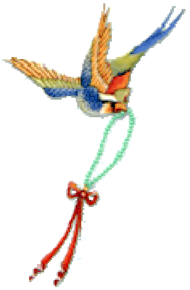


计算机数学建模

# 第七讲 统计回归模型(1)


周毓明

南京大学计算机科学与技术系

## 数学建模的基本方法

机理分析　　测试分析

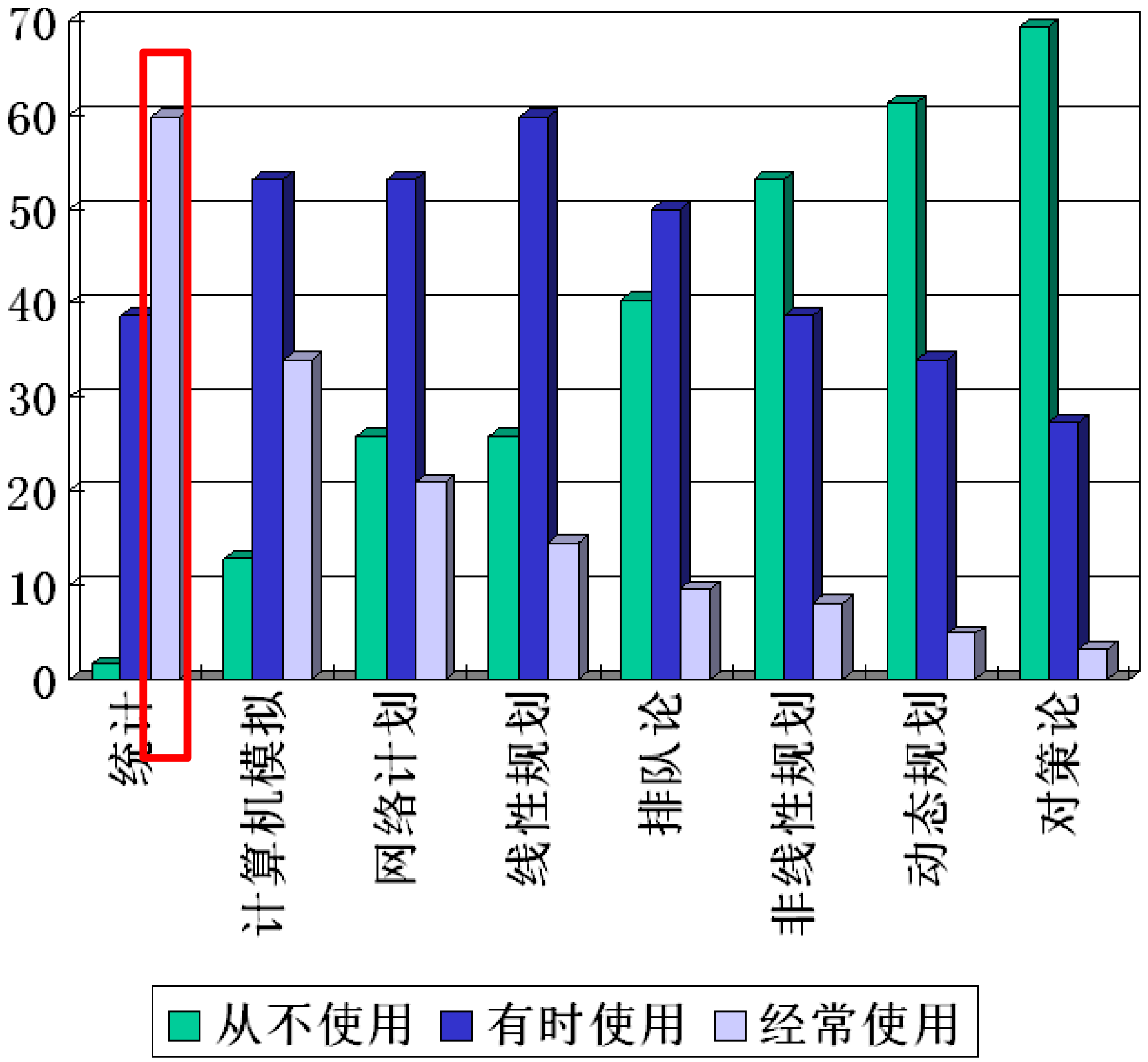
由于客观事物内部规律的复杂及人们认识程度的限制，无法分析实际对象内在的因果关系，建立合乎机理规律的数学模型。

通过对数据的统计分析，找出与数据拟合最好的模型

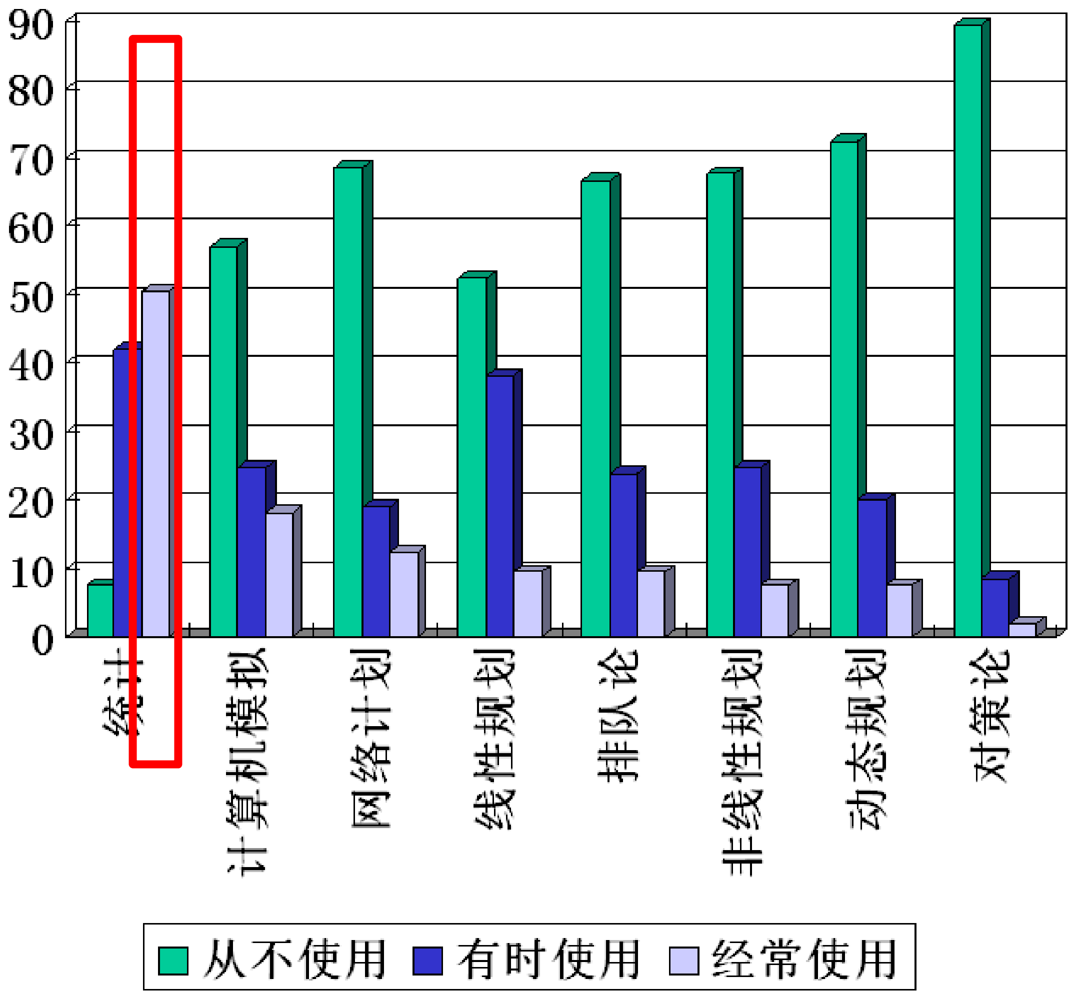
回归模型是用统计分析方法建立的最常用的一类模型

- 只简单涉及回归分析的数学原理和方法
- 通过实例讨论如何选择不同类型的模型
- 对软件得到的结果进行分析，对模型进行改进

# 运筹学方法使用情况(美1983)(%)



# 运筹学方法在中国使用情况(随机抽样)(%)



# 课程内容

1. 数学概念与模型
2. 实际案例与分析
3. 计算机典型应用

# 1. 数学概念与模型

① 描述性统计

② 线性回归

③ Logistic回归

# 描述性统计: Location

## 1. Mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

## 2. Median

$$Md = \begin{cases} X_{(\frac{n+1}{2})} & , n \in odd \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}] & , n \in even \end{cases}$$

## 3. Mode

Example:

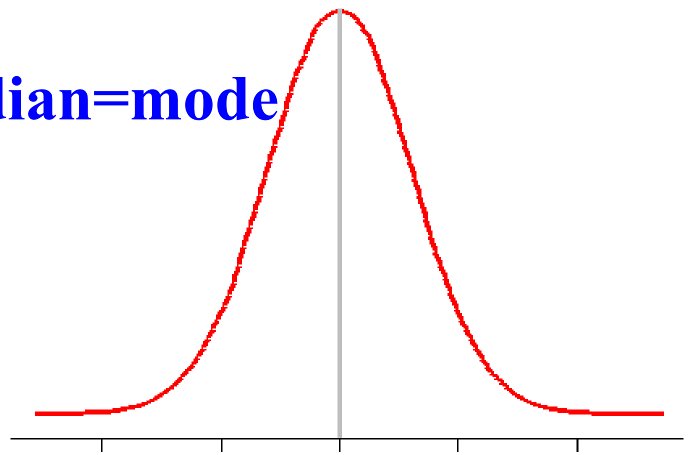
Observations : ( 1, 11, 10, 2, 7, 5 )

Mean : (1+11+10+2+7+5)/6 = 6

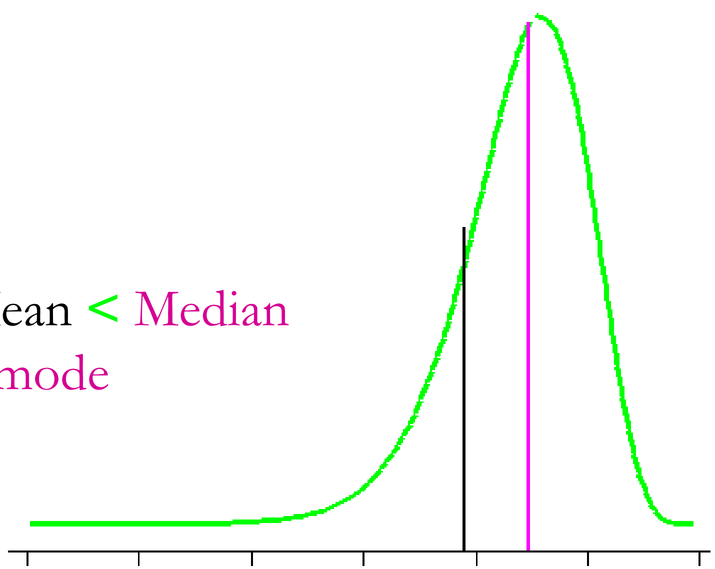
Median : $(X_{(3)}+X_{(4)})/2 = (5+7)/2 = 6$

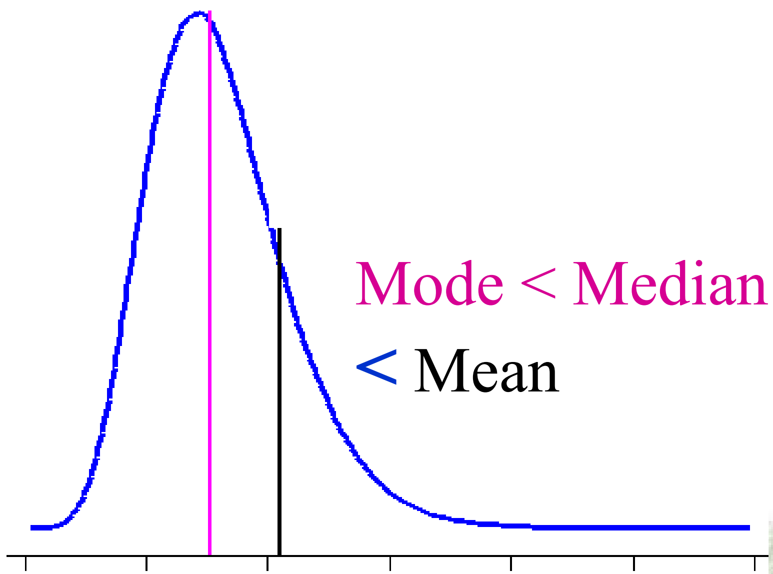
# 描述性统计: Mean v.s. Median

**Mean = Median=mode**



Mean < Median
<mode

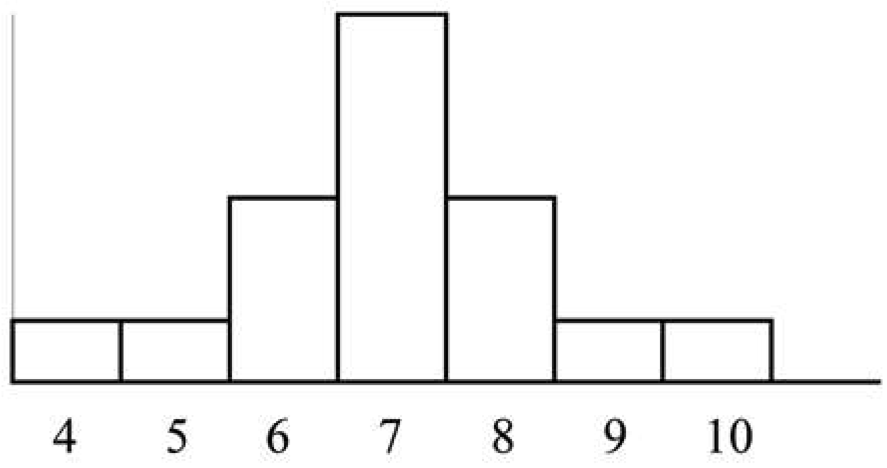




Mode < Median
< Mean

# 描述性统计: Mean v.s. Median

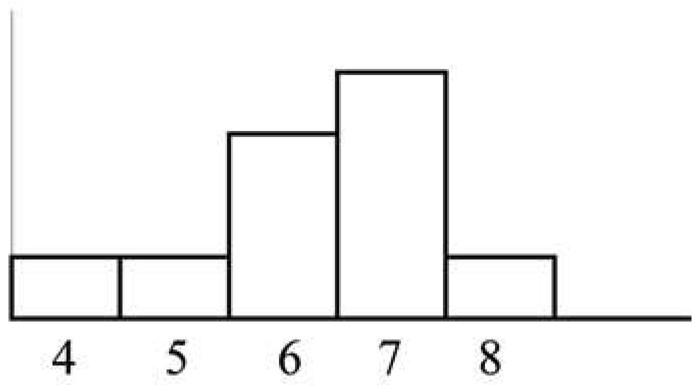
4 ; 5 ; 6 ; 6 ; 6 ; 7 ; 7 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 9 ; 10




