



UNIVERSITAS GADJAH MADA
DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING
BACHELOR IN CIVIL ENGINEERING

Statistics and Probability

Regression

Curve fitting

- A line or curve that represents a number of data points
- There are two methods to find such line or curve
 - Regression
 - Interpolation
- Engineering applications
 - Trend analysis
 - Hypothesis testing

Regression vs interpolation

Regression

The data show significant errors or noise

To find a single curve that represent general trend of the data

Regression line (curve) does not need to pass every data point

Interpolation

The data are accurate

To find a curve or curves that encompass(es) every data point

To estimate values between data points

Regression and interpolation

- Extrapolation
 - Similar to interpolation but applied to outside range of data points
 - Not recommended

Curve fitting to measured data

- Trend analysis
 - Use of data trend (measurements, experiments) to estimate values
 - If the data are accurate, use interpolation technique
 - If the data show noise, use regression technique
- Hypothesis testing
 - Comparison between theoretical values with computed ones

Recall on statistical parameters

- Arithmetic mean

$$\Rightarrow \bar{Y} = \frac{1}{n} \sum y_i$$

- Standard deviation

$$\Rightarrow s_Y = \sqrt{\frac{S_t}{n-1}} \quad S_t = \sum (y_i - \bar{Y})^2$$

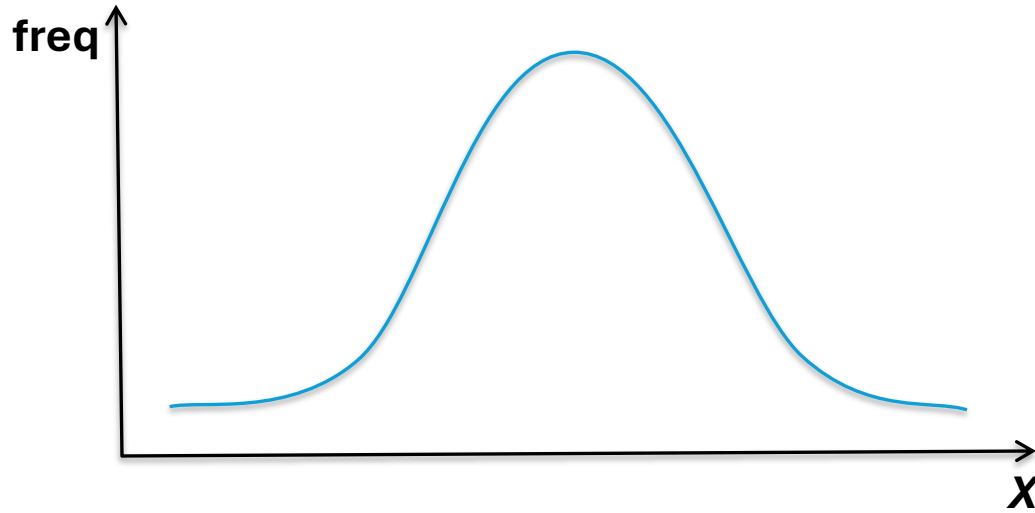
- Variance

$$\Rightarrow s_Y^2 = \frac{S_t}{n-1}$$

- Coefficient of variation

$$\Rightarrow c_v = \frac{s_Y}{\bar{Y}} 100\%$$

Probability distribution



Normal Distribution
one of data distributions
that is frequently
encountered in engineering

Regression

Simple Linear Regression

Regression: least-square method

- To find a single curve or function (approximate) that represents the general trend of the data
 - The data show significant error
 - The curve does not need to pass every data point
- Methods
 - Linear regression (simple linear regression)
 - Linearized expressions
 - Polynomial regression
 - Multiple linear regression
 - Non-linear regression

Regression: least-square method

■ How

- Spreadsheet (Microsoft Excel)
- Computer program
 - MatLab
- Freeware
 - Octave
 - Scilab
 - Freemat
- Self-made computer program

Simple linear regression

- To find a straight line that represents the general trend of data points: $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$
 - $y_{reg} = a_0 + a_1x$
 - a_0 intercept
 - a_1 slope
- Microsoft Excel
 - =INTERCEPT(y,x)
 - =SLOPE(y,x)

