Software Metrics

Lecture 5

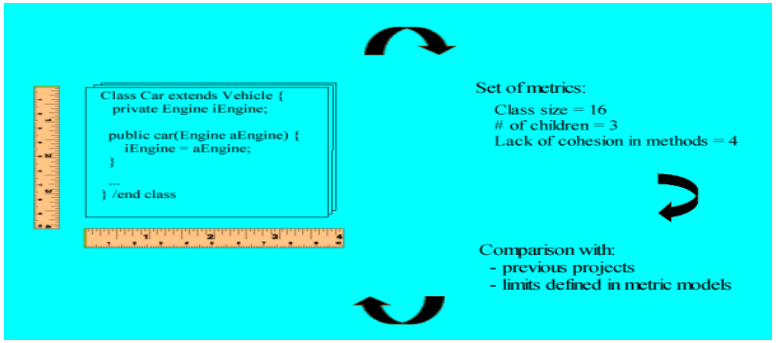
Data analysis

Yuming Zhou

Contents

- 通用的数据分析过程
- 受控实验的数据分析

数据集



(Copied from *staff.cs.utu.fi/opin**not**/kurssit/SemSE/05/SES-Malminen.ppt*)

No	WMC	LCOM	…	SLOC	Fault
1	35	96	…	1255	2
2	73	100		2828	0
…	…	…	…	…	…

No	**WMC**	**LCOM**	**…**	**SLOC**	**Fault**
1	15	9	…	135	?
2	45	10		282	?
…	…				

软件缺陷预测：问题

问题描述

① 模块中包含多少个缺陷？

② 系统中哪些模块包含缺陷？

③ 系统中哪些模块最有可能包含缺陷？

预测数量

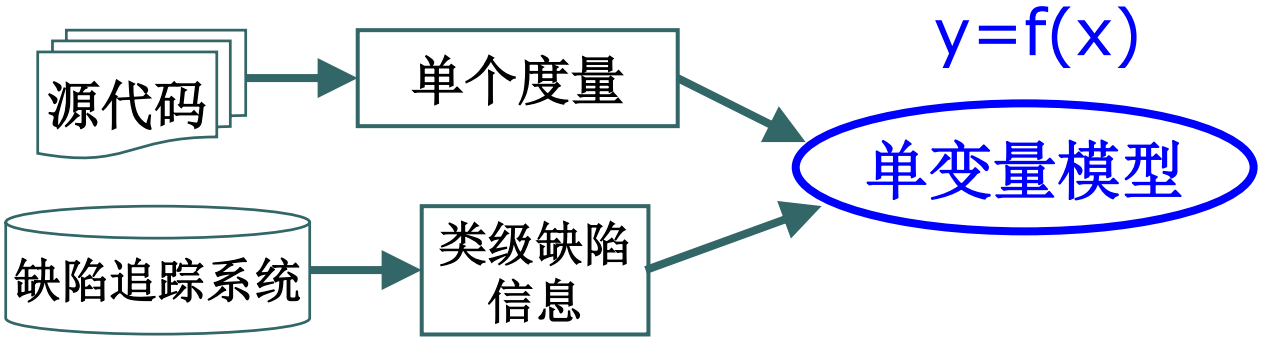
预测类别

预测序

软件缺陷预测：方法

预测方法

第一步：
单变量分析



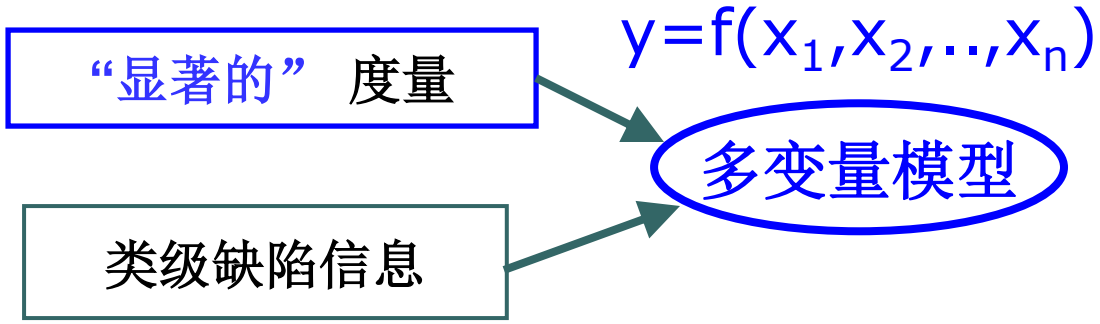
分析单个度量与缺陷的统计相关性(alpha = 0.05)。

f(x)可为线性回归模型、logistic回归模型或者其他模型

软件缺陷预测：方法

预测方法

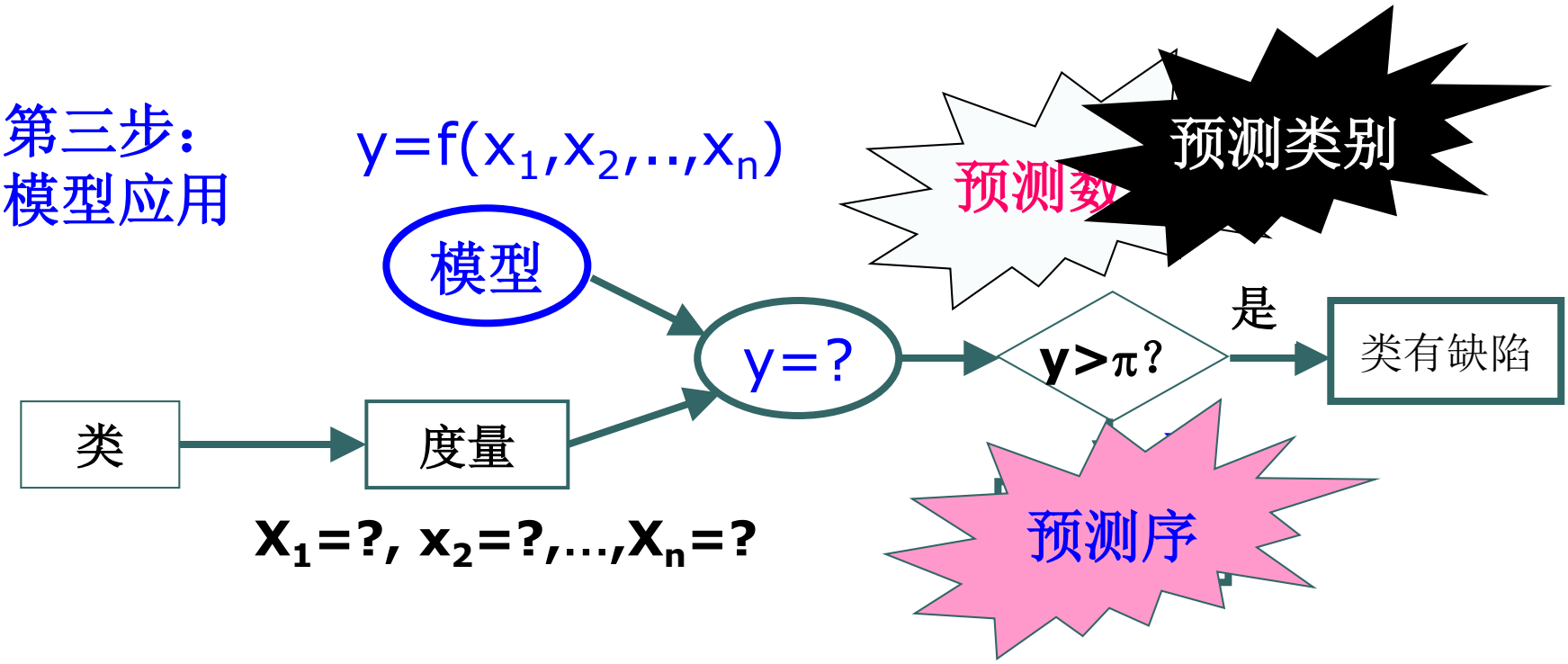
第二步：
多变量分析



仅选择第1步中统计相关的度量建立多变量模型

软件缺陷预测：方法

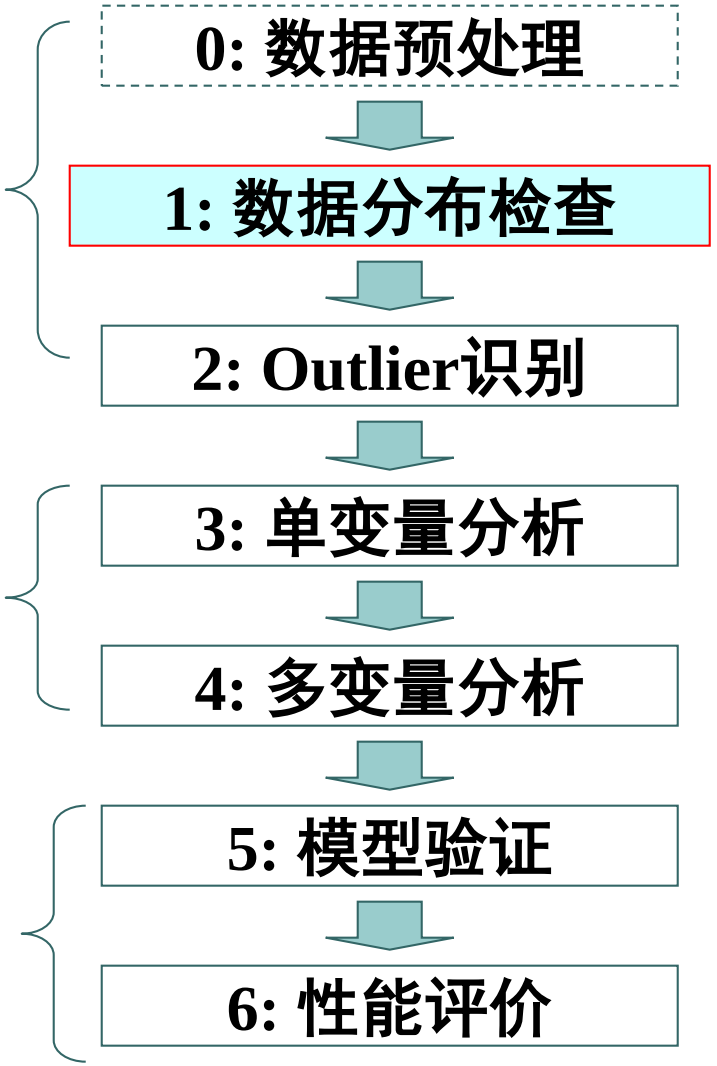
预测方法



预处理

模型构建

模型评价



位置：25%＋50%+75%

标准差

偏度 + 峰度

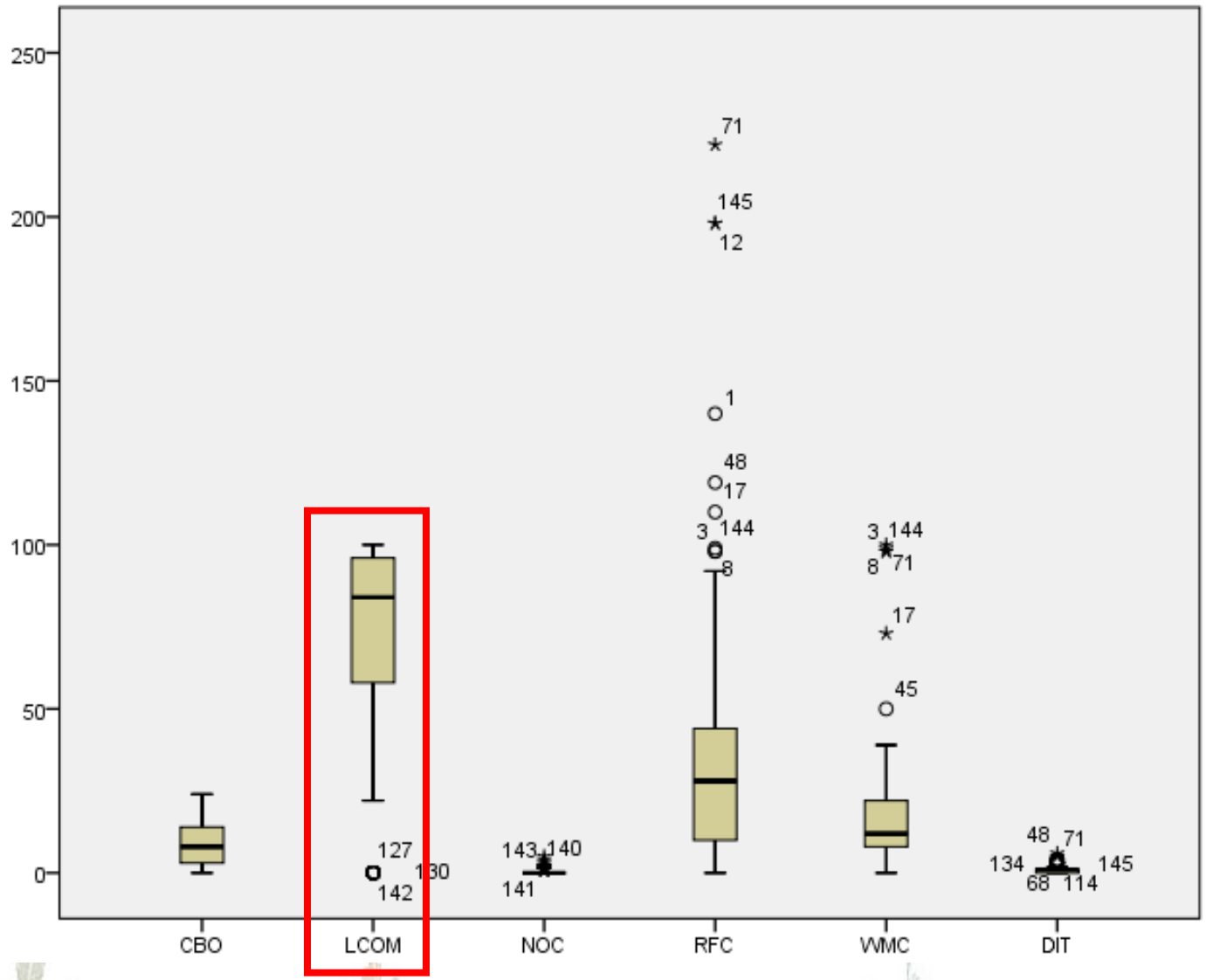
箱线图

Example

统计量

		CBO	LCOM	NOC	RFC	WMC	SLOC	DIT
N	有效	145	145	145	145	145	145	145
	缺失	0	0	0	0	0	0	0
均值		8.32	68.72	.21	34.38	17.42	258.8966	1.00
标准差		6.377	36.889	.699	36.203	17.449	378.36622	1.258
偏度		.377	-1.107	4.379	2.672	3.092	3.371	1.378
偏度的标准误		.201	.201	.201	.201	.201	.201	.201
峰度		-.818	-.404	22.271	9.566	11.640	14.487	1.580
峰度的标准误		.400	.400	.400	.400	.400	.400	.400
百分位数	25	3.00	56.50	.00	10.00	8.00	43.0000	.00
	50	8.00	84.00	.00	28.00	12.00	146.0000	1.00
	75	14.00	96.00	.00	44.50	22.00	285.0000	1.50

Example


250
200
150
100
50
0
71
145
12
1
48
17
3
144
8
3
144
8
71
17
45
127
130
142
143
140
141
48
71
134
145
68
114
CBO
LCOM
NOC
RFC
VMC
DIT

位置统计量

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

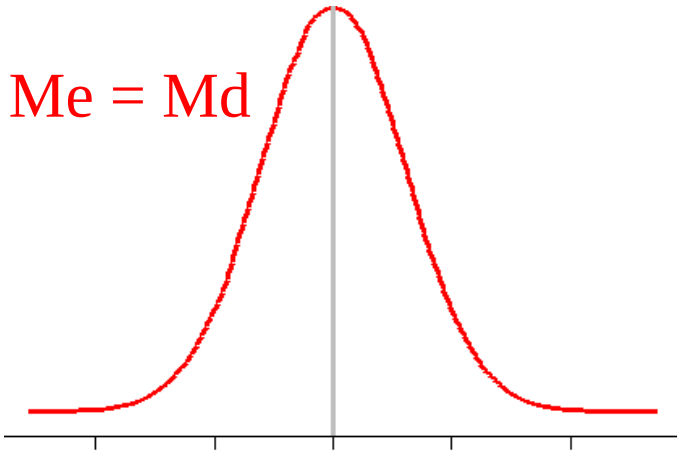
Example:

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

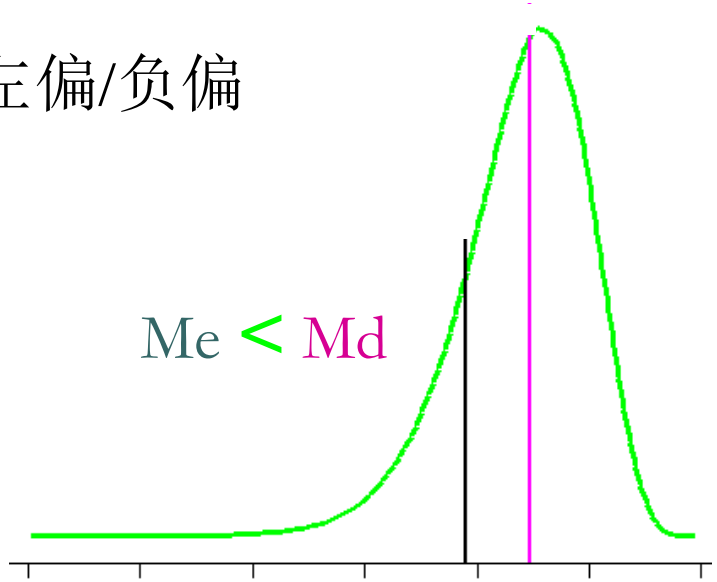
Mean : (1+11+10+2+7+5)/6 = 6

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

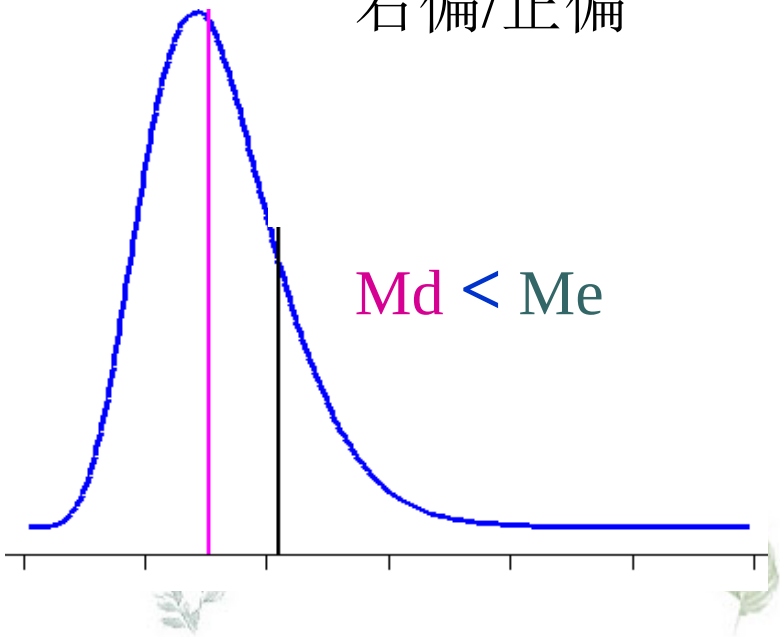
Mean v.s. Median



左偏/负偏



右偏/正偏



位置统计量 (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)

