Linear Regression (II)

Zhijian He

2020-11-25

Anscombe 在 1973 年构造了 4 组数据，每组数据都是由 11 对点(x_i, y_i) 组成，试分析 4 组数据是否通过回归方程的检验。

Table 1: Anscombe 前 6 组数据

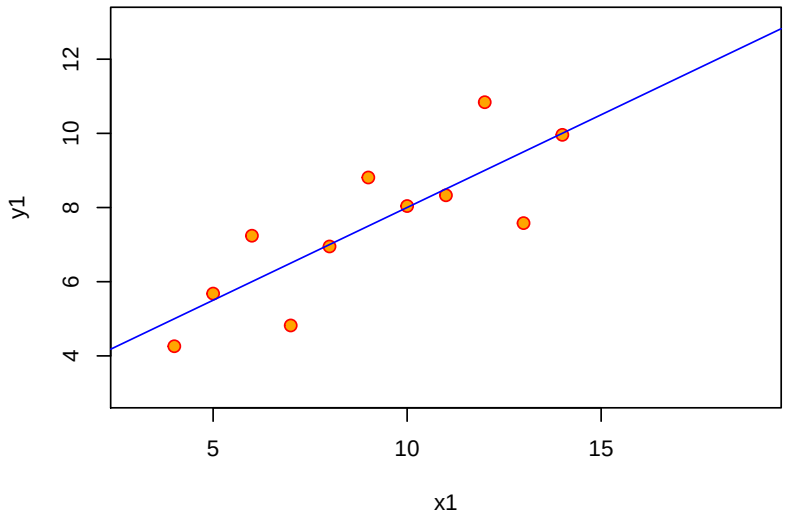
x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.10	8.84	7.04

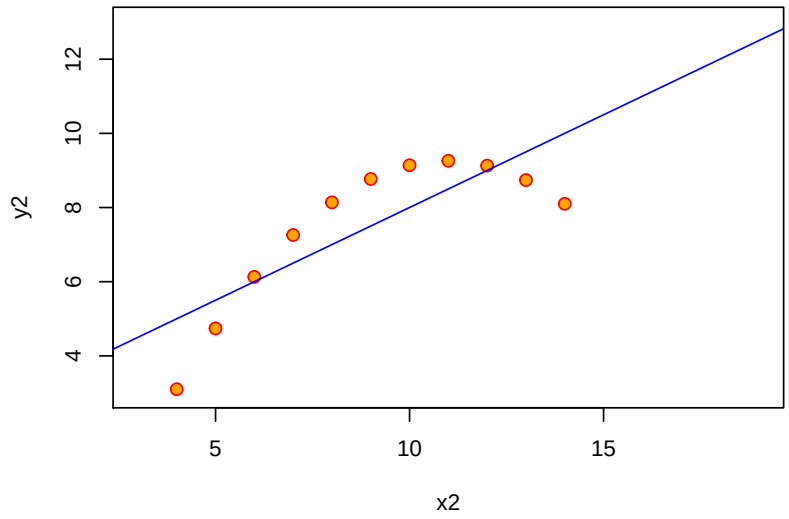
回归结果

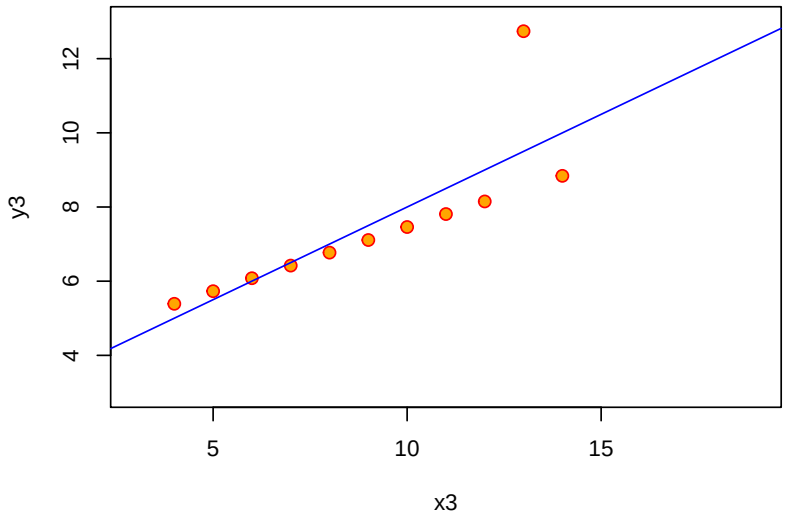
Table 2: 四种情况的回归结果，最后一列为 F 检验 p 值

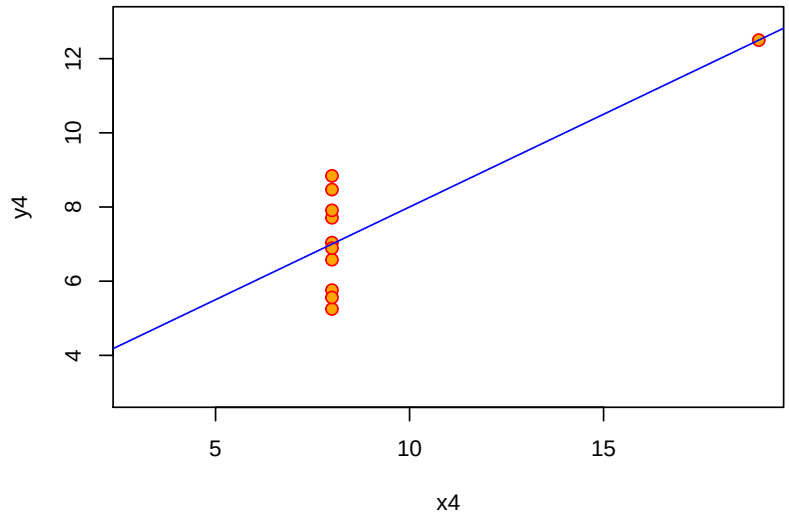
	Estimate	Std. Error	t value	Pr(>|t|)	p_value
(Intercept)1	3.0000909	1.1247468	2.667348	0.0257341	0.0021696
x1	0.5000909	0.1179055	4.241455	0.0021696	NA
(Intercept)2	3.0009091	1.1253024	2.666758	0.0257589	0.0021788
x2	0.5000000	0.1179637	4.238590	0.0021788	NA
(Intercept)3	3.0024545	1.1244812	2.670080	0.0256191	0.0021763
x3	0.4997273	0.1178777	4.239372	0.0021763	NA
(Intercept)4	3.0017273	1.1239211	2.670763	0.0255904	0.0021646
x4	0.4999091	0.1178189	4.243028	0.0021646	NA

