

# C -- 题

## 颜色与物质浓度辨识

崔 恒 建

首都师范大学

昆明，2017.11.25



- 甲醛测量，试纸读颜色
  - 随着照相技术和颜色分辨率的提高，希望建立颜色读数和物质浓度的数量关系
- 只要输入照片中颜色读数就能够获得待测物质浓度

# 二氧化硫

0 水

10

20

40

60

100



- 试根据附件所提供的有关颜色读数和物质浓度数据，请你完成下列问题：
- 1. 附件**Data1.xls**中分别给出了**5**种物质的在不同浓度下的颜色读数，讨论从这**5**组数据能否确定颜色读数和物质浓度之间的关系，并给出一些准则来评价这**5**组数据的优劣。
- 2. 对附件**Data2.xls**中所给数据，建立颜色读数和物质浓度的数学模型，并给出模型的误差分析。
- 3. 探讨数据量和颜色维度对模型的影响。

- 希望用颜色预测浓度构建模型：

$$Y \sim f(R, G, B, H, S)$$

这里  $f$  一般未知，根据机理，可对每个变量单调。



## 思路:

- 用 Data1 探索建模方法（数据质量评估）
- 用 Data2 验证上述建模方法
- 影响建模的变量选择与分析



1. 建模过程就是选择  $f$  的过程。  
通常  $f$  选择的类型（可用于预测）：

参数或半参数函数形式，如：

$$f = f_0(b_0 + b_1R + b_2G + b_3B + b_4S + b_5H)$$

$f_0$  单调，S型。

- 线性模型（大多数学生用）
- 线性模型的单调变换（建议）

$$Y \sim f_0(b_0 + b_1R + b_2G + b_3B + b_4S + b_5H)$$

广义线性模型（ $f_0$  形式已知，  
如 **logistic** 变换等）



## 2. 数据质量评估（新）：

数据质量是建模的基础，本题主要考虑  
基于模型的

误差分析与评判：

 残差图、MSE、MSCV、 $R^2$

 异常或离群点识别：

 3sigma准则、Boxplot



### 3. 样本大小和颜色维数对模型的影响。

逐步回归，变量选择

误差分析与比较

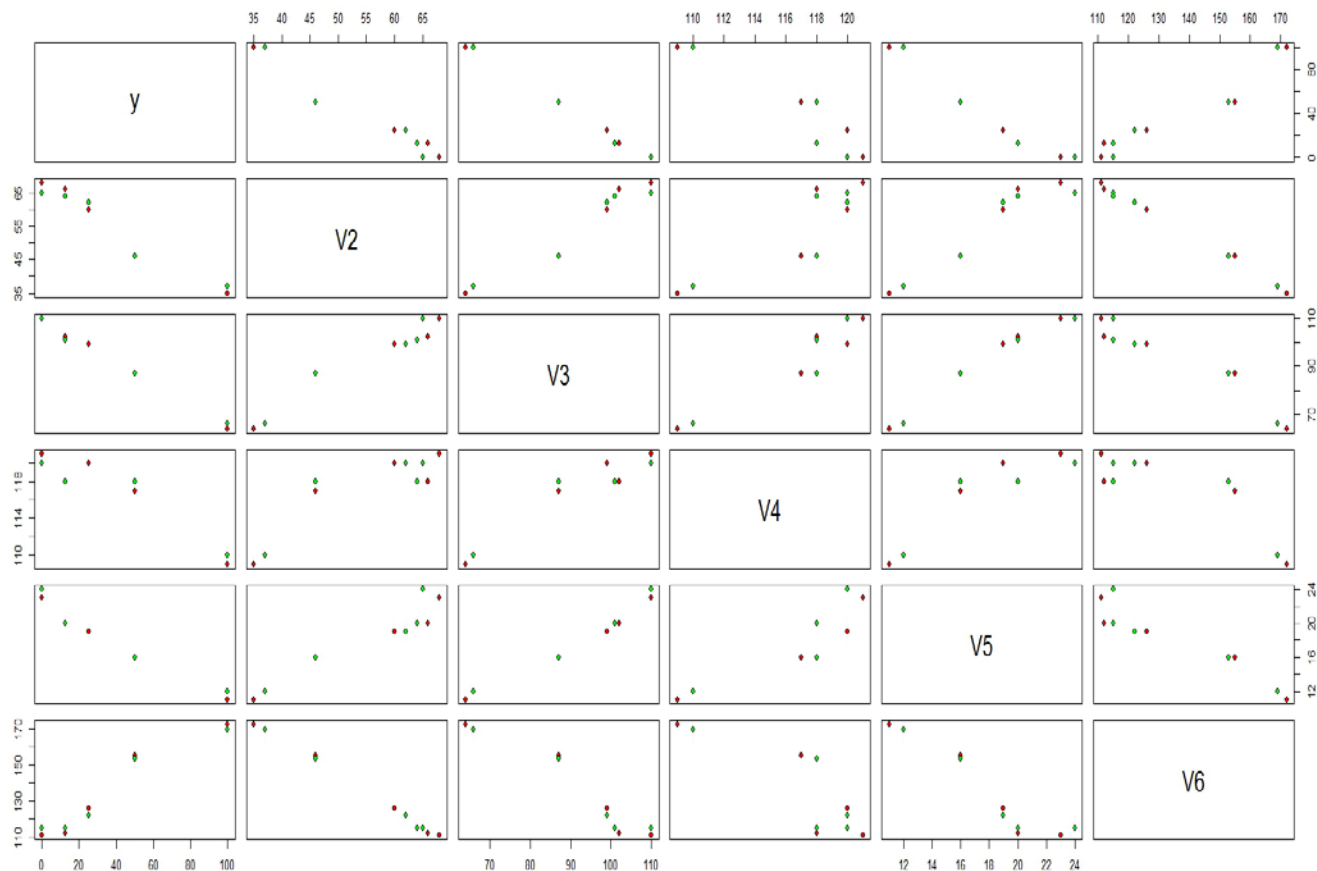
### 4. 使用R、Matlab语言



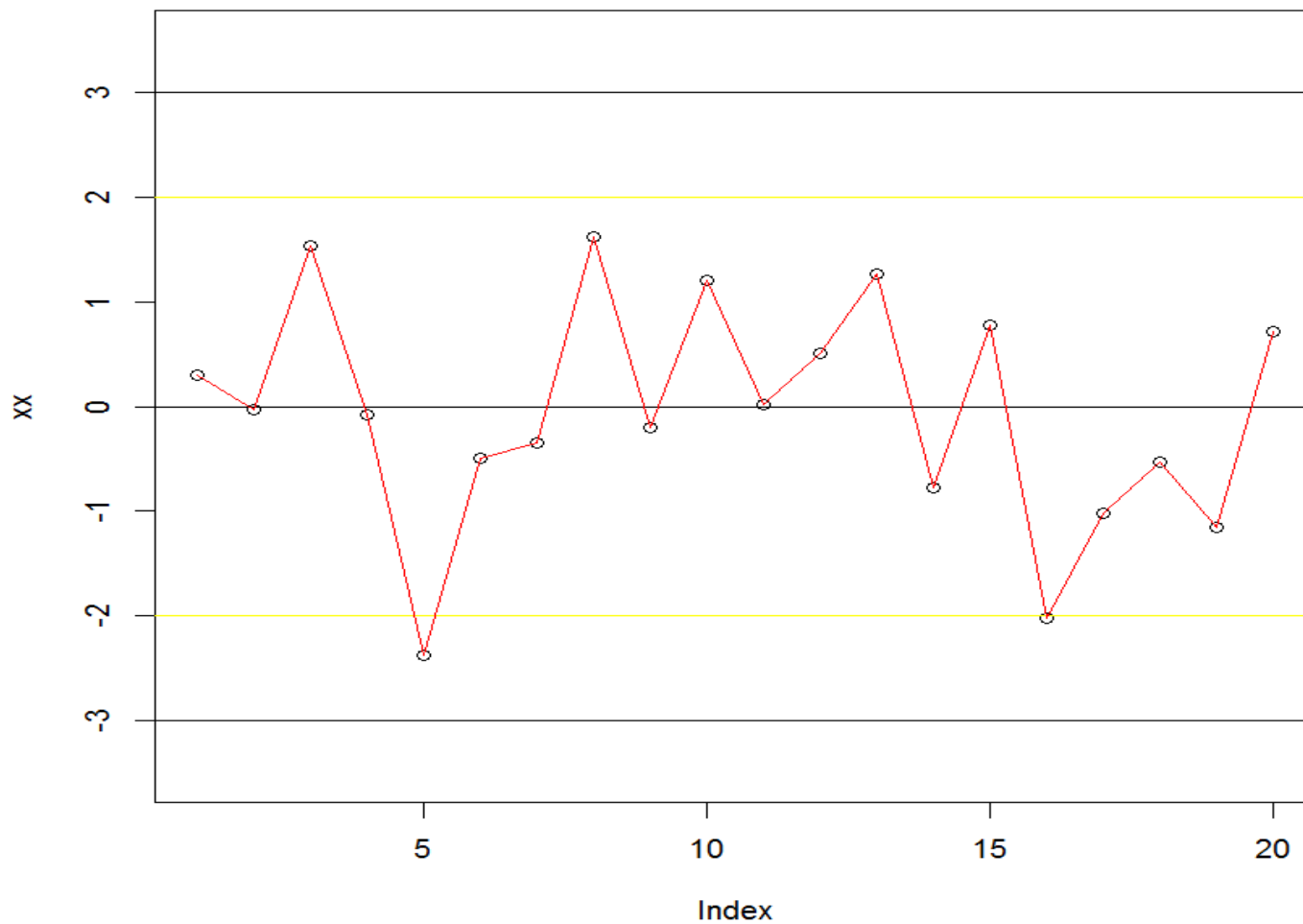
# 基本数据可视化

## 1. 矩阵散点图

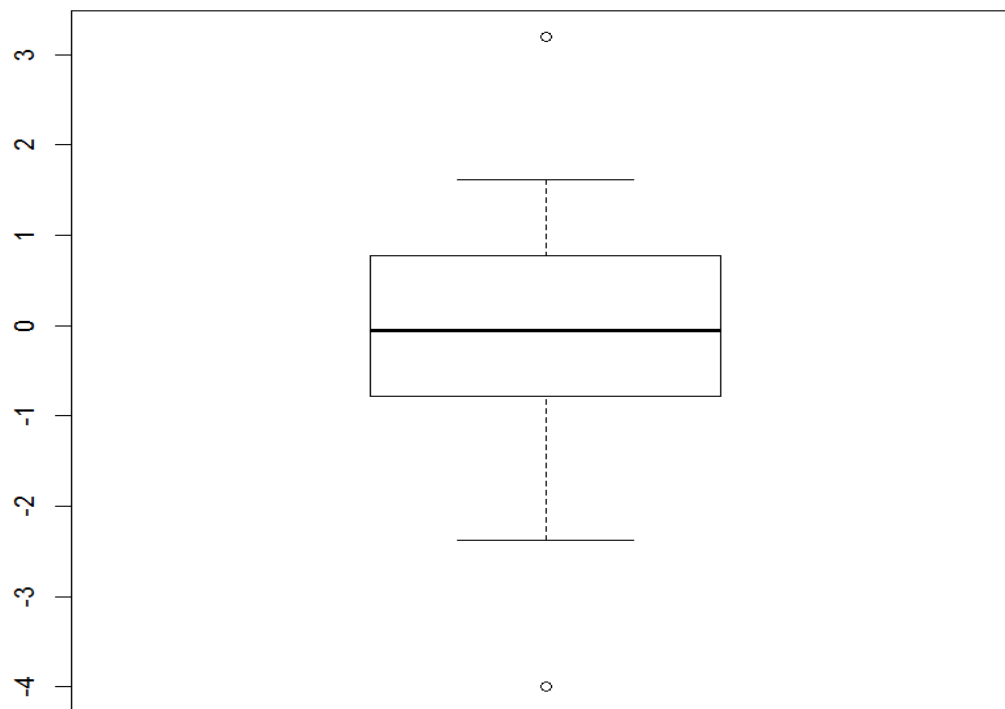
矩阵散点图



## 2. 3-sigma控制图:



### 3. 箱线图（异常点识别）



### 4. 其他可视化方法

## 二、线性回归模型

$$Y = b_0 + b_1 R + b_2 G + b_3 B + b_4 S + b_5 H + \varepsilon$$
$$= X^T b + \varepsilon$$