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

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

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

位置统计量 (cont'd)

Remarks:

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

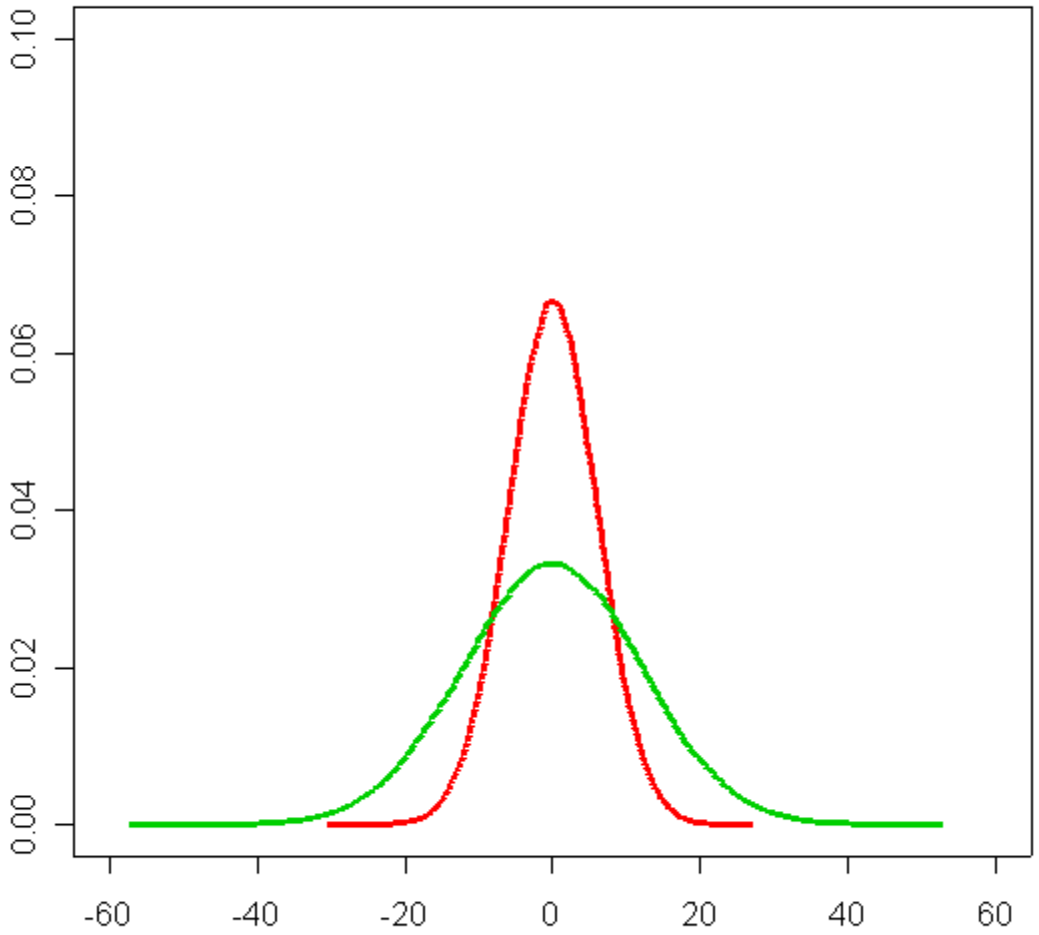
2^0. Quartile: 3 cut points

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

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

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

离散程度



离散程度 (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)

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

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

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

离散程度（cont'd)

4. Mean Absolute Deviation

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

5. Variance

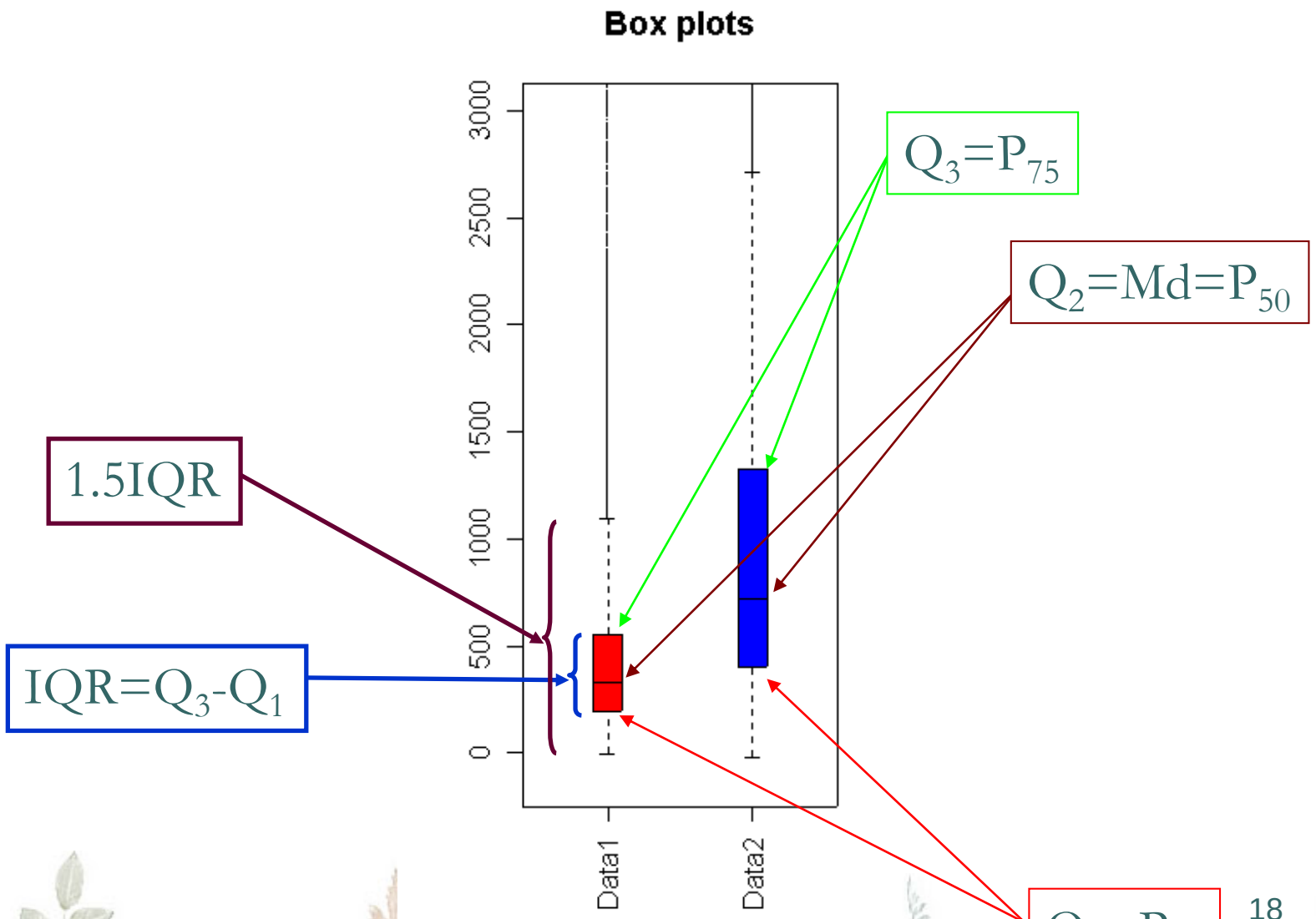
$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \text{ (统计量)}, \qquad \sigma^2=\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\mu\right)^2 \text{ (参数)}$$

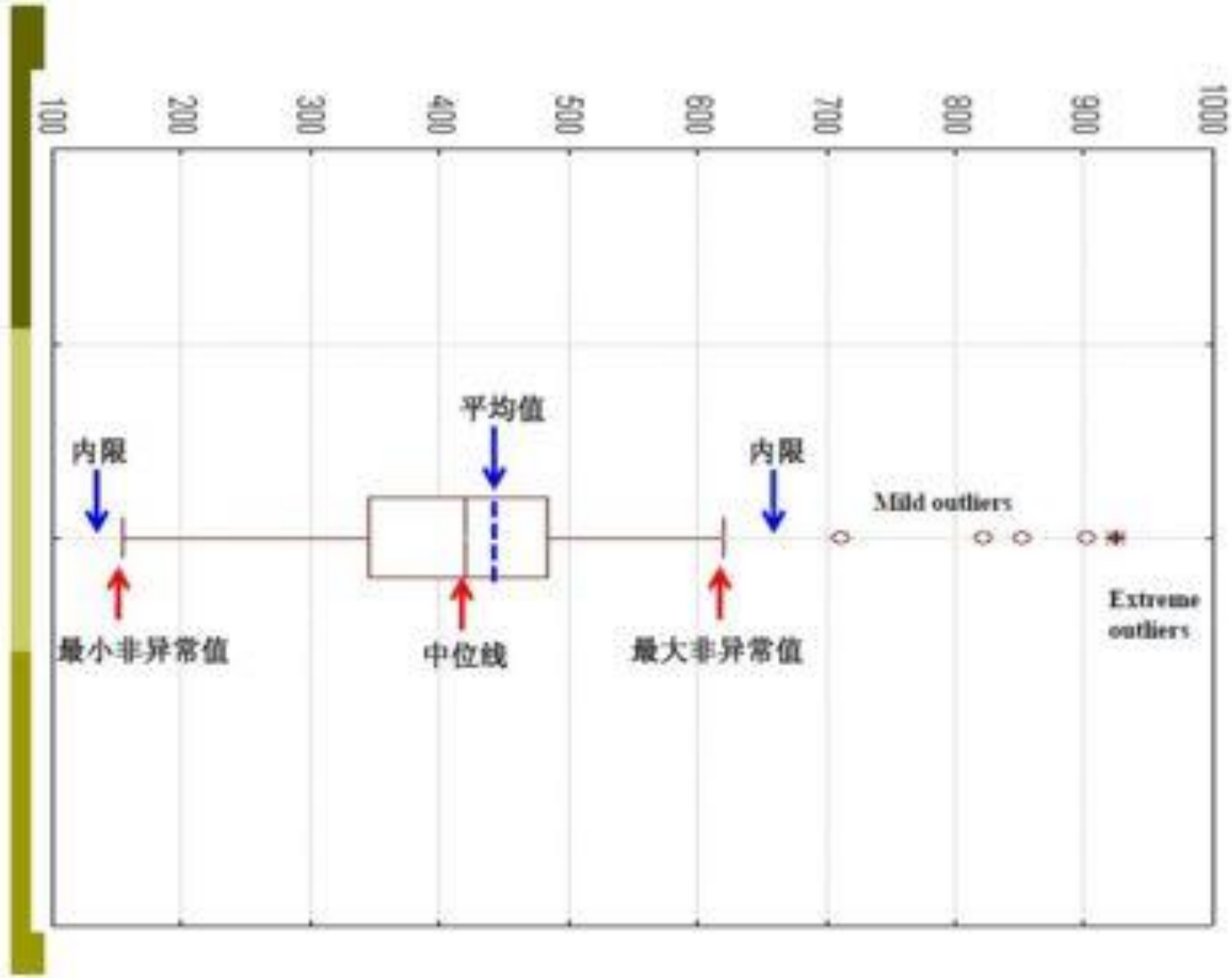
6. Standard Deviation

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

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

箱线图


Box plots
3000
2500
2000
1500
1000
500
0
Data1
Data2
$Q_3=P_{75}$
$Q_2=Md=P_{50}$
1.5IQR
$IQR=Q_3-Q_1$
$Q_1=P_{25}$


100
200
300
400
500
600
700
800
900
1000
平均值
内限
内限
Mild outliers
Extreme outliers
最小非异常值
中位线
最大非异常值

偏度和峰度

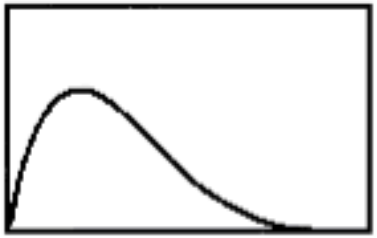
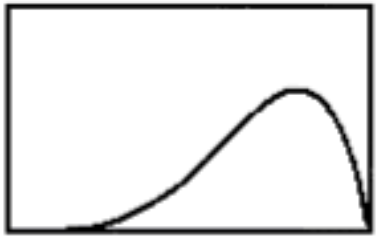
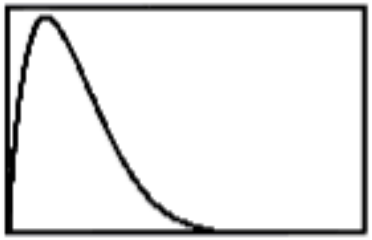
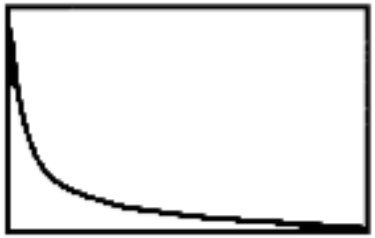
1. Skewness: a measure of symmetry, or more precisely, the lack of symmetry

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

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

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

非正态分布--->正态分布？

Form	Transformation	Form	Transformation
	Square Root newx = sqrt(x)		Reflect and Square Root newx =sqrt(k-x)
	Logarithm newx = lg10(x)		Reflect and Logarithm newx =lg10(k-x)
	Inverse newx =1/x		Reflect and Inverse newx =1/(k-x)

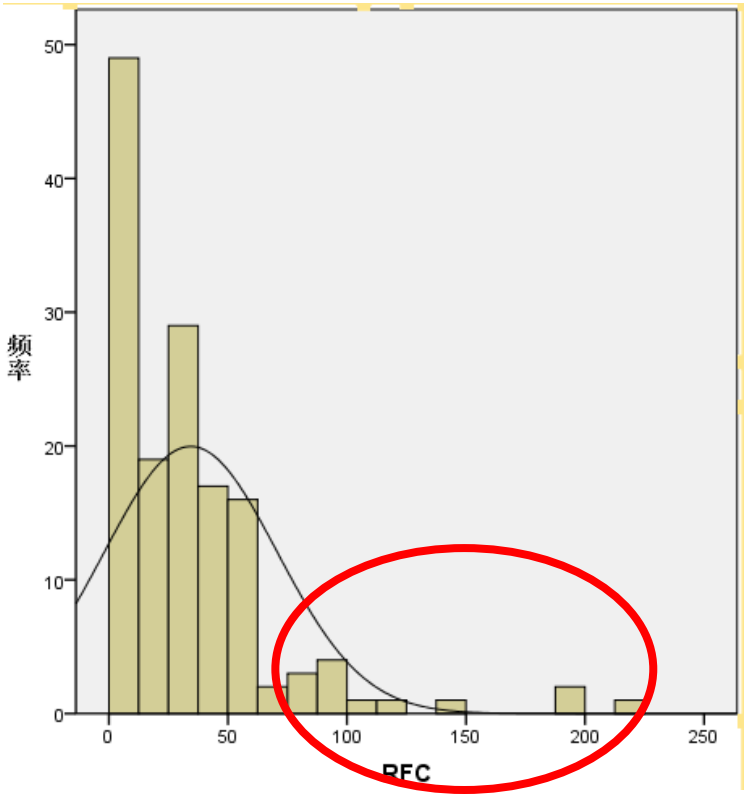
近似正态：偏度和峰度都在[-1，1]之间

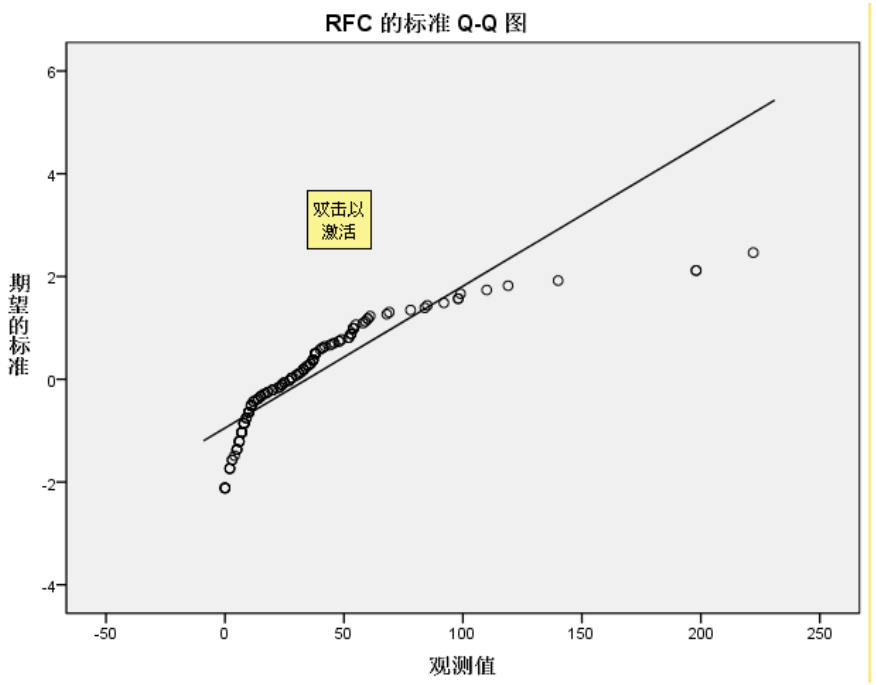
统计量

		CBO	LCOM	NOC	RFC	WMC	SLOC	DIT
N	有效	145	145	145	145	145	145	145
	缺失	0	0	0	0	0	0	0
均值		8.32	68.72	.21	34.38	17.42	258.8966	1.00
标准差		6.377	36.889	.699	36.203	17.449	378.36622	1.258
偏度		.377	-1.107	4.379	2.672	3.092	3.371	1.378
偏度的标准误		.201	.201	.201	.201	.201	.201	.201
峰度		-.818	-.404	22.271	9.566	11.640	14.487	1.580
峰度的标准误		.400	.400	.400	.400	.400	.400	.400
极小值		0	0	0	0	0	1.00	0
极大值		24	100	5	222	100	2495.00	6
百分位数	25	3.00	56.50	.00	10.00	8.00	43.0000	.00
	50	8.00	84.00	.00	28.00	12.00	146.0000	1.00
	75	14.00	96.00	.00	44.50	22.00	285.0000	1.50