回归直线


y1
x1
4
6
8
10
12
5
10
15


y2
x2
4
6
8
10
12
5
10
15


y3
x3
4
6
8
10
12
5
10
15


y4
x4
4
6
8
10
12
5
10
15

回归诊断

回归分析都是基于误差项的假定进行的，最常见的假设

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

- 如何考察数据基本上满足这些假设？自然从残差的角度来解决问题，这种方法叫残差分析。
- 研究那些数据对统计推断（估计、检验、预测和控制）有较大影响的点，这样的点叫做影响点。剔除那些有较强影响的异常/离群 (outlier) 数据，这就是所谓的影响分析 (influence analysis).

残差的定义为

$$\hat{\epsilon} = Y - \hat{Y}$$

残差的性质

在假设$\epsilon \sim N(0, \sigma^2 I_n)$ 下，

1. $\hat{\epsilon} \sim N(0, \sigma^2(I_n - P))$
2. $Cov(\hat{Y}, \hat{\epsilon}) = 0$
3. $1^\top \hat{\epsilon} = 0$

从中可以看出，$Var[\hat{\epsilon}_i] = \sigma^2(1 - p_{ii})$, 其中$p_{ij}$ 为投影矩阵的元素。该方差与σ^2 以及p_{ii} 有关，因此直接比较残差$\hat{\epsilon}_i$ 是不恰当的。

为此，将残差标准化：

$$\frac{\hat{\epsilon}_i - E[\hat{\epsilon}_i]}{\sqrt{Var[\hat{\epsilon}_i]}} = \frac{\hat{\epsilon}_i}{\sigma\sqrt{1 - p_{ii}}}, \ i = 1, \ldots, n$$

学生化残差

由于σ 是未知的，所以用$\hat{\sigma}$ 来代替，其中$\hat{\sigma}^2 = S_e^2/(n-p)$. 于是得到学生化 (studentized residuals)

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-p_{ii}}}$$

- t_i 虽然是$\hat{\epsilon}_i$ 的学生化，但它的分布并不服从t 分布，它的分布通常比较复杂
- $t_1, \ldots, t_n$ 通常是不独立的
- 在实际应用中，可以近似认为：$t_1, \ldots, t_n$ 是相互独立，服从$N(0,1)$ 分布
- 在实际应用中使用的残差图就是根据上述假定来对模型合理性进行诊断的。

残差图

残差图：以残差为纵坐标，其他的量（一般为拟合值$\hat{y}_i$）为横坐标的散点图。

由于可以近似认为：$t_1, \ldots, t_n$ 是相互独立，服从$N(0,1)$ 分布，所以可以把它们看作来自$N(0,1)$ 的 iid 样本

根据标准正态的性质，大概有95% 的t_i 落入区间$[-2,2]$ 中。由于$\hat{Y}$ 与$\hat{\epsilon}$ 不相关，所以$\hat{y}_i$ 与学生化残差t_i 的相关性也很小。

这样在残差图中，点$(\hat{y}_i, t_i), i = 1, \ldots, n$ 大致应该落在宽度为 4 的水平带$|t_i| \leq 2$ 的区域内，且不呈现任何趋势。