其中:  $X^T = (1, R, G, B, S, H)$

$$E(\varepsilon) = 0,$$

$$Var(\varepsilon) = \sigma^2$$

$$Y_i = X_i^T b + \varepsilon_i (1 \leq i \leq n)$$

■ 最小二乘估计:

$$\hat{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$$E(\hat{b}) = b, \quad Cov(\hat{b}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$$

■ 残差:

$$r_i = y_i - X_i^T \hat{b}$$

■ T-化残差:

$$r_i^* = r_i / s(r)$$

- 复相关系数（决定系数）：

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

- 调整的复相关系数：

$$R^{*2} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- **MSE:**  $MSE = \frac{1}{n} \sum_{i=1}^n r_i^2$



- **MSCV**（平均平方交叉核实预测误差）：

$$MSCV = \frac{1}{n} \sum_{i=1}^n \left( y_i - X_i^T \hat{b}_{(-i)} \right)^2$$

- **显著性检验**：  $H_0 : b_j = 0, H_1 : b_j \neq 0$

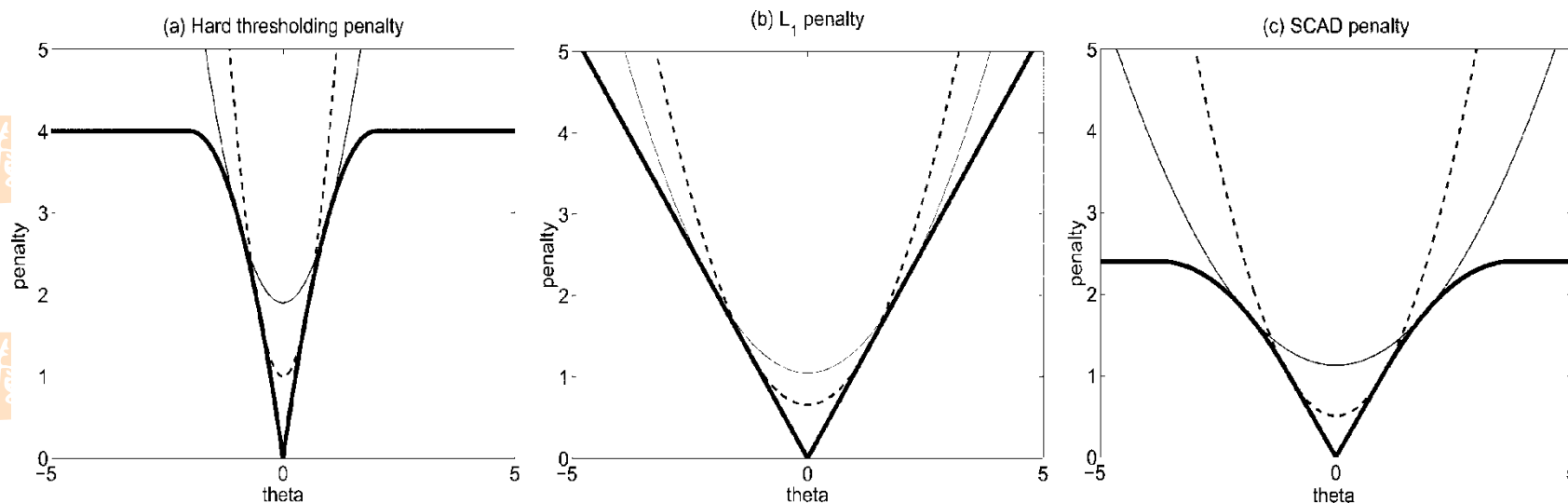
$$T = \frac{(\mathbf{X}^T \mathbf{X})_i^{1/2} \hat{b}_i}{\sqrt{nMSE/(n-p-1)}} \approx t(n-p-1)$$

**近似t检验。**

## ■ 变量选择

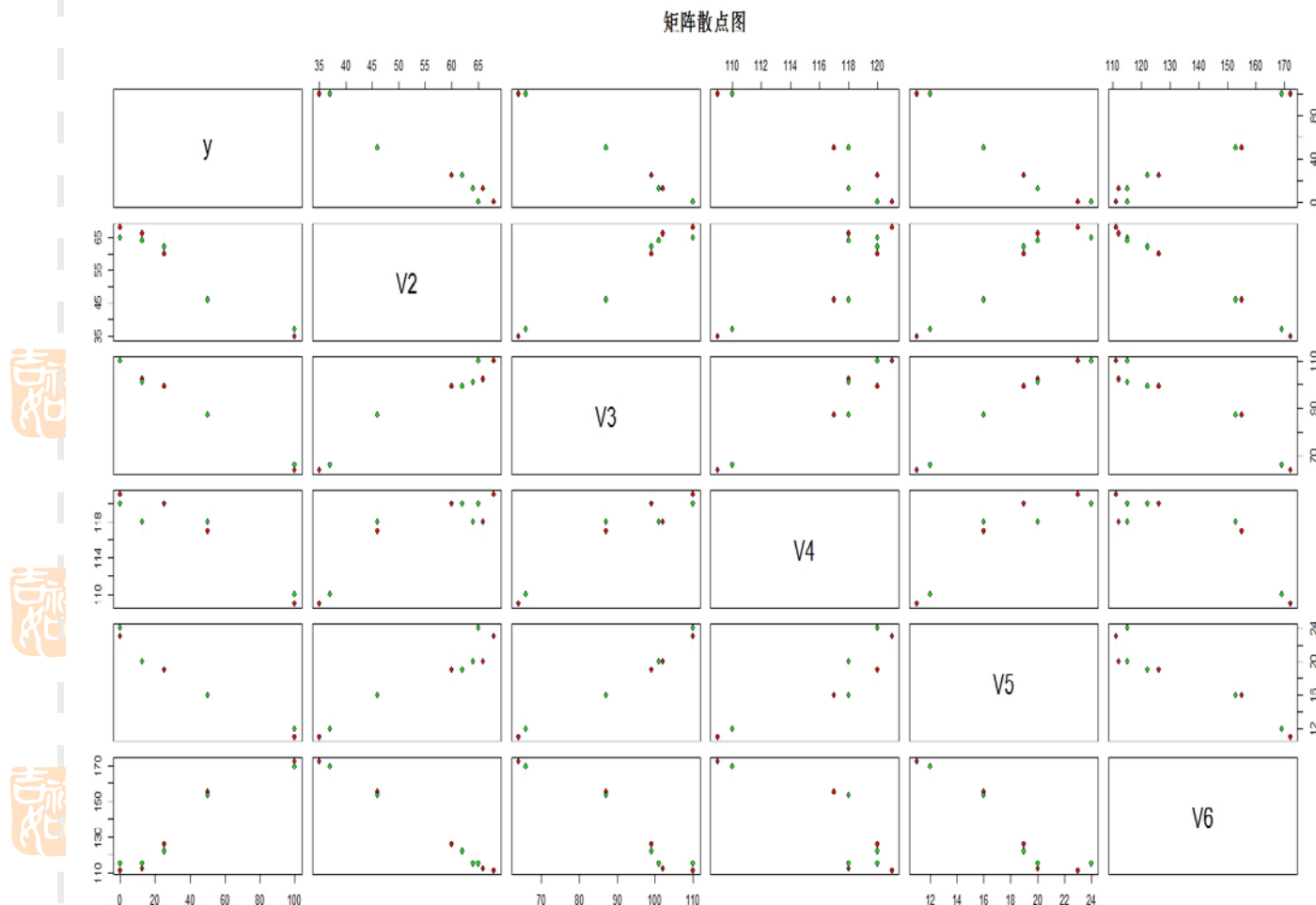
$$\frac{1}{n} \sum_{i=1}^n (y_i - X_i^T b)^2 + \sum_{j=1}^p P_{\lambda} (|b_j|) = \min$$

■  $P_{\lambda}(\cdot)$  是惩罚函数:



# 例如：组胺 (n=10)

## 一、矩阵散点图



## 二、线性回归模型:

Call:

```
lm(formula = y ~ x[, 1] + x[, 2] + x[, 3] + x[, 4] + x[, 5])
```

Residuals:

1	2	3	4	5	6	7	8	9
-0.99313	-0.08324	-0.18405	-0.08707	0.92020	0.73337	0.14180	-0.22235	1.05651
10								
-1.28202								

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-212.7650	84.1714	-2.528	0.064819	.
x[, 1]	2.8548	0.9089	3.141	0.034819	*
x[, 2]	-4.4873	0.4470	-10.039	0.000554	***
x[, 3]	2.3213	0.6533	3.553	0.023733	*
x[, 4]	4.5932	0.8663	5.302	0.006078	**
x[, 5]	1.1415	0.4298	2.656	0.056641	.