RFC偏度=2.672　正偏(右偏)

如何变换，使得更贴近正态？





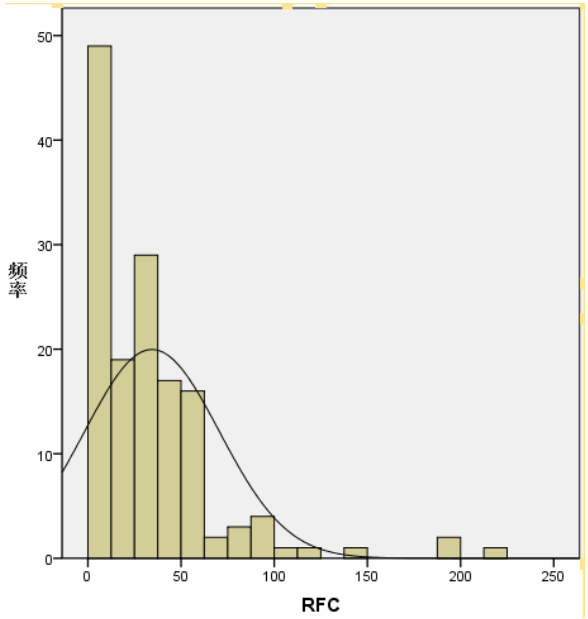
NOC偏度=4.379　正偏(右偏)

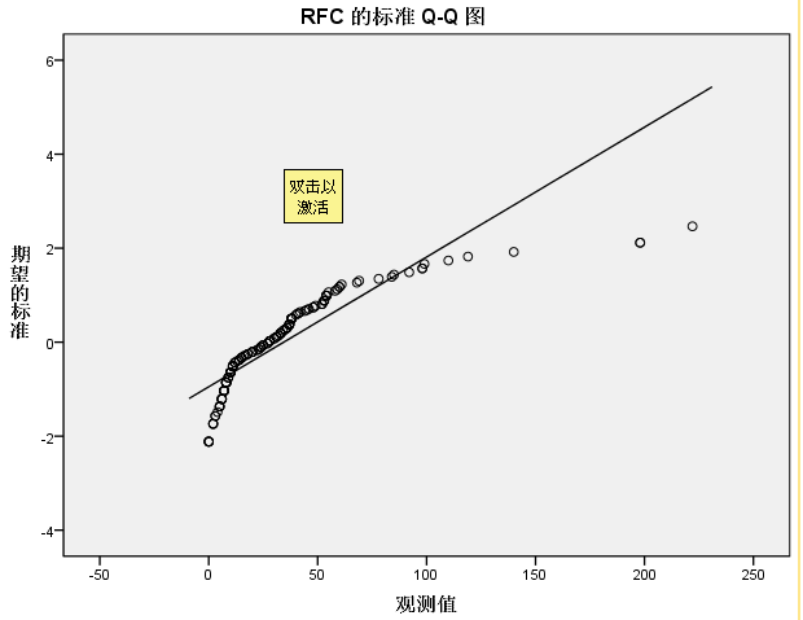
如何变换，使得更贴近正态？

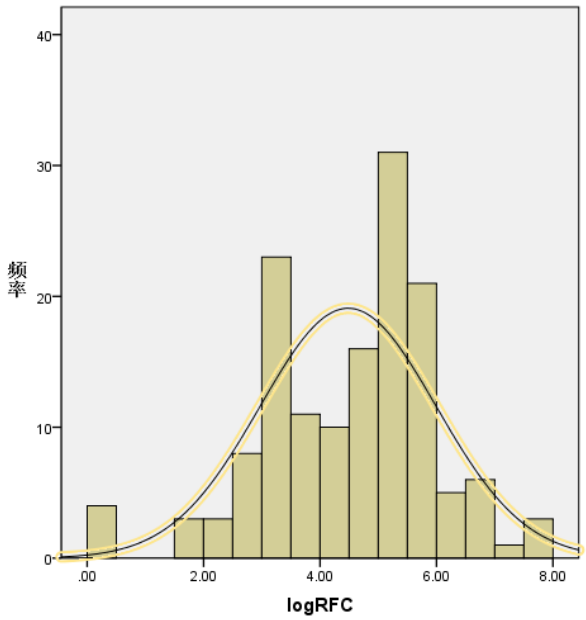
Sqrt(RFC+1)　　Log2(RFC+1)　　-1/(RFC+1)

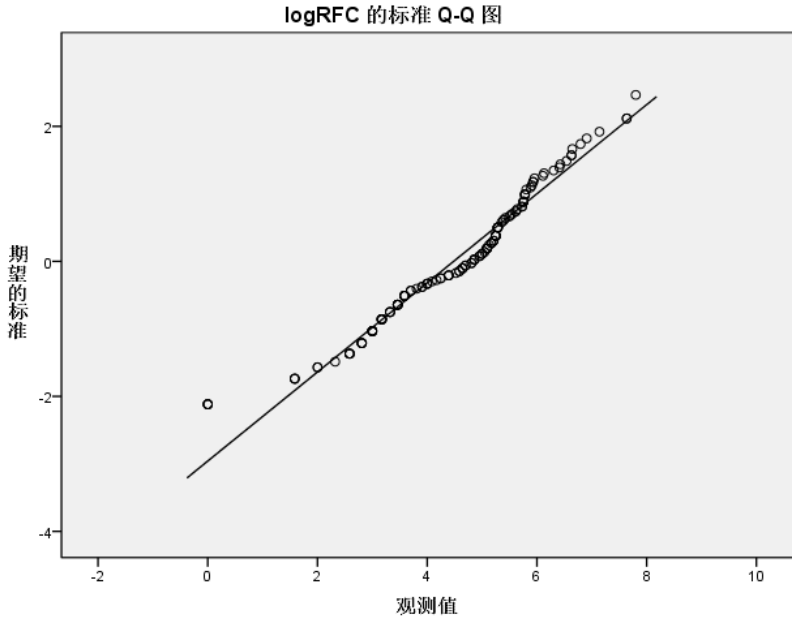
统计量

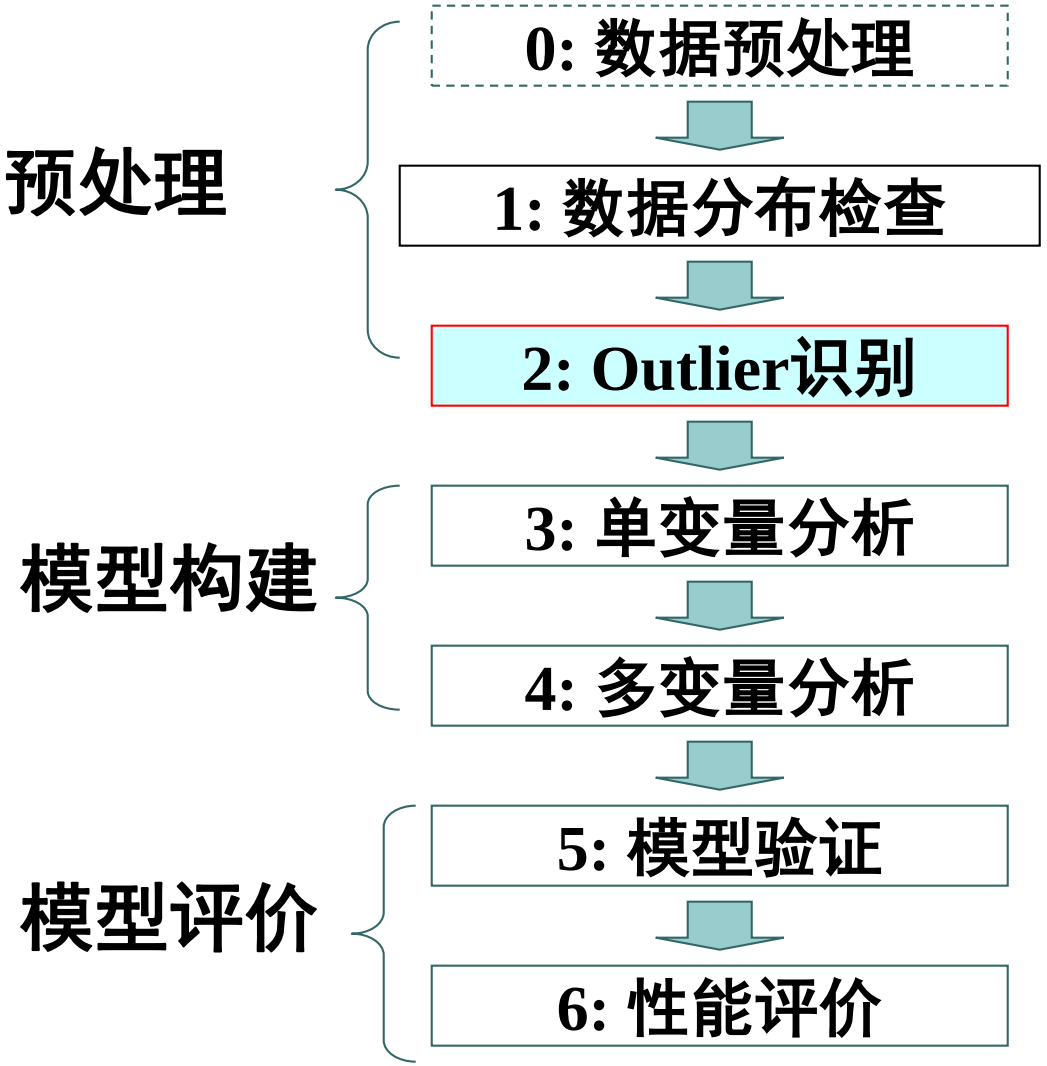
		RFC	sqrtRFC	logRFC	inverseRFC
N	有效	145	145	145	145
	缺失	0	0	0	0
偏度		2.672	.995	-.614	-4.686
偏度的标准误		.201	.201	.201	.201


频率
RFC


RFC 的标准 Q-Q 图
双击以
激活
期望的标准
观测值

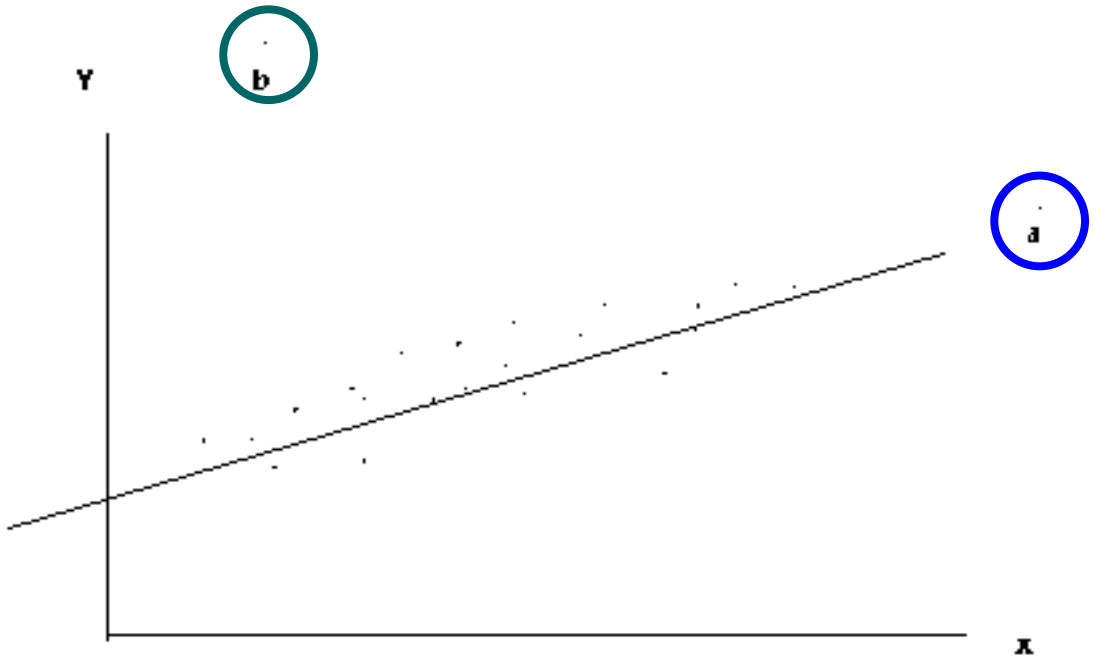

频率
logRFC


logRFC 的标准 Q-Q 图
期望的标准
观测值



排除有影响的: Cook's d > 1

有影响的outlier识别



Outlier a does not distort and outlier b does

有影响的outlier识别

Cook's 距离 (常用的阈值为1)

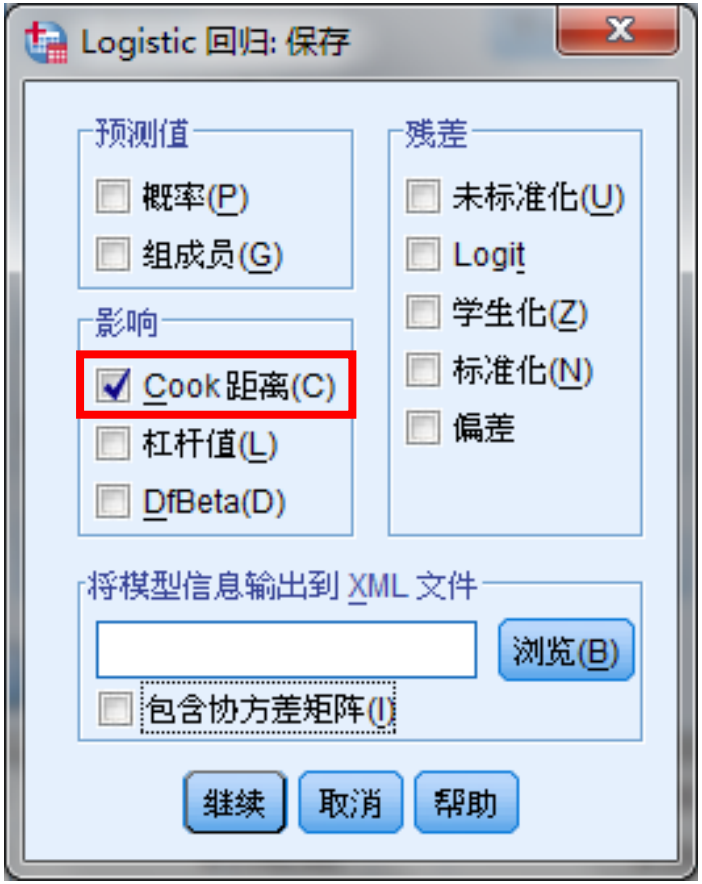
$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\ \mathrm{MSE}}$$

$\hat{Y}_j$ the prediction from the full regression model for observation j

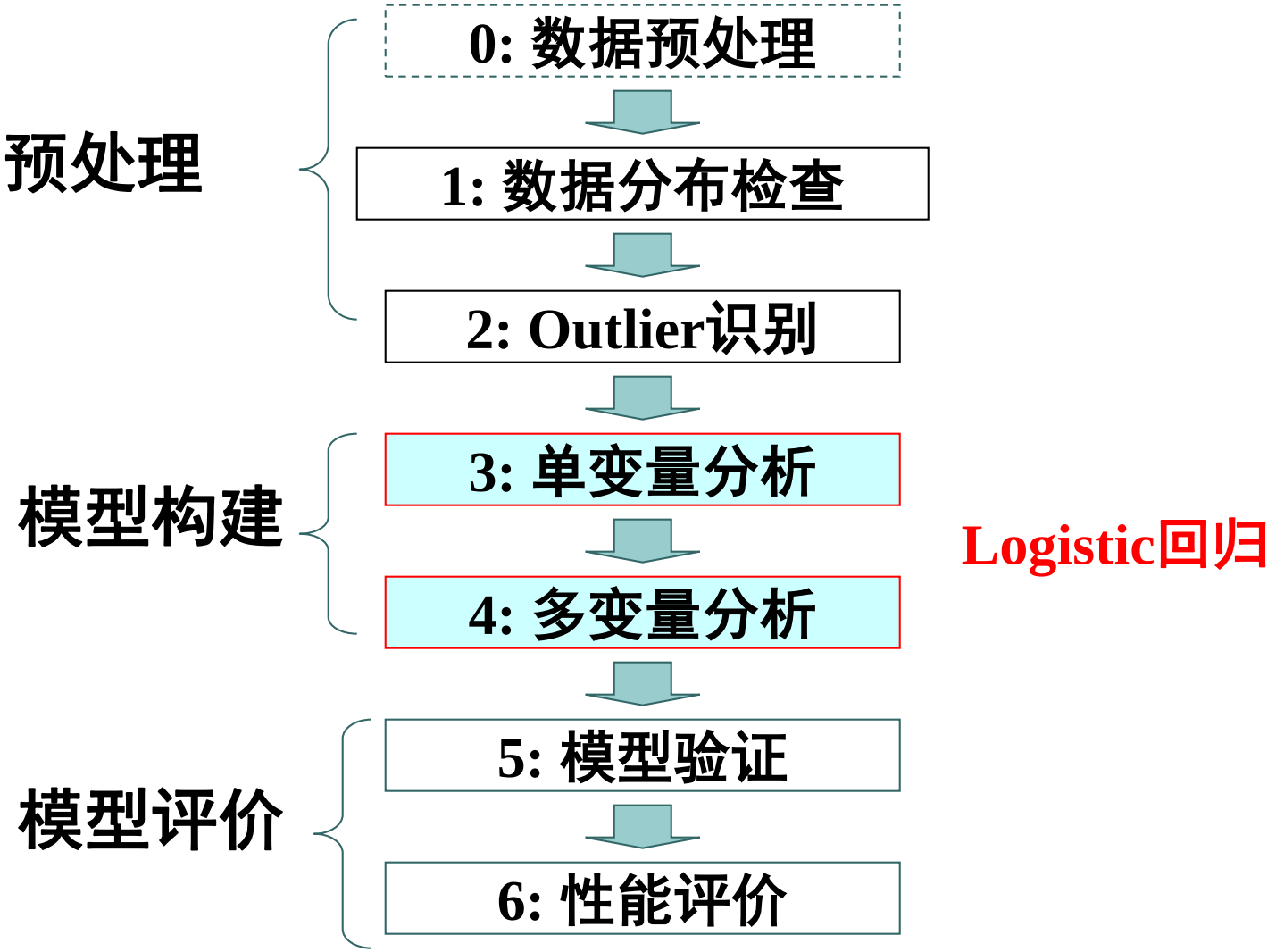
$\hat{Y}_{j(i)}$ the prediction for observation *j* from a refitted regression model in which observation *i* has been omitted