Simple Linear Regression

■ Error or residual

- Discrepancies between actual value of y (y data) and approximate value of y (y_{reg}) according to linear expression ($a_0 + a_1x$)

$$e = y - y_{reg} = y - (a_0 + a_1x)$$

- Minimize the sum of squared residues

$$\min[S_r] = \min \left[\sum e_i^2 \right] = \min \left[\sum (y - a_0 - a_1x)^2 \right]$$

Simple linear regression

- How to find a_0 and a_1 ?
 - Differentiate the equation of S_r twice; firstly with respect to a_0 and lastly with respect to a_1
 - Set each of the two equations to zero
 - Solve the equations for a_0 and a_1

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum (y_i - a_0 - a_1 x_i) x_i$$

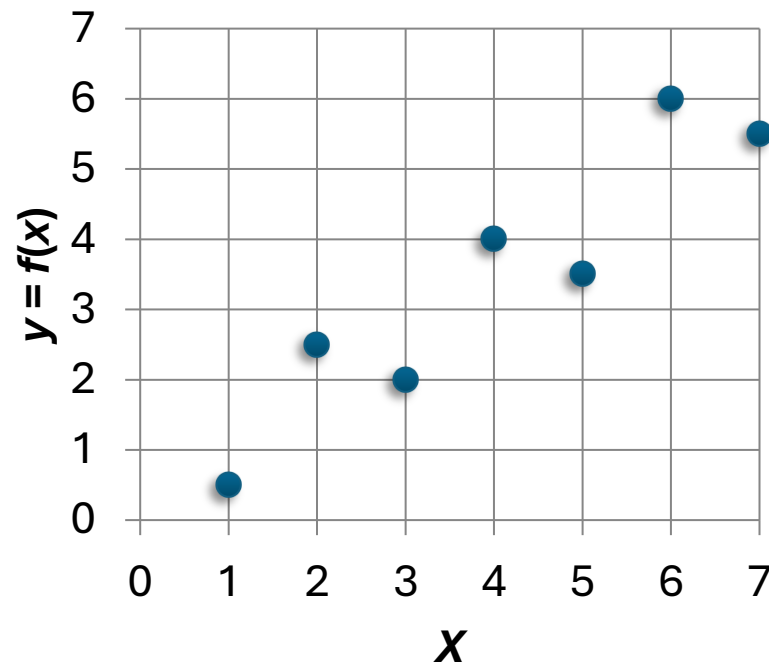
$$\frac{\partial S_r}{\partial a_0} = 0 \qquad \frac{\partial S_r}{\partial a_1} = 0$$

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

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

Example #1

i	x_i	$y_i = f(x_i)$
0	1	0.5
1	2	2.5
2	3	2
3	4	4
4	5	3.5
5	6	6
6	7	5.5



Example #1

i	x_i	y_i	$x_i y_i$	x_i^2	y_{reg}	$(y_i - y_{reg})^2$	$(y_i - y_{mean})^2$
0	1	0.5	0.5	1	0.910714	0.168686	8.576531
1	2	2.5	5	4	1.75	0.5625	0.862245
2	3	2.0	6	9	2.589286	0.347258	2.040816
3	4	4.0	16	16	3.428571	0.326531	0.326531
4	5	3.5	17.5	25	4.267857	0.589605	0.005102
5	6	6.0	36	36	5.107143	0.797194	6.612245
6	7	5.5	38.5	49	5.946429	0.199298	4.290816
Σ	28	24.0	119.5	140	Σ	2.991071	22.71429