Mean = Median = Mode =7

# 描述性统计: Mean v.s. Median

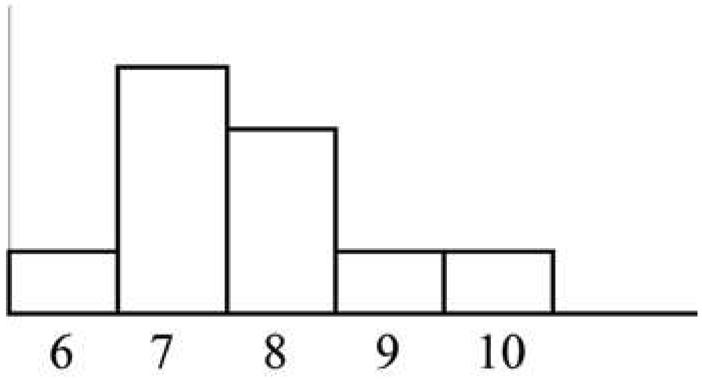
4 ; 5 ; 6 ; 6 ; 6 ; 7 ; 7 ; 7 ; 7 ; 8




Mean = 6.3
Median = 6.5
Mode = 7

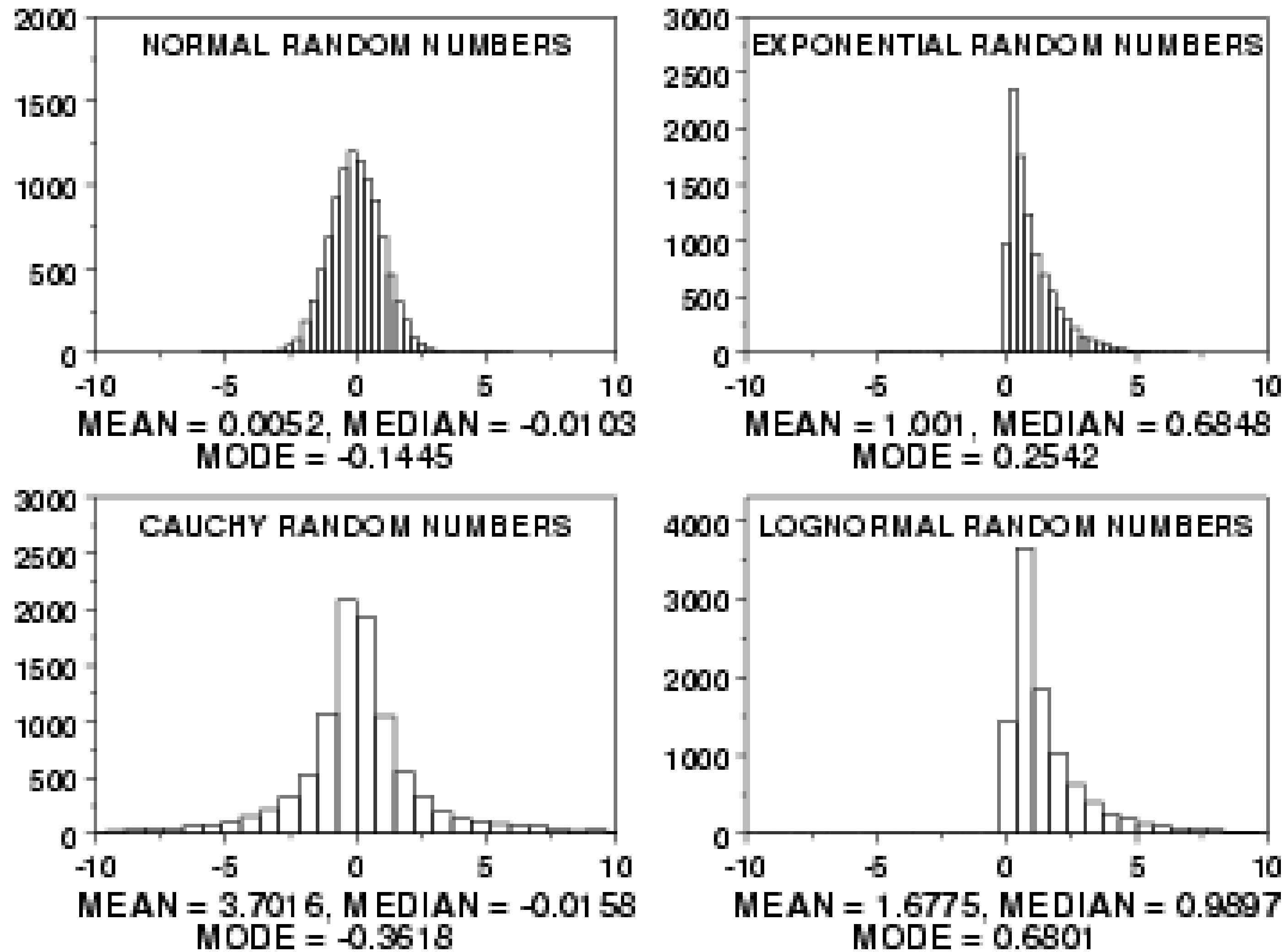
# 描述性统计: Mean v.s. Median

6 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 9 ; 10




Mean = 7.7
Median = 7.5
Mode = 7

# 描述性统计: Mean v.s. Median



# 描述性统计: Location (cont'd)

## 4. k-th Percentile

$$P_k = \begin{cases} X_{([|i|]+1)} & , i \notin Z \\ \frac{1}{2}[X_{(i)} + X_{(i+1)}] & , i \in Z \end{cases} \quad \text{where } i = \frac{k}{100} n$$

Example:

Observations : ( 1, 11, 10, 2, 7, 5 )

Order statistics : ( 1, 2, 5, 7, 10, 11 )

$$P_{25} = X_{(1+1)} = X_{(2)} = 2 , \qquad i = \frac{25}{100} 6 = 1.5$$

$$P_{50} = \frac{1}{2}[X_{(3)} + X_{(4)}] = \frac{1}{2}[5 + 7] = 6 , \qquad i = \frac{50}{100} 6 = 3$$

$$P_{75} = X_{(4+1)} = X_{(5)} = 10 , \qquad i = \frac{75}{100} 6 = 4.5$$

# 描述性统计: Location (cont'd)

Remarks:

$1^0$. $P_{50}$=Md (median)

$2^0$. Quartile: 3 cut points

$Q_1 = P_{25}$ ($25^{th}$-percentile),

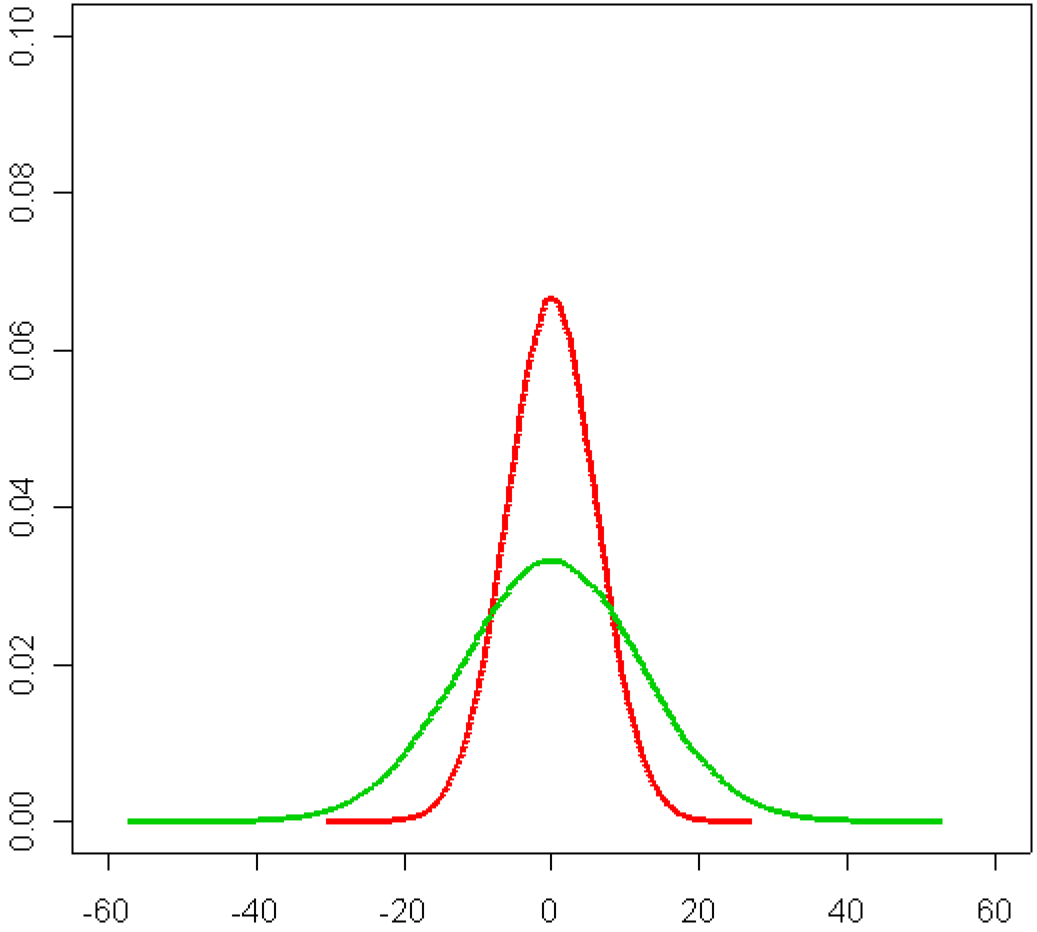
$Q_2 =$ Md (Median) $= P_{50}$ ($50^{th}$-percentile) ,

$Q_3 = P_{75}$ ($75^{th}$-percentile)

# 描述性统计: Location (cont'd)

| Order | Value | Boundary |
|---|---|---|
| 1 | 27.75 | |
| 2 | 37.35 | |
| 3 | 38.35 | |
| 4 | 38.35 | |
| 5 | 38.75 | |
| Second Quartile | | 39.250 |
| 6 | 39.75 | |
| 7 | 40.50 | |
| 8 | 41.00 | |
| 9 | 41.15 | |
| 10 | 42.55 | |
| Third Quartile | | 42.725 |
| 11 | 42.90 | |
| 12 | 43.60 | |
| 13 | 43.85 | |
| 14 | 47.30 | |
| 15 | 47.90 | |
| Fourth Quartile | | 48.025 |
| 16 | 48.15 | |
| 17 | 49.86 | |
| 18 | 51.25 | |
| 19 | 51.50 | |
| 20 | 56.00 | |
| Data Table divided into quartiles | | |

# 描述性统计: Dispersion



# 描述性统计: Dispersion (cont'd)

**1. Range**: $R = X_{(n)} - X_{(1)}$

**2. Interquartile-range**:

$IQR = Q_3 - Q_1 = P_{75} - P_{25}$

**3. Quartile deviation**: $Q.D. = IQR/2$

Example:

Observations : ( 1, 11, 10, 2, 7, 5 )

Order statistics : ( 1, 2, 5, 7, 10, 11 )

$R = X_{(6)} - X_{(1)} = 11 - 1 = 10$

$IQR = Q_3 - Q_1 = 10 - 2 = 8$

$Q.D. = IQR/2 = 8/2 = 4$

# 描述性统计: Dispersion (cont'd)

## 4. Mean Absolute Deviation

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right| \quad (统计量), \qquad MAD = \frac{1}{N}\sum_{i=1}^{N}|X_i - \mu| \quad (参数)$$

## 5. Variance

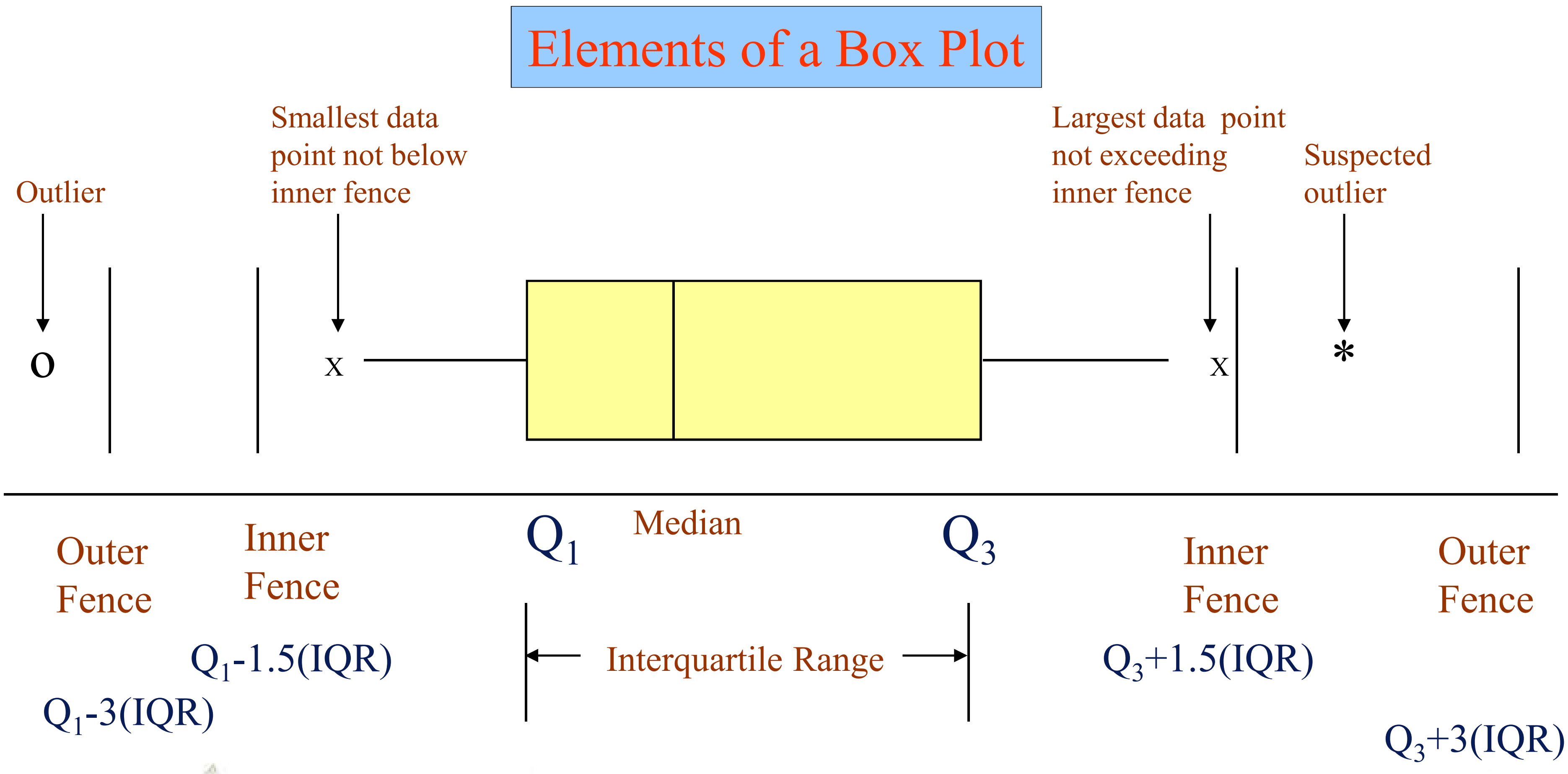
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad (统计量), \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2 \quad (参数)$$