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

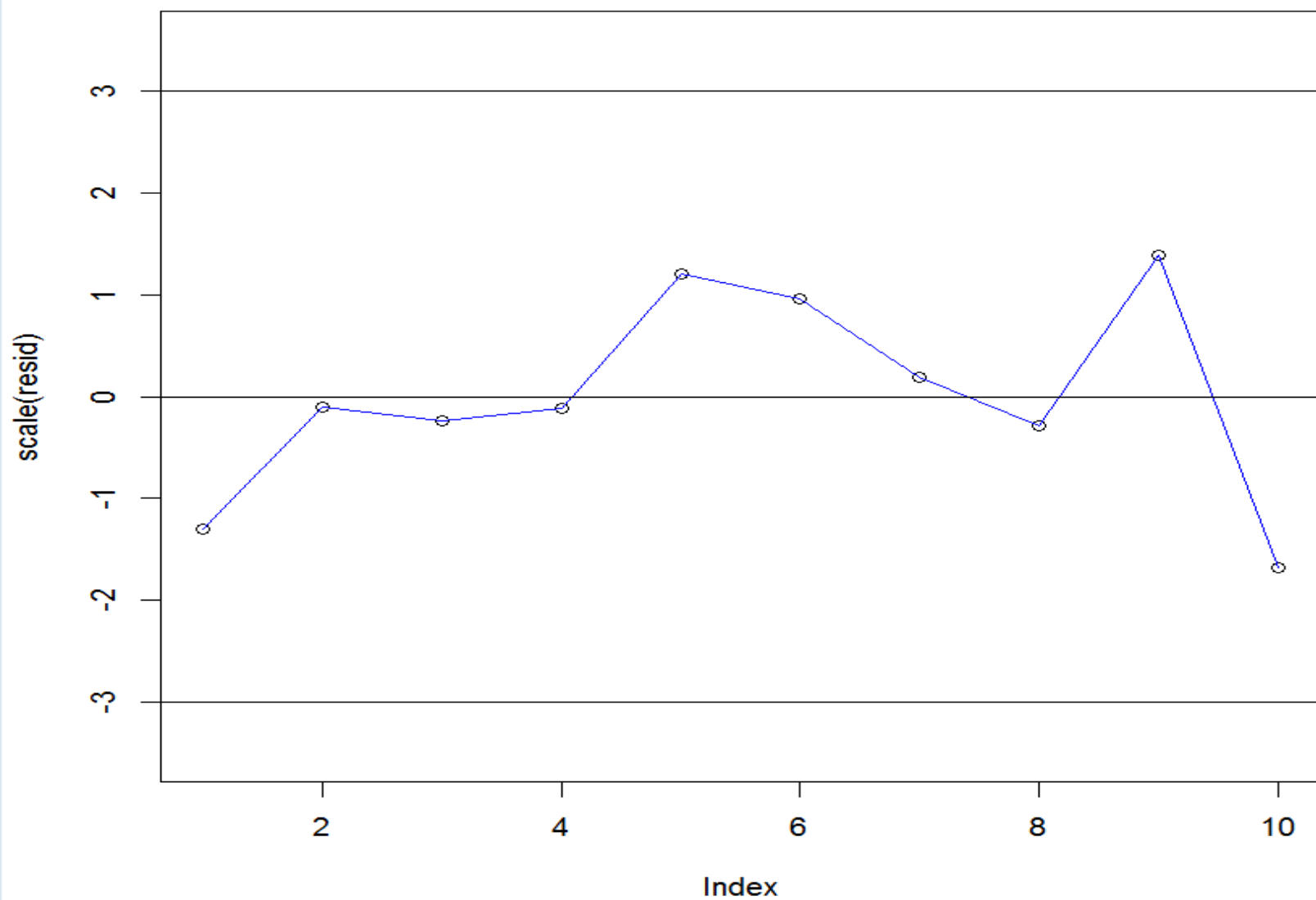
Residual standard error: 1.145 on 4 degrees of freedom

Multiple R-squared: 0.9996, Adjusted R-squared: 0.9991

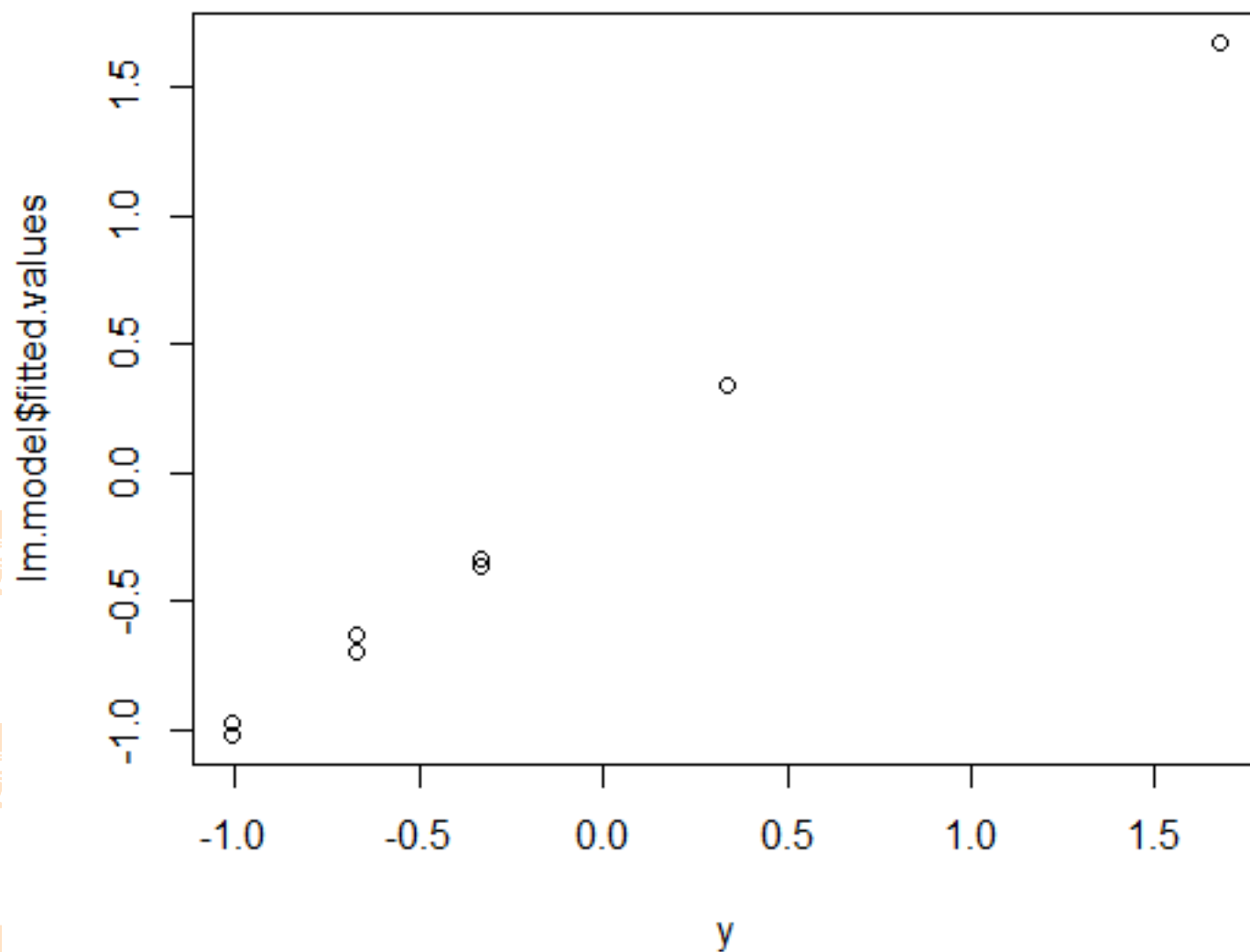
F-statistic: 1904 on 5 and 4 DF, p-value: 7.71e-07



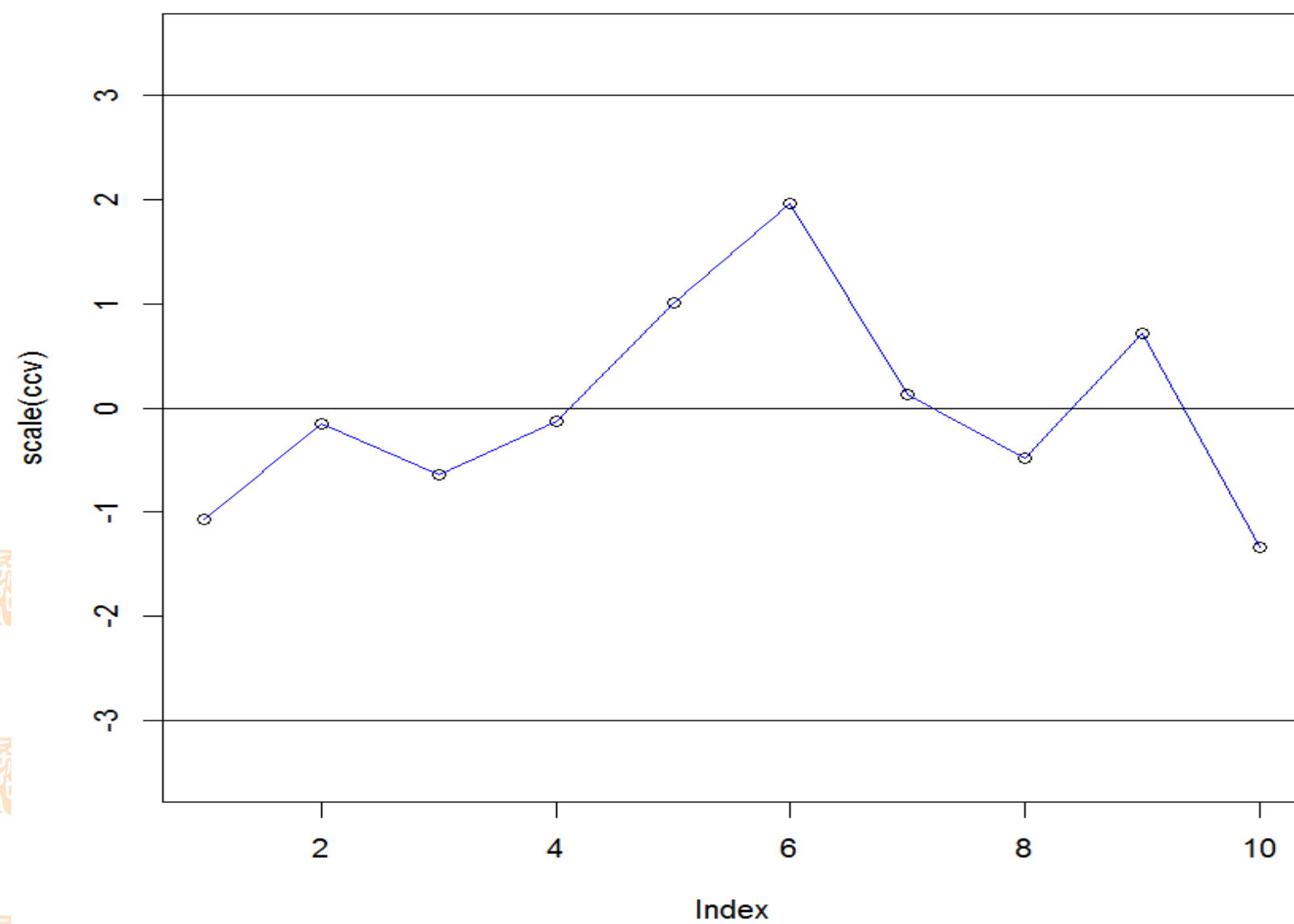
线性回归t-化残差图



## 线性回归真实值与预测值之间的关系



t-化CCV



## 1. MSE和MSCV(标准化)

MSE: 0.525 (0.0004)

MSCV: 3.511 (0.0025)

## 2. 3或 2sigma 点

2sigma: 无,

3sigma: 无



### 三、Logistic回归模型

- 机理：取

$$f_0(x) = \frac{\exp(x)}{1 + \exp(x)}$$

$$Z \approx \frac{\exp(b_0 + b_1 R + b_2 G + b_3 B + b_4 S + b_5 H)}{1 + \exp(b_0 + b_1 R + b_2 G + b_3 B + b_4 S + b_5 H)}$$

■ 这里， $Z = \frac{Y - \min + d}{\max - \min + 2d}$ ， $Y$  是浓度。

Call:

```
lm(formula = Z ~ x)
```

Residuals:

1	2	3	4	5	6	7	8
-0.0036797	-0.0002991	-0.0006495	-0.0003076	0.0033442	0.0027068	0.0005114	-0.0008541
9	10						
0.0039005	-0.0046728						

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-0.956310	0.308917	-3.096	0.036373	*
xV2	0.010391	0.003336	3.115	0.035700	*
xV3	-0.016369	0.001640	-9.978	0.000567	***
xV4	0.008492	0.002398	3.541	0.023981	*
xV5	0.016704	0.003179	5.254	0.006280	**
xV6	0.004144	0.001578	2.627	0.058394	.

