Part 5. Curve Fitting
Chapter 17. Least-Squares Regression

Lecture 15

Linear Regression

17.1

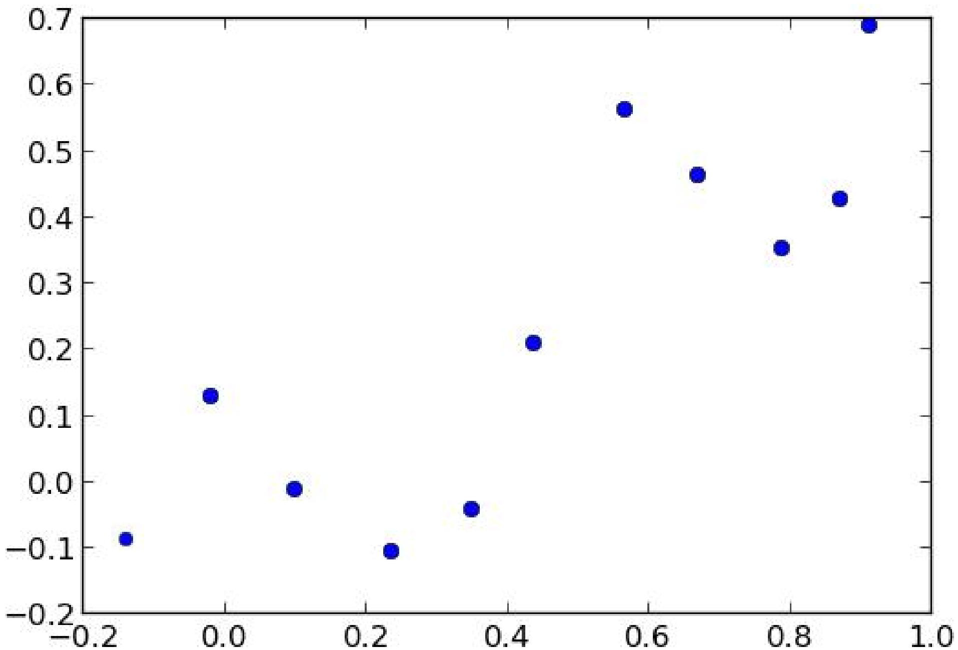
Homeyra Pourmohammadali

Curve Fitting- Motivation

- Data are often given for discrete values along continuum.

- Estimates of points between discrete values may be required.

- Curve fitting techniques can fit curves to discrete data to obtain required intermediate values.



Curve Fitting- Main Engineering Applications

1

Trend Analysis

- Predicting values of dependent variable: extrapolation beyond data points or interpolation between data points.

2

Hypothesis Testing

- Comparing existing mathematical model with measured data

Curve Fitting- Engineering Applications Examples

- Removing measurement noise
- Filling in missing data points (e.g. improper data record)
- Find trajectory of an object (***s***) based on discrete velocity values (***v*** is derivative of ***s*** and ***a*** is the second derivative of ***s***)
- Integrating digital data (e.g. find area under curve with discrete points)
- Differentiating digital data (e.g. modeling the discrete data with a polynomial and differentiating polynomial)

Curve Fitting- General Approaches

Two general approaches:

Data exhibit a significant degree of scatter

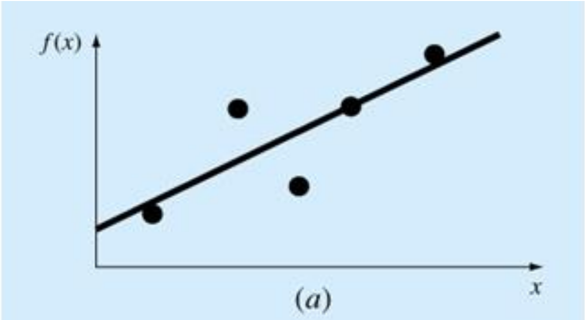
The strategy is to derive a single curve that represents the general trend of the data.