MSE is the mean square error of the regression model

p is the number of fitted parameters in the model



inverseRFC	COO_1
.00	2.23657
-.01	.49119
-.01	.00088
-.01	.00009
-.01	.00444
-.01	.97923


预处理
0: 数据预处理
1: 数据分布检查
2: Outlier识别
模型构建
3: 单变量分析
4: 多变量分析
Logistic回归
模型评价
5: 模型验证
6: 性能评价

Logistic回归模型

An example: faulty or not faulty

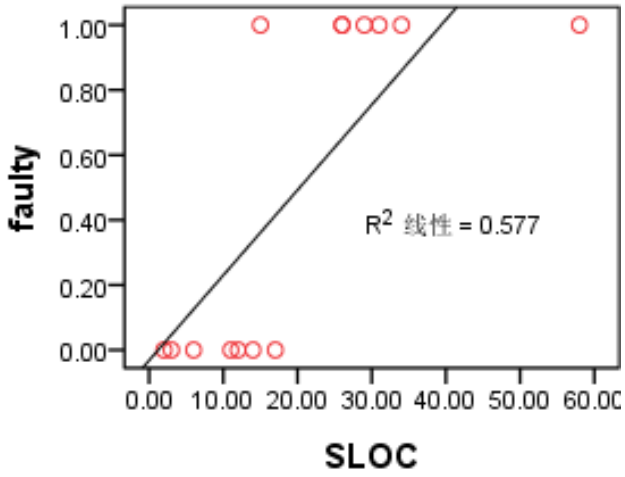
Module id	Faulty?	SLOC
1	0	3
2	1	34
3	0	17
4	0	6
5	0	12
6	1	15
7	1	26
8	1	29
9	0	14
10	1	58
11	0	2
12	1	31
13	1	26
14	0	11

- 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回归模型

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回归模型

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回归模型

The Logit Model

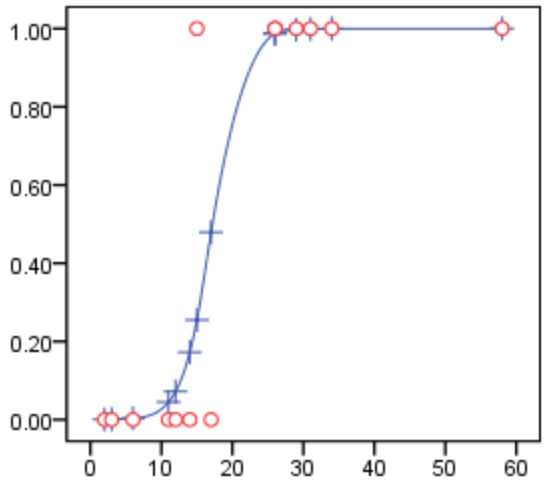
- A Logit Model states that:
  - Prob(Y=1) = F(a + bX)
  - Prob(Y=0) = 1 – F(a + bX)

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

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

Logistic回归模型

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.

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

Logistic回归模型

单变量

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

多变量

$$P(Y=1|X_1,X_2,...,X_k)=\frac{1}{1+e^{-(a+b_1X_1+b_2X_2+...+b_kX_k)}}$$

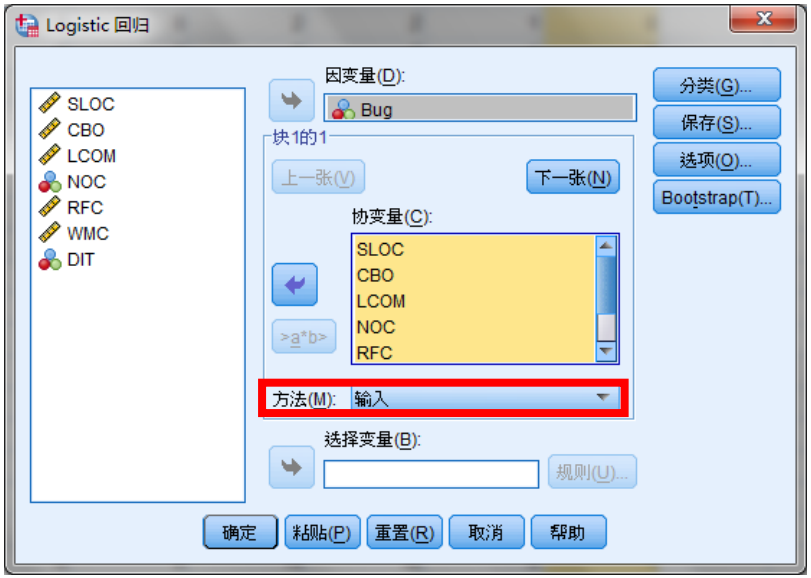
Logistic回归模型

变量选择

	SLOC	CBO	LCOM	NOC	RFC	WMC	DIT	Bug
1	1	0	0	0	60	5	2	0
2	10	0	0	0	10	10	0	0
3	12	0	0	0	10	10	0	0
4	5	1	0	0	2	2	1	0
5	5	1	0	0	2	2	1	0
6	11	1	0	0	3	3	0	0
7	115	3	0	0	8	8	0	1
8	34	4	0	0	6	6	0	0
9	49	4	0	0	5	5	0	0
10	64	4	0	0	5	5	0	1
11	112	4	0	0	5	5	0	1
12	128	4	0	0	8	8	0	0

Logistic回归模型

使用所有变量(1)



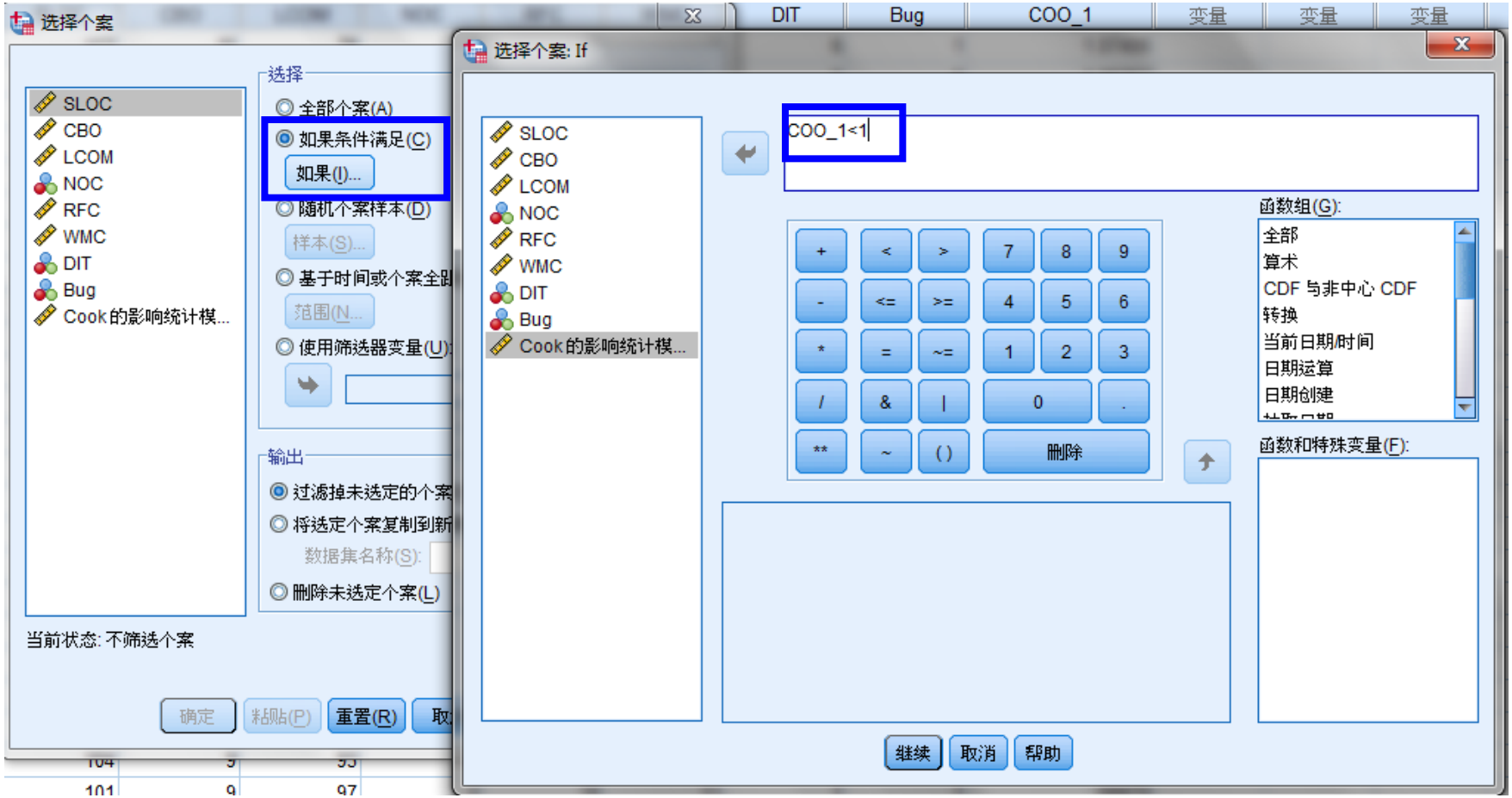
方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1[a]	SLOC	.004	.002	5.192	1	.023	1.004
	CBO	.268	.062	18.583	1	.000	1.307
	LCOM	-.014	.007	3.832	1	.050	.986
	NOC	-.359	.307	1.375	1	.241	.698
	RFC	.024	.017	2.111	1	.146	1.025
	WMC	-.037	.035	1.129	1	.288	.964
	DIT	-1.065	.353	9.113	1	.003	.345
	常量	-1.851	.536	11.927	1	.001	.157

	DIT	Bug	COO_1
7	6	1	1.07406
00	5	0	1.06361
7	0	1	.62902
12	3	0	.45747

Logistic回归模型

使用所有变量(2)



Logistic回归模型

使用所有变量(3)

	SLOC	CBO	LCOM	NOC	RFC	WMC	DIT	Bug	COO_1	filter_$
1	177	15	78	0	119	7	6	1	1.07406	0
2	1016	9	86	0	222	100	5	0	1.06361	0
3	265	9	0	4	7	7	0	1	.62902	1
4	759	14	72	0	27	12	3	0	.45747	1

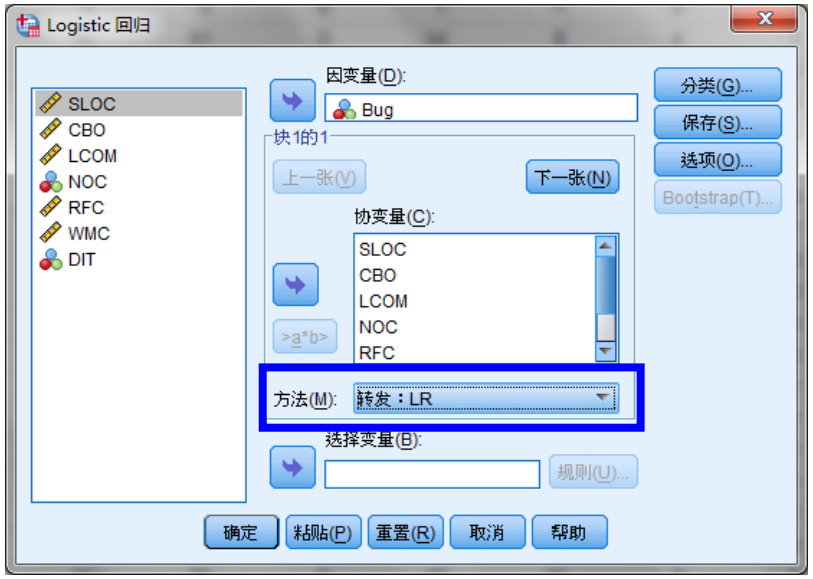
方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1[a]	SLOC	.004	.002	6.221	1	.013	1.004
	CBO	.263	.065	16.107	1	.000	1.301
	LCOM	-.015	.007	4.271	1	.039	.985
	NOC	-.378	.308	1.514	1	.218	.685
	RFC	.023	.018	1.572	1	.210	1.023
	WMC	-.011	.035	.106	1	.745	.989
	DIT	-1.164	.373	9.729	1	.002	.312
	常量	-2.027	.563	12.973	1	.000	.132

	SLOC	CBO	LCOM	NOC	RFC	WMC	DIT	Bug	COO_1	filter_$	COO_2
1	265	9	0	4	7	7	0	1	.62902	1	.67764
2	268	14	91	0	54	8	4	1	.35748	1	.54519
3	759	14	72	0	27	12	3	0	.45747	1	.44986
4	142	3	90	0	11	8	1	1	.36748	1	.42516

Logistic回归模型

前向逐步回归(1)



方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1[a]	CBO	.207	.037	31.435	1	.000	1.229
	常量	-2.238	.395	32.079	1	.000	.107
步骤 2[b]	CBO	.289	.051	32.095	1	.000	1.336
	DIT	-.624	.211	8.751	1	.003	.536
	常量	-2.347	.430	29.777	1	.000	.096
步骤 3[c]	SLOC	.003	.001	6.206	1	.013	1.003
	CBO	.230	.054	18.330	1	.000	1.259
	DIT	-.708	.221	10.317	1	.001	.492
	常量	-2.456	.454	29.235	1	.000	.086

a. 在步骤 1 中输入的变量: CBO.

b. 在步骤 2 中输入的变量: DIT.

c. 在步骤 3 中输入的变量: SLOC.

Bug	COO_1
1	1.03269
1	.24395
0	.20138
0	.18699

Logistic回归模型

前向逐步回归(2)

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1[a]	CBO	.207	.037	31.435	1	.000	1.229
	常量	-2.238	.395	32.079	1	.000	.107
步骤 2[b]	CBO	.289	.051	32.095	1	.000	1.336
	DIT	-.624	.211	8.751	1	.003	.536
	常量	-2.347	.430	29.777	1	.000	.096
步骤 3[c]	SLOC	.003	.001	6.206	1	.013	1.003
	CBO	.230	.054	18.330	1	.000	1.259
	DIT	-.708	.221	10.317	1	.001	.492
	常量	-2.456	.454	29.235	1	.000	.086

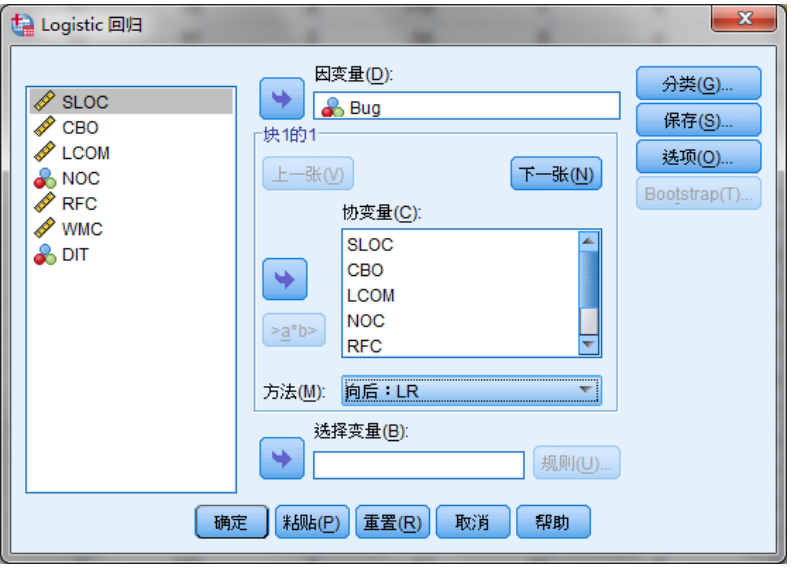
排除有影响的数据点后 → 再次前向逐步回归

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1[a]	CBO	.204	.037	30.772	1	.000	1.227
	常量	-2.228	.394	31.938	1	.000	.108
步骤 2[b]	CBO	.309	.054	32.395	1	.000	1.362
	DIT	-.808	.236	11.665	1	.001	.446
	常量	-2.384	.441	29.238	1	.000	.092
步骤 3[c]	SLOC	.004	.001	7.681	1	.006	1.004
	CBO	.242	.057	18.010	1	.000	1.273
	DIT	-.949	.253	14.117	1	.000	.387
	常量	-2.515	.472	28.429	1	.000	.081

Logistic回归模型

后向逐步回归(1)



Bug	COO_1
1	1.03987
1	.26296
0	.24188

方程中的变量

		B	S.E.	Wals	df	Sig.	Exp (B)
步骤 1[a]	SLOC	.004	.002	5.192	1	.023	1.004
	CBO	.268	.062	18.583	1	.000	1.307
	LCOM	-.014	.007	3.832	1	.050	.986
	NOC	-.359	.307	1.375	1	.241	.698
	RFC	.024	.017	2.111	1	.146	1.025
	WMC	-.037	.035	1.129	1	.288	.964
	DIT	-1.065	.353	9.113	1	.003	.345
	常量	-1.851	.536	11.927	1	.001	.157
步骤 2[a]	SLOC	.003	.001	4.674	1	.031	1.003
	CBO	.274	.062	19.735	1	.000	1.315
	LCOM	-.015	.007	4.427	1	.035	.985
	NOC	-.367	.302	1.477	1	.224	.693
	RFC	.011	.011	1.090	1	.296	1.011
	DIT	-.865	.285	9.234	1	.002	.421
	常量	-1.983	.525	14.263	1	.000	.138
步骤 3[a]	SLOC	.003	.001	7.098	1	.008	1.003
	CBO	.261	.060	19.169	1	.000	1.298
	LCOM	-.013	.007	3.764	1	.052	.987
	NOC	-.385	.297	1.680	1	.195	.680
	DIT	-.689	.224	9.446	1	.002	.502
	常量	-1.914	.517	13.720	1	.000	.148
步骤 4[a]	SLOC	.003	.001	6.957	1	.008	1.003
	CBO	.259	.059	19.241	1	.000	1.296
	LCOM	-.013	.007	3.653	1	.056	.987
	DIT	-.681	.224	9.273	1	.002	.506
	常量	-1.975	.510	15.019	1	.000	.139

a. 在步骤 1 中输入的变量: SLOC, CBO, LCOM, NOC, RFC, WMC, DIT.