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

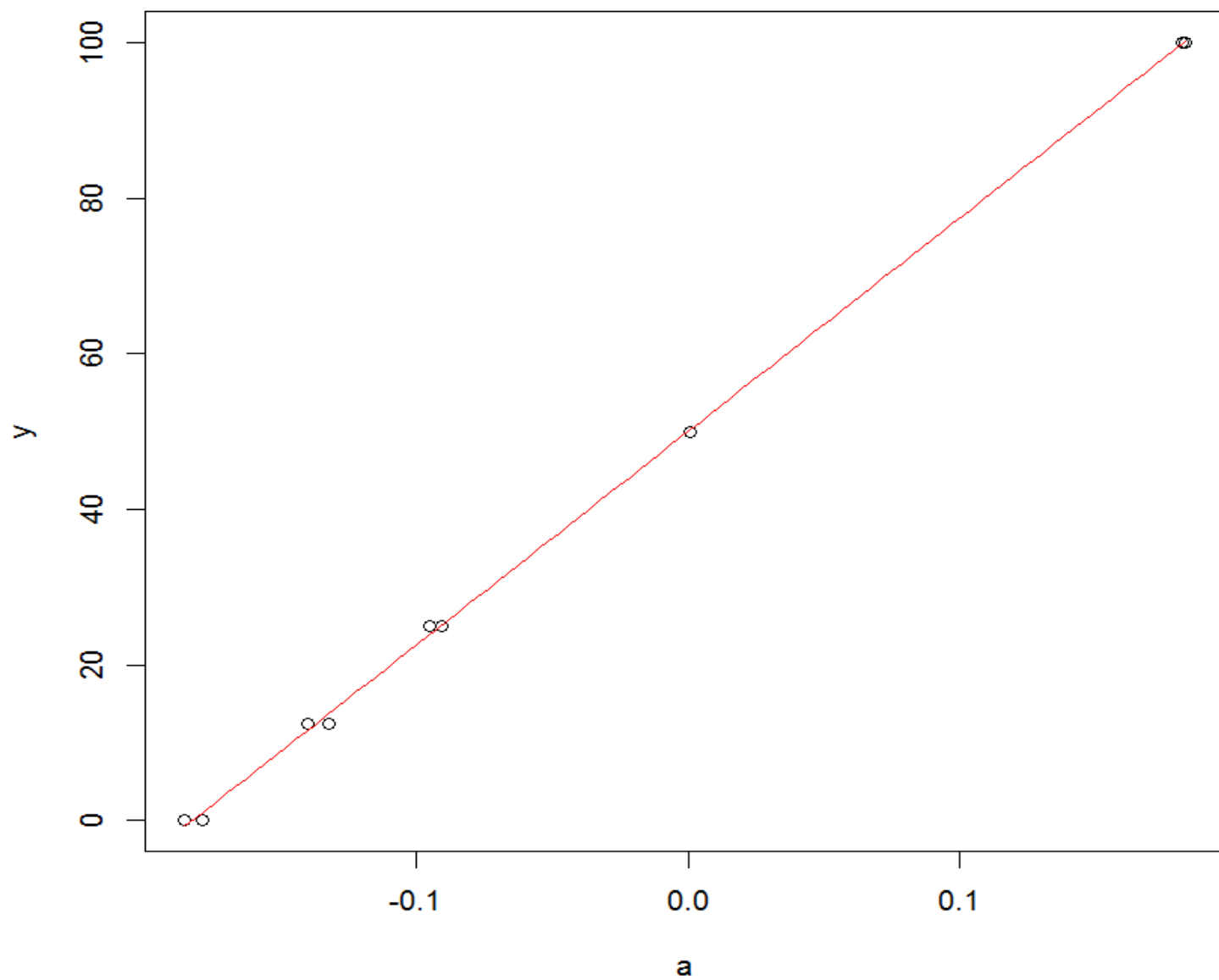
Residual standard error: 0.004204 on 4 degrees of freedom

Multiple R-squared: 0.9996, Adjusted R-squared: 0.999

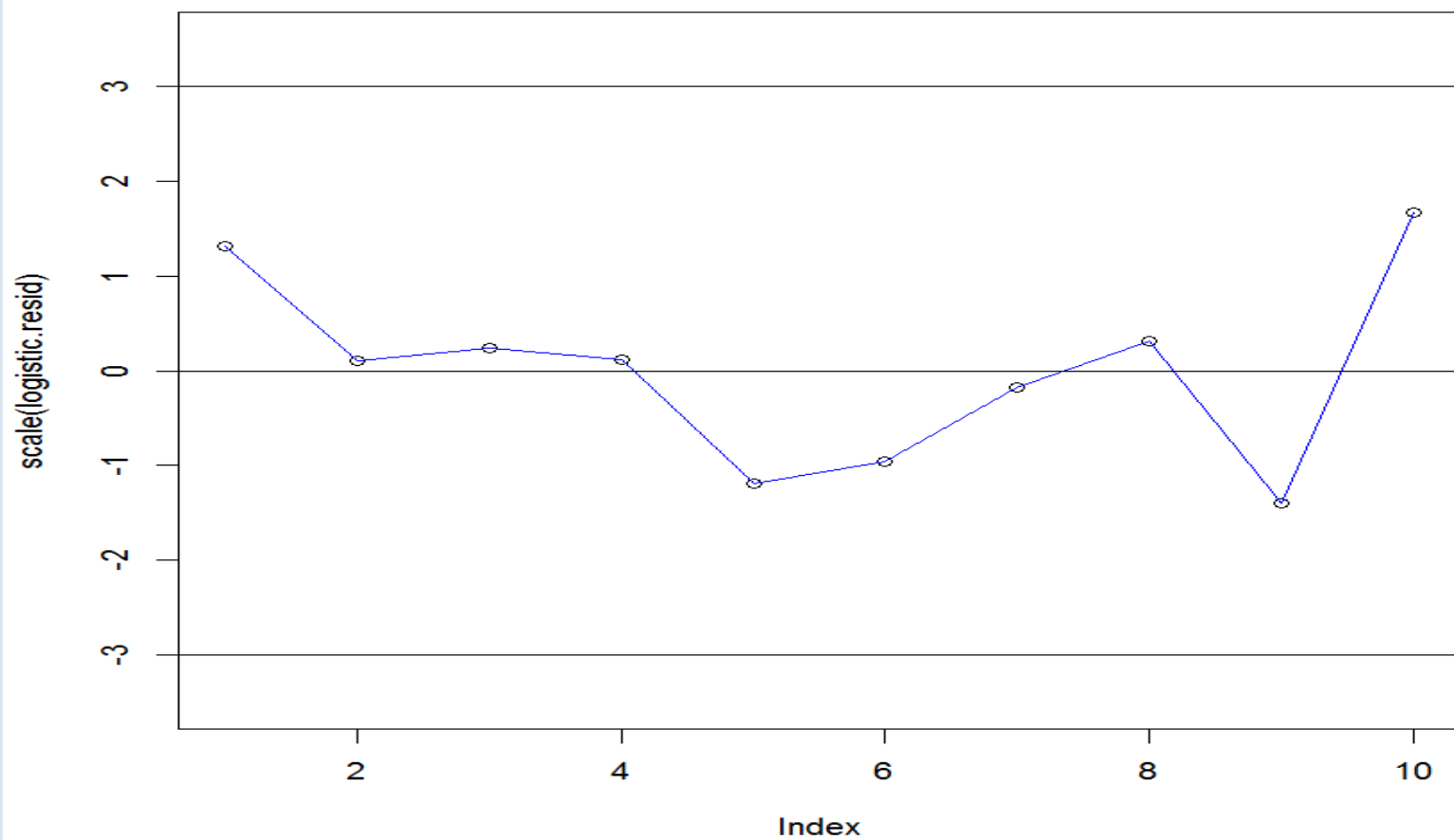
F-statistic: 1879 on 5 and 4 DF, p-value: 7.921e-07