残差图（1）

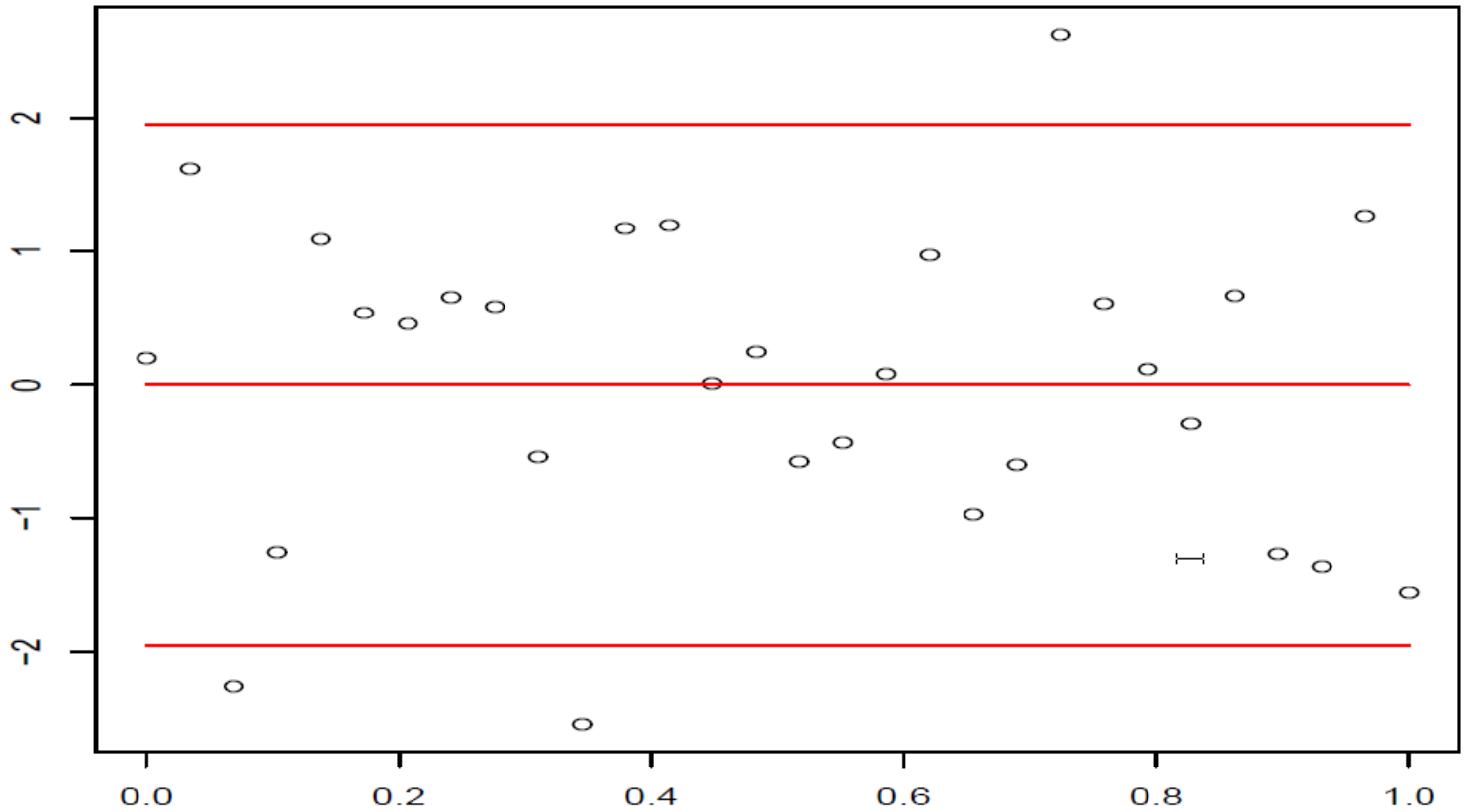


Figure 1: 正常的残差图

残差图（2）

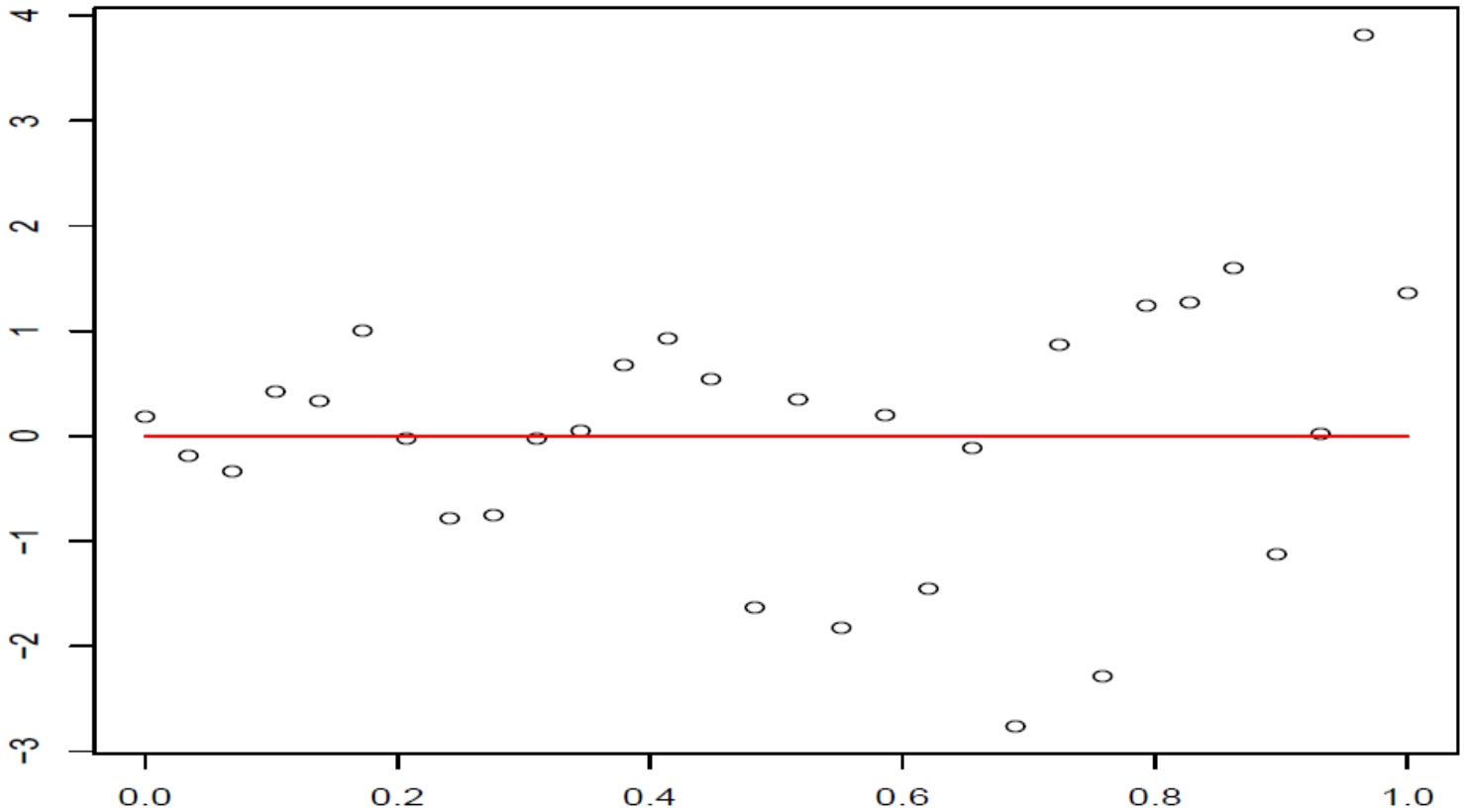


Figure 2: 误差随着横坐标的增加而增加