Data is very precise

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

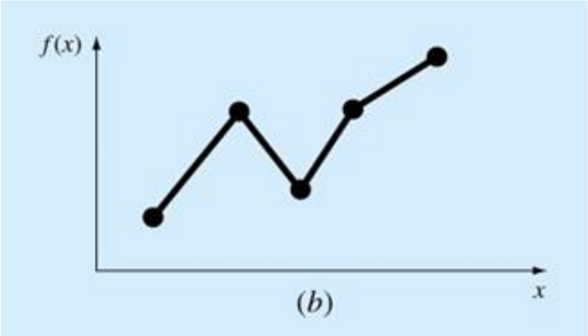
Curve Fitting-Non-Computer Methods

a) Sketch one straight-line that visually conforms to all data



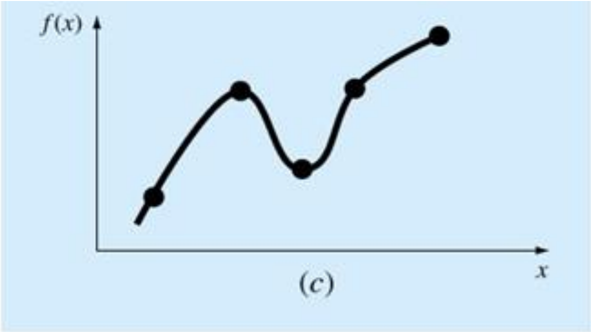
(a)

b) Using straight-line segments to connect the points



(b)

c) Using curves to represent data



(c)

Example 1. Curve fitting. A study investigating household budgeting practices surveyed a random sample of 7 families in a small town, collecting data for the total food expenditures last month vs. gross monthly income:

Income ($100)	55	83	38	61	33	49	67
Food ($100)	14	24	13	16	9	15	17

Least Squares Regression: Linear Regression

Linear Regression

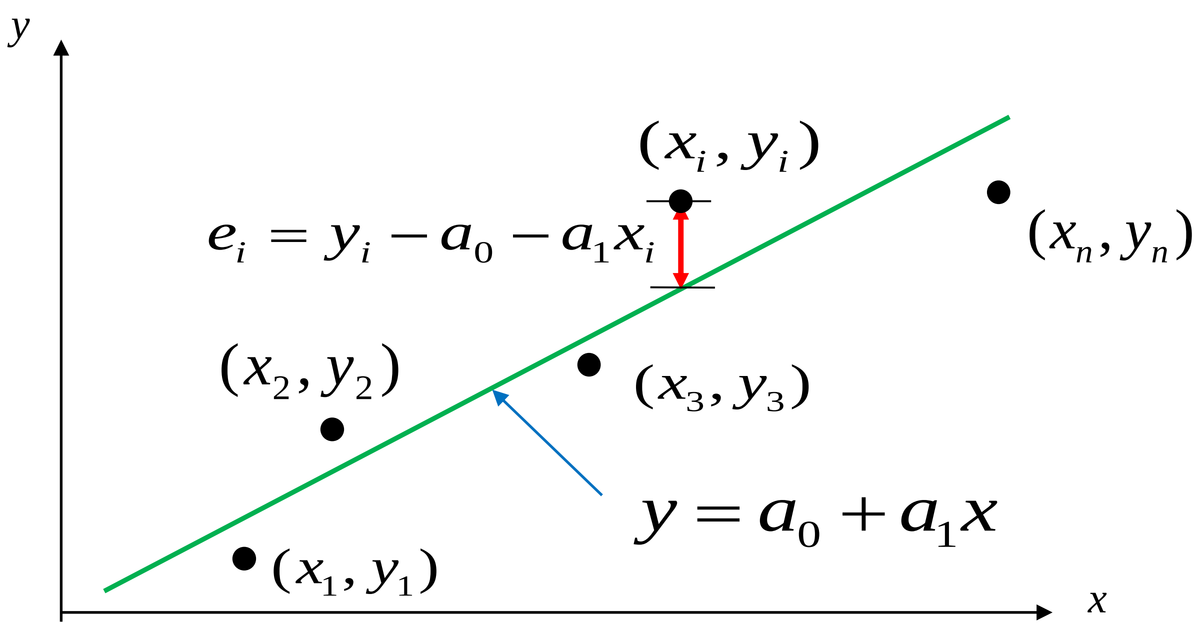
- Fitting a straight line to a set of paired observations:

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

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

a_1 : slope, a_0 : intercept,

e : error, or residual, between model and observations



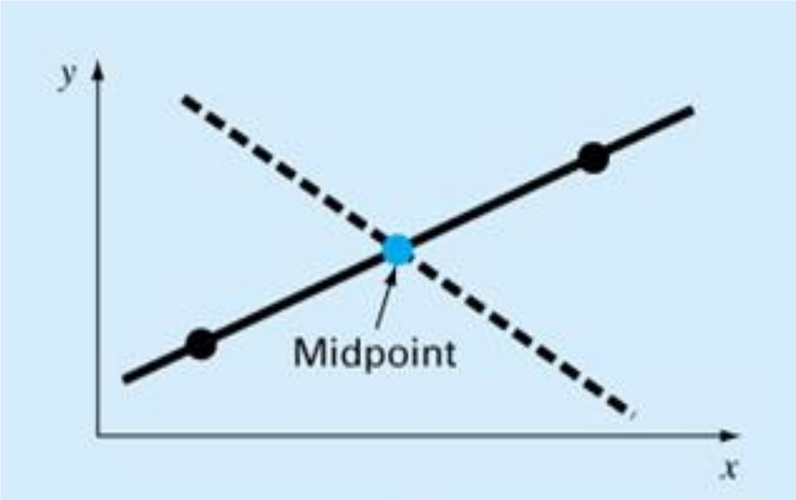
Linear regression of y vs x data showing residuals at a typical point, x_i.

Criteria for a "Best" Fit

Criterion 1. Minimize the sum of the residual errors for all available data (where *n* is total number of points):

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

- Is this an adequate criterion? does it yield a unique best fit?

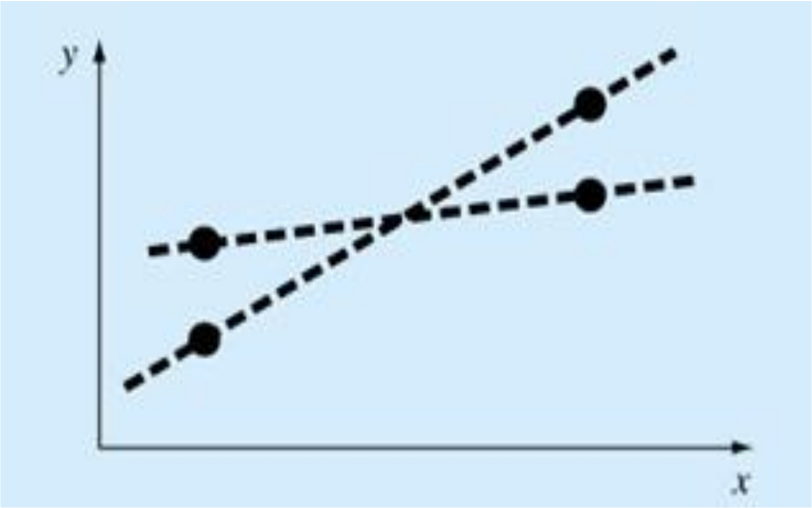


Criteria for a "Best" Fit

Criterion 2. Minimize the sum of the absolute values

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

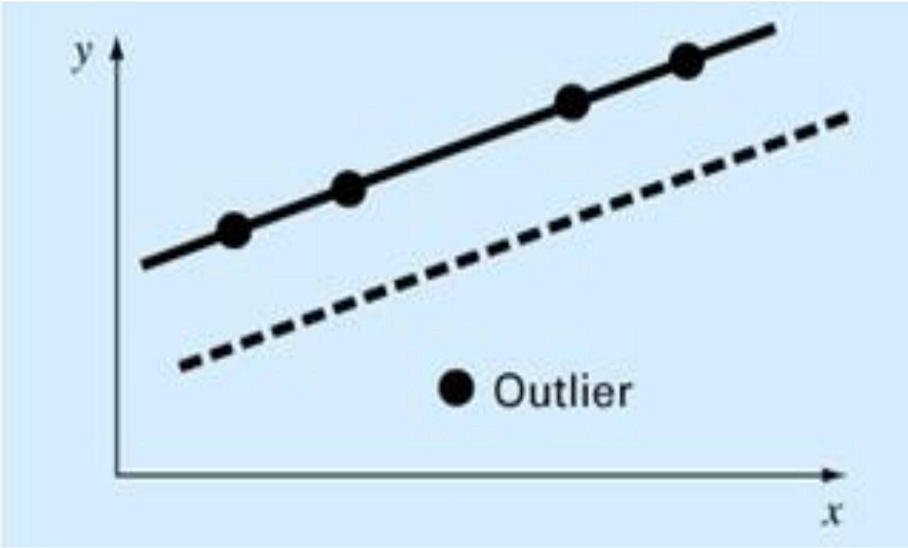
- Is this an adequate criterion? does it yield a unique best fit?