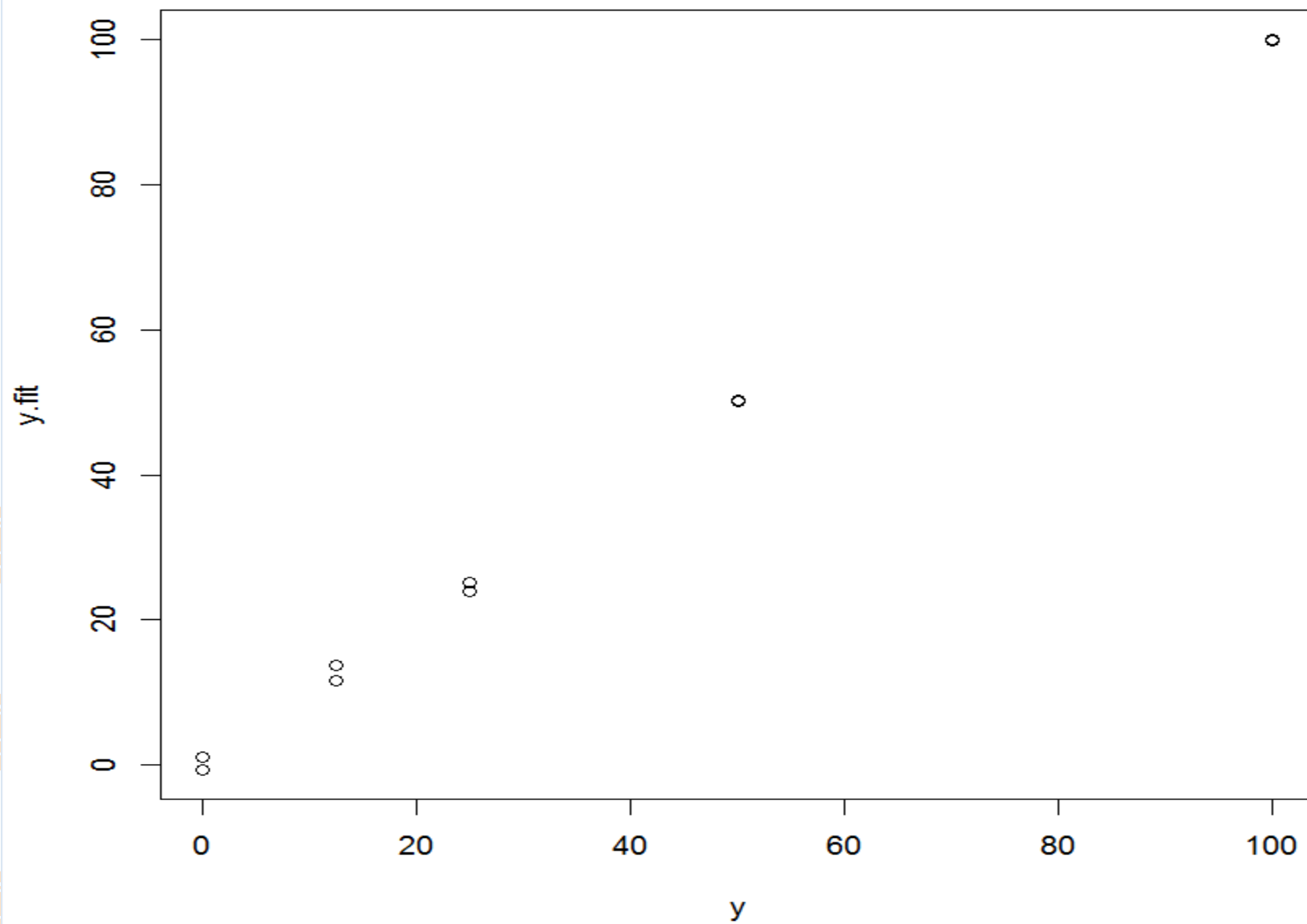
## logistic回归函数形式



logistic回归t-残差图

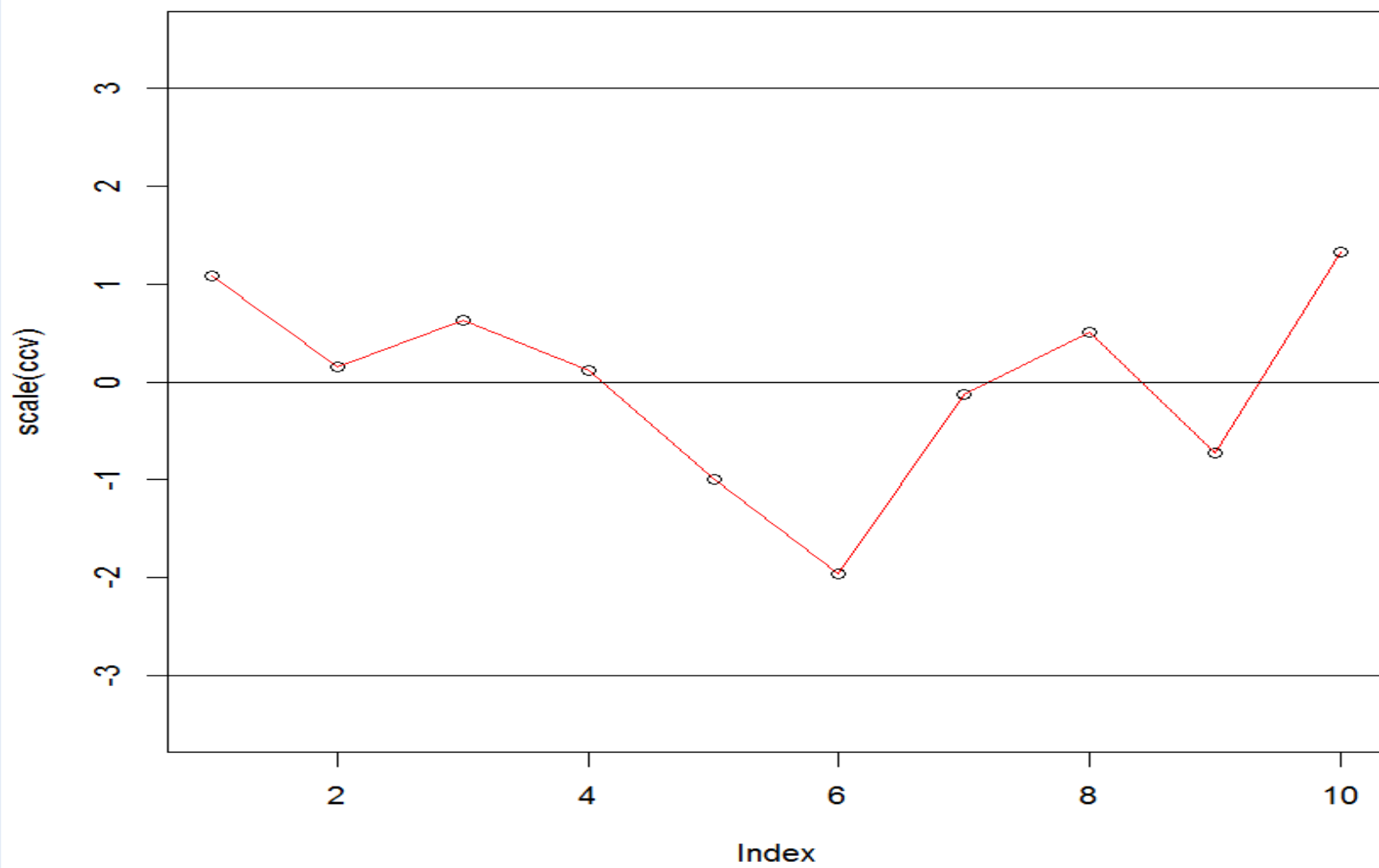


## logistic拟合值与真实值的关系





# t-CCV



## 1. MSE和MSCV(标准化)

MSE: 0.529 (0.0004) ,

MSCV: 3.536 (0.0025)

## 2. 3或 2sigma 点

2sigma: 无, 3sigma: 无

## 四、变量选择

```
> scad<-ncvreg(x+0,y,family="gaussian",penalty="SCAD")
> dof<-apply(abs(scad$beta[2:(pp+1),])>.Machine$double.eps,2,sum)
> bic<-as.vector(n*log(scad$loss)+2*log(n)*dof)  ###BIC准则
> t1<-sort(bic,ind=T)
> sbeta<-as.vector(scad$beta[2:(pp+1),t1$ix[1]])
> sbeta
[1] 0.401292 -4.445134 2.979073 4.578606 0.000000
> |
```

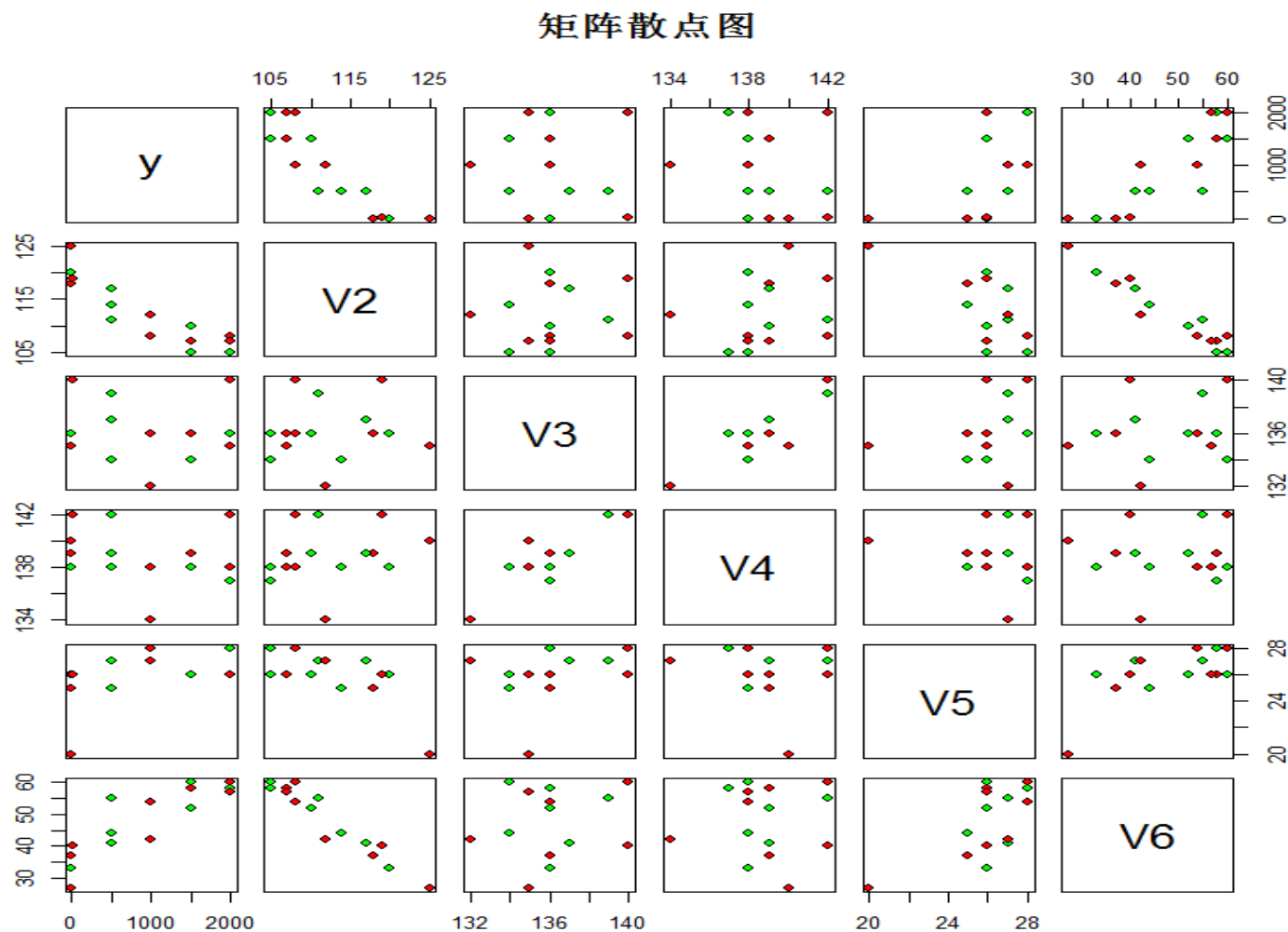


- 可去掉第5个变量，再进行建模，影响不大，与上述结果基本一致。



# 奶中尿素 (n=15)

## 一、矩阵图散点图



## 二、线性回归

Call:

```
lm(formula = y ~ x[, 1] + x[, 2] + x[, 3] + x[, 4] + x[, 5])
```

Residuals:

Min	1Q	Median	3Q	Max
-368.58	-158.55	10.19	71.70	587.03

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	12221.2	10790.0	1.133	0.2866
x[, 1]	280.1	280.5	0.999	0.3440
x[, 2]	495.2	181.6	2.726	0.0234 *
x[, 3]	-811.3	314.9	-2.576	0.0299 *
x[, 4]	-365.9	122.7	-2.981	0.0154 *
x[, 5]	251.1	158.2	1.587	0.1470