残差图（3）

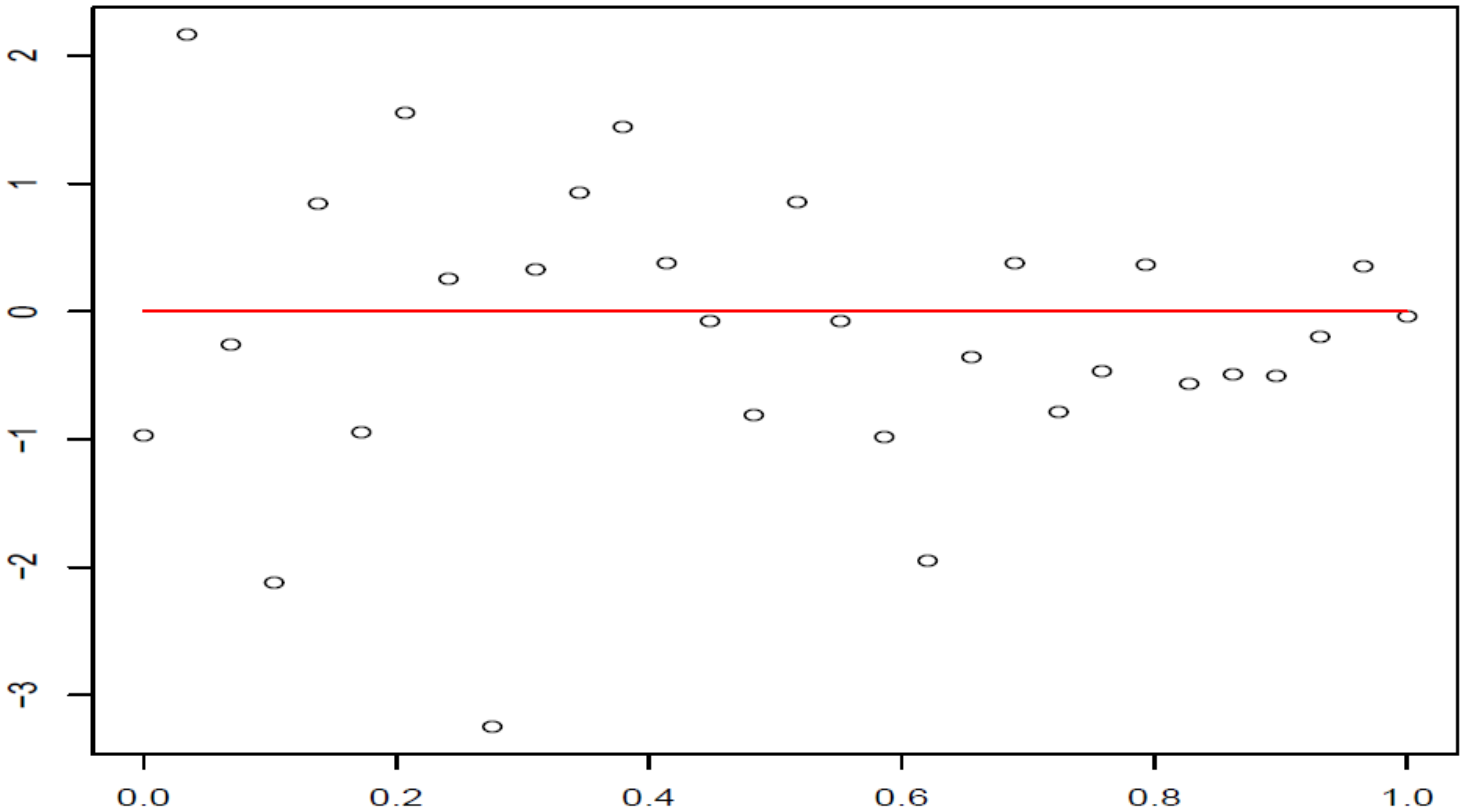


Figure 3: 误差随着横坐标的增加而减少

残差图（4）

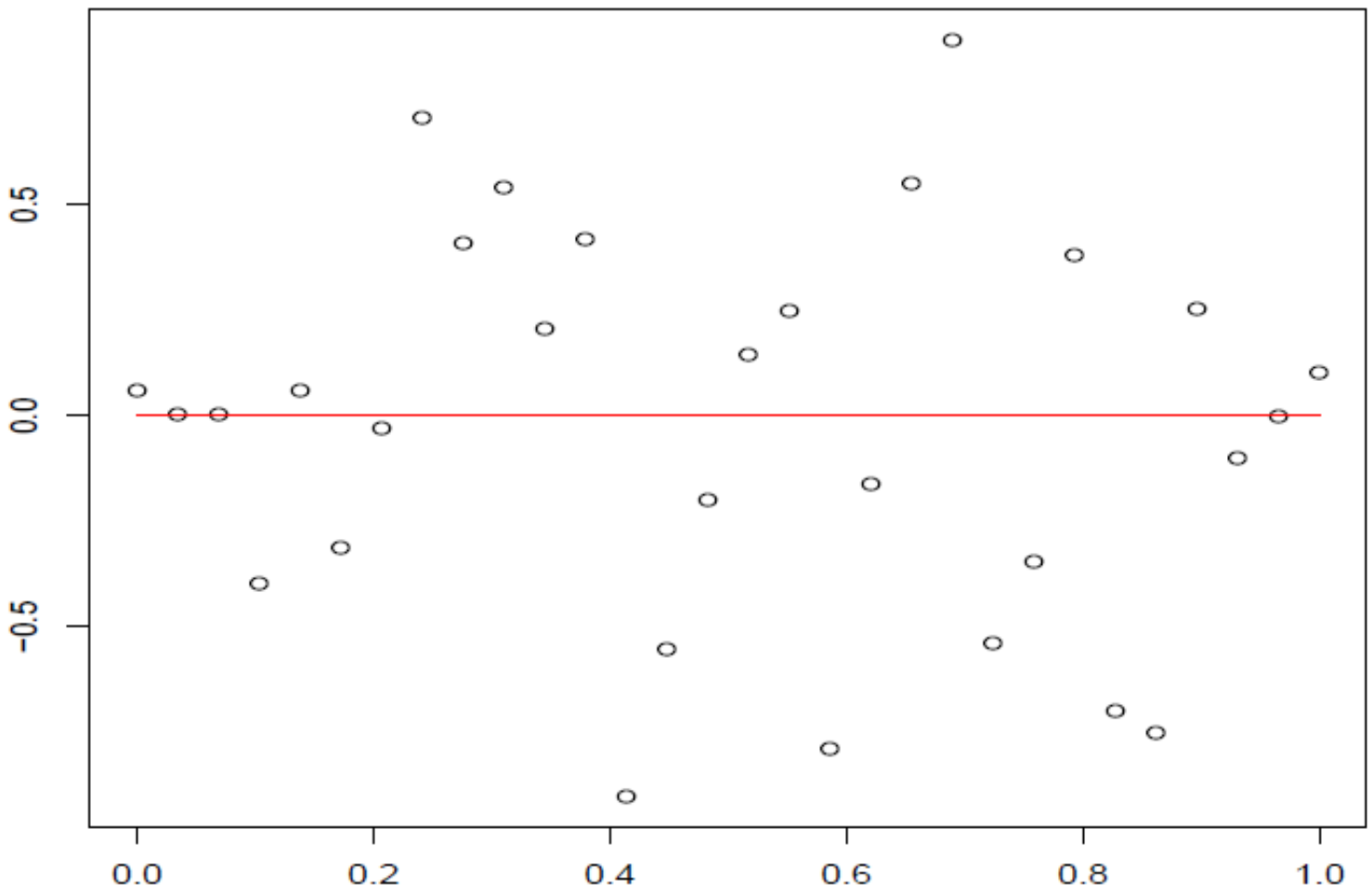