Example #1

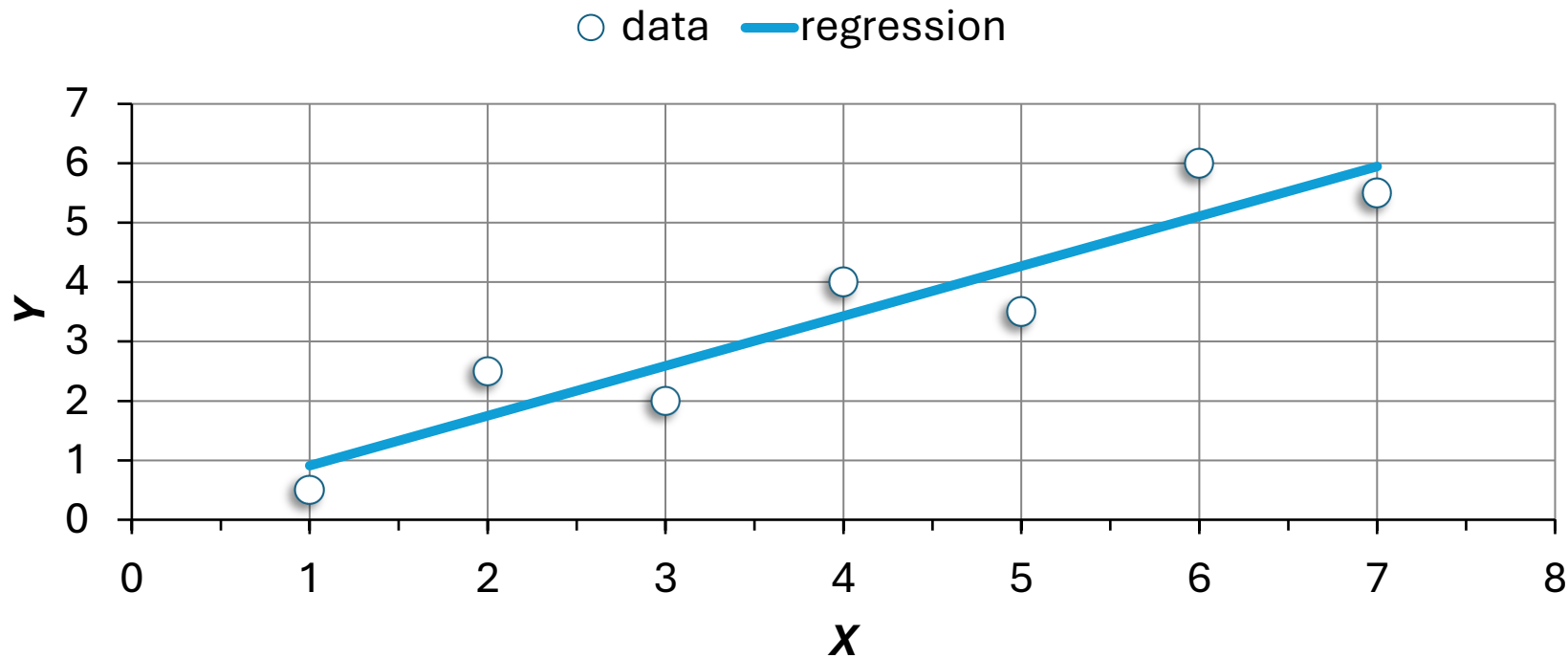
$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{7(119.5) - 28(24)}{7(140) - (28)^2} = 0.839286$$

$$\bar{y} = \frac{24}{7} = 3.4$$

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

$$a_0 = \bar{y} - a_1 \bar{x} = 3.4 - 0.839286(4) = 0.071429$$

Example #1



Error

■ Error

- Standard error magnitude

$$s_{y/x} = \sqrt{\frac{S_r}{n-2}} \quad S_r = \sum (y_i - a_0 - a_1 x_i)^2$$

- Notice its similarity with standard deviation

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

Error

- Difference between the two “errors” signifies an improvement of the prediction or a reduction of error