## 6. Standard Deviation

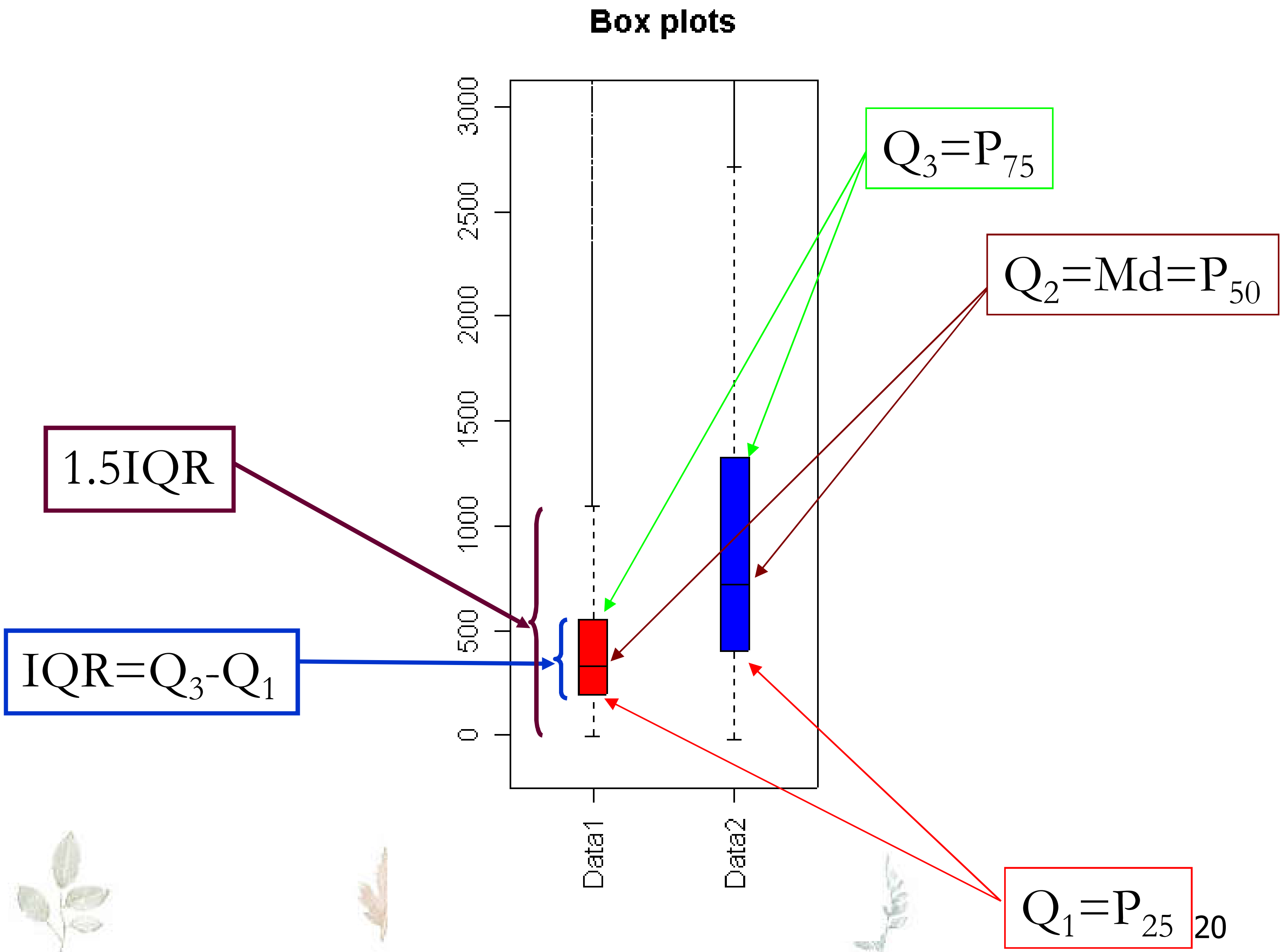
$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right]} \quad (统计量)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \mu^2} \quad (参数)$$

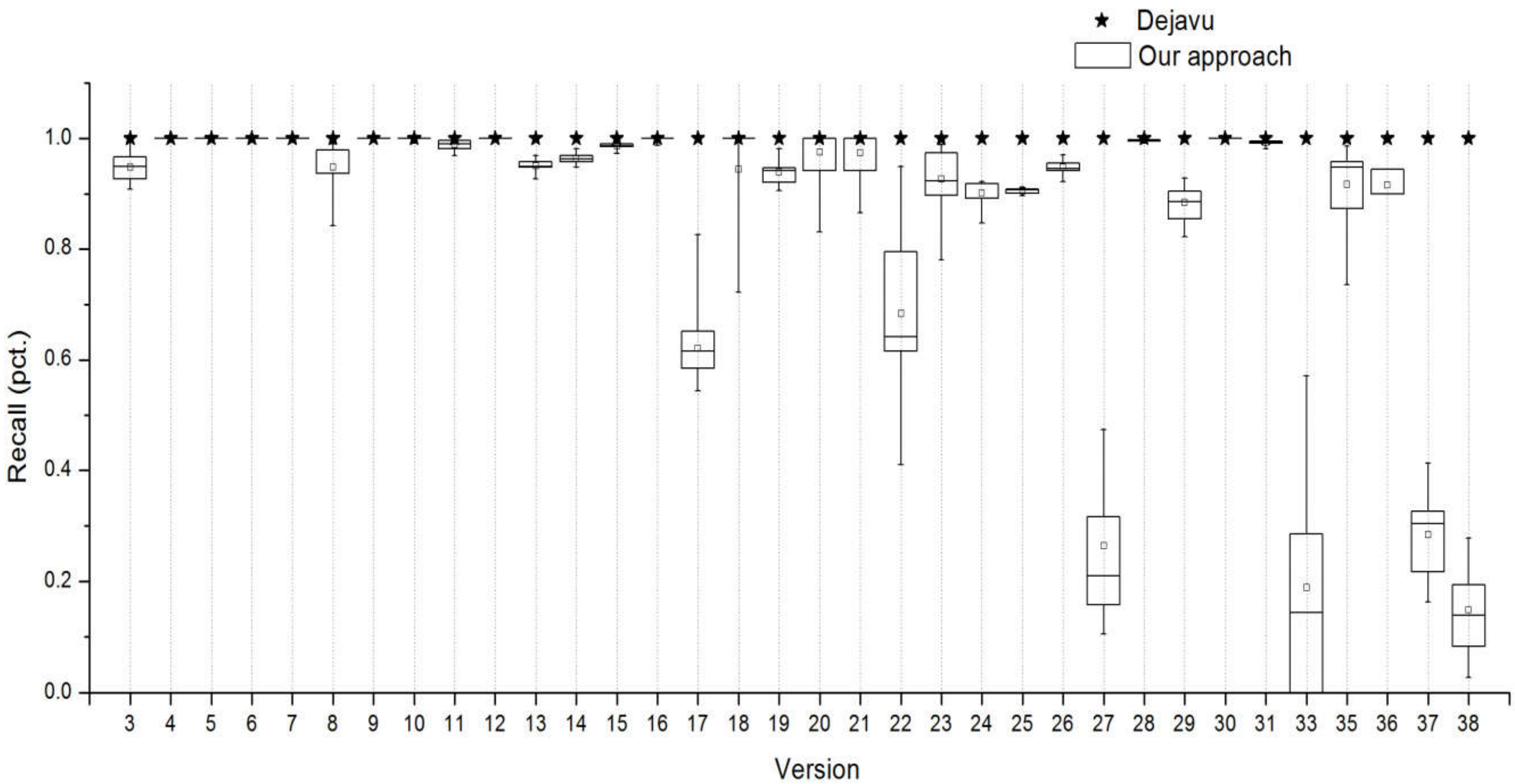
# 描述性统计: Box plot



# 描述性统计: Box plot



# 描述性统计: Box plot




**Comparison of recall between our approach when FR=0.3 and Dejavu**

Y. Duan, et al. Improving Cluster Selection Techniques of Regression Testing by Slice Filtering. SEKE 2010.

# 描述性统计: Skewness and Kurtosis

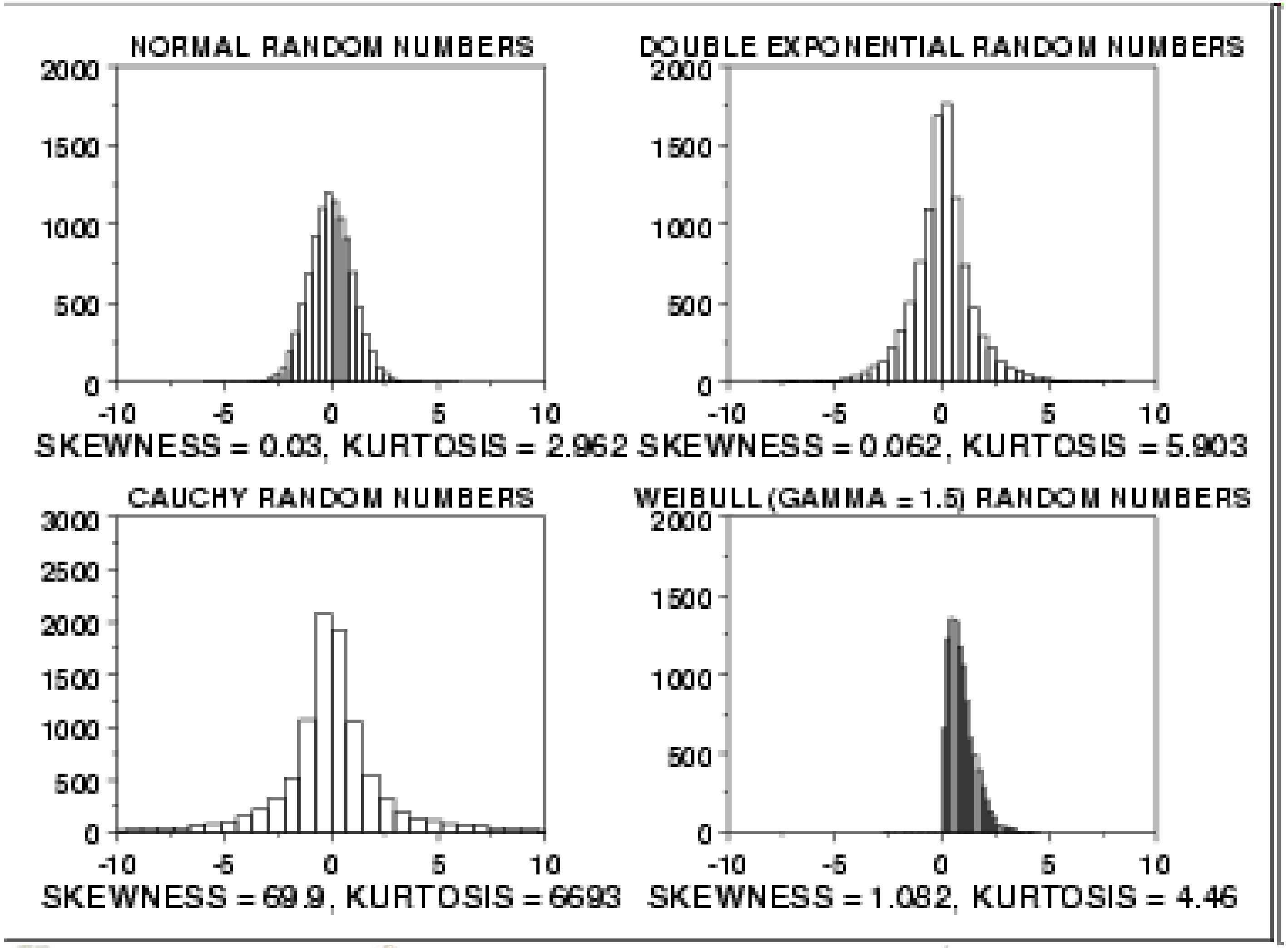
1. **Skewness**: a measure of symmetry, or more precisely, the lack of symmetry

$$skewness = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^3}{(n-1)S^3}$$

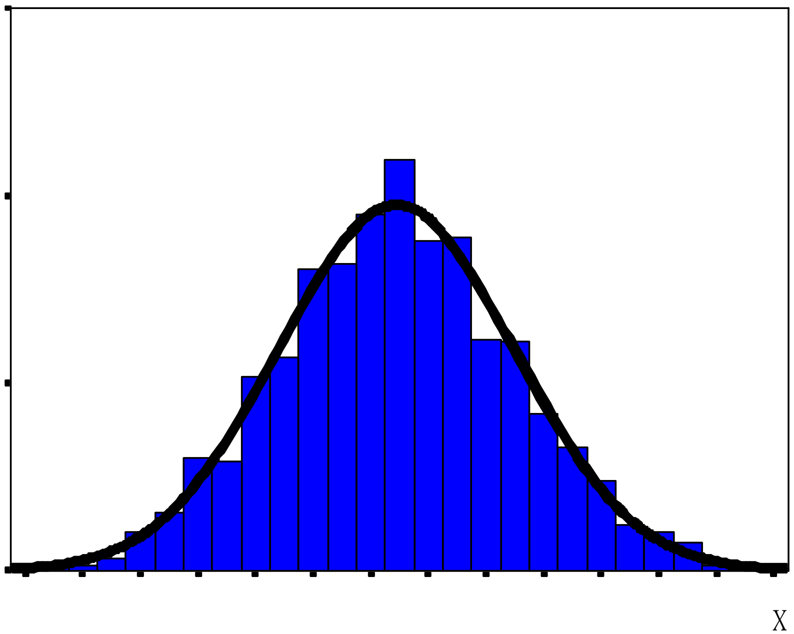
2. **Kurtosis**: a measure of whether the data are peaked or flat relative to a normal distribution