Figure 4: 误差中间大，两端小

残差图（5）

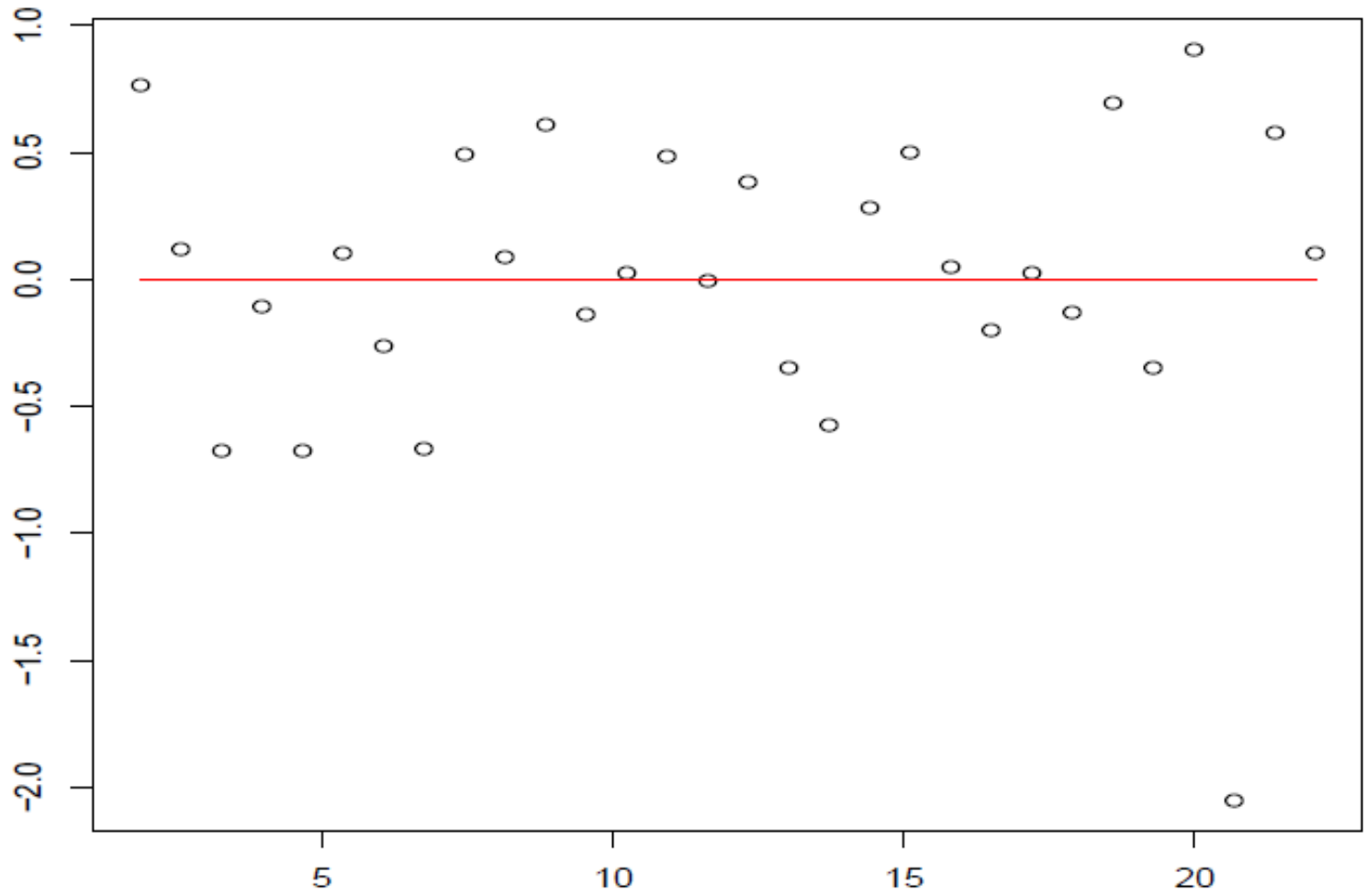


Figure 5: 回归函数可能非线性，或者误差相关或者漏掉重要的自变量

残差图（6）

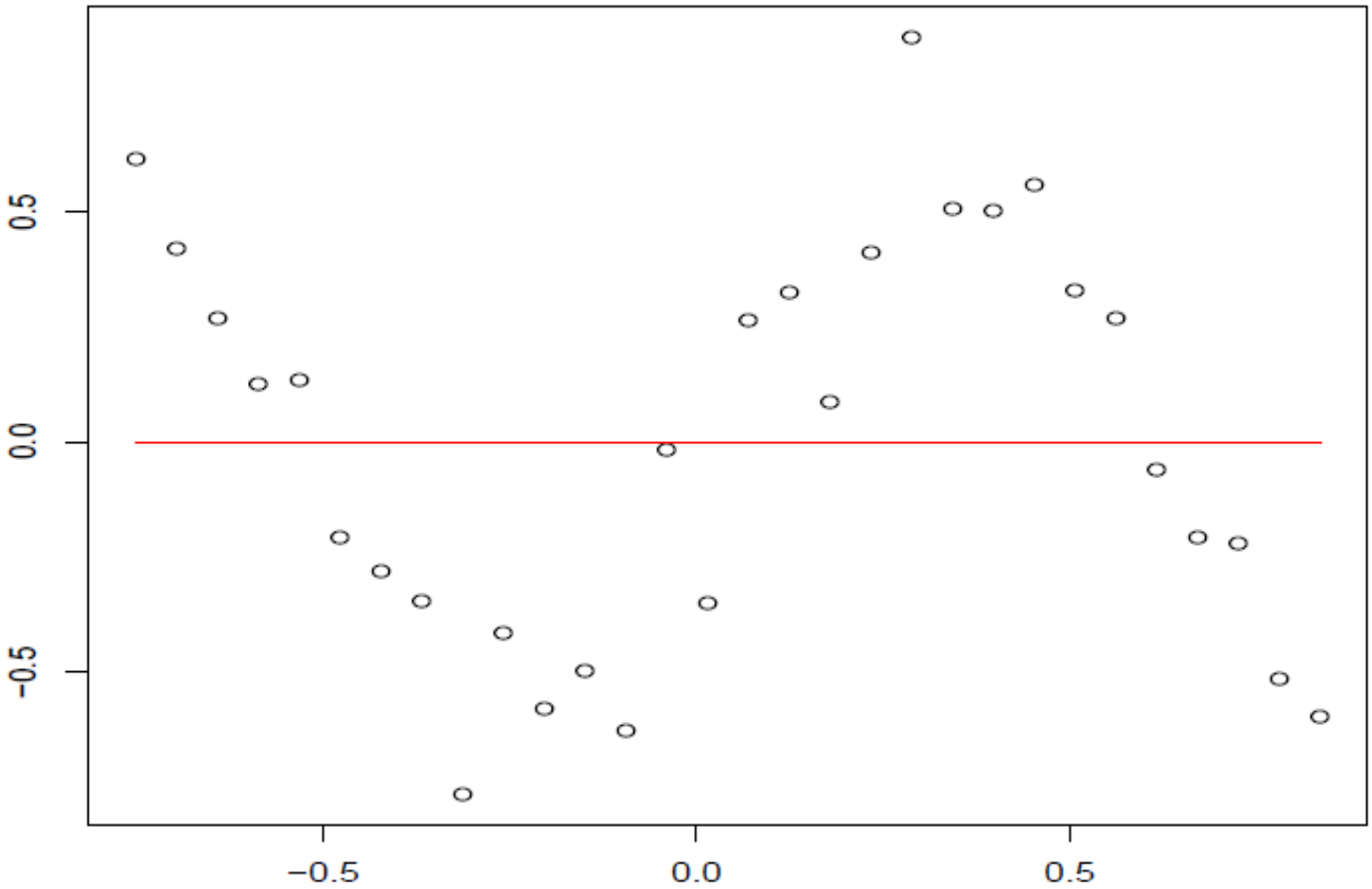