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



coefficient of determination

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



correlation coefficient



$$-1 \leq r \leq +1$$

Error

$$S_r = \sum (y_i - a_0 - a_1 x_i)^2 = 2.991071$$

$$S_t = \sum (y_i - \bar{y})^2 = 22.71429$$

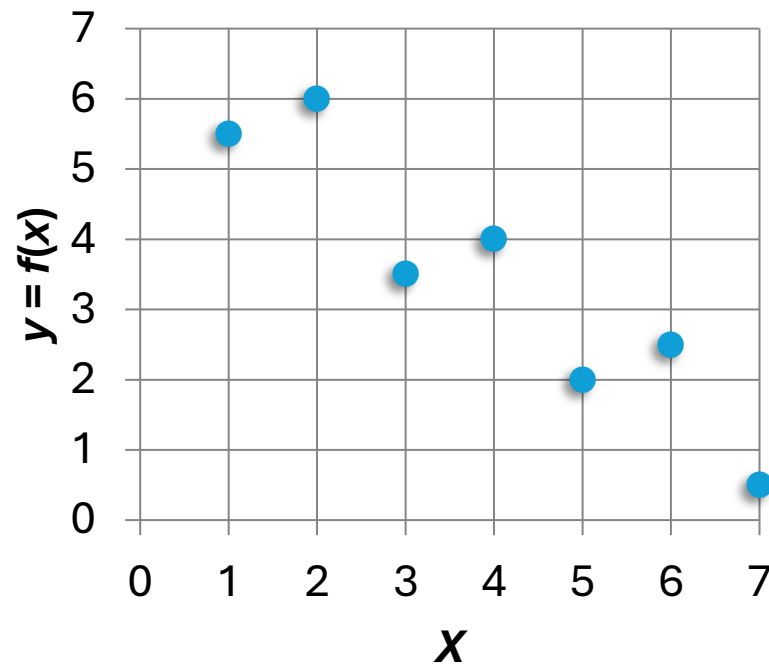
$$r^2 = \frac{S_t - S_r}{S_t} = \frac{22.71429 - 2.991071}{22.71429} = 0.868318$$

$$r = 0.931836$$

$$-1 \leq r \leq +1$$

Example #2

i	x_i	$y_i = f(x_i)$
0	1	5.5
1	2	6
2	3	3.5
3	4	4
4	5	2
5	6	2.5
6	7	0.5



Regression

Polynomial Regression

Polynomial regression

- Some engineering data, although exhibiting a marked pattern, is poorly represented by a straight line
 - Method 1: coordinate transformation (linearized non-linear eq.)
 - Method 2: polynomial regression
 - The m th-degree polynomial

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m$$

- The sum of the squares of the residuals

$$S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1x_i - a_2x_i^2 - \cdots - a_mx_i^m)^2$$

Polynomial regression

- The least-square method extended to fit the data to an m th-degree polynomial
- These equations can be set equal to zero and rearranged to develop a set of normal equations

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x_i^m)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x_i^m)$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x_i^m)$$

$$\cdot$$
$$\cdot$$
$$\cdot$$

$$\frac{\partial S_r}{\partial a_m} = -2 \sum x_i^m (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_m x_i^m)$$

Polynomial regression

$$a_0 n + a_1 \sum_{i=1}^n x_i + a_2 \sum_{i=1}^n x_i^2 + \cdots + a_m \sum_{i=1}^n x_i^m = \sum_{i=1}^n y_i$$

$$a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 + a_2 \sum_{i=1}^n x_i^3 + \cdots + a_m \sum_{i=1}^n x_i^{m+1} = \sum_{i=1}^n x_i y_i$$

$$a_0 \sum_{i=1}^n x_i^2 + a_1 \sum_{i=1}^n x_i^3 + a_2 \sum_{i=1}^n x_i^4 + \cdots + a_m \sum_{i=1}^n x_i^{m+2} = \sum_{i=1}^n x_i^2 y_i$$

⋮

$$a_0 \sum_{i=1}^n x_i^m + a_1 \sum_{i=1}^n x_i^{m+1} + a_2 \sum_{i=1}^n x_i^{m+2} + \cdots + a_m \sum_{i=1}^n x_i^{2m} = \sum_{i=1}^n x_i^m y_i$$

- There are $m + 1$ linear equations having $m + 1$ unknowns, i.e. $a_0, a_1, a_2, \dots, a_m$
- These linear equations can be simultaneously solved by using methods such as
 - Gauss elimination
 - Gauss-Jordan
 - Jacobi iteration
 - Matrix inversion