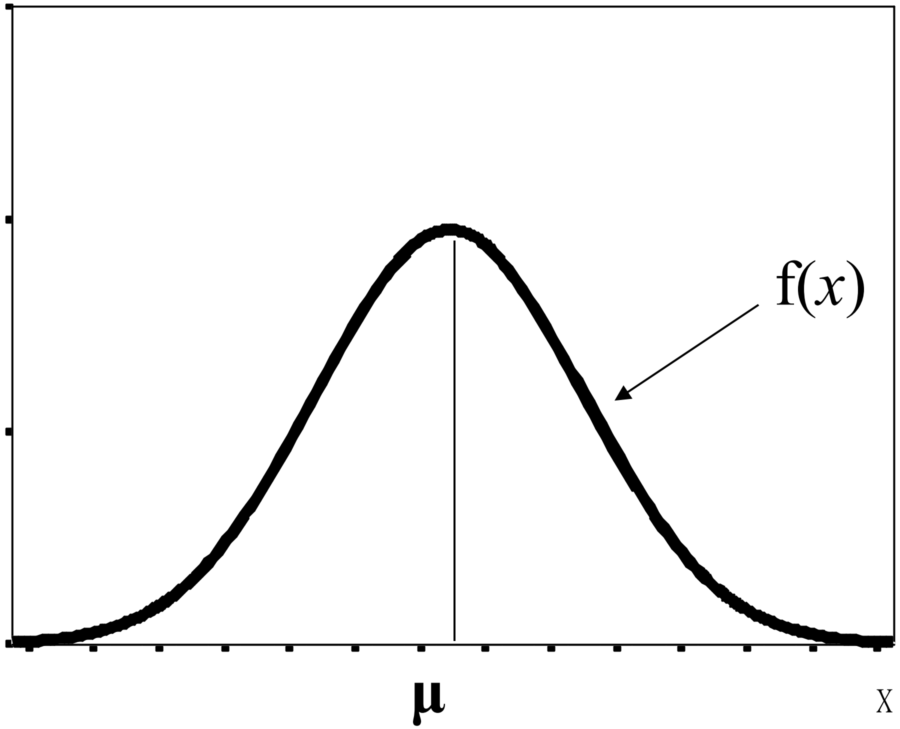
$$kurtosis = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{(n-1)S^4} - 3$$

# 描述性统计: Skewness and Kurtosis



# 描述性统计: Normal distribution



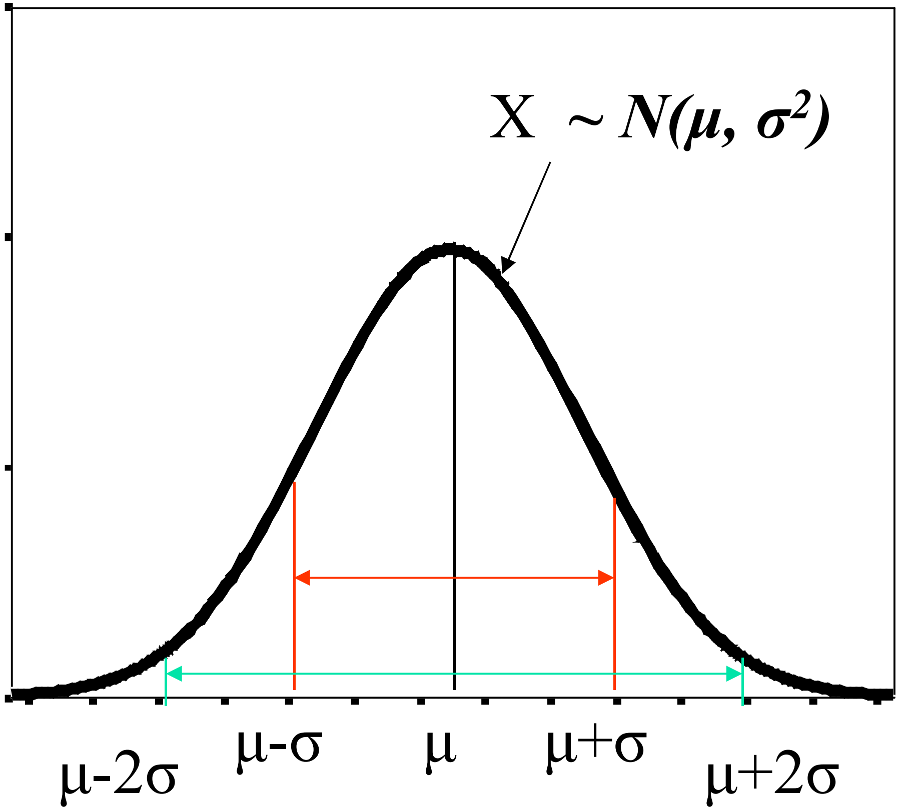



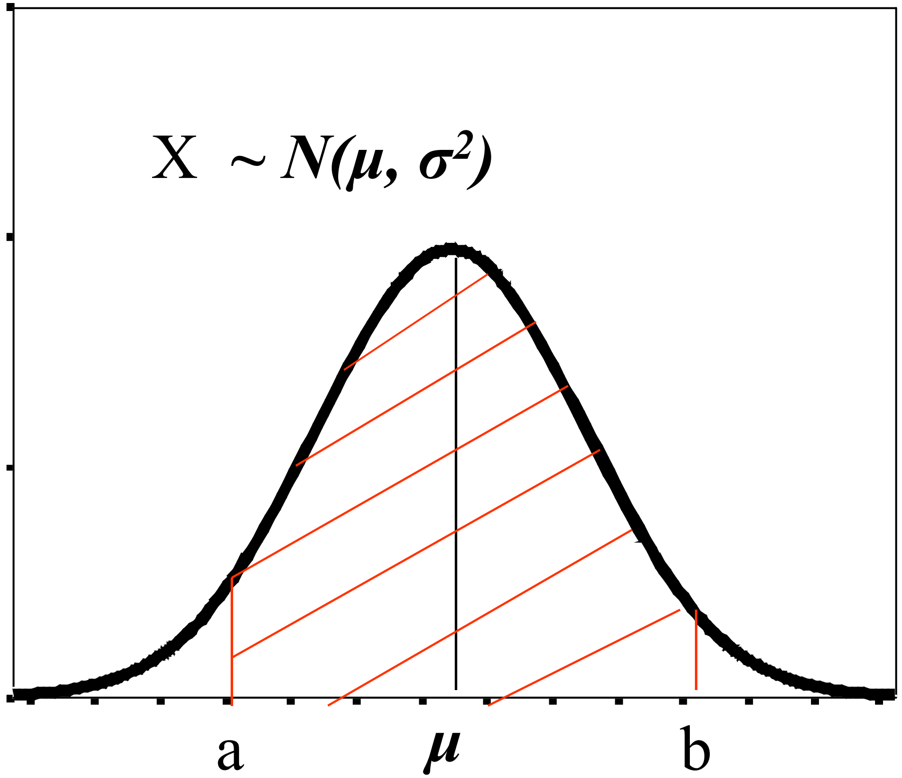

r.v. X $\sim \boldsymbol{N(\mu, \sigma^2)}$

the pdf for X is $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $, x \in R, -\infty < \mu < \infty, \sigma > 0$

$E(X) = \mu$ , $Var(X) = \sigma^2$

# 描述性统计: Normal distribution




$$P(a \le X \le b) = F(b) - F(a)$$

$$= \int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = ??$$

$$P(\mu - \sigma \le X \le \mu + \sigma) = 0.683$$

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.954$$

$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.997$$

# 描述性统计: Normal distribution

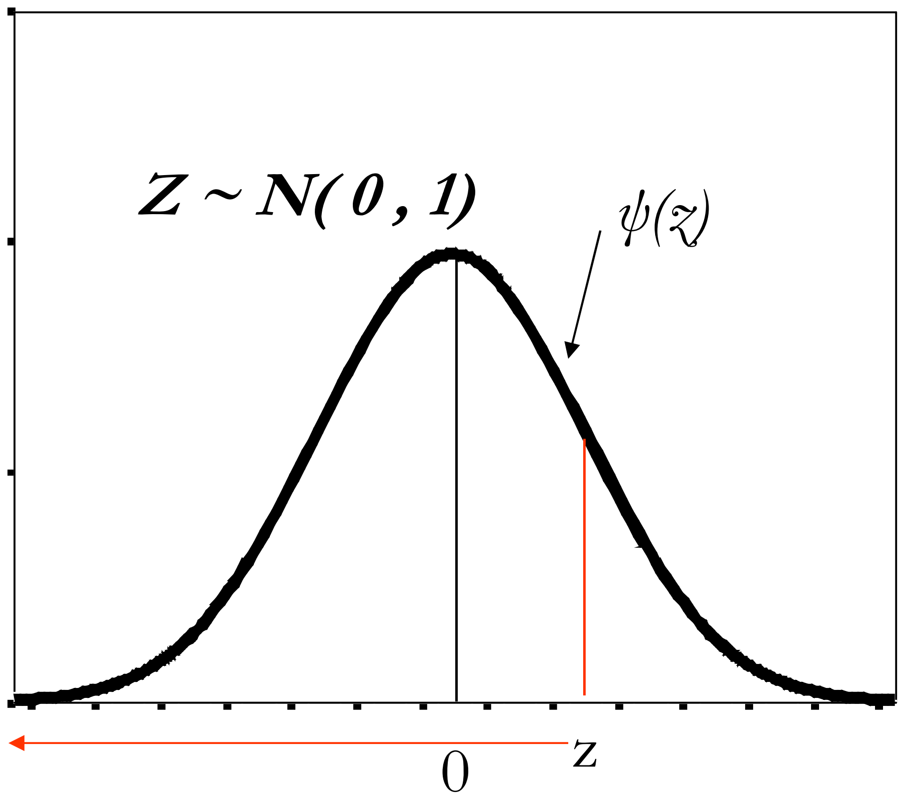
- X $\sim N(\mu, \sigma^2)$ standardized $Z \sim N(0, 1)$

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

- the pdf for Z is

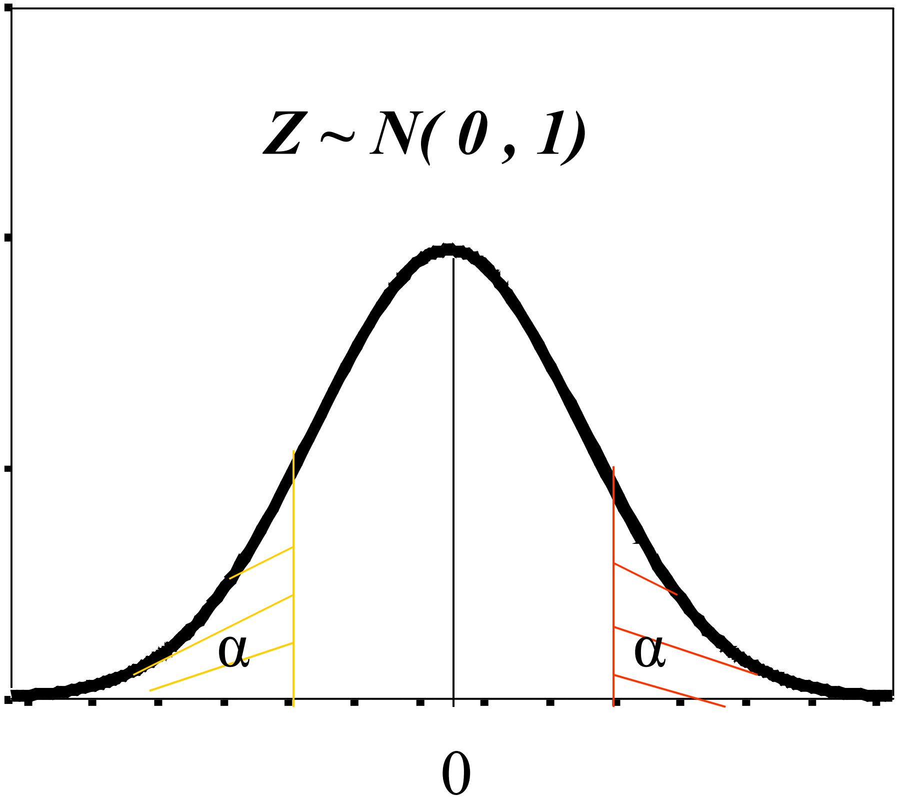
$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

$$-\infty < z < \infty$$




$$P(Z \le z) = \Phi(z) = \int_{-\infty}^{z} \phi(z)dz = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = ??$$ （查表）

# 描述性统计: Normal distribution




- $P(Z \geq z_\alpha) = P(Z \leq -z_\alpha) = \alpha$
- $\Phi(z_\alpha) = 1 - \Phi(-z_\alpha) \quad \Rightarrow \quad \Phi(z_\alpha) + \Phi(-z_\alpha) = 1$

例: $z_{0.025} = 1.96$ , $z_{0.05} = 1.645$

# Example

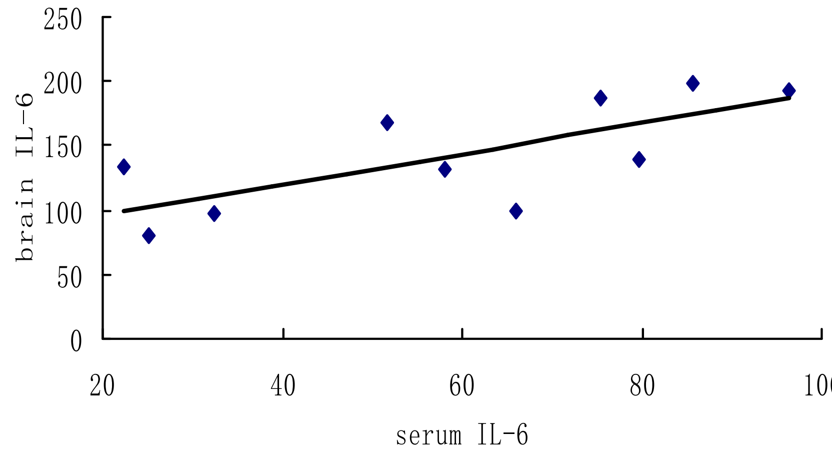
TABLE 1
Descriptive Statistics of the Classes

| Metric | N | Max. | 75% | Median | 25% | Min. | Mean | Std. dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|
| LCOM1 | 4830 | 171850 | 87 | 23 | 6 | 0 | 174.247 | 2615.348 | 59.225 | 3851.002 |
| LCOM2 | 4830 | 166390 | 60 | 13 | 1 | 0 | 151.022 | 2508.804 | 60.757 | 3997.301 |
| LCOM3 | 4830 | 492 | 7 | 4 | 2 | 1 | 6.250 | 11.641 | 19.300 | 674.415 |
| LCOM4 | 4830 | 282 | 4 | 2 | 1 | 1 | 3.300 | 6.796 | 24.413 | 825.727 |
| Co | 4830 | 1 | 0.333 | 0.089 | −0.017 | −2 | 0.026 | 0.609 | −1.195 | 2.946 |
| Co' | 4830 | 1 | 0.5 | 0.306 | 0.167 | 0 | 0.338 | 0.306 | 0.922 | −0.149 |
| LCOM5 | 3735 | 2 | 0.933 | 0.833 | 0.667 | 0 | 0.764 | 0.294 | −0.603 | 2.360 |
| Coh | 3735 | 1 | 0.458 | 0.267 | 0.150 | 0 | 0.338 | 0.249 | 1.085 | 0.600 |
| TCC | 4417 | 1 | 1 | 0.5 | 0.167 | 0 | 0.503 | 0.410 | 0.071 | −1.642 |
| LCC | 4417 | 1 | 1 | 0.672 | 0.2 | 0 | 0.562 | 0.425 | −0.195 | −1.688 |
| ICH | 4938 | 2976 | 17 | 4 | 0 | 0 | 16.115 | 73.573 | 22.495 | 722.809 |

Copied from: Yuming Zhou, et al. Examining the potentially confounding effect of class size on Associations between object-oriented metrics. IEEE Transactions on Software Engineering, 2009, 35(5): 607-623.