Criteria for a “Best” Fit

Criterion 3. (called Minimax Criterion) Minimize the maximum distance that an individual point falls from the line

- Is this an adequate criterion? does it yield a unique best fit?

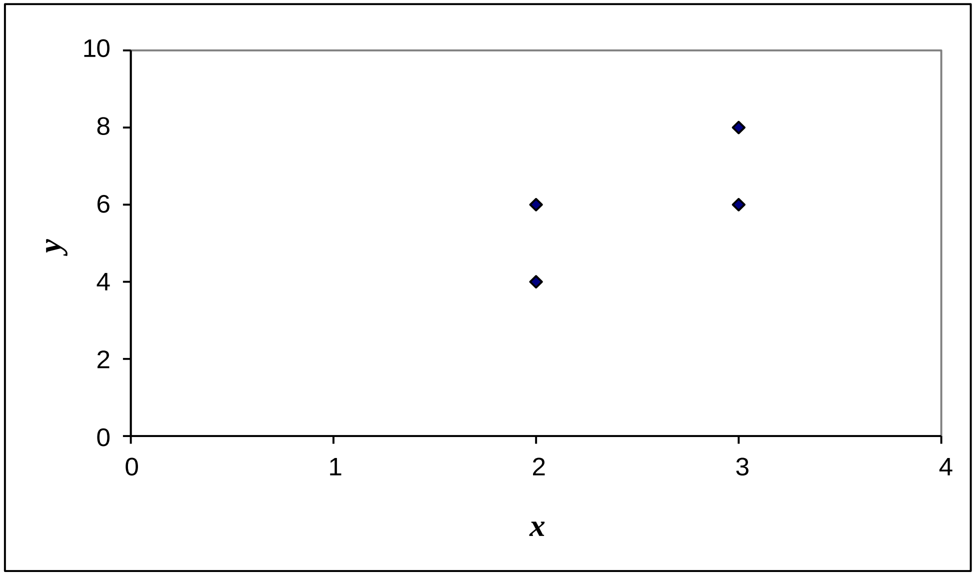


Example 2. Appropriate Criterion. Given the data points (2,4), (3,6), (2,6) and (3,8), best fit the data to a straight line. Use Criterion#1 and 2.

Minimize $\sum_{i=1}^{n} e_i$ or $\sum_{i=1}^{n} |e_i|$

Data Points

x	y
2.0	4.0
3.0	6.0
2.0	6.0
3.0	8.0



Data points for y vs x data.

Criteria for a "Best" Fit

Criterion 4: Minimize the sum of the squares of the residuals between the measured y and the y calculated with the linear model:

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

- Is this an adequate criterion?

Yields a unique line for a given set of data.

Criteria for a "Best" Fit

Criterion 4: Need to find a_0 and a_1 coefficients in such a way that minimize S_r.