Example

- Fit a second-order polynomial to the data in the table on the right

$$y = a_0 + a_1x + a_2x^2$$

- Answer

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

$$r^2 = 1 - \frac{S_r}{S_t} = 1 - \frac{3.74657}{2513.39} = 0.9985$$

$$r = 0.9993$$

x_i	y_i
0	2.1
1	7.7
2	13.6
3	27.2
4	40.9
5	61.1

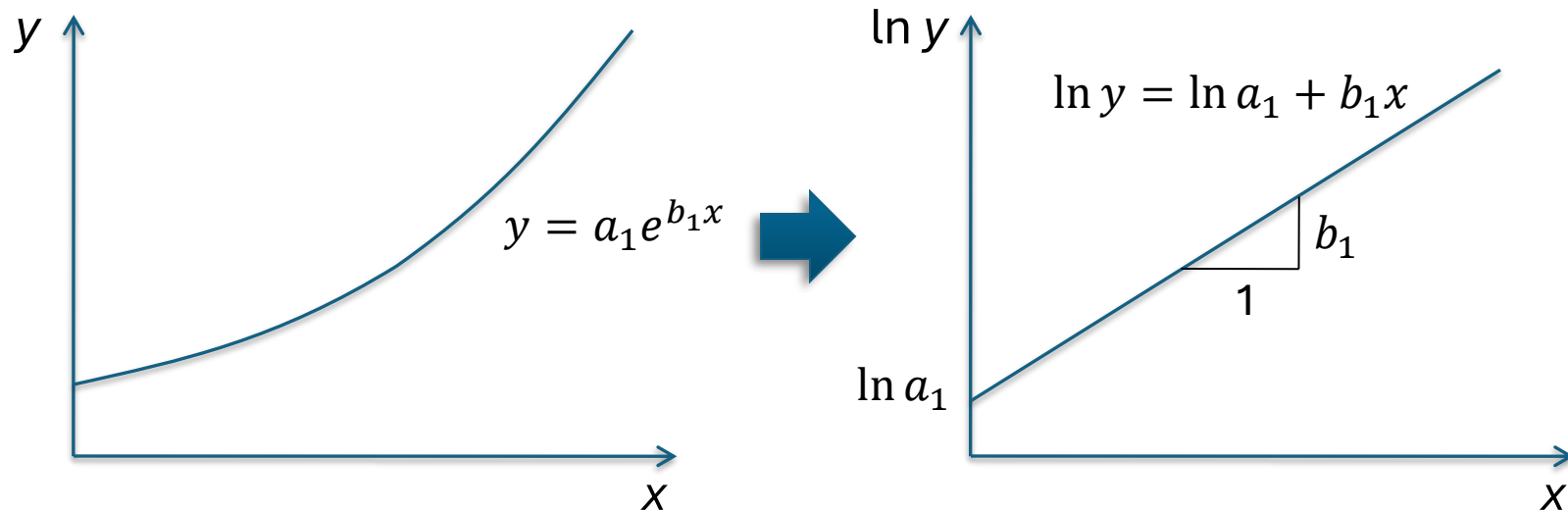
Regression

Regression of Linearized Expression

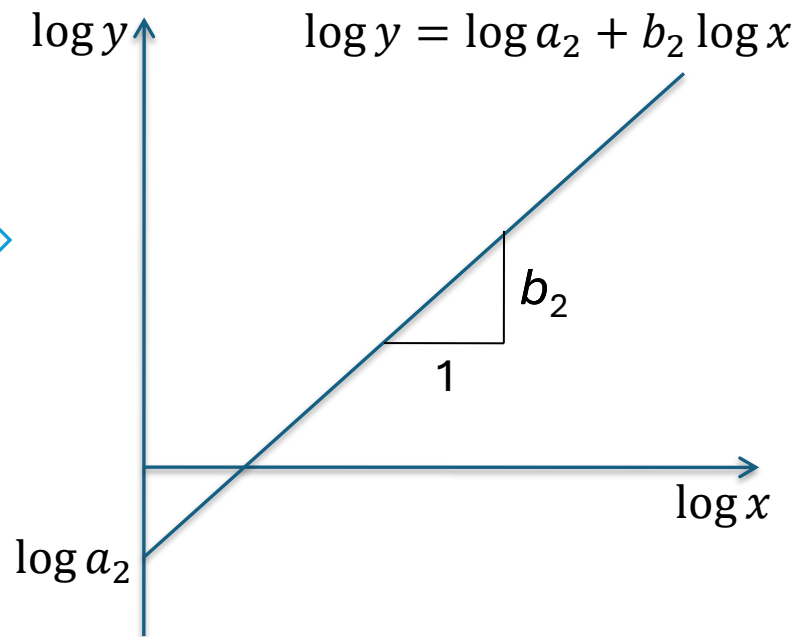
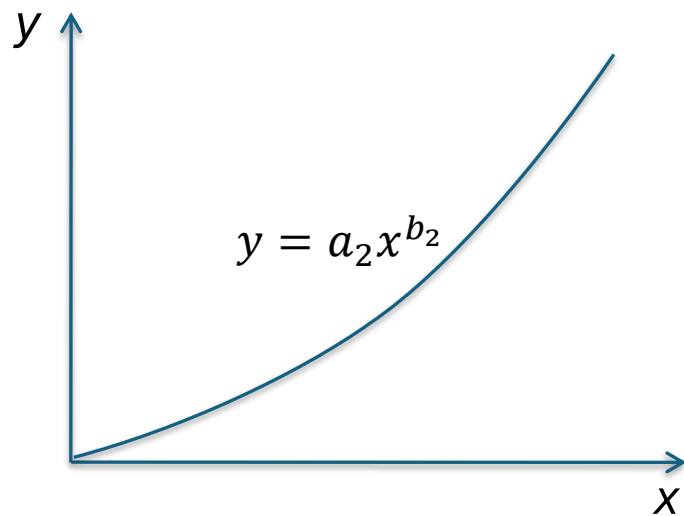
Linear Regression