双击以激活

Logistic回归模型

后向逐步回归(2)

步骤 4[a]	SLOC	.003	.001	6.957	1	.008	1.003
	CBO	.259	.059	19.241	1	.000	1.296
	LCOM	-.013	.007	3.653	1	.056	.987
	DIT	-.681	.224	9.273	1	.002	.506
	常量	-1.975	.510	15.019	1	.000	.139

排除有影响的数据点后
再次后向逐步回归

	常量	-1.988	.534	13.861	1	.000	.137
步骤 4[a]	SLOC	.004	.001	8.494	1	.004	1.004
	CBO	.272	.063	18.796	1	.000	1.313
	LCOM	-.013	.007	3.459	1	.063	.987
	DIT	-.925	.256	13.072	1	.000	.396
	常量	-2.048	.526	15.137	1	.000	.129

Logistic回归模型

前向和后向模型比较

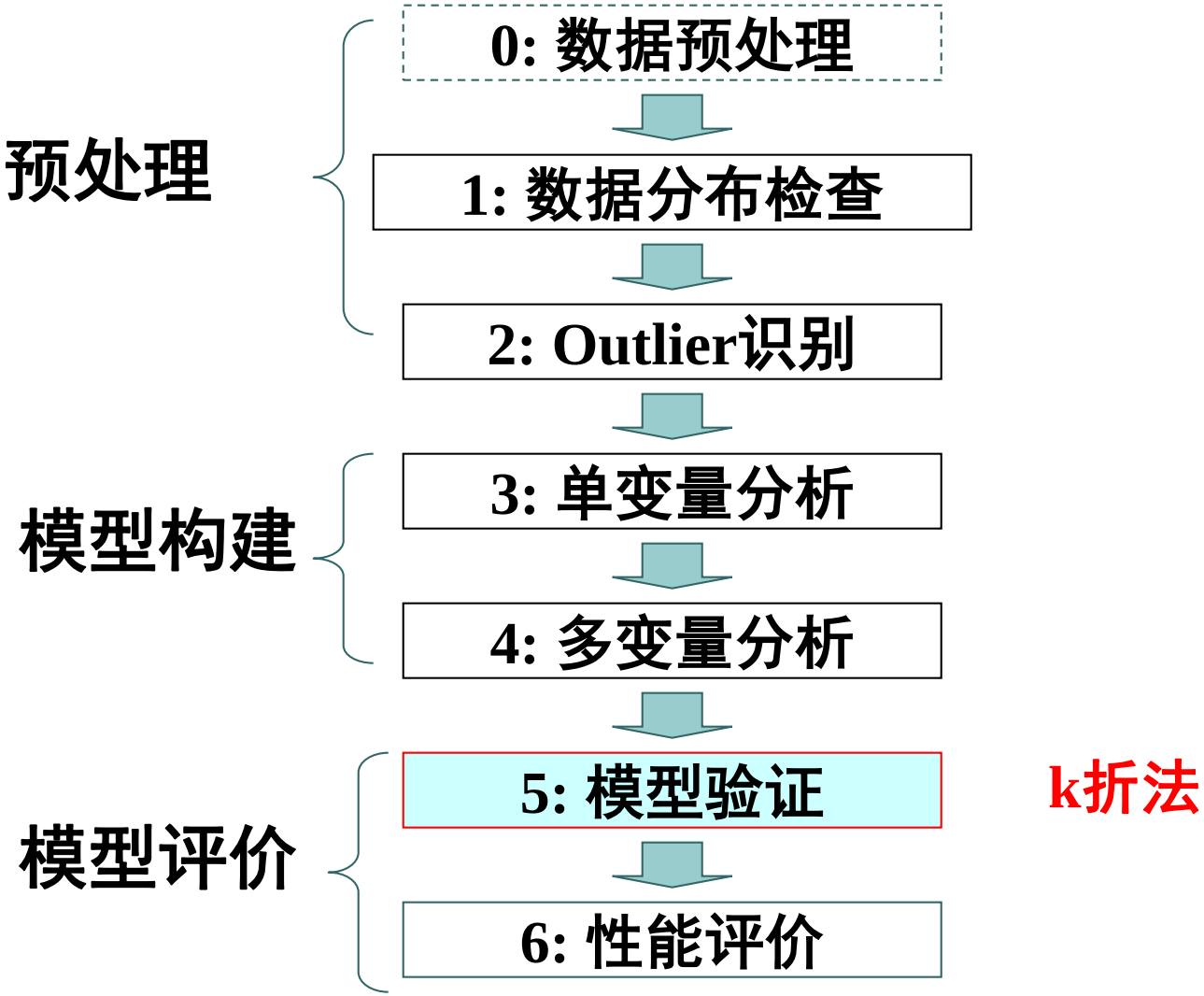
前向

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤 1[a]	CBO	.204	.037	30.772	1	.000	1.227
	常量	-2.228	.394	31.938	1	.000	.108
步骤 2[b]	CBO	.309	.054	32.395	1	.000	1.362
	DIT	-.808	.236	11.665	1	.001	.446
	常量	-2.384	.441	29.238	1	.000	.092
步骤 3[c]	SLOC	.004	.001	7.681	1	.006	1.004
	CBO	.242	.057	18.010	1	.000	1.273
	DIT	-.949	.253	14.117	1	.000	.387
	常量	-2.515	.472	28.429	1	.000	.081

后向

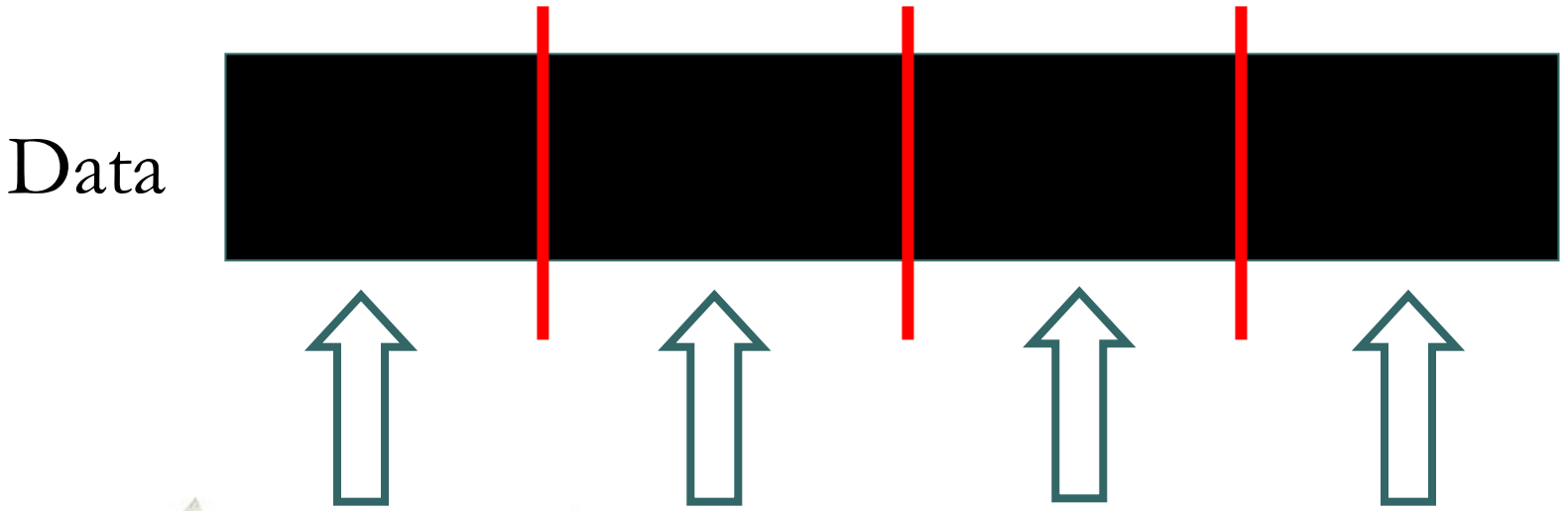
	常量	-1.988	.534	13.861	1	.000	.137
步骤 4[a]	SLOC	.004	.001	8.494	1	.004	1.004
	CBO	.272	.063	18.796	1	.000	1.313
	LCOM	-.013	.007	3.459	1	.063	.987
	DIT	-.925	.256	13.072	1	.000	.396
	常量	-2.048	.526	15.137	1	.000	.129


预处理
0: 数据预处理
1: 数据分布检查
2: Outlier识别
模型构建
3: 单变量分析
4: 多变量分析
模型评价
5: 模型验证
6: 性能评价
k折法

k-fold cross-validation

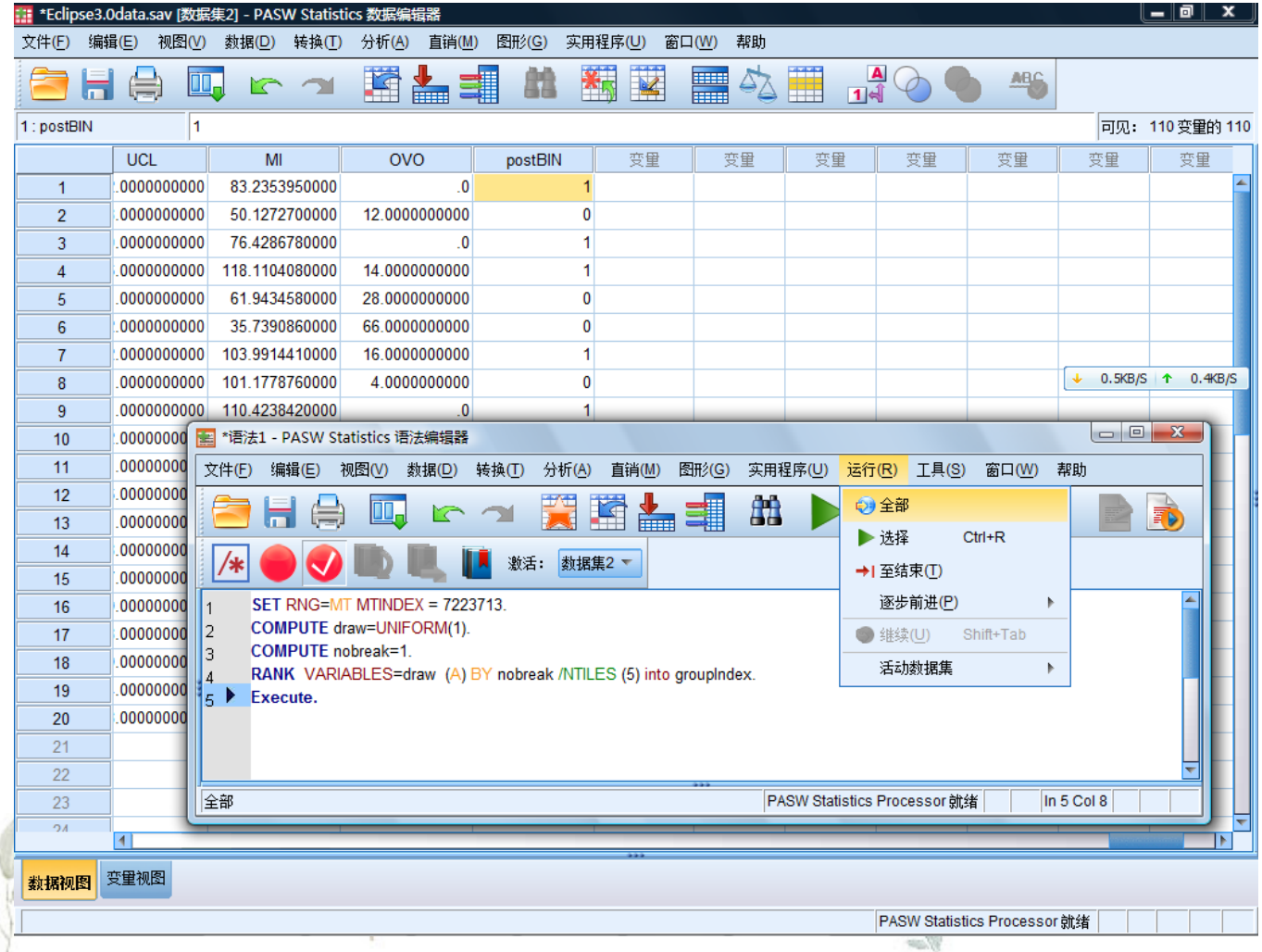
将数据集D分割成规模相等的k份;

```
for fold = 1 to k do{
    testData = D[fold]; trainData = D – D[fold];
    model = buildModel(trainData, ML);
    result [fold] = evaluate(model, testData);
}
```