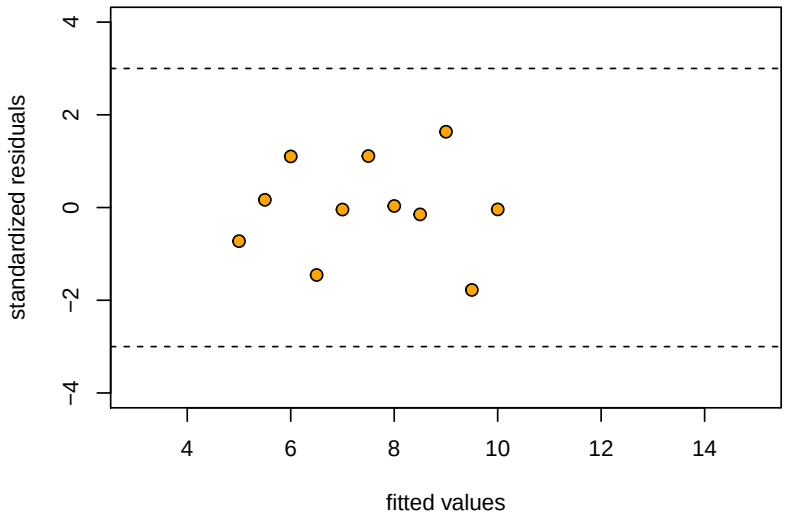
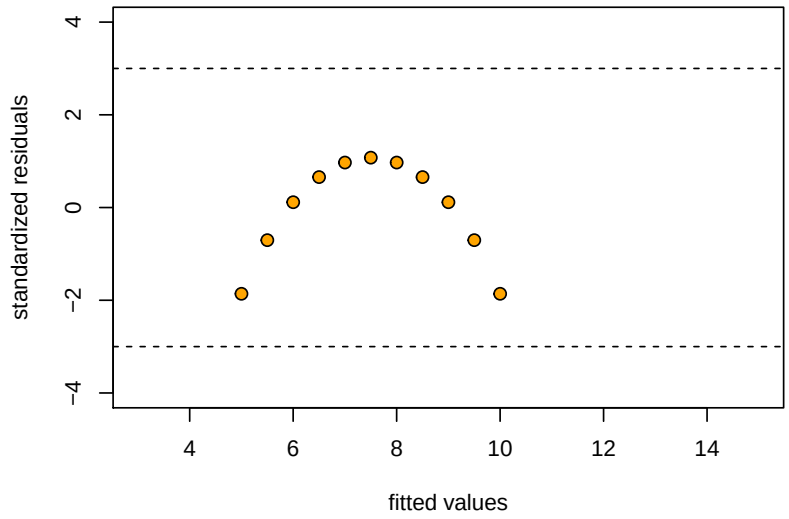
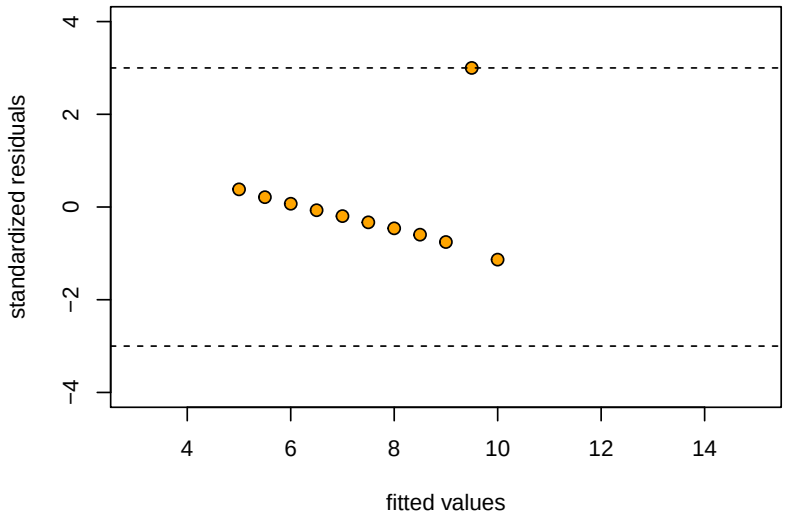