- Linearized non-linear equations
 - Logarithmic eq. \rightarrow linear eq.
 - Exponential eq. \rightarrow linear eq.
 - n -th order polynomial eq. ($n > 1$) \rightarrow linear eq.
 - etc.

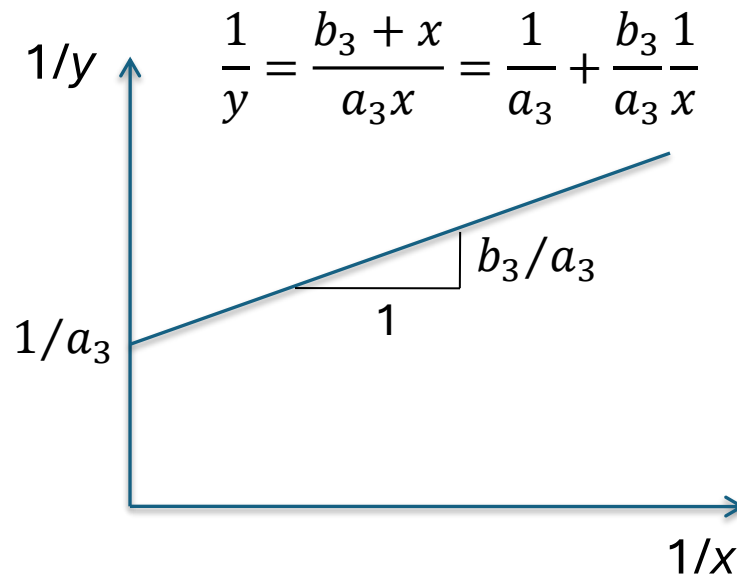
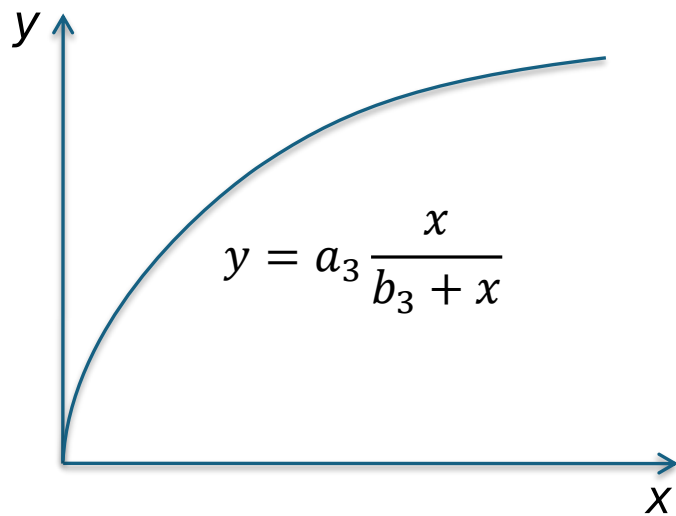
Linear Regression