# 线性回归: 单变量回归

| Patient $i$ | Serum IL-6 (pg/ml) $x$ | Brain IL-6 (pg/ml) $y$ |
|---|---|---|
| 1 | 22.4 | 134.0 |
| 2 | 51.6 | 167.0 |
| 3 | 58.1 | 132.3 |
| 4 | 25.1 | 80.2 |
| 5 | 65.9 | 100.0 |
| 6 | 79.7 | 139.1 |
| 7 | 75.3 | 187.2 |
| 8 | 32.4 | 97.2 |
| 9 | 96.4 | 192.3 |
| 10 | 85.7 | 199.4 |



# 线性回归: 单变量回归

The population simple linear regression model:

$$y = \underbrace{\alpha + \beta x}_{\text{Nonrandom or Systematic Component}} + \underbrace{\varepsilon}_{\text{Random Component}} \qquad or \qquad m_{y|x} = \alpha + \beta x$$

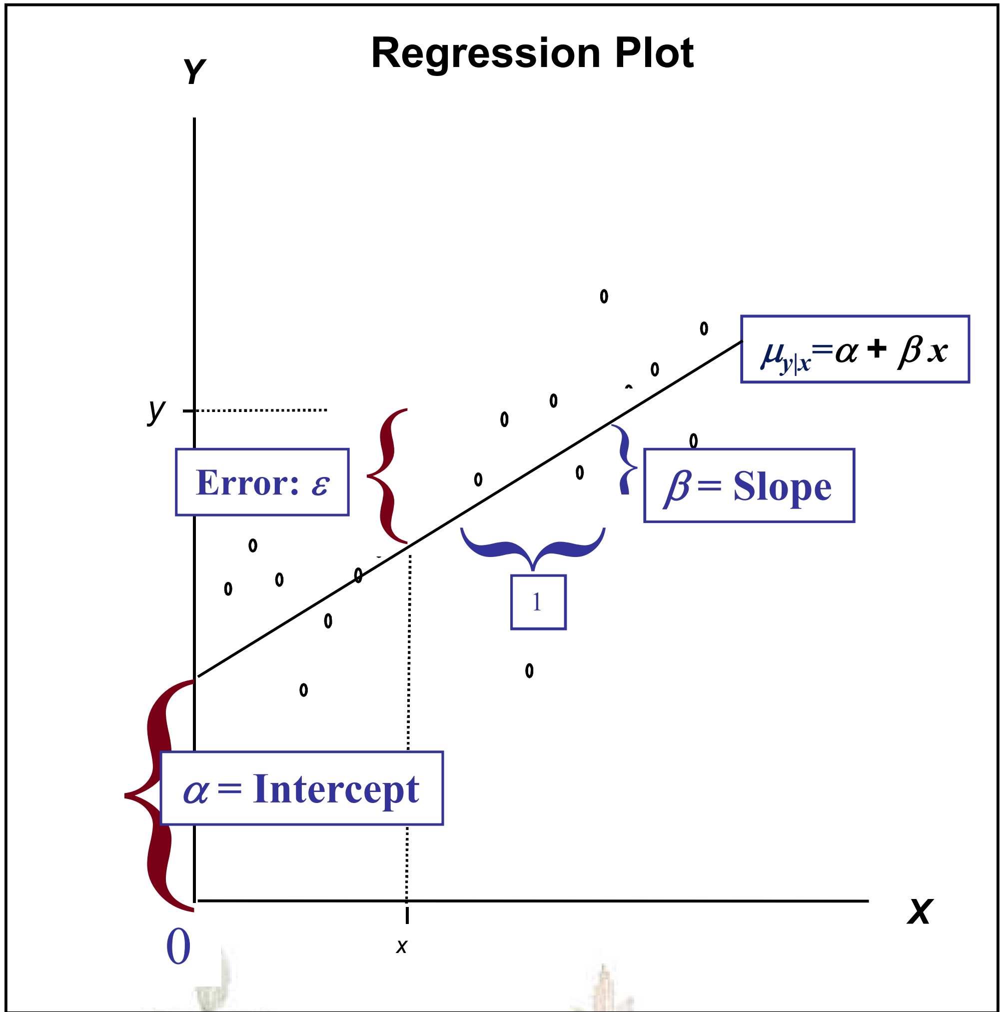
Where $y$ is the **dependent** (response) **variable**, the variable we wish to explain or predict; $x$ is the **independent** (explanatory) **variable**, also called the **predictor variable**; and $\varepsilon$ is the **error term**, the only random component in the model, and thus, the only source of randomness in $y$.

$m_{y/x}$ is the mean of y when $x$ is specified, all called the **conditional mean** of Y.

$\alpha$ is the **intercept** of the systematic component of the regression relationship.

$\beta$ is the **slope** of the systematic component.

# 线性回归: 单变量回归




The simple linear regression model posits an exact linear relationship between the **expected** or average value of Y, the dependent variable Y, and X, the independent or predictor variable:

$$m_{y/x} = \alpha + \beta x$$

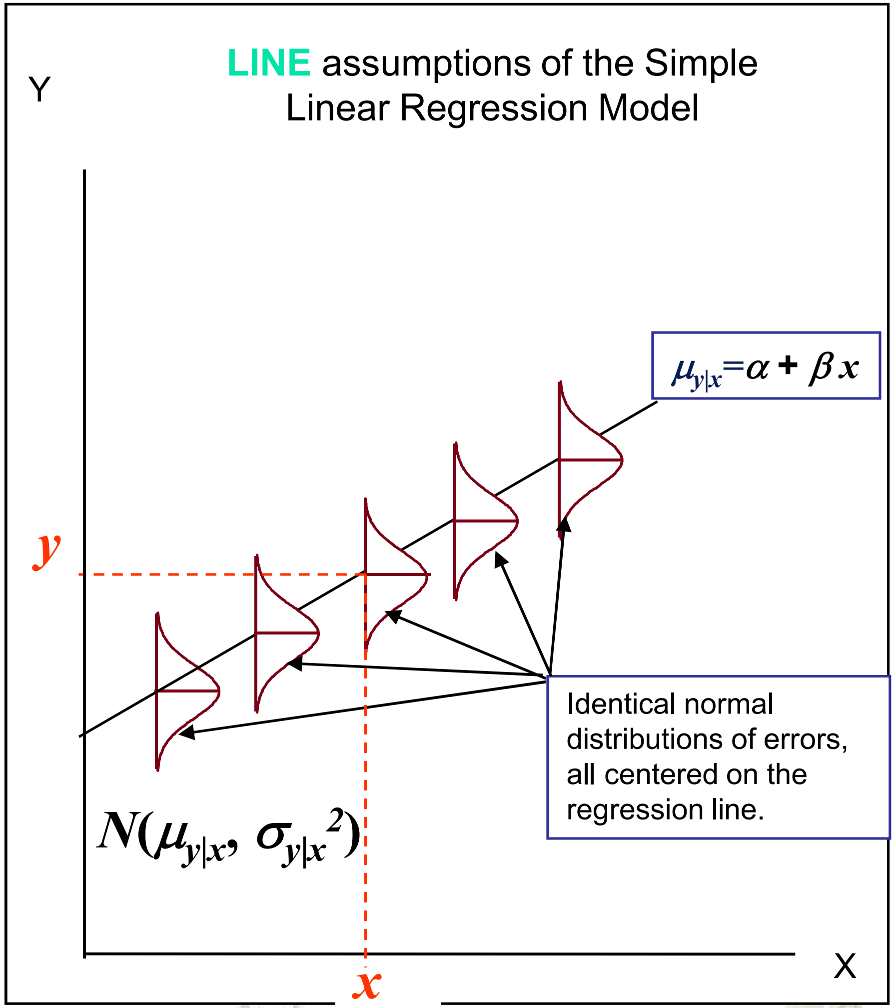
Actual observed values of Y ($y$) differ from the expected value ($m_{y|x}$) by an unexplained or random error($\varepsilon$):

$$y = m_{y|x} + \varepsilon$$
$$= \alpha + \beta x + \varepsilon$$

# 线性回归: 单变量回归

## Assumptions of the Simple Linear Regression Model

- The relationship between $X$ and $Y$ is a straight-Line (linear) relationship.
- The values of the independent variable $X$ are assumed fixed (not random); the only randomness in the values of $Y$ comes from the error term $\varepsilon_i$.
- The errors $\varepsilon_i$ are uncorrelated (i.e. **Independent**) in successive observations. The errors $\varepsilon_i$ are **Normally** distributed with mean 0 and variance $\sigma^2$(**Equal variance**). That is: $\varepsilon_i \sim N(0, \sigma^2)$



# 线性回归: 单变量回归

## Estimation: The Method of Least Squares

Estimation of a simple linear regression relationship involves finding estimated or predicted values of the intercept and slope of the linear regression line.

The **estimated regression equation:**

$$y = a + bx + e$$

where ***a*** estimates the ***intercept*** of the population regression line, $\alpha$;
***b*** estimates the ***slope*** of the population regression line, $\beta$;
and ***e*** stands for the observed errors ------- the residuals from fitting the estimated regression line $a + bx$ to a set of $n$ points.

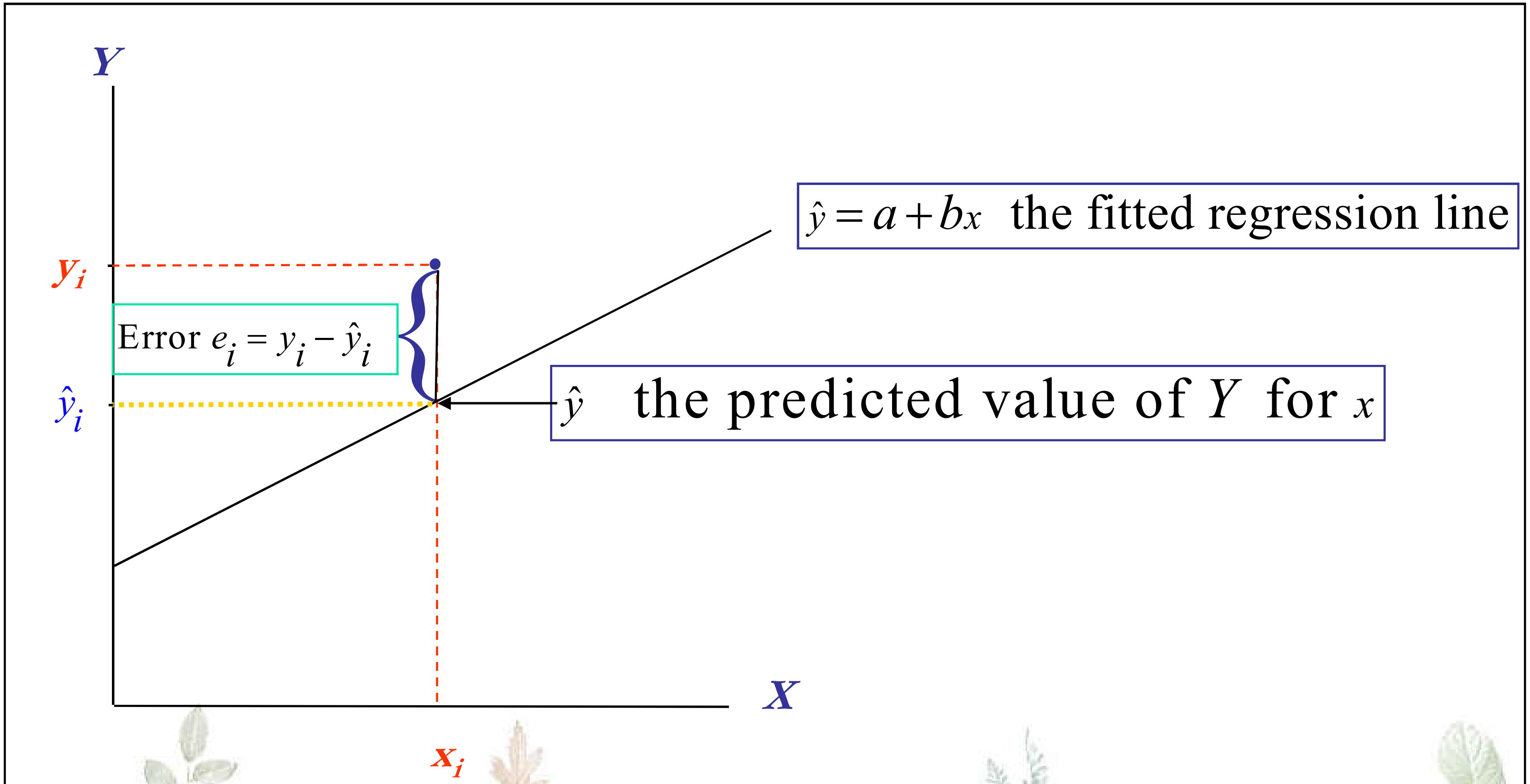
The estimated regression line:

$$\hat{y} = a + bx$$

where $\hat{y}$ ($y$ - hat) is the value of Y lying on the fitted regression line for a given value of $X$.