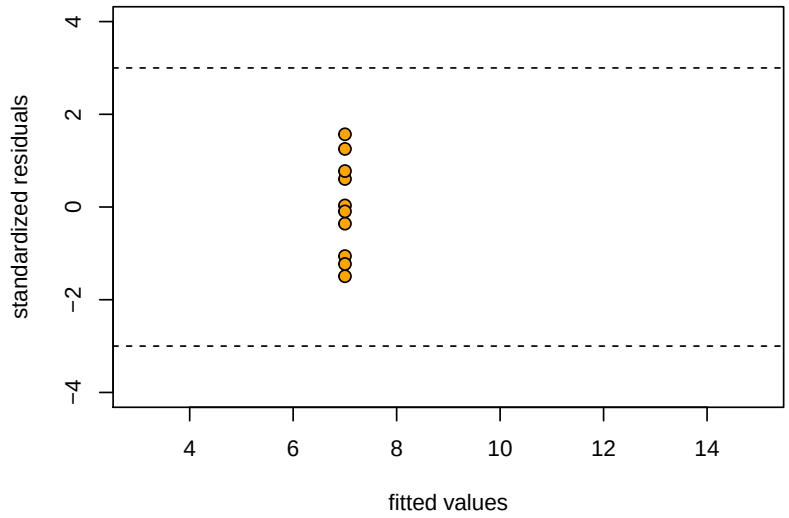
Figure 6: 回归函数可能非线性

案例的残差图


standardized residuals
fitted values
4
6
8
10
12
14
−4
−2
0
2
4


standardized residuals
fitted values
4
6
8
10
12
14
−4
−2
0
2
4


standardized residuals
fitted values
4
6
8
10
12
14
−4
−2
0
2
4


standardized residuals
fitted values
4
6
8
10
12
14
−4
−2
0
2
4

残差图诊断的思路

- 如果残差图中显示非线性，可适当增加自变量的二次项或者交叉项。具体问题具体分析。
- 如果残差图中显示误差方差不相等 (heterogeneity, 方差非齐性)，可以对变量做适当的变换，使得变换后的相应变量具有近似相等的方差 (homogeneity, 方差齐性)。最著名的方法是 **Box-Cox** 变换，见综述论文：

R. M. Sakia. The Box-Cox Transformation Technique: A Review. The Statistician, 41: 169-178, 1992.

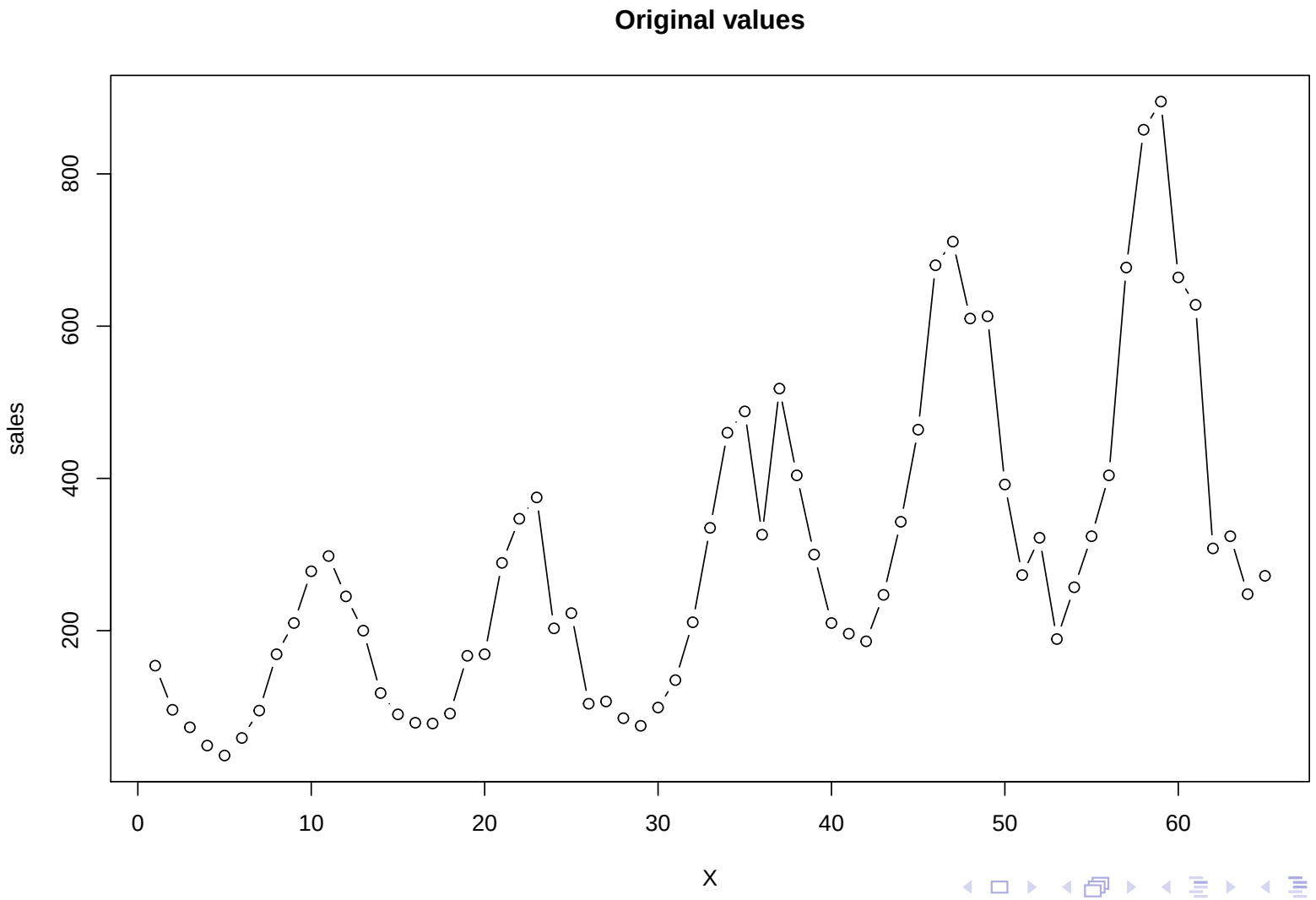
$$Y(\lambda) = \begin{cases} \frac{1}{\lambda}(Y^{\lambda} - 1), & \lambda \neq 0 \\ \log Y, & \lambda = 0 \end{cases}$$