Differentiate with respect to these coefficients

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

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

Least-Squares Fit of a Straight Line

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

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

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

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

Normal equations, can be solved simultaneously

$$\sum a_0 = n a_0$$

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

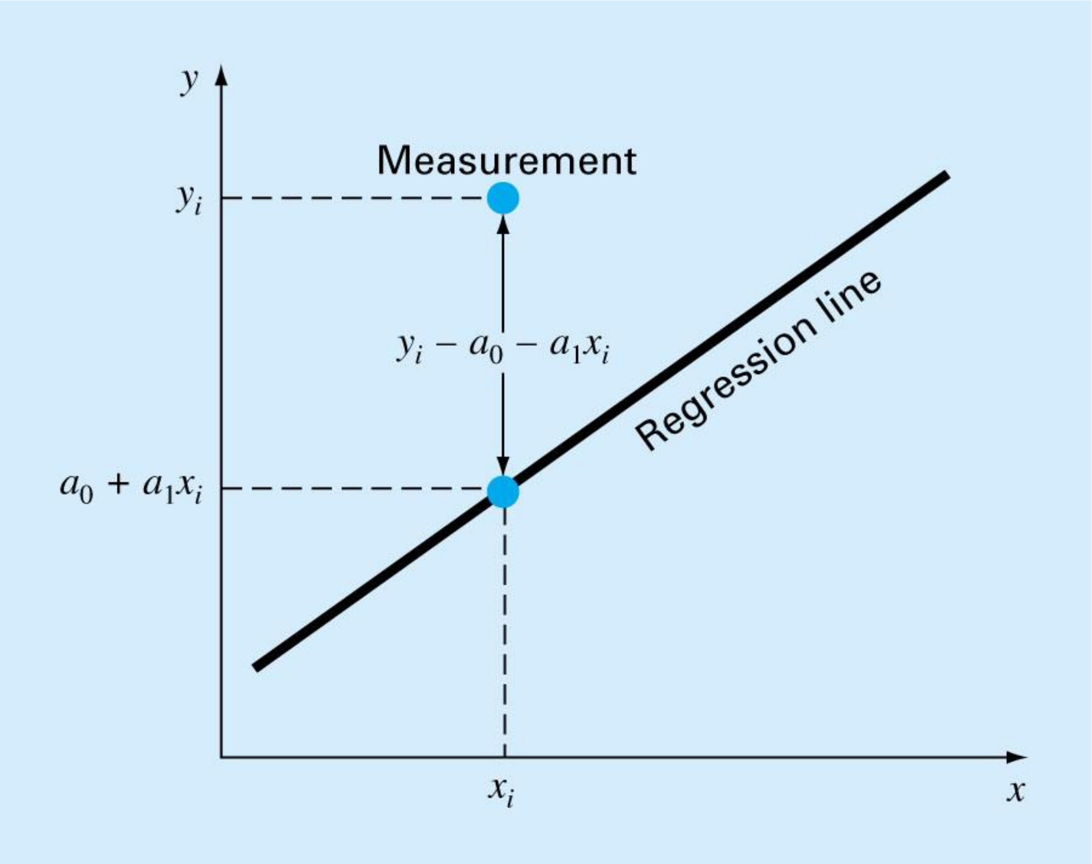
$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2}$$

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

Example 3. Linear Regression. A study wishes to develop an empirical model for the number of calories per single serving of breakfast cereal as a function of the amount of sugar. Thirteen different samples are measured as follows. Find the coefficients of regression line: a_0 and a_1

Sugar (g)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	190	100	120	190	190	110	120

Error of Linear Regression



Residual in linear regression: vertical distance between a data point and the line

“Goodness” of Our Fit

If total sum of the squares around the mean for the dependent variable, y, is S_t

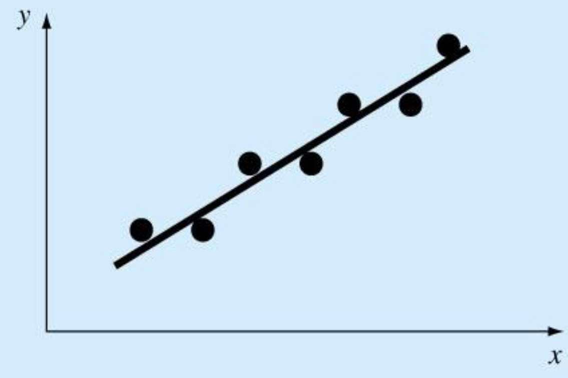
If sum of the squares of residuals around the regression line is S_r

If 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:

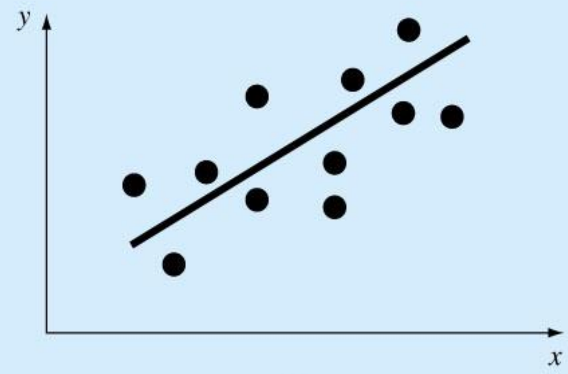
$$r^2 = \frac{S_t - S_r}{S_t}$$

r^2 - coefficient of determination

r – correlation coefficient



(a)



(b)

Error in Linear Regression

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

r : correlation efficient

$$0 < r < 1$$

Poor fit (no fit) Perfect fit of linear data

Special Cases

- For a perfect fit

$$S_r = 0 \quad \& \quad r = r^2 = 1$$

signifying that the line explains 100% of the variability of data.