# 线性回归: 单变量回归

## Errors in Regression



# 线性回归: 单变量回归

## Least Squares Regression

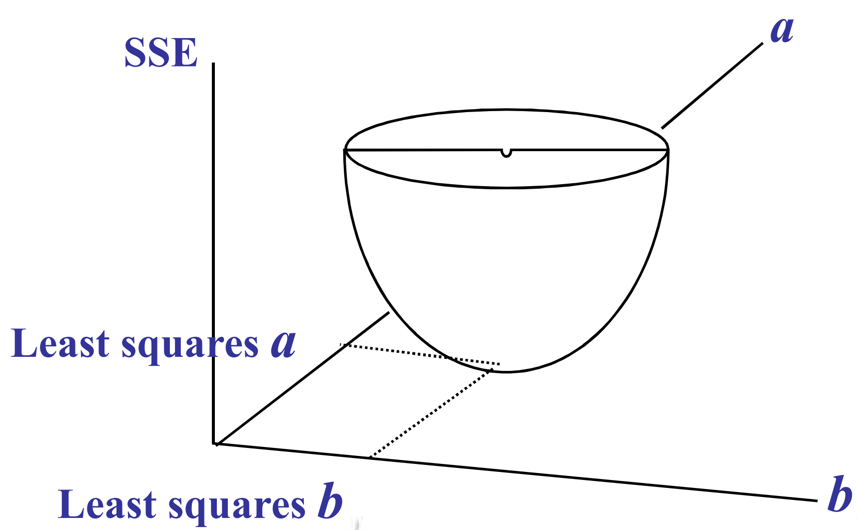
The sum of squared errors in regression is:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

**SSE: sum of squared errors**

The **least squares regression line** is that which *minimizes* the SSE with respect to the estimates ***a*** and ***b***.

**Parabola function**



# 线性回归: 单变量回归

## Sums of Squares, Cross Products, and Least Squares Estimators

Sums of Squares and Cross Products:

$$l_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

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

$$l_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

Least − squares regression estimators:

$$b = \frac{l_{xy}}{l_{xx}} \qquad \hat{y} = a + bx$$

$$a = \bar{y} - b\bar{x}$$

# 线性回归: 单变量回归

| Patient | $x$ | $y$ | $x^2$ | $y^2$ | $x \times y$ |
|---|---|---|---|---|---|
| 1 | 22.4 | 134.0 | 501.76 | 17956.0 | 3001.60 |
| 4 | 25.1 | 80.2 | 630.01 | 6432.0 | 2013.02 |
| 8 | 32.4 | 97.2 | 1049.76 | 9447.8 | 3149.28 |
| 2 | 51.6 | 167.0 | 2662.56 | 27889.0 | 8617.20 |
| 3 | 58.1 | 132.3 | 3375.61 | 17503.3 | 7686.63 |
| 5 | 65.9 | 100.0 | 4342.81 | 10000.0 | 6590.00 |
| 7 | 75.3 | 187.2 | 5670.09 | 35043.8 | 14096.16 |
| 6 | 79.7 | 139.1 | 6352.09 | 19348.8 | 11086.27 |
| 10 | 85.7 | 199.4 | 7344.49 | 39760.4 | 17088.58 |
| 9 | 96.4 | 192.3 | 9292.96 | 36979.3 | 18537.72 |
| Total | 592.6 | 1428.7 | 41222.14 | 220360.5 | 91866.46 |

$$l_{xx}=\sum x_i^2-\frac{\left(\sum x_i\right)^2}{n}=41222.14-\frac{592.6^2}{10}=6104.66$$

$$l_{yy}=\sum y_i^2-\frac{\left(\sum y_i\right)^2}{n}=220360.47-\frac{1428.70^2}{10}=16242.10$$

$$l_{xy}=\sum x_i y_i-\frac{\sum x_i \sum y_i}{n}=91866.46-\frac{592.6\times 1428.70}{10}=7201.70$$

$$b=\frac{l_{xy}}{l_{xx}}=\frac{7201.70}{6104.66}=1.18$$

$$a=\bar{y}-b\bar{x}=\frac{1428.7}{10}-(1.18)\left(\frac{592.6}{10}\right)$$

$$=72.96$$

regression equation:

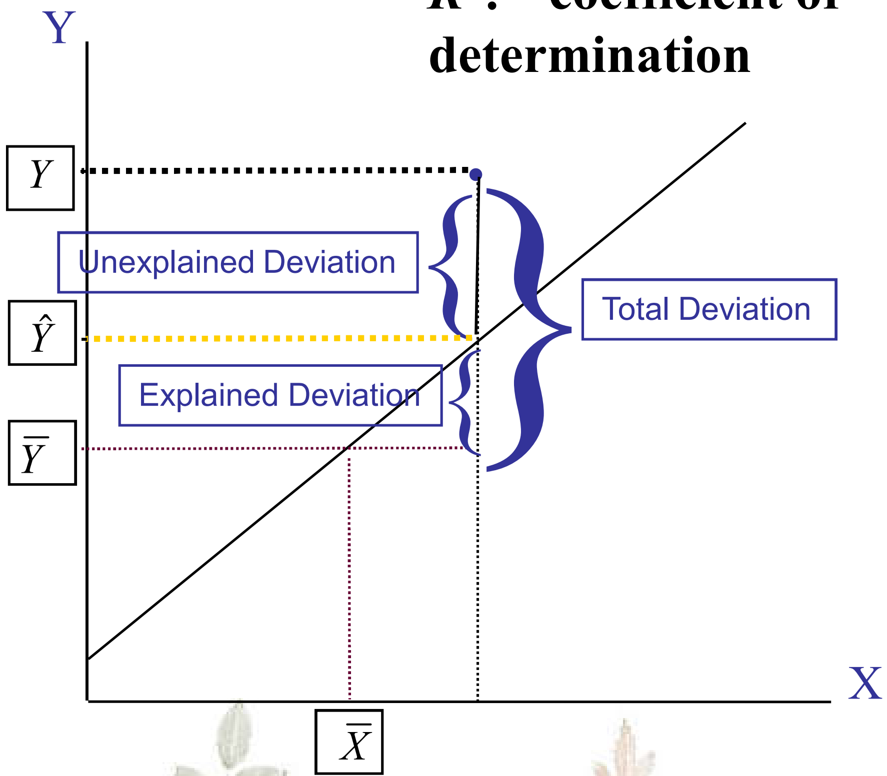
$$\hat{y}=72.96+1.18x$$

# 线性回归: 单变量回归

## How Good is the Regression?

The **coefficient of determination,** $R^2$, is a descriptive measure of the strength of the regression relationship, a measure how well the regression line fits the data.

**$R^2$: coefficient of determination**




| $(y-\overline{y})$ | = | $(y-\hat{y})$ | + | $(\hat{y}-\overline{y})$ |
|---|---|---|---|---|
| Total Deviation | = | Unexplained Deviation (Error) | | Explained Deviation (Regression) |

$$\sum(y-\overline{y})^2 = \sum(y-\hat{y})^2 + \sum(\hat{y}-\overline{y})^2$$

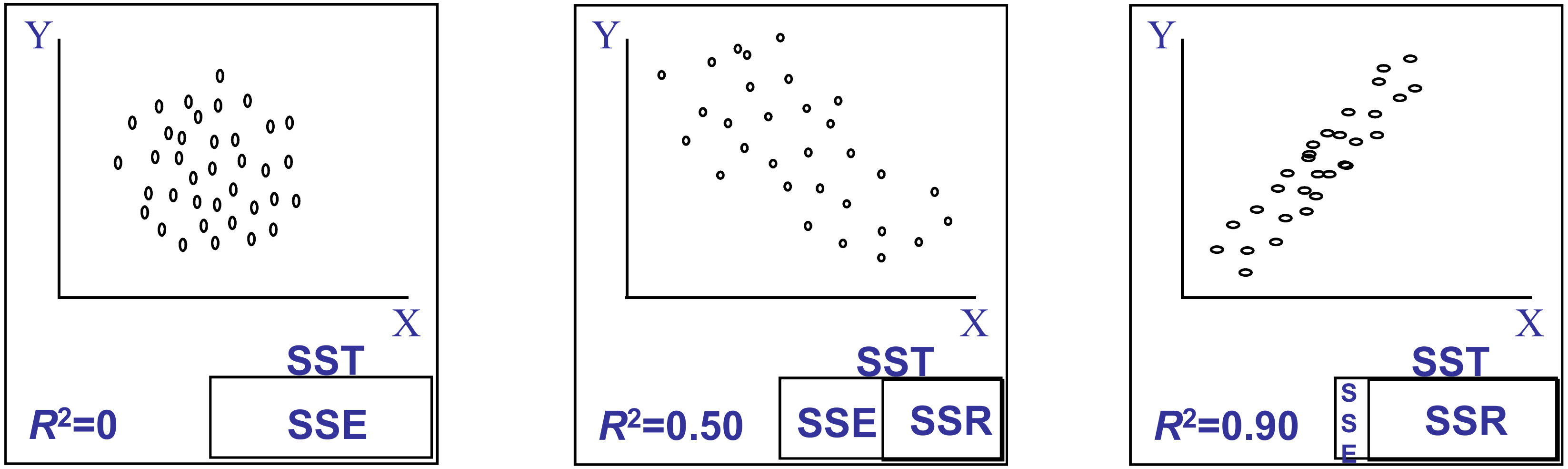
SST = SSE + SSR

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Percentage of total variation explained by the regression.

# 线性回归: 单变量回归

## The Coefficient of Determination




$$R^2 = \frac{SSR}{SST} = \frac{bl_{xy}}{l_{yy}} = \frac{1.180 \times 7201.70}{16242.10}$$

$$= 0.5231 = 52.31\%$$

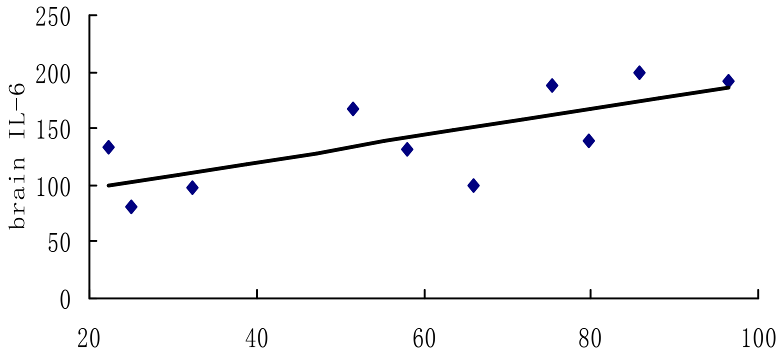



Figure18.1 Regression line between serum IL-6 and brain IL-6

# 线性回归: 单变量回归

## Assumptions of Regression

- Homoscedasticity(等方差)
  - The probability distribution of the errors has constant variance
- Independence of Errors
  - Error values are statistically independent
- Normality of Error
  - Error values (ε) are normally distributed for any given value of X

# 线性回归: 单变量回归

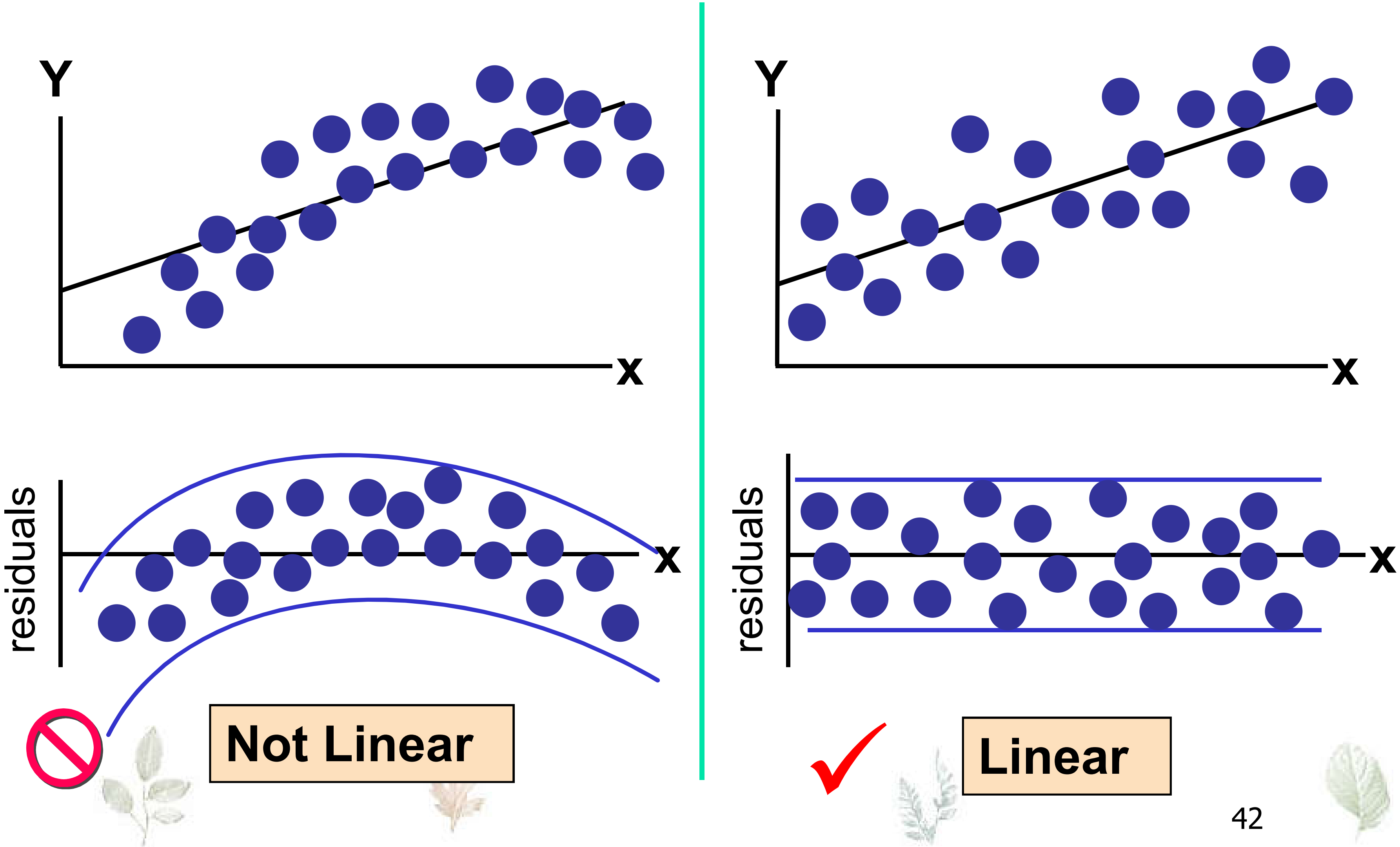
## Residual Analysis

$$\boxed{e_i = Y_i - \hat{Y}_i}$$

- The residual for observation i, $e_i$, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - Examine for **linearity assumption**
  - Examine for **constant variance** for all levels of X (homoscedasticity)
  - Evaluate **independence assumption**
  - Evaluate **normal distribution assumption**
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

