{R}$ is an upper triangular matrix.
  - Solve $\mathbf{Z}^T\mathbf{Z}\mathbf{a} = \mathbf{Z}^T\mathbf{y}$ as $\mathbf{R}^T\mathbf{R}\mathbf{a} = \mathbf{Z}^T\mathbf{y}$

- QR decomposition
- Singular value decomposition

Solving $\mathbf{Z}^T\mathbf{Z}\mathbf{a} = \mathbf{Z}^T\mathbf{y}$ (Cholesky decomposition) **

- Given a nxm matrix $\mathbf{Z}$.
- Suppose we have computed $\mathbf{R}_{m\text{x}m}$ from $\mathbf{Z}^T\mathbf{Z}$ using Cholesky decomposition
- If we add an additional column to $\mathbf{Z}$, then the new $\mathbf{R}$ will be in the form

$$\begin{bmatrix} & & & r_{1,m+1} \\ & \mathbf{R}_{m\times m} & & r_{2,m+1} \\ & & & \vdots \\ 0 & 0 & \cdots & r_{m+1,m+1} \end{bmatrix}$$

i.e., we only need to compute the $(m+1)^{\text{th}}$ column of $\mathbf{R}$.

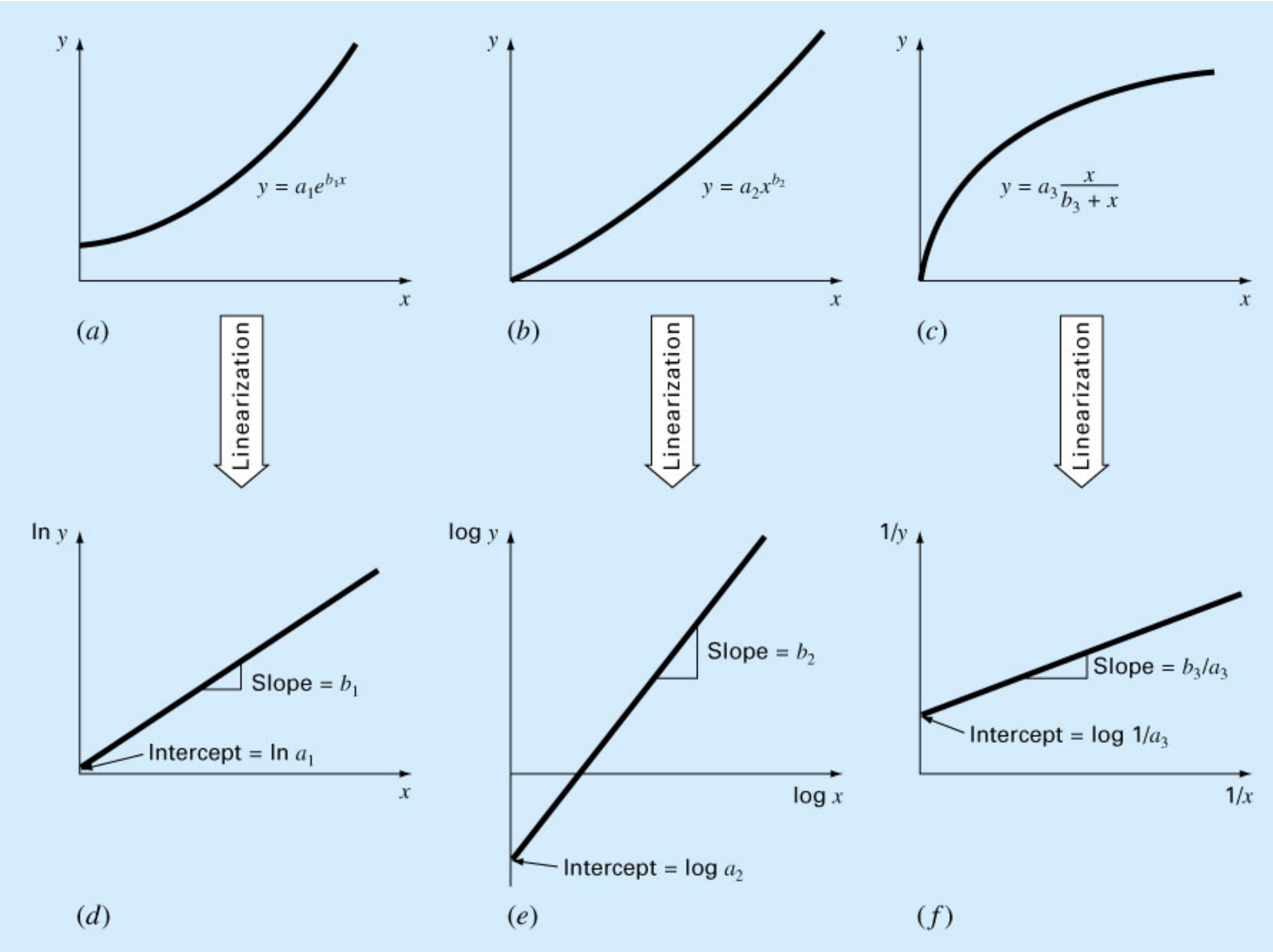
- Suitable for testing how much improvement in terms of least-square fit a polynomial of one degree higher can provide