- For:

$$r = r^2 = 0 \qquad \& \qquad S_r = S_t$$

the fit represents no improvement.

Example 4. Error of Linear Regression. A study wishes to develop an empirical model for the number of calories per single serving of breakfast cereal as a function of the amount of sugar. Thirteen different samples are measured as follows. Find the correlation coefficient related directly to residual error.

Sugar (g)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	190	100	120	190	190	110	120

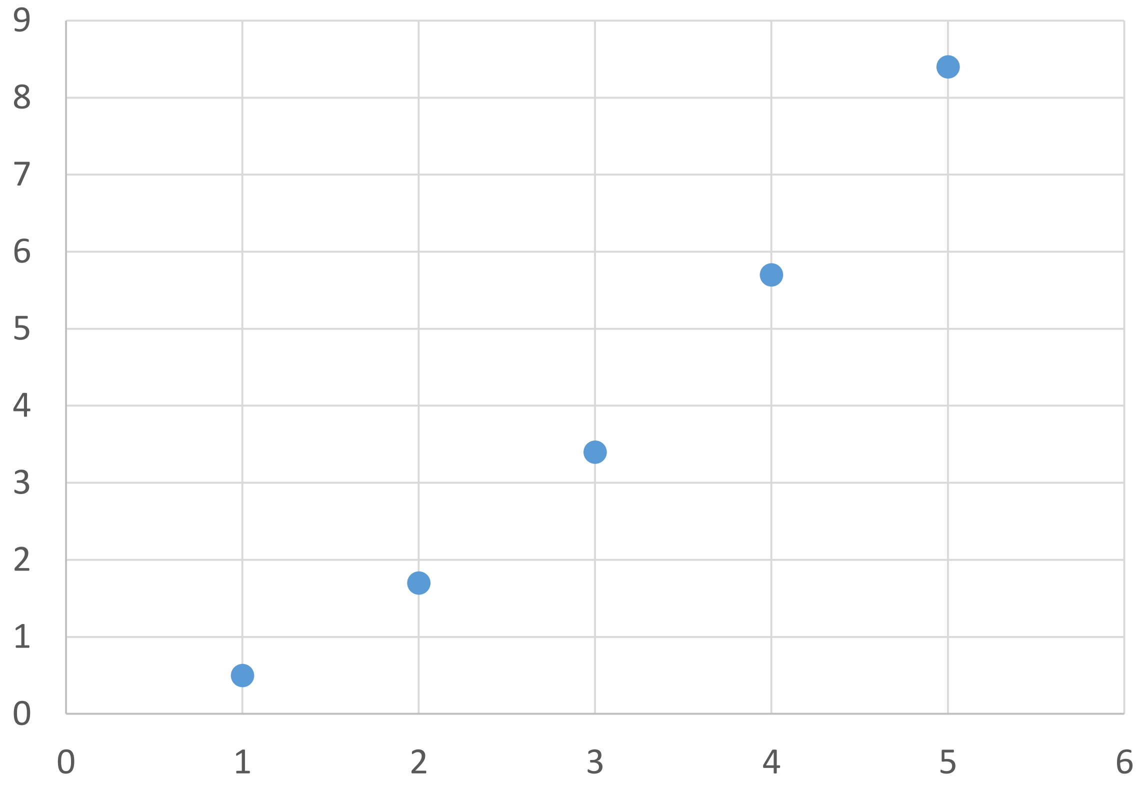
What about <u>non-linear</u> relationships?

Example 5. Non-linear relationship $y = a\, e^{bx}$

Example 5 continued. Non-linear relationship.

Example 5 continued. Non-linear relationship.

x	y
1	0.5
2	1.7
3	3.4
4	5.7
5	8.4



Example 5 continued. Non-linear relationship.

x	**y**	**log(x)**	**log(y)**
1	0.5	0	-0.301
2	1.7	0.301	0.230
3	3.4	0.477	0.531
4	5.7	0.602	0.756
5	8.4	0.699	0.924

Recall
Mathematics & Statistic
Self-Study

Recall: Mathematics- Mean & StDev

Arithmetic Mean. The sum of the individual data points (y_i) divided by the number of points (n).

$$\bar{y} = \frac{\sum y_i}{n}$$

$$i = 1, \ldots, n$$

Standard Deviation (StDev). The most common measure of a spread for a sample.

$$S_y = \sqrt{\frac{S_t}{n-1}}$$

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

or

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

Recall: Mathematics-Variance & c.v.