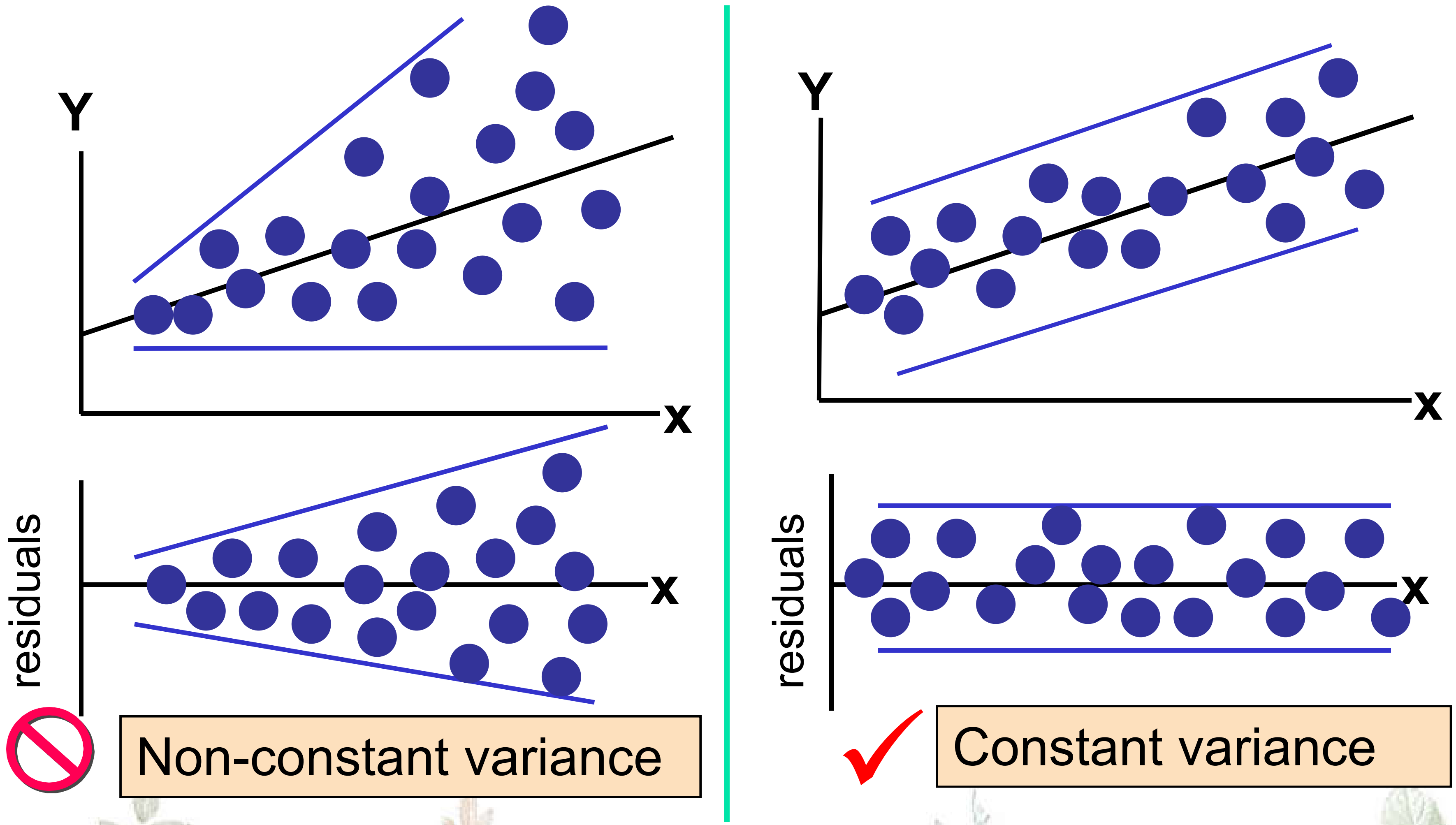
# 线性回归: 单变量回归

Residual Analysis for Linearity



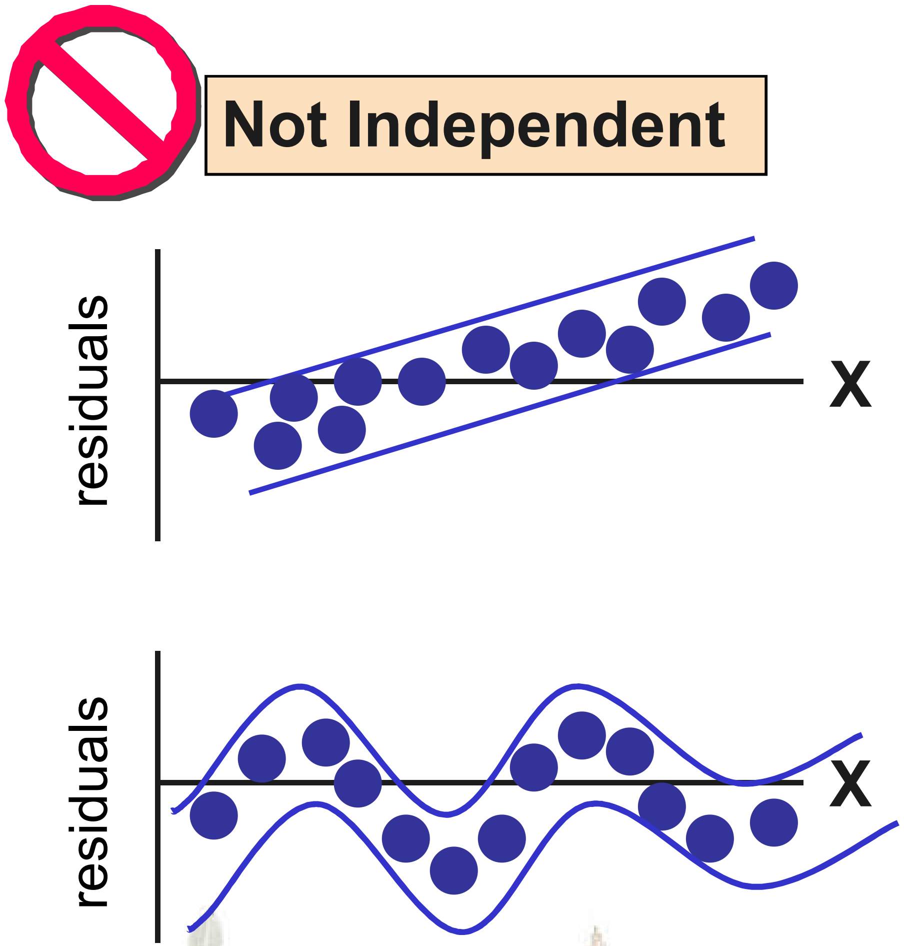
# 线性回归: 单变量回归

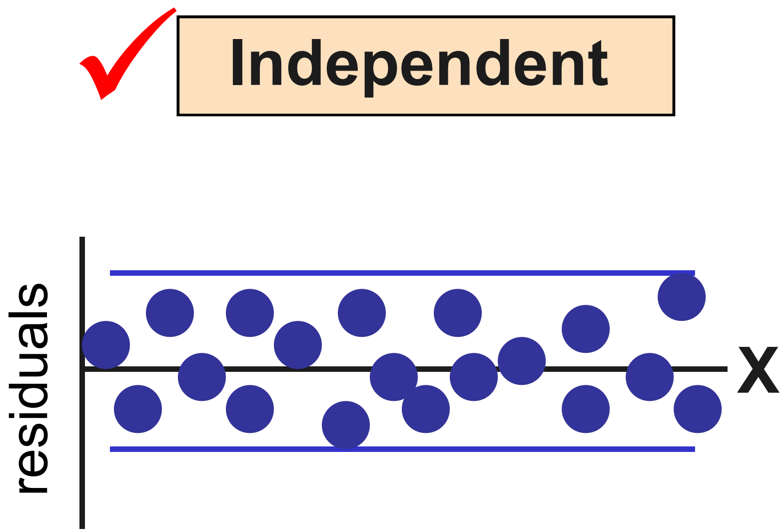
## Residual Analysis for Homoscedasticity



# 线性回归: 单变量回归

## Residual Analysis for Independence

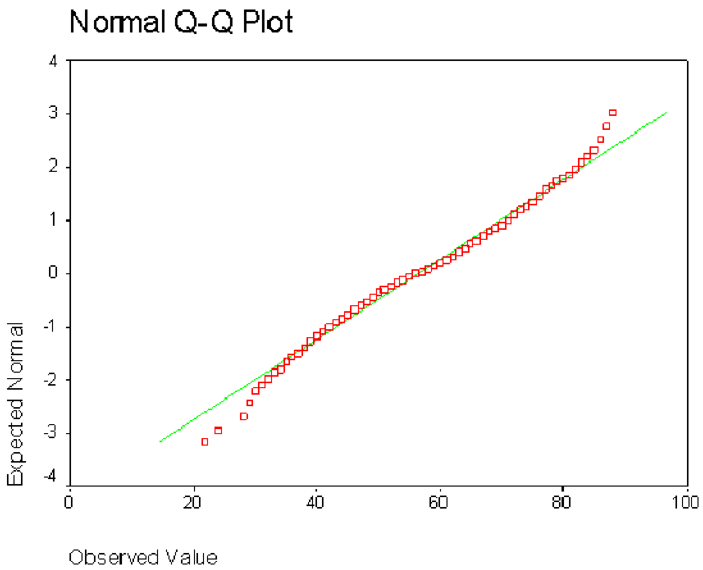


# 线性回归: 单变量回归

## Residual Analysis for Normality

**How do you know the residuals are normally distributed?**
Plot the residuals in Q-Q plot, which is a way to test normality. The idea of the Q-Q plot is that it plots the actual data along the y-axis, and the values that the data would have if they were exactly the percentiles of a normal curve (bell curve). So if the data is approximately like that of a bell curve, the line should look fairly close to straight. If not, it should be off.




Solution to non-normality: Transformation of the dependent variable.

**R语言：QQ图/PP图**
**SPSS：QQ图/PP图**

## R语言：QQ图/PP图

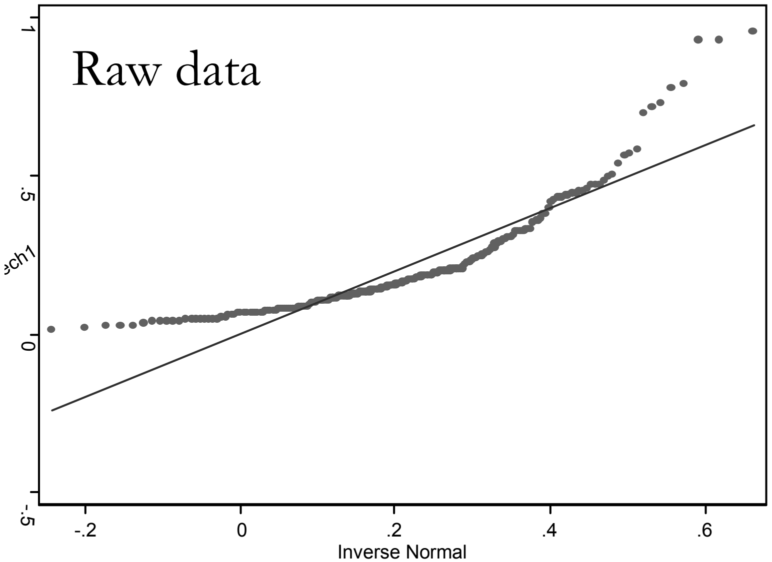
```
n = 100
a = rnorm(n)       #产生100个正态随机变量

t = rank(a) / n    #求观察累积概率，即百分位
q = qnorm(t)       #求百分位对应的数值（正态分布下）

plot(a, q)         #画Q-Q图


########################
p = pnorm(a)       #求正态分布函数值（正态累积概率）
plot(p, t)         #画P-P图
```

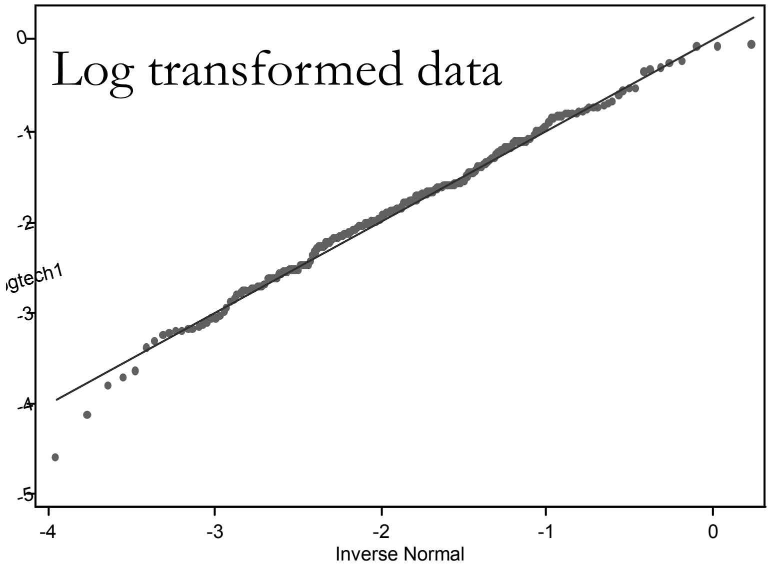
# 线性回归: 单变量回归

## Residual Analysis for Normality

Solution to non-normality:
Transformation of the dependent variable.