Linear Regression



Linear Regression



Regression

Multiple Linear Regression

Multiple linear regression

- Suppose the dependent variable y is a linear function of two independent variables x_1 and x_2

$$y = a_0 + a_1x_1 + a_2x_2$$

- The best values of the coefficients are determined by setting up the sum of the squares of the residuals

$$S_r = \sum_{i=1}^n (y_i - a_0 - a_1x_{1i} - a_2x_{2i})^2$$

Multiple linear regression

- Differentiating this equation with respect to each of the unknown coefficients

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{i=1}^n x_{1i} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum_{i=1}^n x_{2i} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

Multiple linear regression

- Equating the differentials to zero and expressing the resulted equation as a set of simultaneous linear equations yield

$$a_0 n + a_1 \sum_{i=1}^n x_{1i} + a_2 \sum_{i=1}^n x_{2i} = \sum_{i=1}^n y_i$$

$$a_0 \sum_{i=1}^n x_{1i} + a_1 \sum_{i=1}^n x_{1i}^2 + a_2 \sum_{i=1}^n x_{1i} x_{2i} = \sum_{i=1}^n x_{1i} y_i$$

$$a_0 \sum_{i=1}^n x_{2i} + a_1 \sum_{i=1}^n x_{1i} x_{2i} + a_2 \sum_{i=1}^n x_{2i}^2 = \sum_{i=1}^n x_{2i} y_i$$

Multiple linear regression

- Written in matrix form

$$\begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{1i}y_i \\ \sum_{i=1}^n x_{2i}y_i \end{Bmatrix}$$

Example

- Find the best linear equation that fits to the data in the table on the right
- Answer

$$y = 5 + 4x_1 - 3x_2$$

$$r^2 = 1$$

x_1	x_2	y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
7	2	27

Multiple linear regression

- Multiple linear regression can be useful in the derivation of power equations of the general form

$$y = a_0 x_1^{a_1} x_2^{a_2} \dots x_m^{a_m}$$

- Such equations are extremely useful when fitting experimental data
- In order to use the multiple linear regression, the equation is transformed by taking its logarithm to yield

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

$$\Rightarrow y' = a_0' + a_1' x_1' + a_2' x_2' + \dots + a_m' x_m'$$

Regression

General Linear Least Squares

General linear least squares

- The three types of regression that have been presented, i.e. simple linear, polynomial, and multiple linear can be expressed in a general least-squares model

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

- where z_0, z_1, \dots, z_m are $m + 1$ different functions
- $m + 1$ is the number of independent variables
- $n + 1$ is the number of data points

- A special case is when $z_0 = 1$

$$\Rightarrow y = a_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m$$

General linear least squares

- The previous expression can be written in a matrix form

$$\{Y\} = [Z]\{A\}$$

General linear least squares

$$\{Y\} = [Z]\{A\} \Rightarrow [Z]^T [Z]\{A\} = [Z]^T \{Y\}$$

$$[Z] = \begin{bmatrix} a_{01} & a_{11} & \cdot & \cdot & \cdot & a_{m1} \\ a_{02} & a_{12} & \cdot & \cdot & \cdot & a_{m2} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ a_{0n} & a_{1n} & \cdot & \cdot & \cdot & a_{mn} \end{bmatrix}$$

- $\{Y\}$ contains the observed values of the dependent variables
- $[Z]$ is a matrix of the observed values of the independent variables
- $\{A\}$ contains the unknown coefficients

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

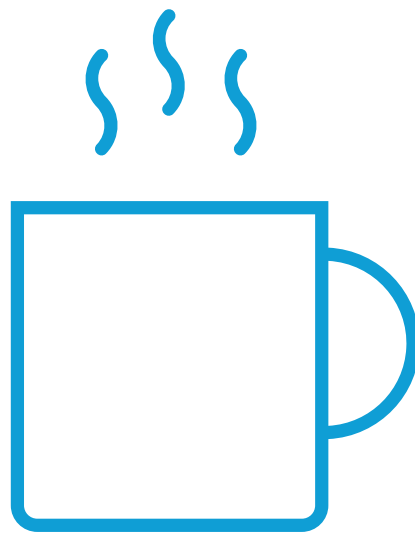
General Linear Least Squares

$$[Z]^T [Z] \{A\} = [Z]^T \{Y\}$$

■ Solution strategy

- LU decomposition
- Cholesky's method

- Matrix inverse approach $\longrightarrow \{A\} = \left[[Z]^T [Z] \right]^{-1} [Z]^T \{Y\}$



Statistics and Probability

Regression