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

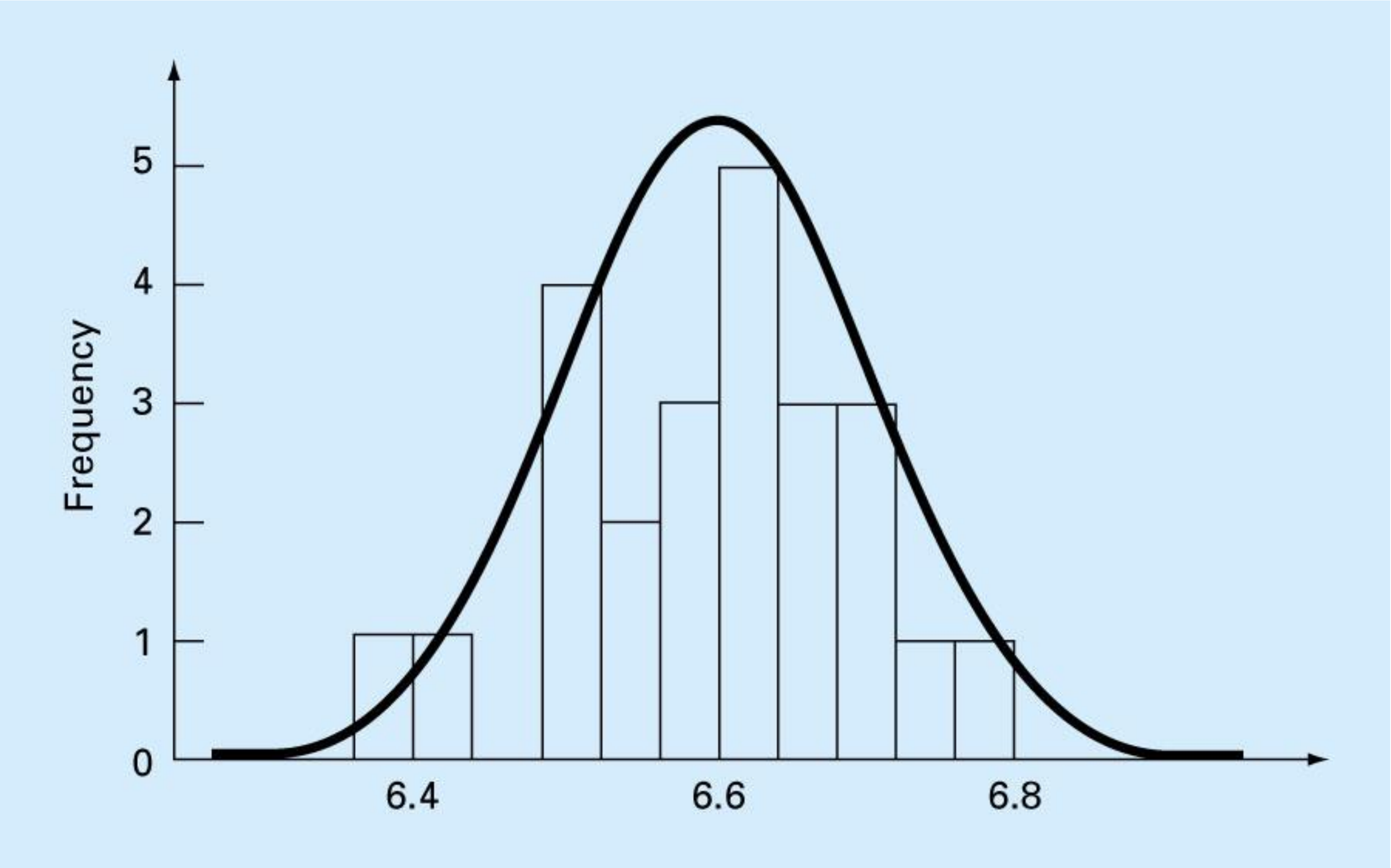
$$S_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

Degrees of freedom

Coefficient of Variation. Has the utility to quantify the spread of data.

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

Recall: Mathematics-Normal Distribution



Recall: Mathematics-Confidence Interval