Example: log transformation

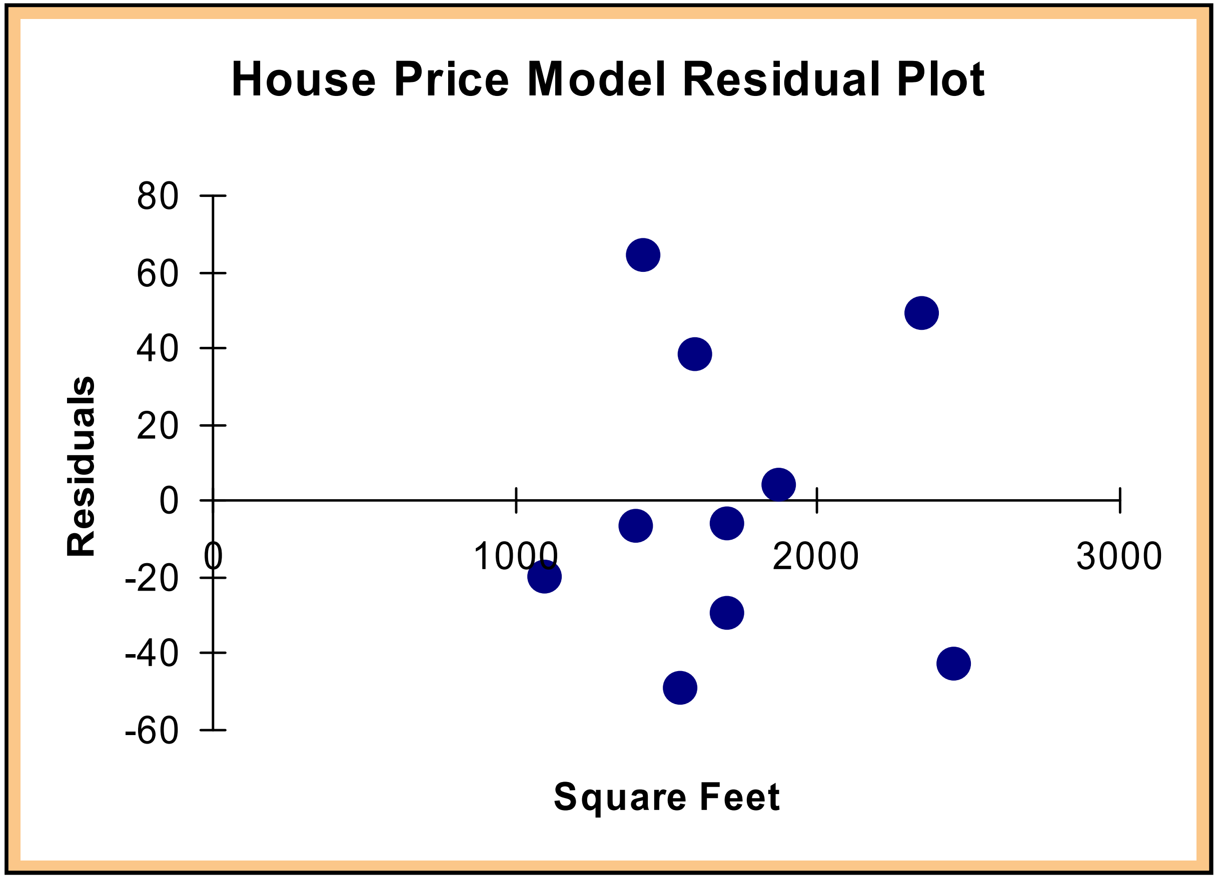


# 线性回归: 单变量回归

## Excel Residual Output

| RESIDUAL OUTPUT | | |
|---|---|---|
| | *Predicted House Price* | *Residuals* |
| 1 | 251.92316 | -6.923162 |
| 2 | 273.87671 | 38.12329 |
| 3 | 284.85348 | -5.853484 |
| 4 | 304.06284 | 3.937162 |
| 5 | 218.99284 | -19.99284 |
| 6 | 268.38832 | -49.38832 |
| 7 | 356.20251 | 48.79749 |
| 8 | 367.17929 | -43.17929 |
| 9 | 254.6674 | 64.33264 |
| 10 | 284.85348 | -29.85348 |




Does not appear to violate
any regression assumptions

# 线性回归: 单变量回归

## Measuring Autocorrelation:
## The Durbin-Watson Statistic

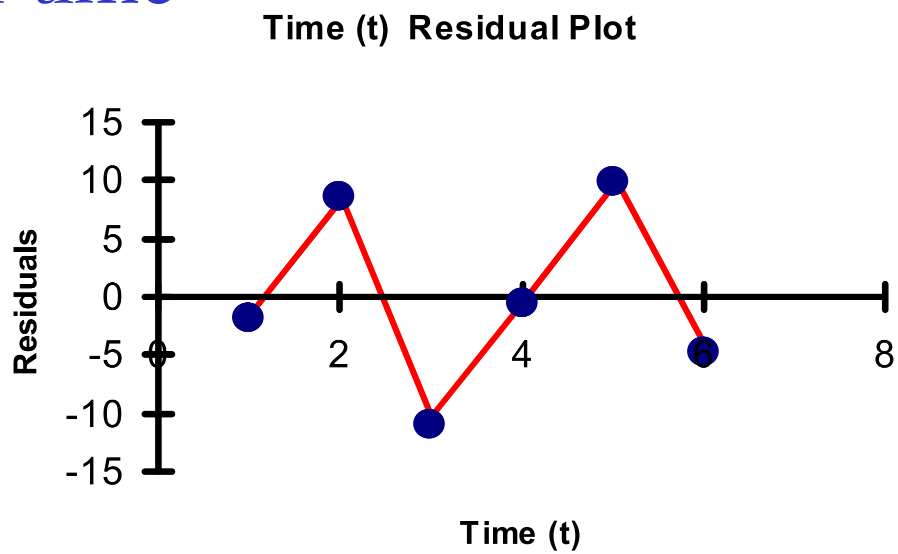
- Used when data are collected over time to detect if autocorrelation is present
- Autocorrelation exists if residuals in one time period are related to residuals in another period

# 线性回归: 单变量回归

## Autocorrelation

- Autocorrelation is correlation of the errors (residuals) over time

- Here, residuals show a cyclic pattern, not random




- Violates the regression assumption that residuals are random and independent

# 线性回归: 单变量回归

## The Durbin-Watson Statistic

- The Durbin-Watson statistic is used to test for autocorrelation

$H_0$: residuals are not correlated

$H_1$: autocorrelation is present

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

- The possible range is $0 \le D \le 4$
- D should be close to 2 if $H_0$ is true
- D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation

# 线性回归: 单变量回归

## Summary

1. **Regression analysis** is applied for prediction while control effect of independent variable X.
2. The principle of least squares in solution of regression parameters is to **minimize the residual sum of squares**.
3. The coefficient of determination, $R^2$, is a descriptive measure of the strength of the regression relationship.
4. There are two confidence bands: one for mean predictions and the other for individual prediction values
5. **Residual analysis** is used to check goodness of fit for models

# Logistic回归: 单变量回归

- Relate one or more independent (predictor) variables to a dependent (outcome) variable
  - Ordinary linear regression
    - **Continuous outcome variable**
    - Determine the relationship between a continuous outcome variable and the predictor variable(s)
  - Logistic regression
    - **Binary outcome variable**
    - Determine the relationship between **the probability of the outcome occurring** and the predictor variable(s)

# Logistic回归: 单变量回归

An example: faulty or not faulty

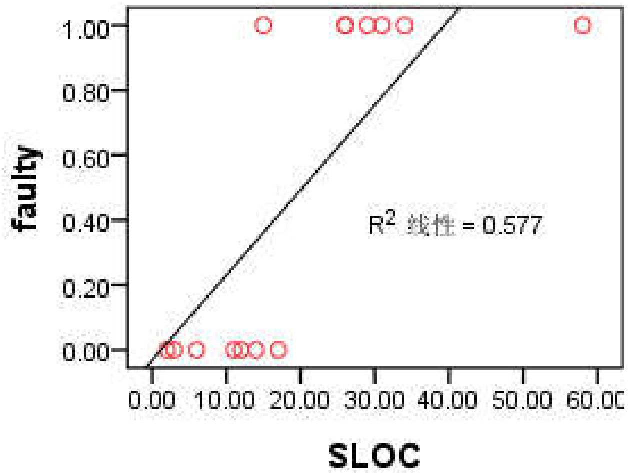
| Module id | Faulty? | SLOC |
|---|---|---|
| 1 | 0 | 3 |
| 2 | 1 | 34 |
| 3 | 0 | 17 |
| 4 | 0 | 6 |
| 5 | 0 | 12 |
| 6 | 1 | 15 |
| 7 | 1 | 26 |
| 8 | 1 | 29 |
| 9 | 0 | 14 |
| 10 | 1 | 58 |
| 11 | 0 | 2 |
| 12 | 1 | 31 |
| 13 | 1 | 26 |
| 14 | 0 | 11 |

- We will be interested then in inference about the probability of having faults
- Were we to use linear regression, we would postulate:

  $Prob\ (Faulty=1) = \alpha + \beta * SLOC + u$

# Logistic回归: 单变量回归

## Linear Probability Models




系数[a]

| 模型 | | 非标准化系数 | | 标准系数 | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | 标准 误差 | 试用版 | | |
| 1 | (常量) | -.032 | .162 | | -.197 | .847 |
| | SLOC | .026 | .006 | .759 | 4.044 | .002 |

a. 因变量: faulty

*Prob (Faulty=1) = -0.32 + 0.026*SLOC*

- **The results suggest that an increase in 1 SLOC increases the probability of having faults, on average, by approx. 0.026 or 2.6%.**
- **So what would the model predict if a module has 100 SLOC?**

*Prob (Faulty=1) = -0.32 + 0.026*100*
*= 2.28*

# Logistic回归: 单变量回归

Linear Probability Models: What is wrong?

- **Basically, the linear relation we had postulated before between X and Y is not appropriate when our dependent variable is dichotomic. Predictions for the probability of the event occurring would lie outside the [0,1] interval, which is unacceptable.**
- **Other two subtle problems:**
  - Distribution of $u_i$ is not normal as we wished it to be
  - The variance of $u_i$ is not constant (problem of heteroscedasticity)

# Logistic回归: 单变量回归

## Non Linear Probability Models

- We want to be able to model the probability of the event occurring with an explanatory variable 'X', but we want the predicted probability to remain within the [0,1] bounds.
  - There is a threshold above which the probability hardly increases as a reaction to changes in the explanatory variable
- Many functions meet these requirements (non-linearity and being bounded within the [0,1] interval)
- We will focus on the Logistic

# Logistic回归: 单变量回归

## The Logit Model

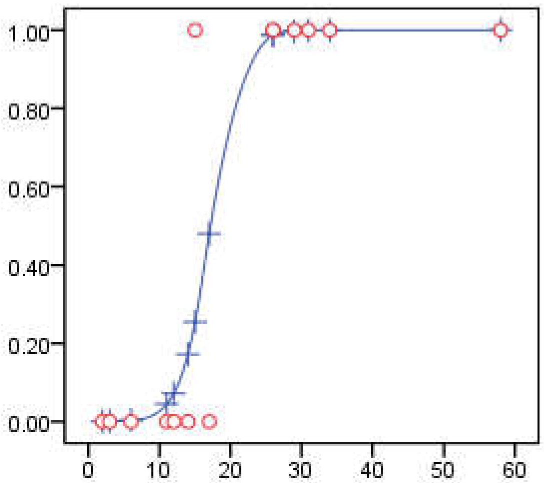
- A Logit Model states that:
  - Prob(Y=1) = F(a + bX)
  - Prob(Y=0) = 1 – F(a + bX)

$$F(a+bX) = \mathrm{P}(Y=1 \mid X) = \frac{1}{1+e^{-(a+bX)}}$$

  - Where F(.) is the 'Logistic Function'.
  - So, the probability of the event occurring is a *logistic function* of the independent variables

# Logistic回归: 单变量回归

The Logit Model




方程中的变量

| | | B | S.E, | Wals | df | Sig. | Exp (B) |
|---|---|---|---|---|---|---|---|
| 步骤 1[a] | SLOC | .495 | .384 | 1.660 | 1 | .198 | 1.640 |
| | 常量 | -8.496 | 6.016 | 1.994 | 1 | .158 | .000 |

a. 在步骤 1 中输入的变量: SLOC.

$$P(faulty = 1 \mid SLOC) = \frac{1}{1 + e^{-(-8.496 + 0.495 * SLOC)}}$$

# Logistic回归: 单变量回归

## Evaluating Logit Regressions

Statistics for comparing alternative logit models:

- Percent Correct Predictions
- Pseudo-$R^2$

# Logistic回归: 单变量回归

## Percent Correct Predictions

分类表[a]

| 已观测 | | | 已预测 | | |
|---|---|---|---|---|---|
| | | | Faulty | | 百分比校正 |
| | | | .00 | 1.00 | |
| 步骤 1 | Faulty | .00 | 7 | 0 | 100.0 |
| | | 1.00 | 1 | 6 | 85.7 |
| | 总计百分比 | | | | 92.9 |

a. 切割值为 .500

- The "Percent Correct Predictions" statistic assumes that if the estimated p is greater than or equal to .5 then the event is expected to occur and not occur otherwise.
- By assigning these probabilities 0s and 1s and comparing these to the actual 0s and 1s, the % correct Yes, % correct No, and overall % correct scores are calculated.
- Note: subgroups for the % correctly predicted is also important, especially if most of the data are 0s or 1s

# Logistic回归: 单变量回归

## Pseudo-$R^2$ Values

模型汇总

| 步骤 | -2 对数似然值 | Cox & Snell R 方 | Nagelkerke R 方 |
|---|---|---|---|
| 1 | 4.729[a] | .650 | .866 |

a. 因为参数估计的更改范围小于 .001，所以估计在迭代次数 9 处终止。

- There are psuedo-$R^2$ statistics that make adjustment for the (0,1) nature of the actual data: two are listed above
- Their computation is somewhat complicated but yield measures that vary between 0 and (somewhat close to) 1 much like the $R^2$ in a LP model.

# Logistic回归: 单变量回归

Odds Ratio

- Interpretation of Regression Coefficient ($b$):
  - In linear regression, the slope coefficient is the change in the mean response as $x$ increases by 1 unit
  - In logistic regression, we can show that:

$$Odds(Y=1|X) = \frac{\Pr(Y=1|X)}{1-\Pr(Y=1|X)} = e^{a+bX}$$

$$OR_{X,X+1}(Y=1) = \frac{Odds(Y=1|X+1)}{Odds(Y=1|X)} = e^{b}$$

$e^b$ represents the change in the odds of the outcome (multiplicatively) by increasing $x$ by 1 unit

# Logistic回归: 单变量回归

Magnitude of association

$$\Pr(Y=1)=\frac{e^{a+bX}}{1+e^{a+bX}}$$

$$OR_{X,X+1}(Y=1)=\frac{Odds(Y=1\mid X+1)}{Odds(Y=1\mid X)}=e^{b}$$

$$OR_{X,X+\sigma}(Y=1)=\frac{Odds(Y=1\mid X+\sigma)}{Odds(Y=1\mid X)}=e^{b\sigma}$$

# Logistic回归: 单变量回归

## Assumptions

- The only "real" limitation on logistic regression is that the outcome must be discrete
- Linearity in the logit – the regression equation should have a linear relationship with the logit form of the DV. There is no assumption about the predictors being linearly related to each other
- No outliers
- Independence of errors

# 小结

- 描述性统计
- 线性回归
- Logistic回归

# Thanks for your time and attention!