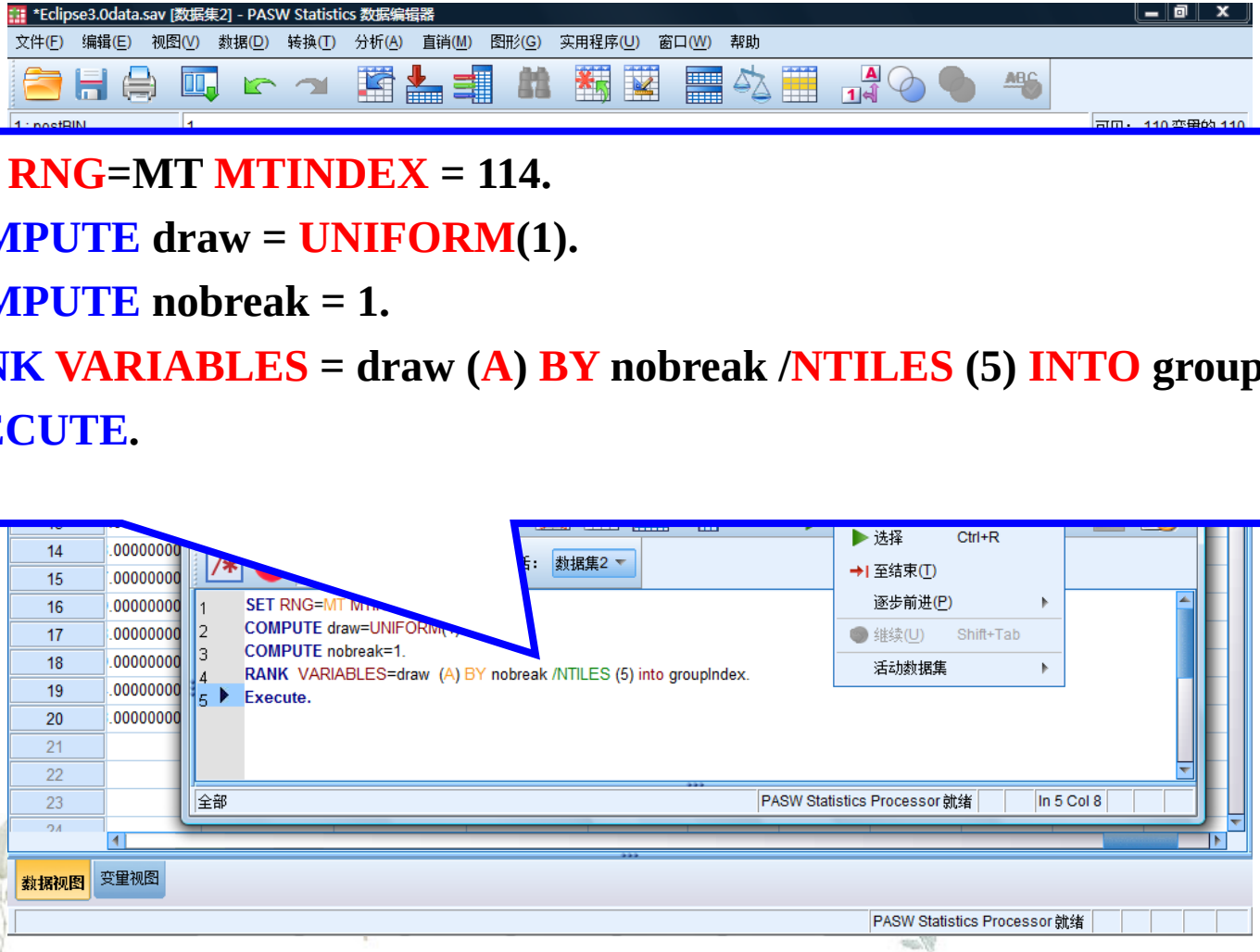
k-fold cross-validation

Example: 5-fold cross-validation



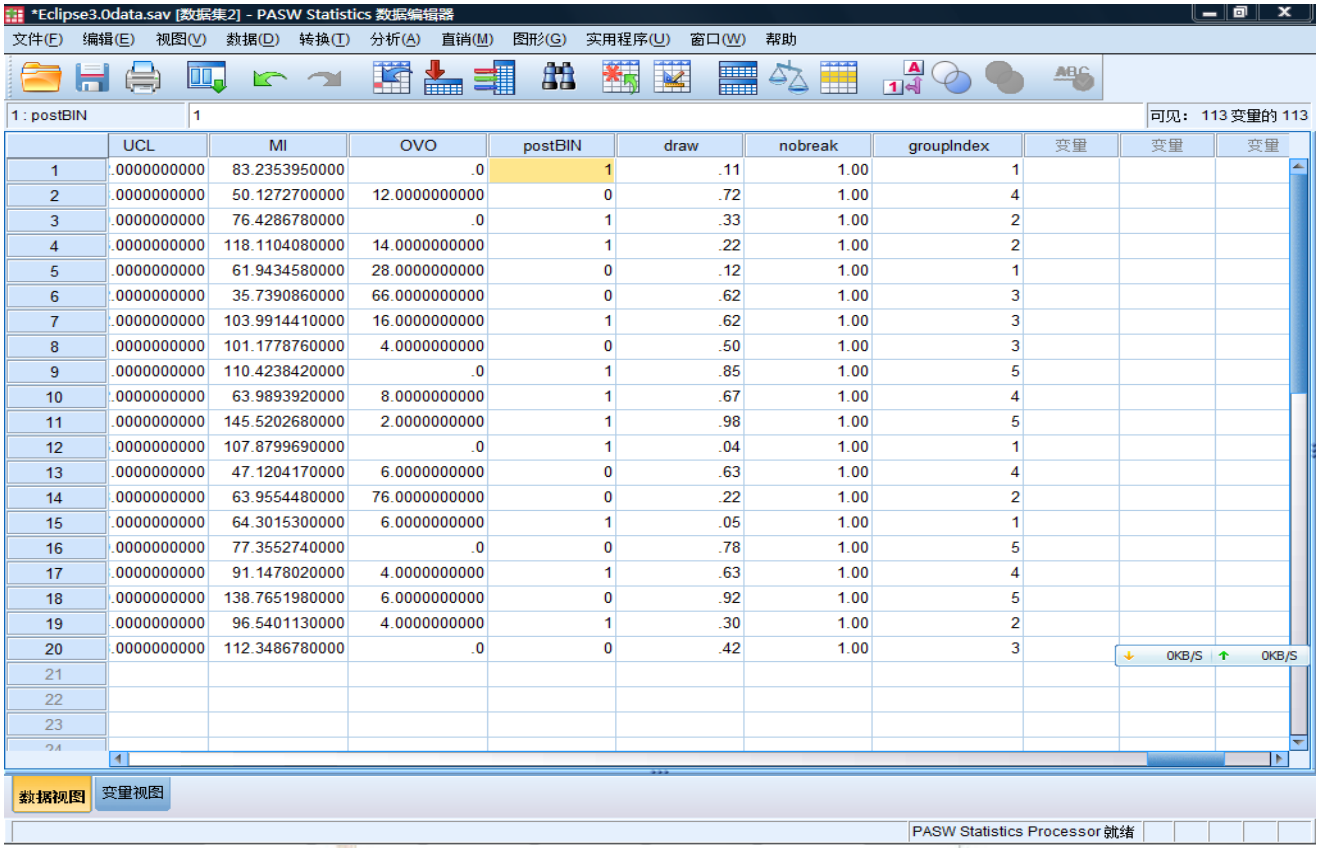
k-fold cross-validation

Example: 5-fold cross-validation



k-fold cross-validation

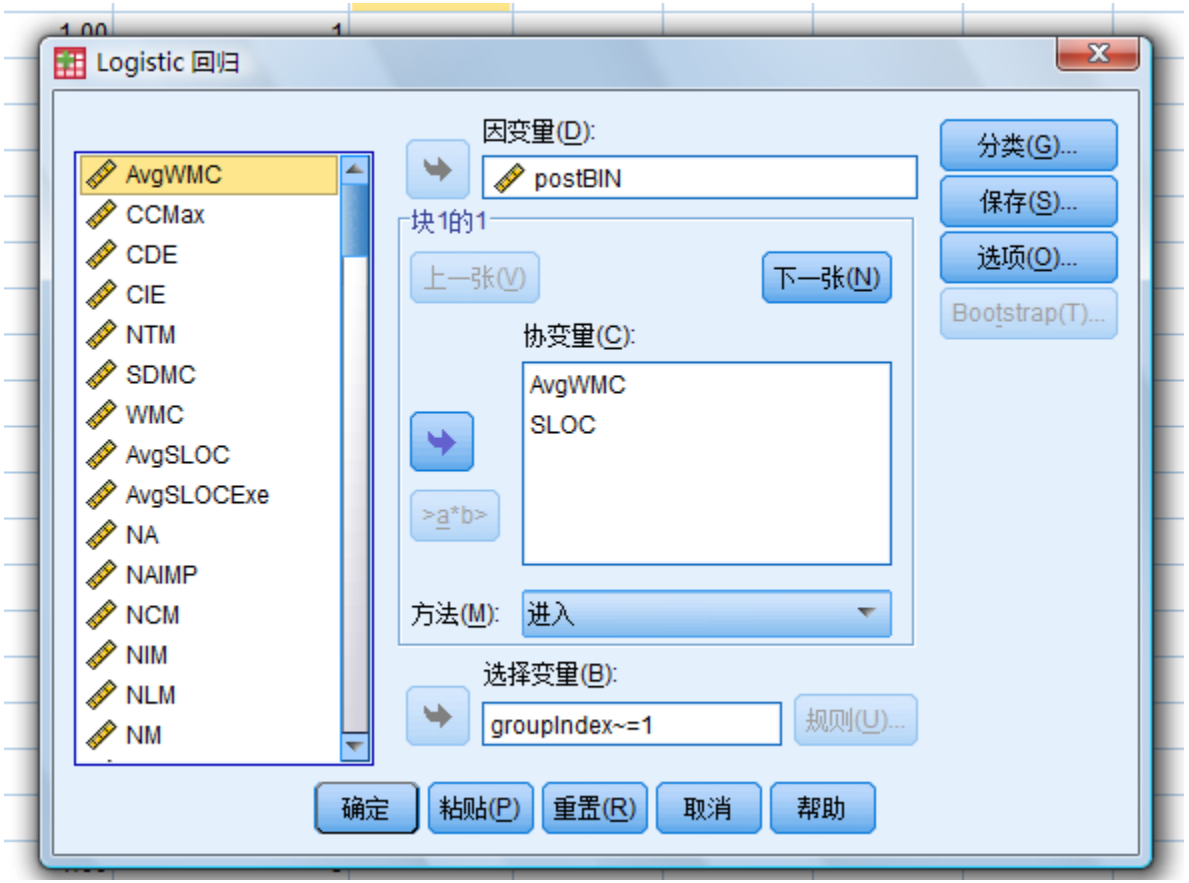
Example: 5-fold cross-validation



	UCL	MI	OVO	postBIN	draw	nobreak	groupIndex	变量	变量	变量
1	.0000000000	83.2353950000	.0	1	.11	1.00	1			
2	.0000000000	50.1272700000	12.0000000000	0	.72	1.00	4			
3	.0000000000	76.4286780000	.0	1	.33	1.00	2			
4	.0000000000	118.1104080000	14.0000000000	1	.22	1.00	2			
5	.0000000000	61.9434580000	28.0000000000	0	.12	1.00	1			
6	.0000000000	35.7390860000	66.0000000000	0	.62	1.00	3			
7	.0000000000	103.9914410000	16.0000000000	1	.62	1.00	3			
8	.0000000000	101.1778760000	4.0000000000	0	.50	1.00	3			
9	.0000000000	110.4238420000	.0	1	.85	1.00	5			
10	.0000000000	63.9893920000	8.0000000000	1	.67	1.00	4			
11	.0000000000	145.5202680000	2.0000000000	1	.98	1.00	5			
12	.0000000000	107.8799690000	.0	1	.04	1.00	1			
13	.0000000000	47.1204170000	6.0000000000	0	.63	1.00	4			
14	.0000000000	63.9554480000	76.0000000000	0	.22	1.00	2			
15	.0000000000	64.3015300000	6.0000000000	1	.05	1.00	1			
16	.0000000000	77.3552740000	.0	0	.78	1.00	5			
17	.0000000000	91.1478020000	4.0000000000	1	.63	1.00	4			
18	.0000000000	138.7651980000	6.0000000000	0	.92	1.00	5			
19	.0000000000	96.5401130000	4.0000000000	1	.30	1.00	2			
20	.0000000000	112.3486780000	.0	0	.42	1.00	3			
21										
22										
23										

k-fold cross-validation

Example: 5-fold cross-validation



k-fold cross-validation

Example: 5-fold cross-validation

案例处理汇总

未加权的案例[a]		N	百分比
选定案例	包括在分析中	16	80.0
	缺失案例	0	.0
	总计	16	80.0
未选定的案例		4	20.0
总计		20	100.0

a. 如果权重有效，请参见分类表以获得案例总

分类表[c]

已观测			已预测					
			选定案例[a]			未选定的案例[b]		
			postBIN			postBIN		
			0	1	百分比校正	0	1	百分比校正
步骤 1	postBIN	0	4	4	50.0	0	1	.0
		1	3	5	62.5	1	2	66.7
	总计百分比				56.3			50.0

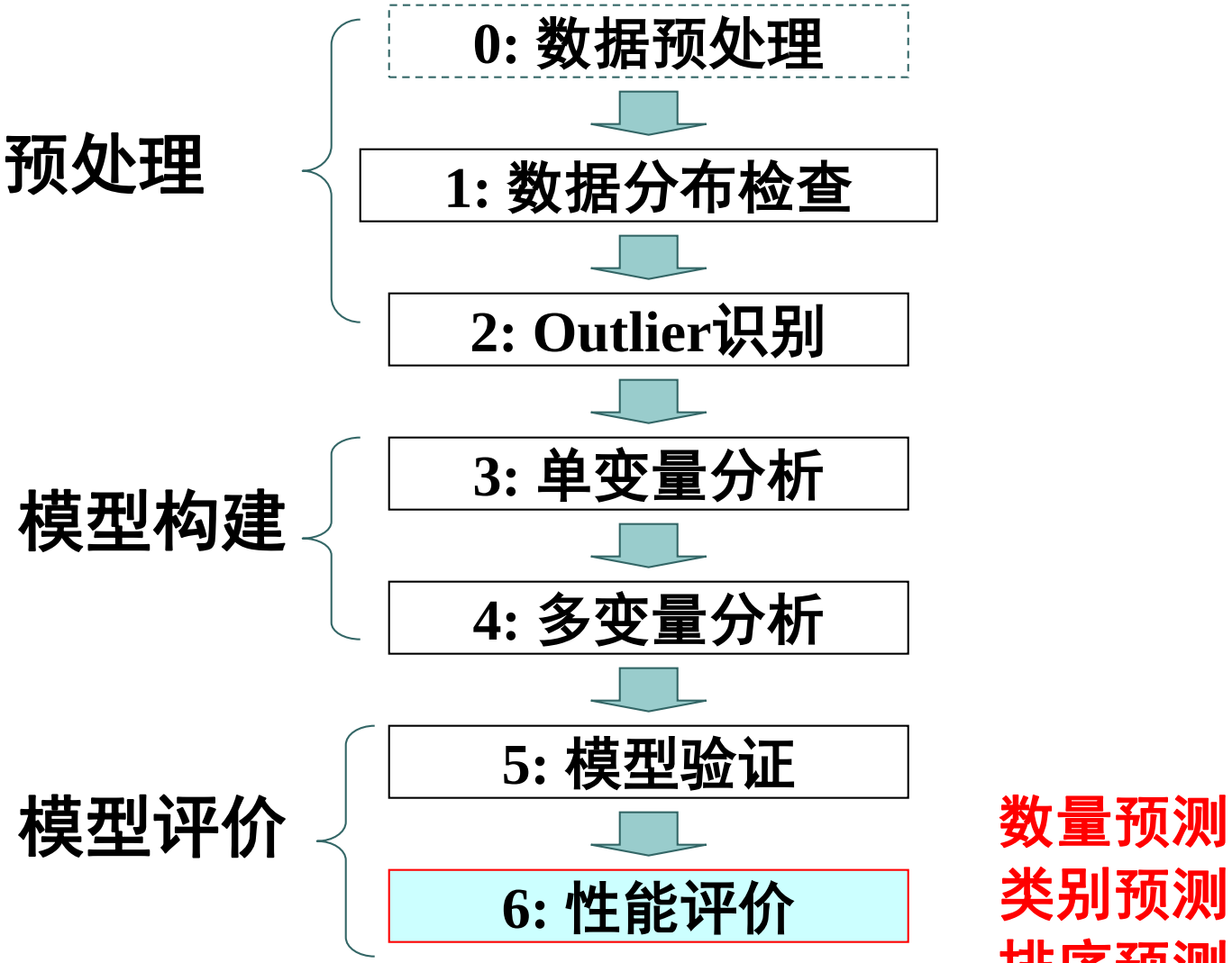
a. 已选定的案例 groupIndex NE 1

b. 未选定的案例 groupIndex EQ 1

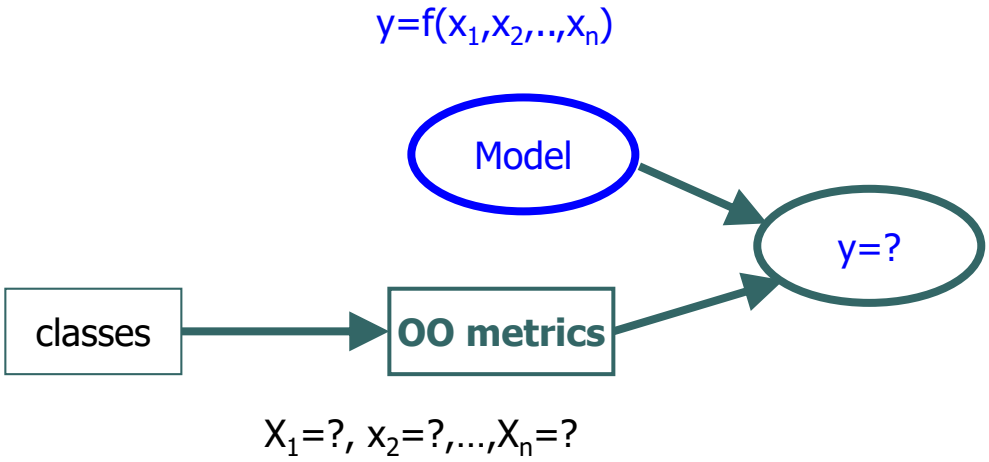
c. 切割值为 .500

10 times 10-fold cross-validation

```
for run = 1 to 10 do {
   D' = randomize(D);
   D'' = strategyDivide(D', 10);
   for fold = 1 to 10 do{
      index = (run – 1)*10 +  fold;
      testData = D''[fold]; trainData = D'' – D''[fold];
      model = buildModel(trainData, ML);
      result [index] = evaluate(model, testData);
    }
}
```


预处理
模型构建
模型评价
0: 数据预处理
1: 数据分布检查
2: Outlier识别
3: 单变量分析
4: 多变量分析
5: 模型验证
6: 性能评价
数量预测
类别预测
排序预测

数量预测



For each case in the data set:

Magnitude of Relative Error (MRE)

$$\mathrm{MRE}_i = \frac{|y_i - \hat{y}_i|}{y_i}$$

数量预测

For the model, compare actual and estimated quantity for n cases in the dataset:

Mean Magnitude of Relative Error (MMRE)

$$\text{MMRE} = \frac{1}{n}\sum_{i=1}^{n} MRE_i$$

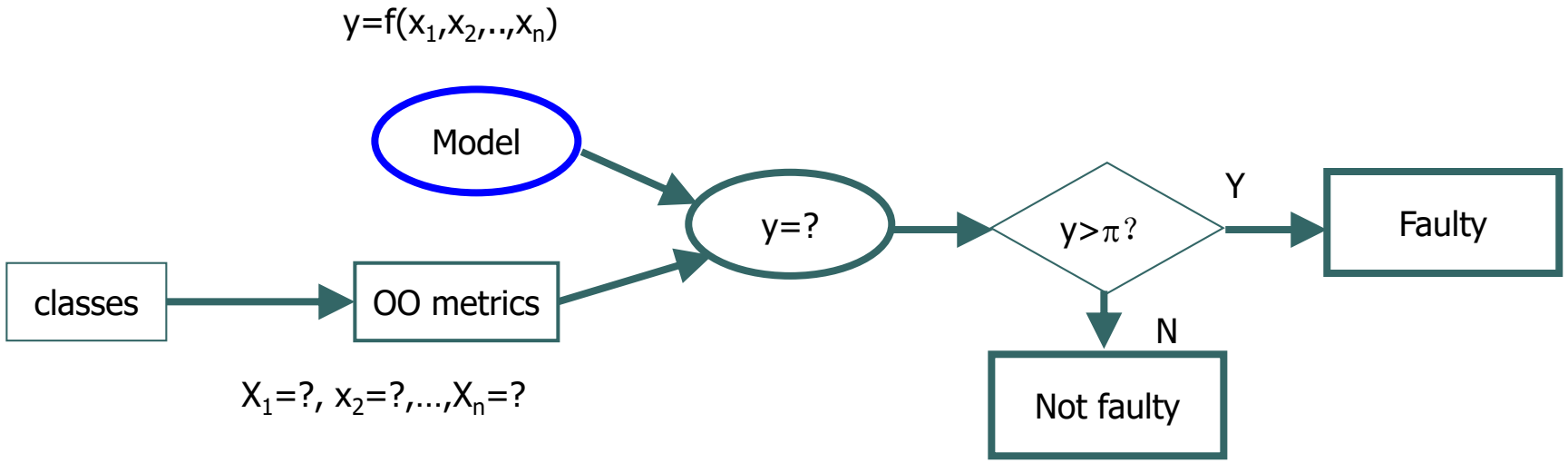
Prediction level

$$\text{Pred}(q) = \frac{k}{n}$$

where k = the number cases in a set of n cases whose MRE <= q

good: $\text{Pred}(0.25) \geq 0.75$

类别预测



Actual	Predicted	
	fault(y>π)	No fault(y≤π)
fault	**a**	**c**
No fault	**b**	**d**

a: # True Positives (TP)

b: # False Positives (FP)

c: # False Negatives (FN)

d: # True Negatives (TN)

类别预测

Actual	Predicted	
	Fault($y>\pi$)	No fault($y\leq\pi$)
fault	**a**	**c**
No fault	**b**	**d**

Disadvantage:
depend on π

Sensitivity = a/(a+c) = TP/(TP+FN) = **Recall**

Specificity = d/(b+d) = TN/(FP+TN)

Precision = a/(a+b) = TP/(TP+FP)

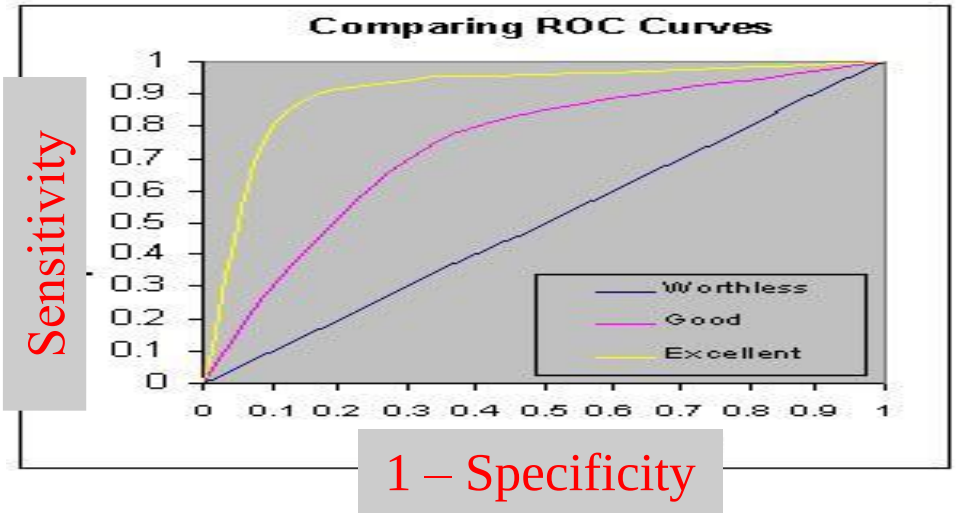
Accuracy = (a+d)/(a+b+c+d)