Linearization of Nonlinear Relationships

- Some non-linear relationships can be transformed so that in the transformed space the data exhibit a linear relationship.

- For examples,

Exponential equation	$y = a_1 e^{b_1 x}$	$\Rightarrow \ln y = \ln a_1 + b_1 x$
Power equation	$y = a_2 x^{b_2}$	$\Rightarrow \log y = \log a_2 + b_2 \log x$
Saturation Growth-rate equation.	$y = a_3 \dfrac{x}{b_3 + x}$	$\Rightarrow \dfrac{1}{y} = \dfrac{b_3}{a_3} \dfrac{1}{x} + \dfrac{1}{a_3}$



y
$y = a_1 e^{b_1 x}$
x
(a)
Linearization
y
$y = a_2 x^{b_2}$
x
(b)
Linearization
y
$y = a_3 \frac{x}{b_3 + x}$
x
(c)
Linearization
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)

Example

X	1	2	3
Y	4	1	4

Find the saturation growth rate equation $y = a_1 \frac{x}{b_1 + x}$

that best fit the data in least-square sense.

Solution: Step 1: Linearize the curve as

$$y = a_1 \frac{x}{b_1 + x} \Rightarrow \frac{1}{y} = \frac{b_1}{a_1}\frac{1}{x} + \frac{1}{a_1} \Rightarrow y' = c_1 x' + c_2$$

$$\text{where } y' = \frac{1}{y}, x' = \frac{1}{x}, c_1 = \frac{b_1}{a_1}, c_2 = \frac{1}{a_1}$$

Example

Step 2: Transform data from original space to "linearized space".

X	1	2	3
Y	4	1	4
X' = 1/X	1	1/2	1/3
Y' = 1/Y	1/4	1	1/4

Step 3: Perform linear least square fit for $y' = c_1 x' + c_2$

From the data we have $\mathbf{Z} = \begin{bmatrix} 1 & 1 \\ 1/2 & 1 \\ 1/3 & 1 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 1/4 \\ 1 \\ 1/4 \end{bmatrix}$

Solving $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \mathbf{Z}^T\mathbf{y}$ yields $\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} -0.3462 \\ 0.7115 \end{bmatrix}$

$c_2 = 1/a_1 \Rightarrow a_1 = 1.4055, c_1 = b_1/a_1 \Rightarrow b_1 = -0.4866$

Thus $y = 1.4055x/(-0.4866 + x)$ is an "accetably good" curve that fits the data (It is not optimal in least square sense).

- Best least square fit in the transformed space ≠best least square fit in the original space
  - For many applications, however, the parameters obtained from performing least square fit in the transformed space are acceptable.

- Linearization of Nonlinear Relationships
  - Sub-optimal result
  - Easy to compute

Non-Linear Regression **

- The relationship among the parameters, a_i's, is non-linear and cannot be linearized using direct method.
- For example, $y = a_0(1 - e^{-a_1 x})$
- Objective: Find a_0 and a_1 that minimizes

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - a_0(1 - e^{-a_1 x_i}) \right]^2$$

- Possible approaches to find the solution:
  - Applying minimization of non-linear function
  - Set partial derivatives to zero and solve non-linear equation.
  - Gauss-Newton Method

Other Notes

- When performing least square fit,
  - The order of the data in the table is not important
  - The order in which you arrange the basis functions is not important.
  - e.g., Least square fit of

    $y = a_0 + a_1 x$ or $y = b_0 x + b_1$ to

X	3	5	6
Y	4	1	4

or

X	6	3	5
Y	4	4	1

or

X	5	6	3
Y	1	4	4

would yield the same straight line.