---

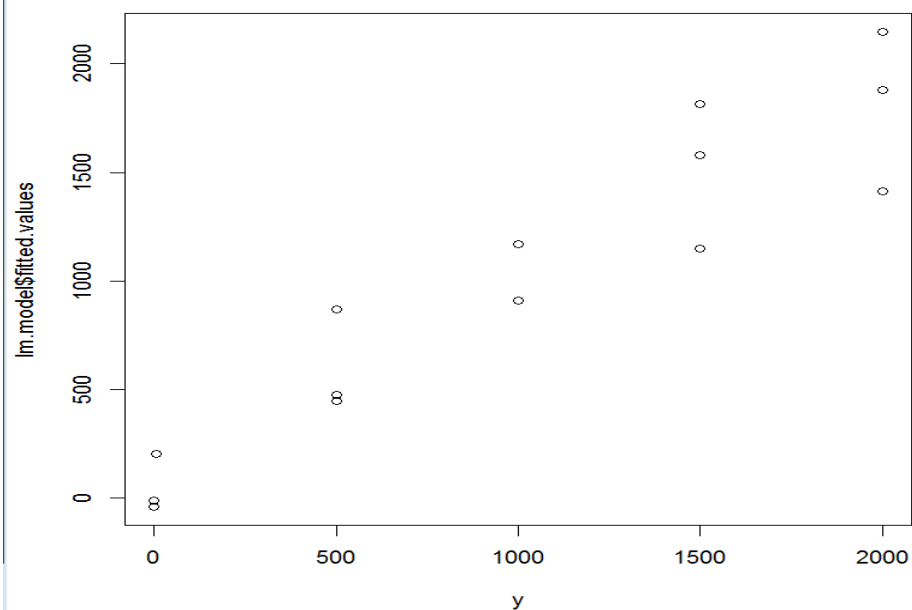
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303.1 on 9 degrees of freedom  
(2 observations deleted due to missingness)

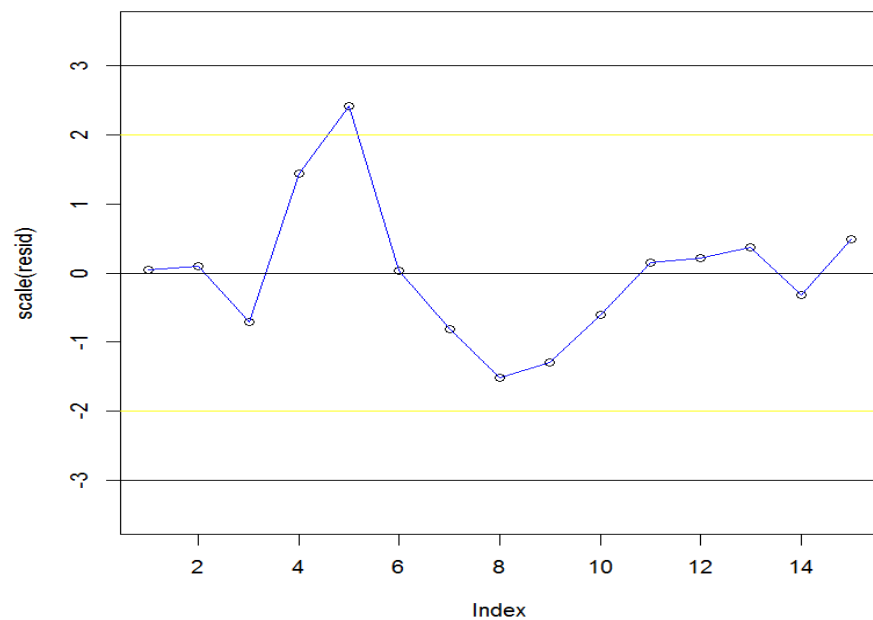
Multiple R-squared: 0.9019, Adjusted R-squared: 0.8473

F-statistic: 16.54 on 5 and 9 DF, p-value: 0.0002653

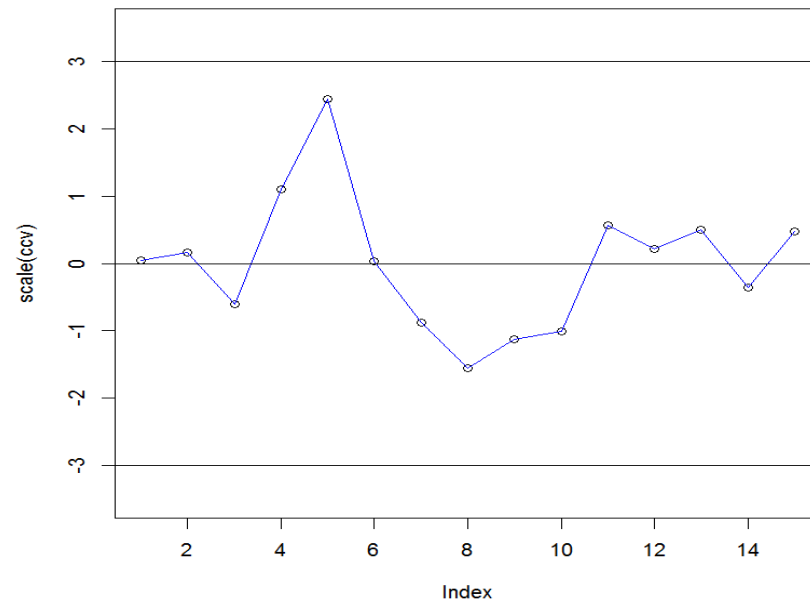
线性回归真实值与预测值之间的关系



线性回归t-化残差图



t-化CCV



## 1. MSE和MSCV(标准化)

MSE: 55118 (0.1910)

MSCV: 125338 (0.2083)

## 2. 3或 2sigma 点

2sigma: 5, 3sigma: 无

## 3. 点7, 8不匹配。

# 三、Logistic回归模型

Call:

```
lm(formula = Z ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.57775	-0.22457	0.03125	0.15490	0.93991

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	13.1882	17.1554	0.769	0.4617
xV2	0.5379	0.4460	1.206	0.2585
xV3	0.7686	0.2888	2.662	0.0260 *
xV4	-1.3352	0.5007	-2.667	0.0258 *
xV5	-0.5507	0.1952	-2.822	0.0200 *
xV6	0.4466	0.2516	1.775	0.1096

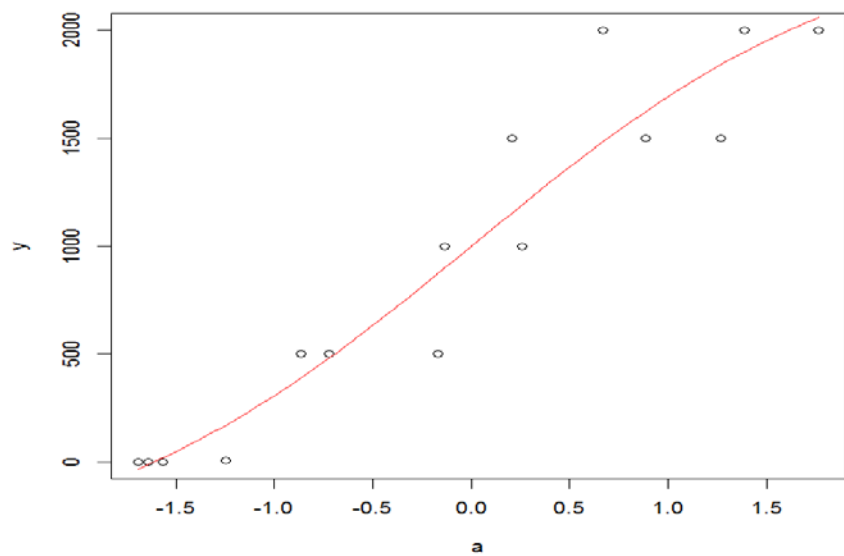
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4819 on 9 degrees of freedom

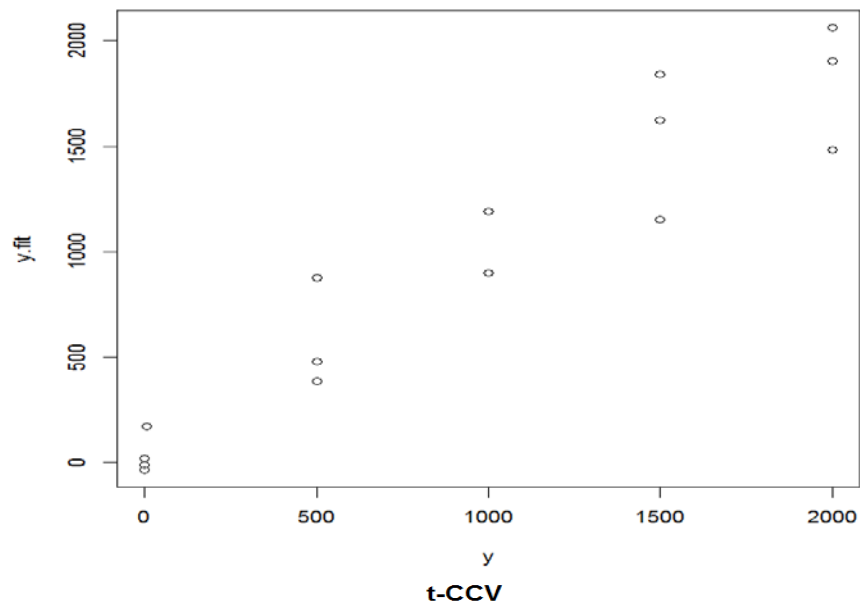
Multiple R-squared: 0.8995, Adjusted R-squared: 0.8437

F-statistic: 16.12 on 5 and 9 DF, p-value: 0.0002936

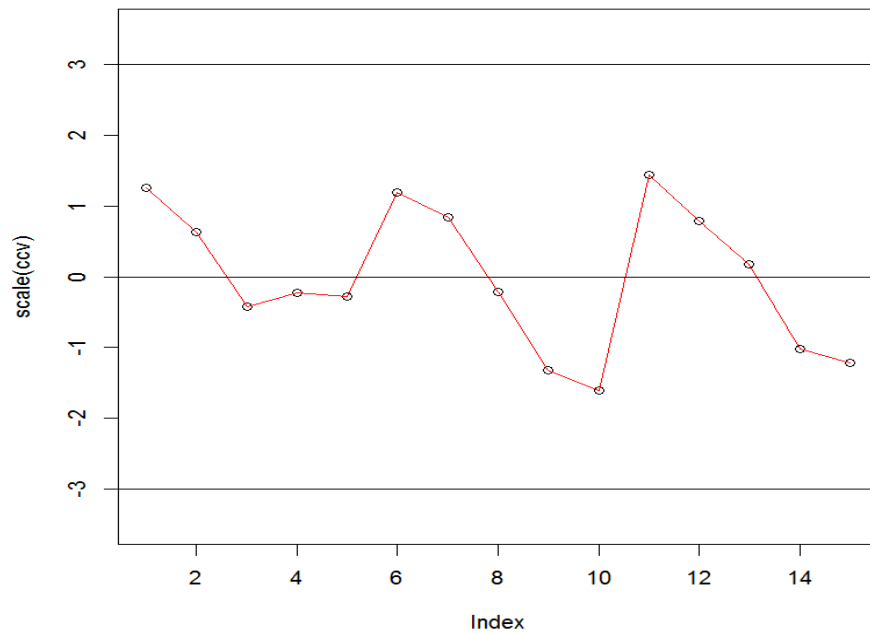
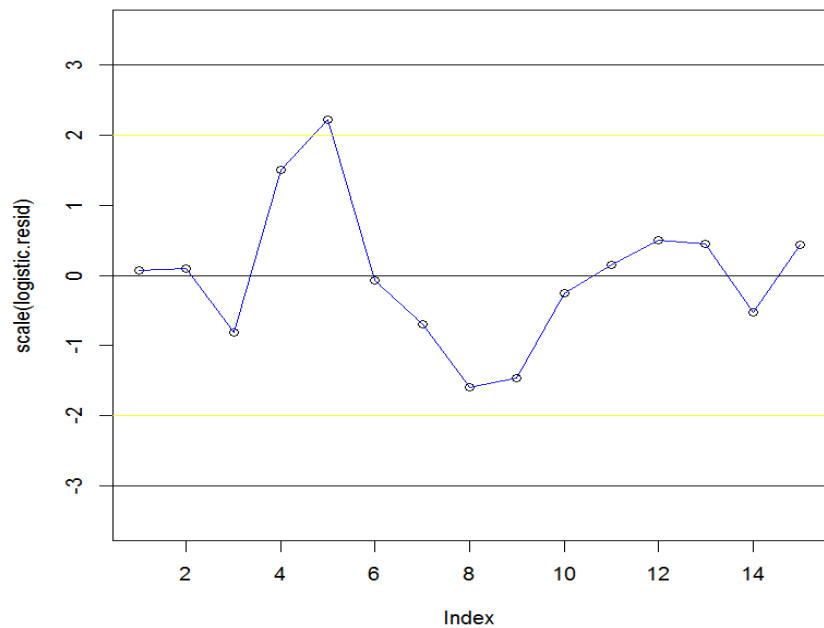
logistic回归函数形式



logistic拟合值与真实值的关系



logistic回归t-残差图



## 1. MSE和MSCV(标准化)

MSE: 50788 (0.0916)

MSCV: 117778 (0.2083)

## 2. 3或 2sigma 点

2sigma: 5, 3sigma: 无

## 3. 点7, 8不匹配。



## 四、变量选择