= (TP+TN)/(TP+FP+FN+TN)

F-measure = 2 * Recall * Precision / (Recall + Precision)

类别预测

AUC (area under ROC, Receiver Operating Characteristic curves)



poor:	[0.5, 0.7)
moderate :	[0.7, 0.9)
very good :	[0.9, 1.0]

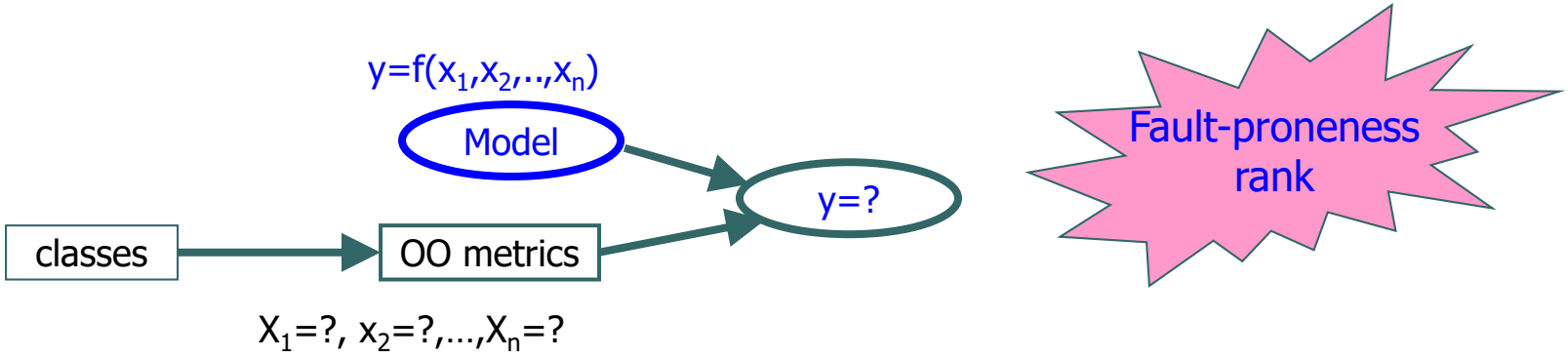
Advantages:

① Does not depend on the threshold π

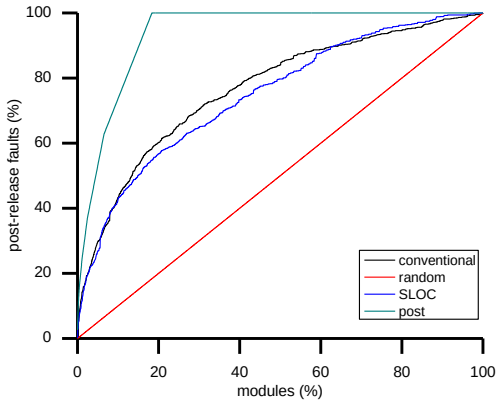
② Does not depend on the prior probabilities of positive and negative cases

③ can be interpreted as the probability that a randomly chosen positive observation is (correctly) rated or ranked with greater suspicion than a randomly chosen negative observation

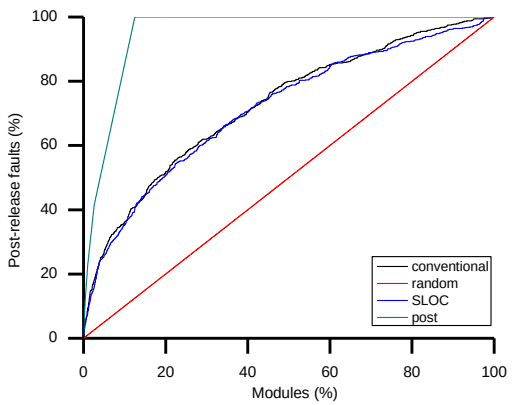
排序预测



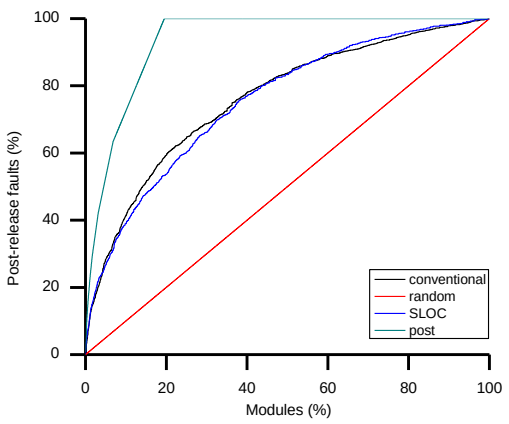
Alberg diagram: x is modules%, y is faults % (if top x% modules are selected to be tested/inspected, y% faults will be found)



Eclipse 2.0

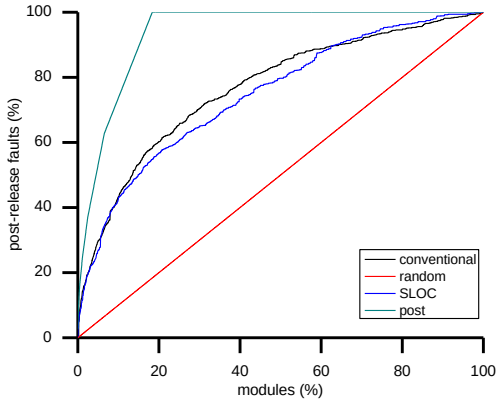


Eclipse 2.1



Eclipse 3.0

排序预测

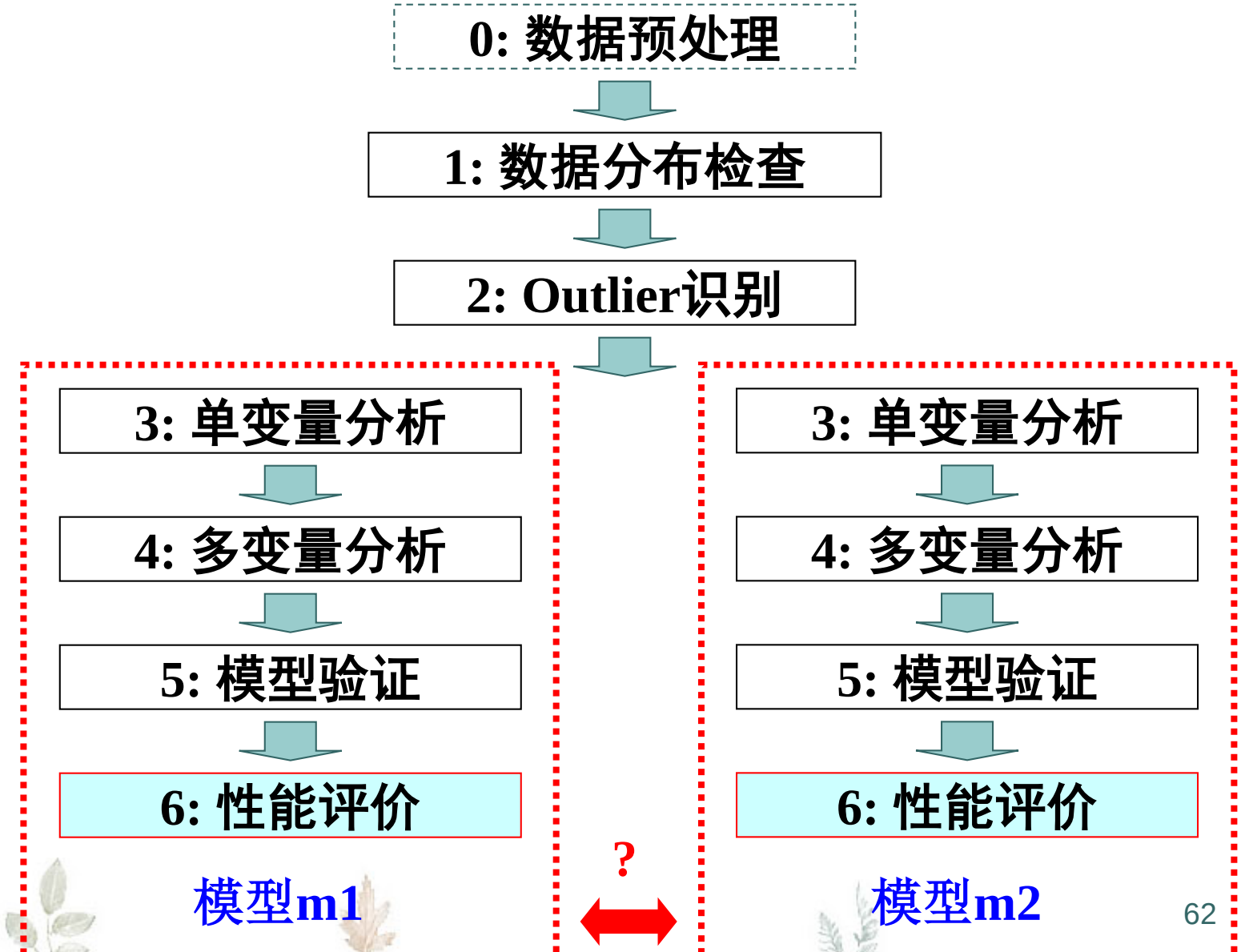


Eclipse 2.0

Cost effectiveness:

$$CE_{\pi}(\text{model}) = \frac{Area_{\pi}(\text{model}) - Area_{\pi}(\text{Random})}{Area_{\pi}(\text{optimal}) - Area_{\pi}(\text{Random})}$$

哪种预测模型效果好？


0: 数据预处理
1: 数据分布检查
2: Outlier识别
3: 单变量分析
4: 多变量分析
5: 模型验证
6: 性能评价
模型m1
?
3: 单变量分析
4: 多变量分析
5: 模型验证
6: 性能评价
模型m2

哪种预测模型效果好？

Given model A and model B, the problem is:

Model A is **statistically significantly better** than model B?

If model A and model B are validated on the same data sets, we use:

- ✓ paired t-test
- ✓ Wilcoxon Signed-Rank Test (paired)

哪种预测模型效果好？

Paired Sample t-test

$H_0 : \mu_d = \mu_0$
$H_1 : \mu_d \neq \mu_0$ (two-tailed).
μ_d: mean of population differences.
α: significant level (e.g., 0.05).
Test Statistic:

$$T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}}$$

$\bar{d}$: average of sample differences.

S_d: standard deviation of sample difference

n: number of pairs.

- Reject H_0 if $|t_d| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ_d: $\bar{d} - t_{\alpha/2}S/\sqrt{n} \leq \mu_d < \bar{d} + t_{\alpha/2}S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_d)$, $\mathbf{T} \sim t_{n-1}$.

哪种预测模型效果好？

Wilcoxon Signed-Rank Test (paired)

- Null hypothesis: the population median from which both samples were drawn is the same.
- The sum of the ranks for the "positive" (up-regulated) values is calculated and compared against a precomputed table to a p-value.
  - Sorting the absolute values of the differences from smallest to largest.
  - Assigning ranks to the absolute values.
  - Find the sum of the ranks of the positive differences.
- If the null hypothesis is true, the sum of the ranks of the positive differences should be about the same as the sum of the ranks of the negative differences

Pair	Before	After	Diff.	Rank
1	89	73	16	15.5
2	83	77	6	7
3	80	58	22	17
4	72	77	−5	5
5	77	70	7	8
6	74	62	12	13.5
7	69	67	2	2
8	65	68	−3	3
9	60	44	16	15.5
10	55	50	5	5
11	54	46	8	9.5
12	50	38	12	13.5
13	42	47	−5	5
14	48	40	8	9.5
15	44	43	1	1
16	38	29	9	11
17	36	25	11	12

The Wilcoxon signed-rank Test:

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$
$T = \min\{\sum_+ \text{Rank}, \sum_- \text{Rank}\}$
At $\alpha = 0.01$, two-tailed test,
reject H_0 if $T \leq 23$ when $N = 17$.
(Table)

(The zero difference is ignored when assigning ranks. $N_{new} = N_{old} - \#\{ties\}$)

$T = \min\{\sum_+ \text{Rank} = 140, \sum_- \text{Rank} = 13\}$
$= 13$

The obtained T=13 is less than the critical value 23, so we reject H_0.

```
p =

    0.0026

h =

     1

stats =

          zval: -3.0089
    signedrank: 13
```

哪种预测模型效果好？

Assumptions of paired t-test

- For paired t-test, it is the distribution of the subtracted data that must be normal

Assumptions of Wilcoxon signed-rank test

- Do not assume that the data is normally distributed.
- Non-parametric methods are robust to outliers and noisy data

Introduction to hypothesis testing

思考题1

对于给定的两个模型A和B，paired t-test或者Wilcoxon signed-rank test只能说明A和B在统计上是否存在显著差别

如何确定A和B在实用上是否存在重要差别？

思考题2

对于给定的两个模型A和B，通常是在一个数据集上

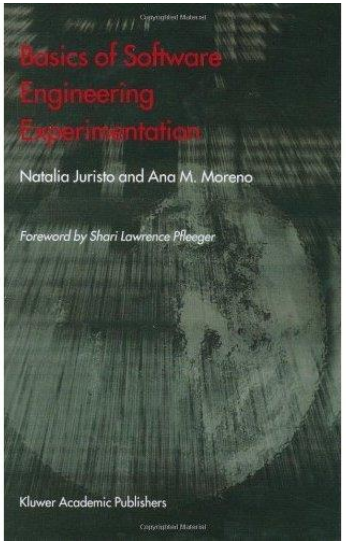
应用paired t-test或者Wilcoxon signed-rank test

来确定A和B在统计上是否存在显著差别

如果在k个数据集上比较m个模型，如何比较这m个模型？

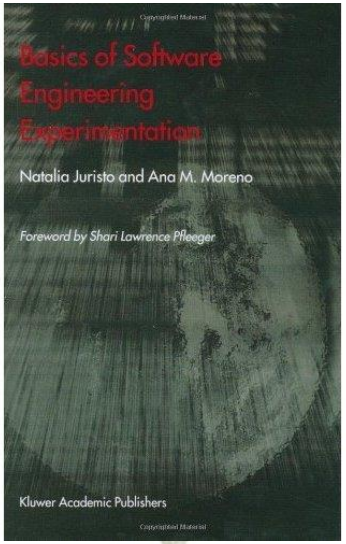
Section 2
Analyzing data for controlled experiments

- One-factor with 2 alternatives
- One-factor with k alternatives
- Analysis for block designs
- …



7 WHICH IS THE BETTER OF TWO ALTERNATIVES?

ANALYSIS OF ONE-FACTOR DESIGNS WITH TWO ALTERNATIVES



8 WHICH OF K ALTERNATIVES IS THE BEST?

ANALYSIS FOR ONE-FACTOR DESIGNS AND K ALTERNATIVES

9 EXPERIMENTS WITH UNDESIRED VARIATIONS:

ANALYSIS FOR BLOCK DESIGNS

10 BEST ALTERNATIVES FOR MORE THAN ONE VARIABLE
ANALYSIS FOR FACTORIAL DESIGNS

11 EXPERIMENTS WITH INCOMPARABLE FACTOR ALTERNATIVES
ANALYSIS FOR NESTED DESIGNS

12 FEWER EXPERIMENTS
ANALYSIS FOR FRACTIONAL FACTORIAL DESIGNS

13 SEVERAL DESIRED AND UNDESIRED VARIATIONS
ANALYSIS FOR FACTORIAL BLOCK DESIGNS

14 NON-PARAMETRIC ANALYSIS METHODS

15 HOW MANY TIMES SHOULD AN EXPERIMENT BE REPLICATED?

Summary

- 通用的数据分析过程
- 受控实验结果的分析

Further reading

- Basics of Software Engineering Experimentation. Natalia Juristo, Ana M. Moreno, Kluwer Academic Publishers, 2001
- Y. Jiang, et al. Techniques for evaluating fault prediction models. Empirical Software Engineering, 13(5), 2008: 561-595
- P.C. Prati. A survey on graphical methods for classification predictive performance evaluation. IEEE TKDE, 2011.

Thanks for your time and attention!