```
> scad<-ncvreg(x+0,y,family="gaussian",penalty="SCAD")
> dof<-apply(abs(scad$beta[2:(pp+1),]))>.Machine$double.eps,2,sum)
> bic<-as.vector(n*log(scad$loss)+2*log(n)*dof)   ###BIC准则
> t1<-sort(bic,ind=T)
> sbeta<-as.vector(scad$beta[2:(pp+1),t1$ix[1]])
> sbeta
[1] -112.4755    0.0000    0.0000    0.0000    0.0000
```

- 只与x1有关。

## ■ 5组数据中，依照模型与数据的拟合和匹配程度评估：

组胺  $\geq$  溴酸钾：基本可以确定关系

奶中尿素：不确定，倾向于可以确定关系

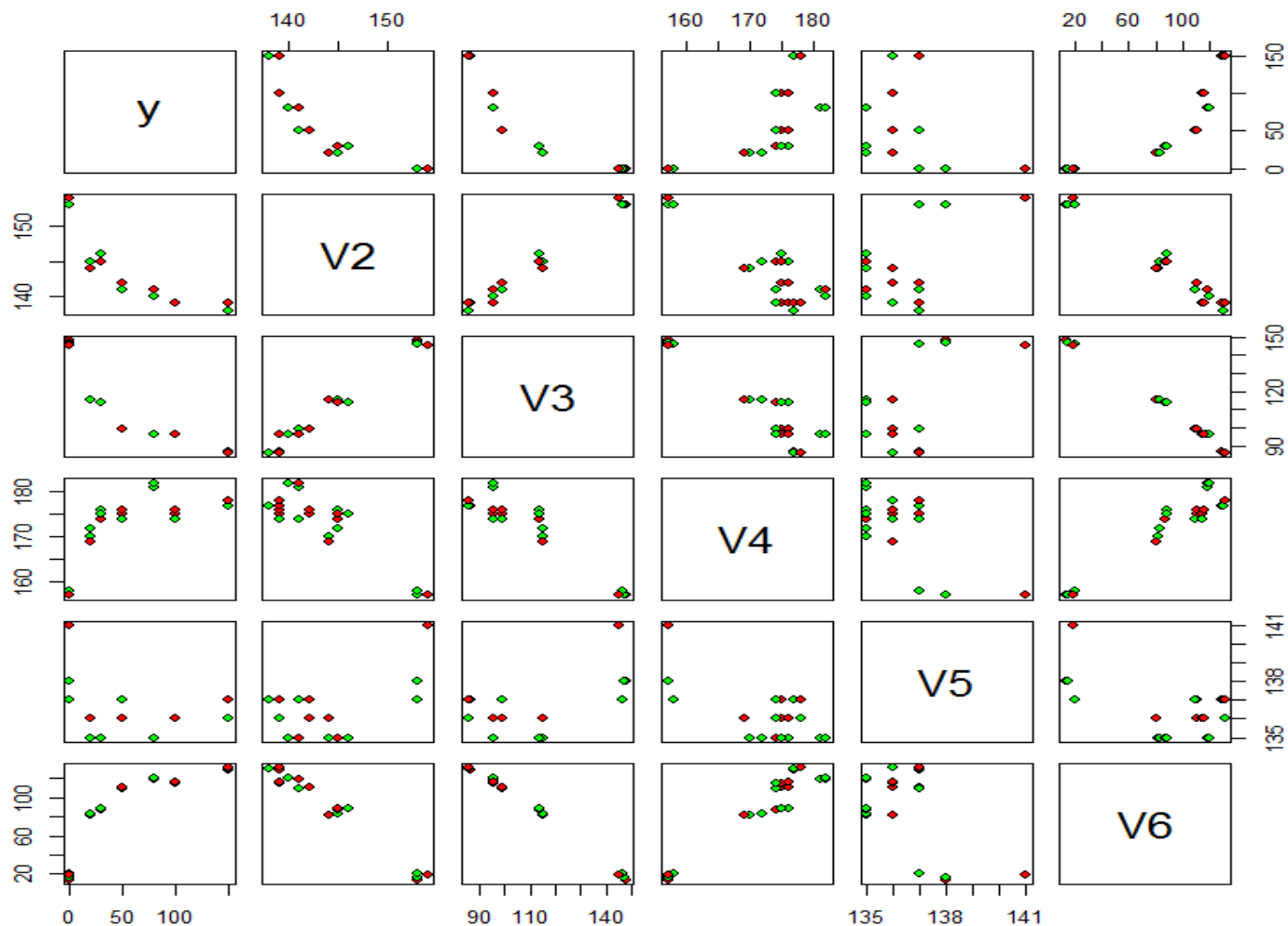
硫酸铝钾  $\geq$  工业碱：不能很好确定关系。



# 二氧化硫 (n=25)

## 一、矩阵图散点图

矩阵散点图



## 二、线性回归

Call:

```
lm(formula = y ~ x[, 1] + x[, 2] + x[, 3] + x[, 4] + x[, 5])
```

Residuals:

Min	1Q	Median	3Q	Max
-38.558	-11.042	5.562	11.209	21.361

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2846.2912	1235.6515	2.303	0.032719	*
x[, 1]	0.6472	5.8138	0.111	0.912533	
x[, 2]	-19.9277	5.1190	-3.893	0.000978	***
x[, 3]	5.2729	3.8861	1.357	0.190735	
x[, 4]	-4.8962	6.0724	-0.806	0.430049	
x[, 5]	-10.3539	3.3588	-3.083	0.006128	**

---

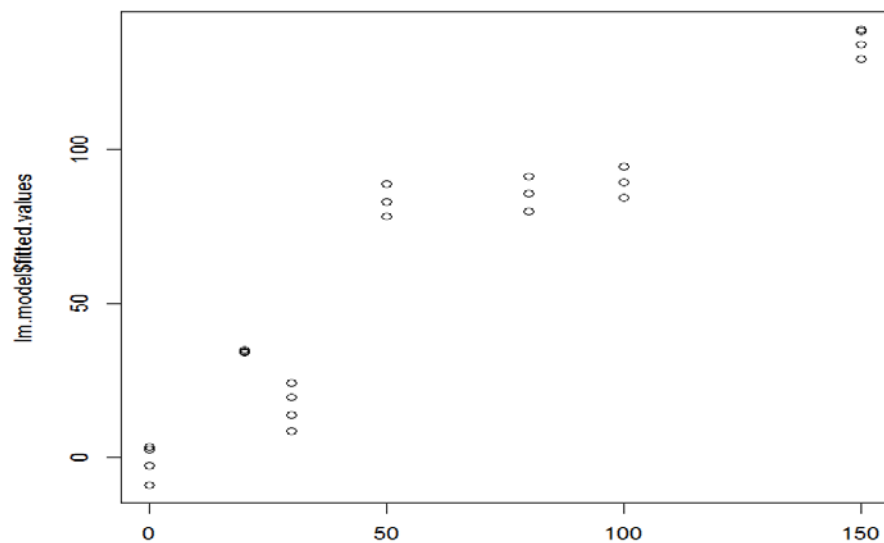
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.55 on 19 degrees of freedom

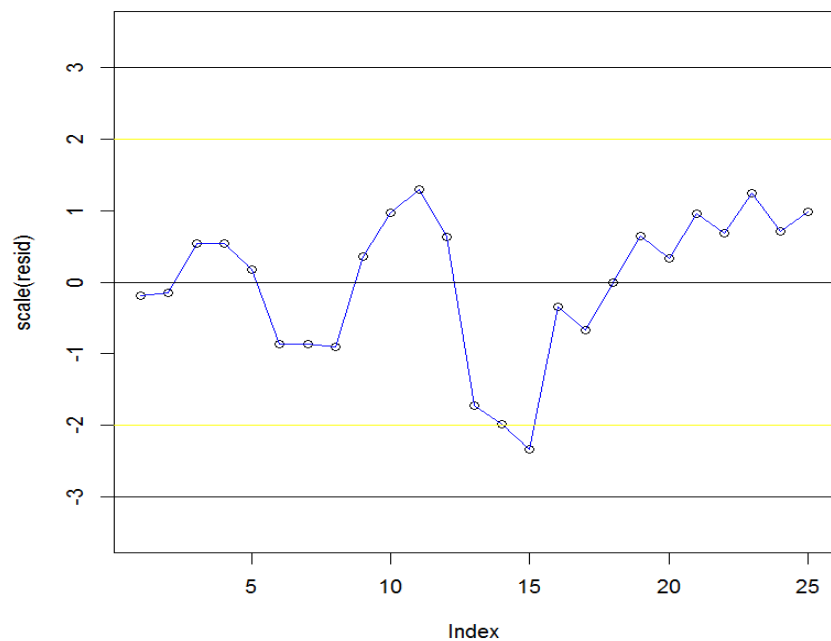
Multiple R-squared: 0.8996, Adjusted R-squared: 0.8731

F-statistic: 34.04 on 5 and 19 DF, p-value: 7.567e-09

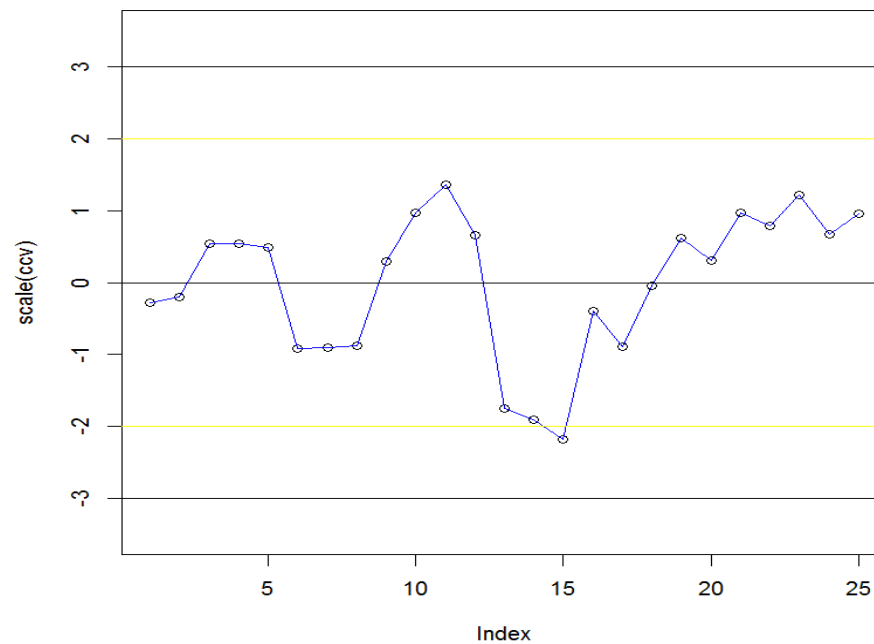
线性回归真实值与预测值之间的关系



线性回归t-化残差图



t-化CCV



## 1. MSE和MSCV(标准化)

MSE: 261.38 (0.0964)

MSCV: 390.38 (0.1440)

## 2. 3或 2sigma 点

2sigma: 第15, 3sigma: 无

### 三、logistic回归

Call:

```
lm(formula = Z ~ x)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.13467	-0.03865	0.01915	0.03913	0.07464

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	9.65521	4.31349	2.238	0.037365	*
xV2	0.00253	0.02030	0.125	0.902103	
xV3	-0.06951	0.01787	-3.890	0.000985	***
xV4	0.01827	0.01357	1.347	0.193812	
xV5	-0.01717	0.02120	-0.810	0.427846	
xV6	-0.03605	0.01172	-3.075	0.006238	**

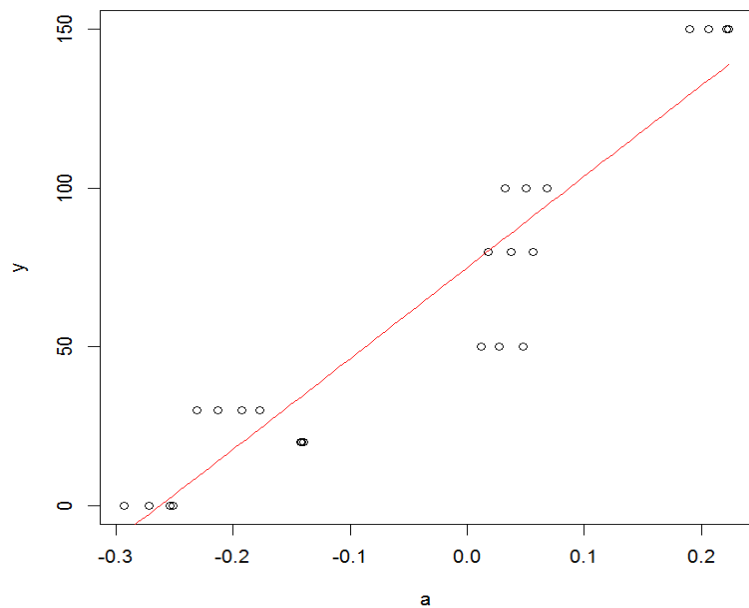
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06474 on 19 degrees of freedom

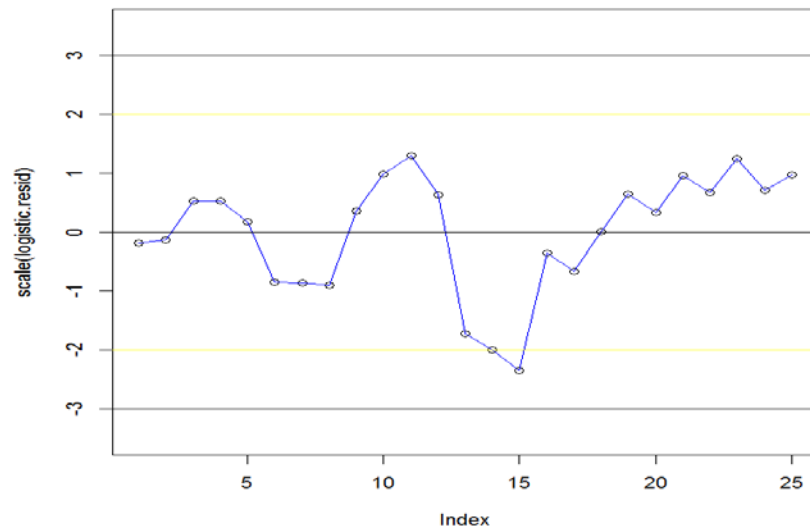
Multiple R-squared: 0.8998, Adjusted R-squared: 0.8734

F-statistic: 34.13 on 5 and 19 DF, p-value: 7.395e-09

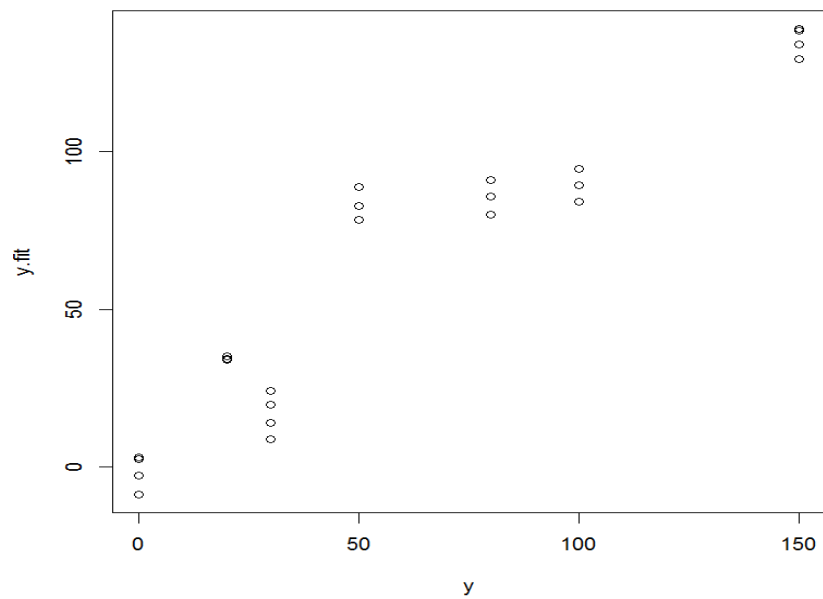
logistic回归函数形式



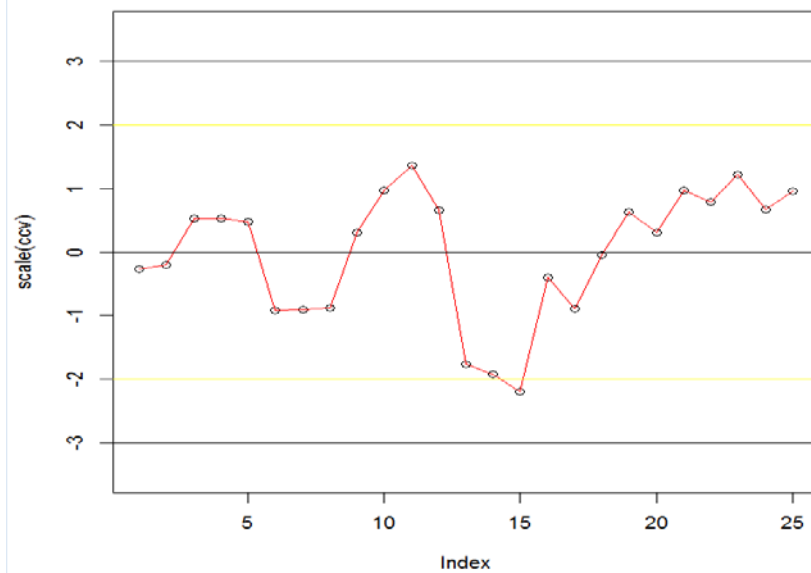
logistic回归t-残差图



logistic拟合值与真实值的关系



t-CCV





## 1. MSE和MSCV(标准化)

MSE: 260.84 (0.0956)

MSCV: 389.02 (0.1389)

## 2. 3或 2sigma 点

2sigma: 第14, 15, 3sigma: 无

## 四、变量选择

```
> scad<-ncvreg(x+0,y,family="gaussian",penalty="SCAD")
> dof<-apply(abs(scad$beta[2:(pp+1),])>.Machine$double.eps,2,sum)
> bic<-as.vector(n*log(scad$loss)+2*log(n)*dof)   ###BIC准则
> t1<-sort(bic,ind=T)
> sbeta<-as.vector(scad$beta[2:(pp+1),t1$ix[1]])
> sbeta
[1] 1.468999 -11.109252 0.000000 0.000000 -4.530689
> |
```

- 可去掉第2, 3个变量, 再进行建模, 与上述结果基本一致。

# 可考虑其他模型：

- 概率变换模型
- 单指标（single-index）模型（ $f_0$  形式未知，用于探索）



# 谢谢!



# C评判标准: 关键是模型的选择与误差分析

一、仅是线性模型+拟合 ( $\leq 5$ ):

i). 无误差分析( $R^2$ , MSE, 残差图, CV等)和异常点分析 (3sigma准则等)、单一变量模型

$\leq 2$ ;

ii). 误差分析和异常点分析至少有一个  $\geq 2$ ;

iii). ii)+逐步回归或变量选择(共线分析)或样本变化分析:  $\geq 3$ 。

二、非线性回归+拟合: i)+0; ii) +1; iii) +1;

三、非线性单调回归+拟合: i)+0; ii)+2; iii)+2.

- 5组数据中，依照模型与数据的拟合和匹配程度排序：

组胺  $\geq$  溴酸钾：基本可以确定关系

奶中尿素：不确定，倾向于可以确定关系

硫酸铝钾  $\geq$  工业碱：不能很好确定关系。