练习：缺陷数据集分析

(1) R介绍

(2)下载xalan2.4数据集:

https://zenodo.org/record/268436/files/xalan-2.4.csv

(3) 编写一个R脚本，分析 wmc, dit, noc，cbo, rfc和lcom的缺陷预测能力

a. 描述性统计：最小值、25%处值、中位值、75%处值、最大值、平均值、偏度(skewness)和峰度(kurtosis)

b. 与bug数据的相关系数: Spearman和Pearson相关系数以及统计显著性

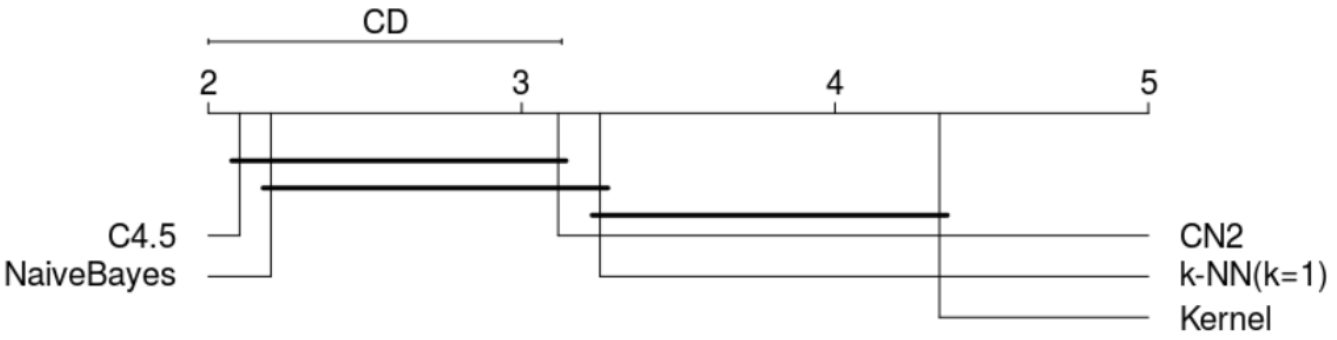
练习：缺陷数据集分析

(3) 编写一个R脚本，…

c. 利用Naïve Bayes和logistic回归等10种机器学习方法建立多变量的缺陷预测模型，不需要进行特征选择

d. 利用10x10交叉验证方法评价上述缺陷预测模型的性能，包括分类性能(评价指标为AUC)和排序性能(评价指标为CE)

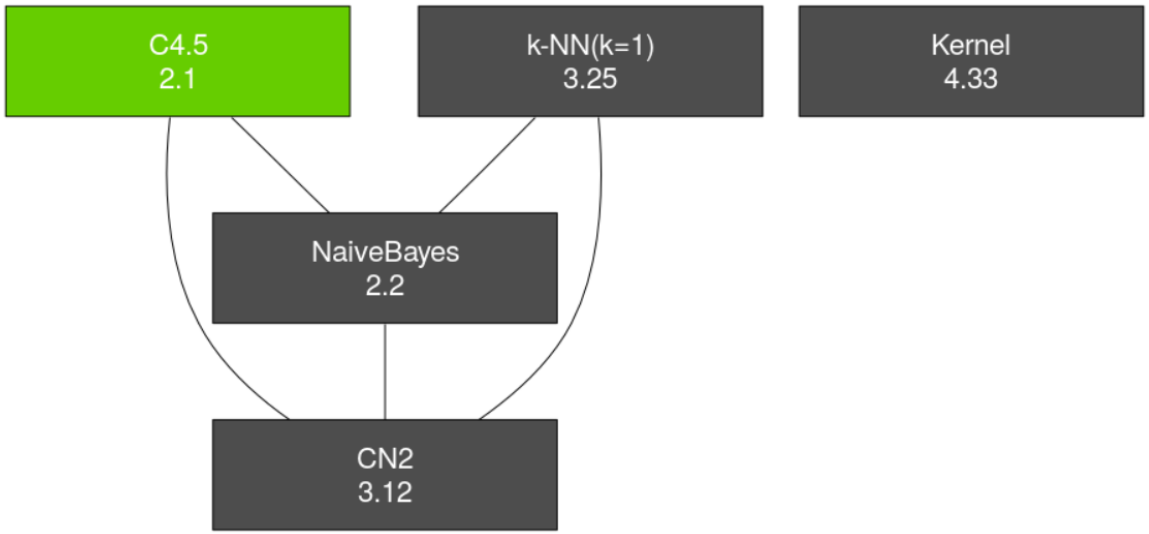
e. 利用CD图比较这10种模型在统计上的差别(plotCD)



练习：缺陷数据集分析

(3) 编写一个R脚本，…

f. 利用Algorithm图比较这10种模型在统计上的差别 (drawAlgorithmGraph)

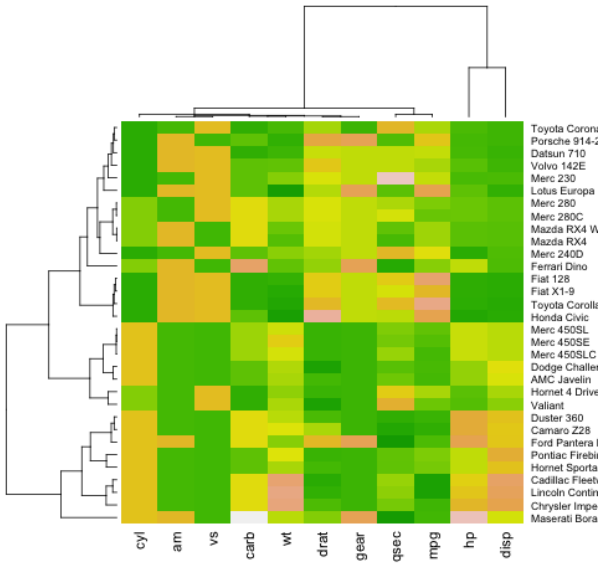


https://cran.r-project.org/web/packages/scmamp/vignettes/Statistical_assessment_of_the_differences.html

练习：缺陷数据集分析

(3) 编写一个R脚本，…

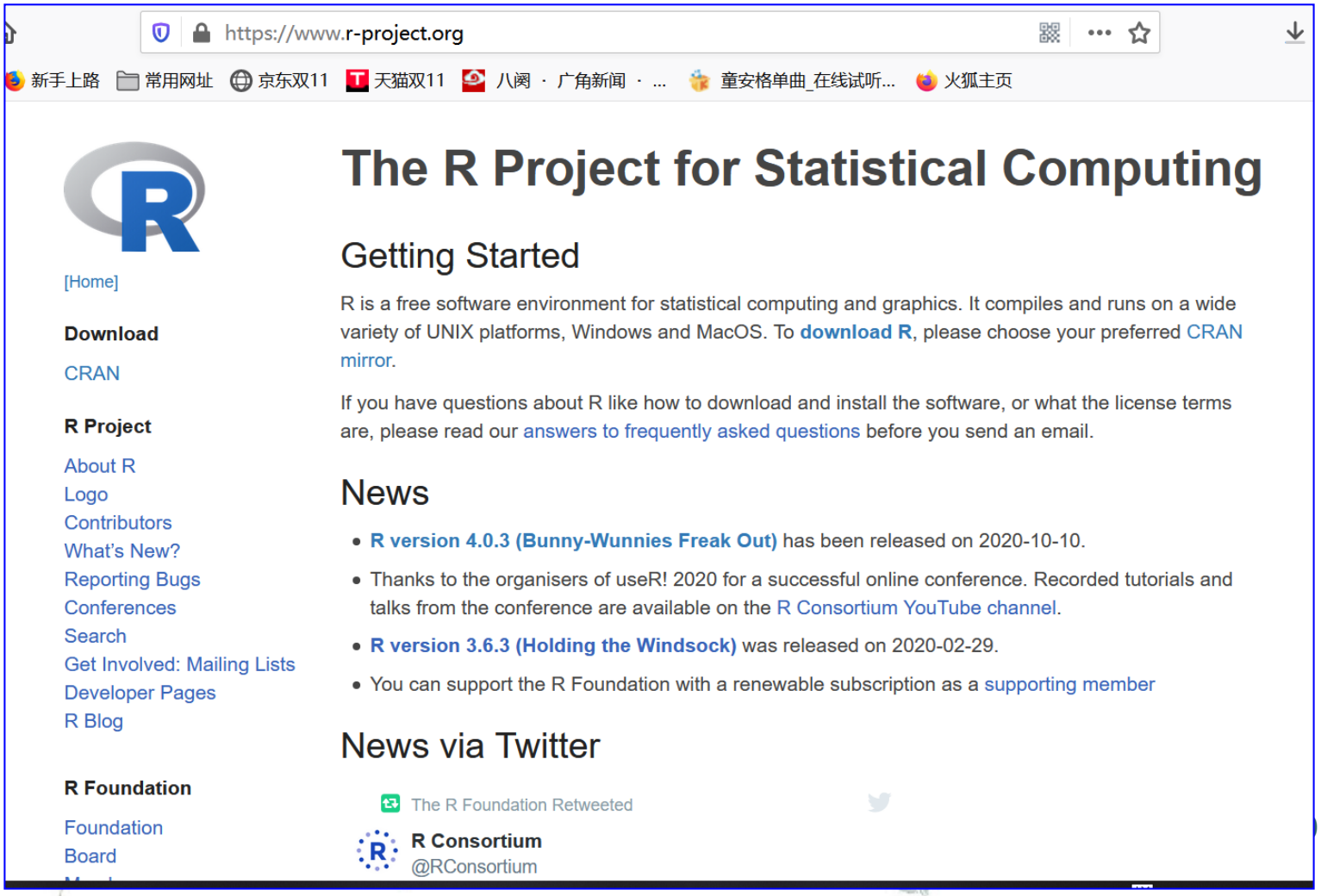
g. 用heatmap展示10个模型在100个测试集上的结果(行为模型，列为100个测试集上的结果



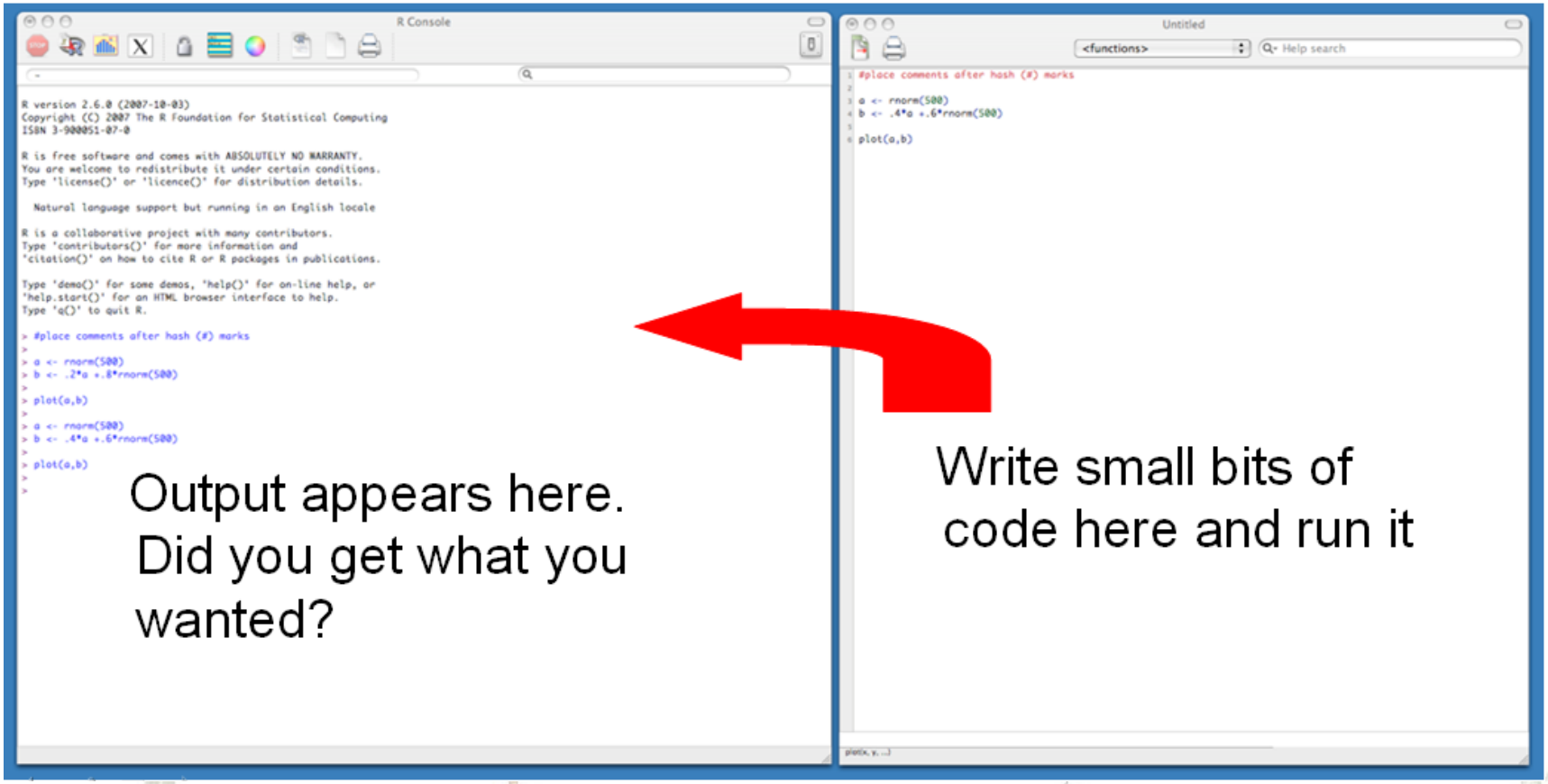
在11月30日前提交：R脚本以及报告

R

http://www.r-project.org/

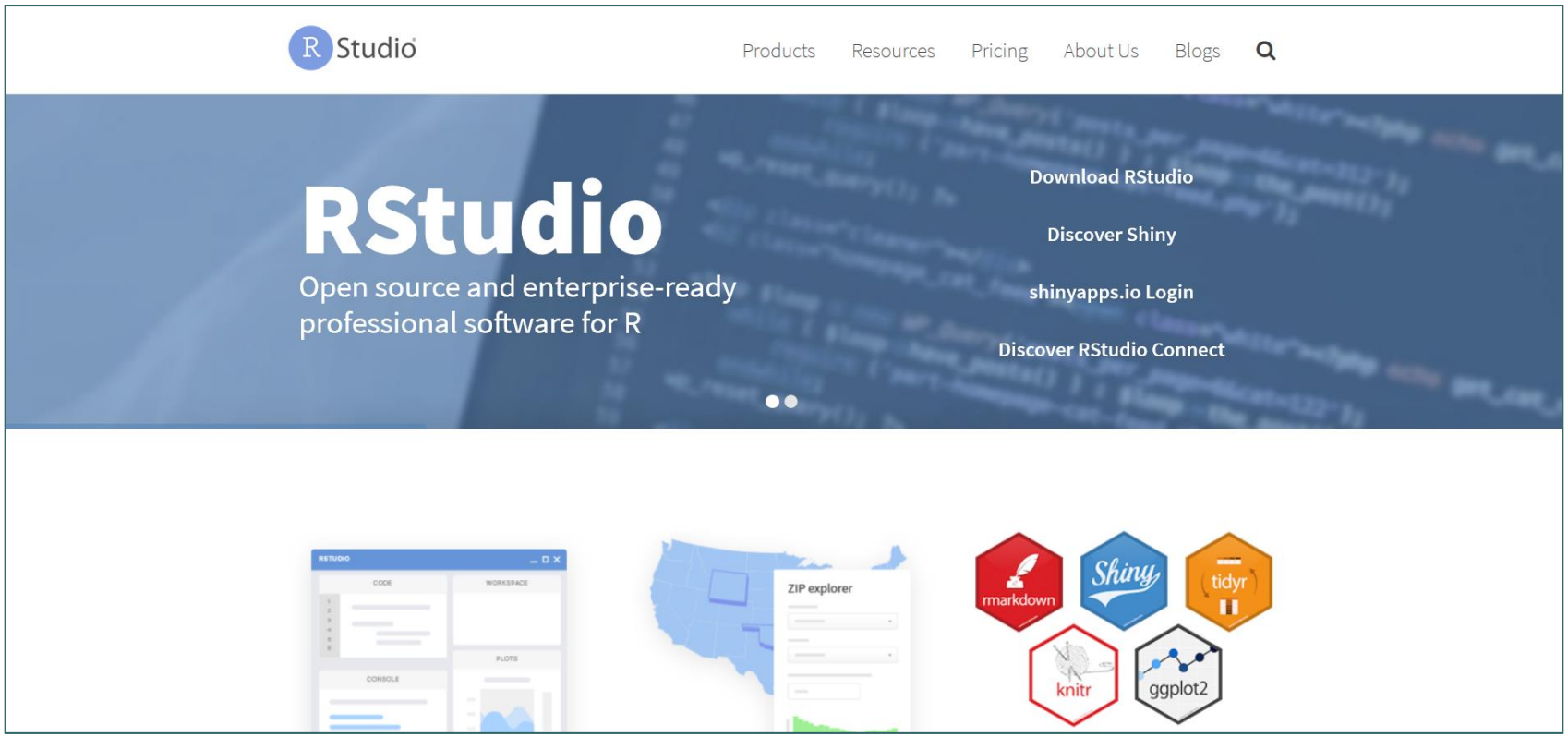


A very brief introduction to R



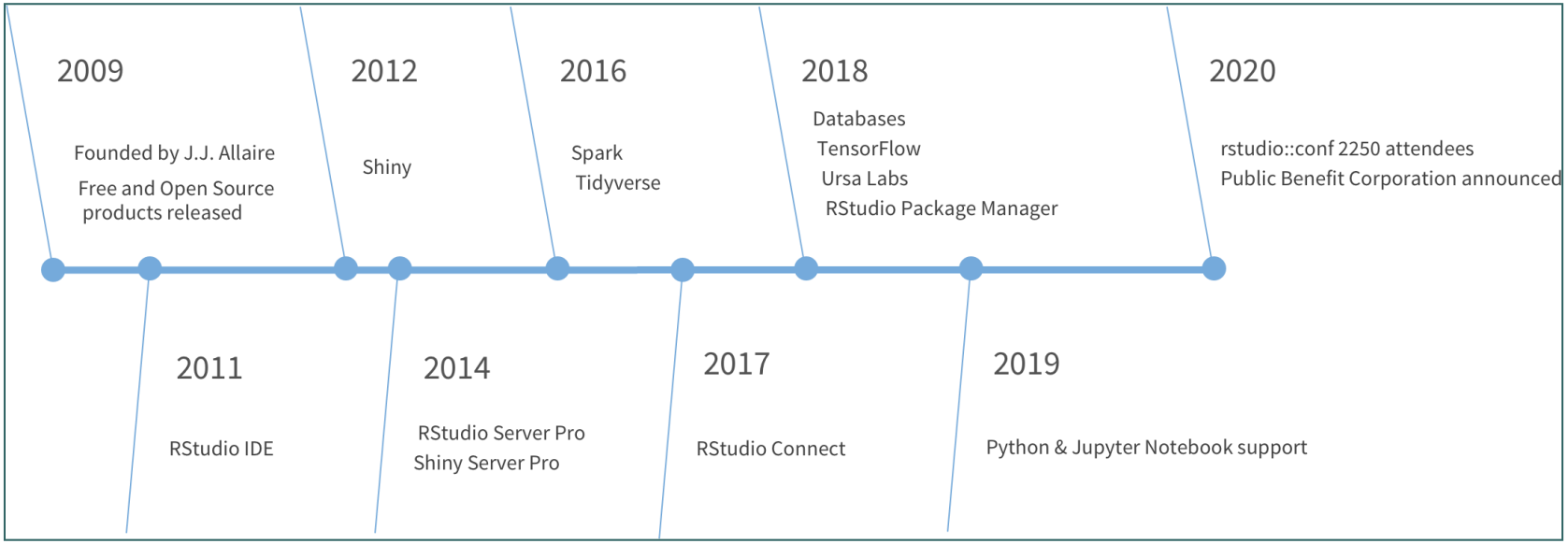
RStudio：R集成开发环境

http://www.rstudio.com/



RStudio：R集成开发环境

http://www.rstudio.com/



RStudio：R集成开发环境

http://www.rstudio.com/