其中λ 是待定的变换参数，可由极大似然法估计。

在 R 中使用命令 `boxcox()` (需要首先运行 `library(MASS)`)

例子

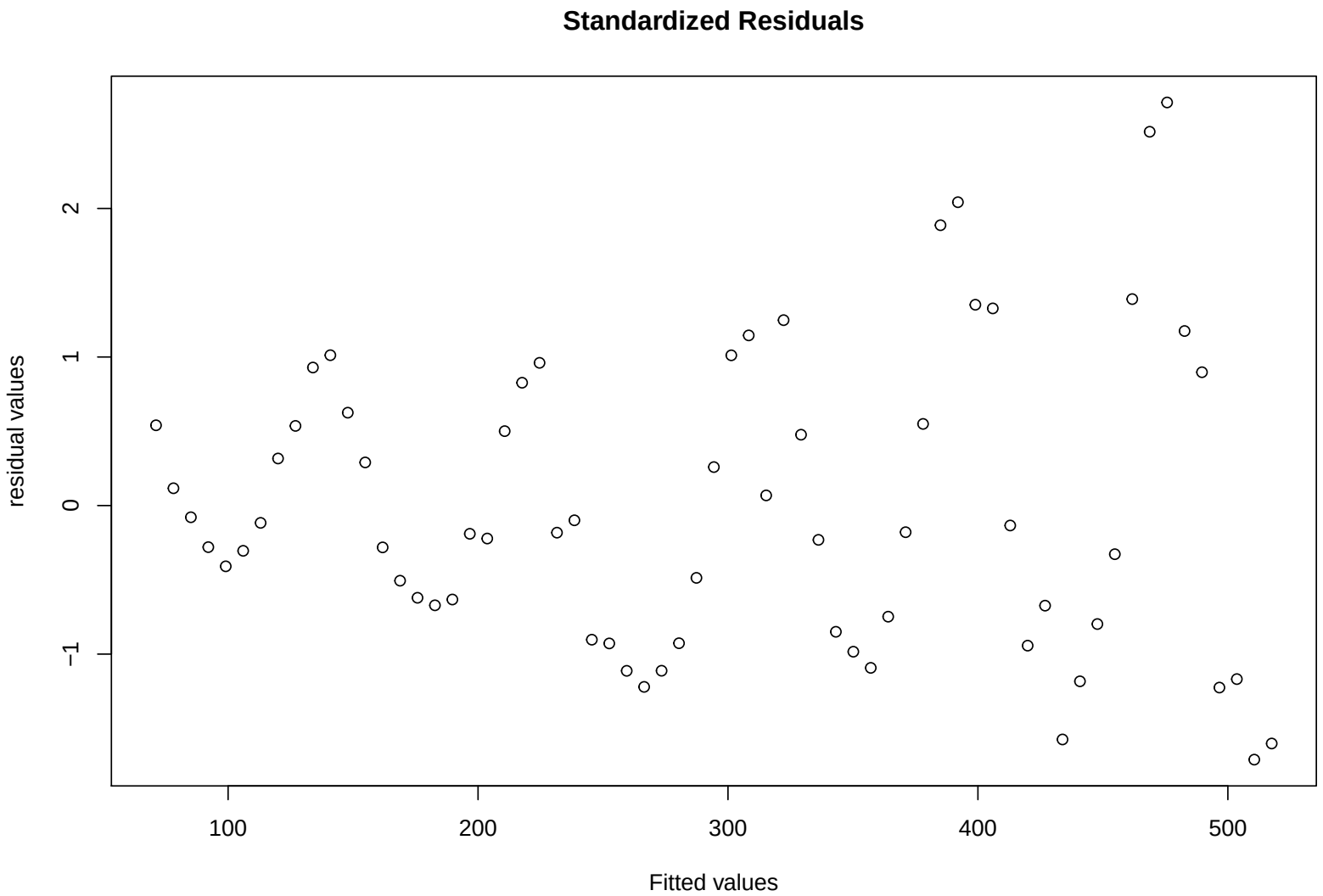
已知某公司从 2000 年 1 月至 2005 年 5 月的逐月销售量，利用所学的统计知识对所建立的模型进行诊断。



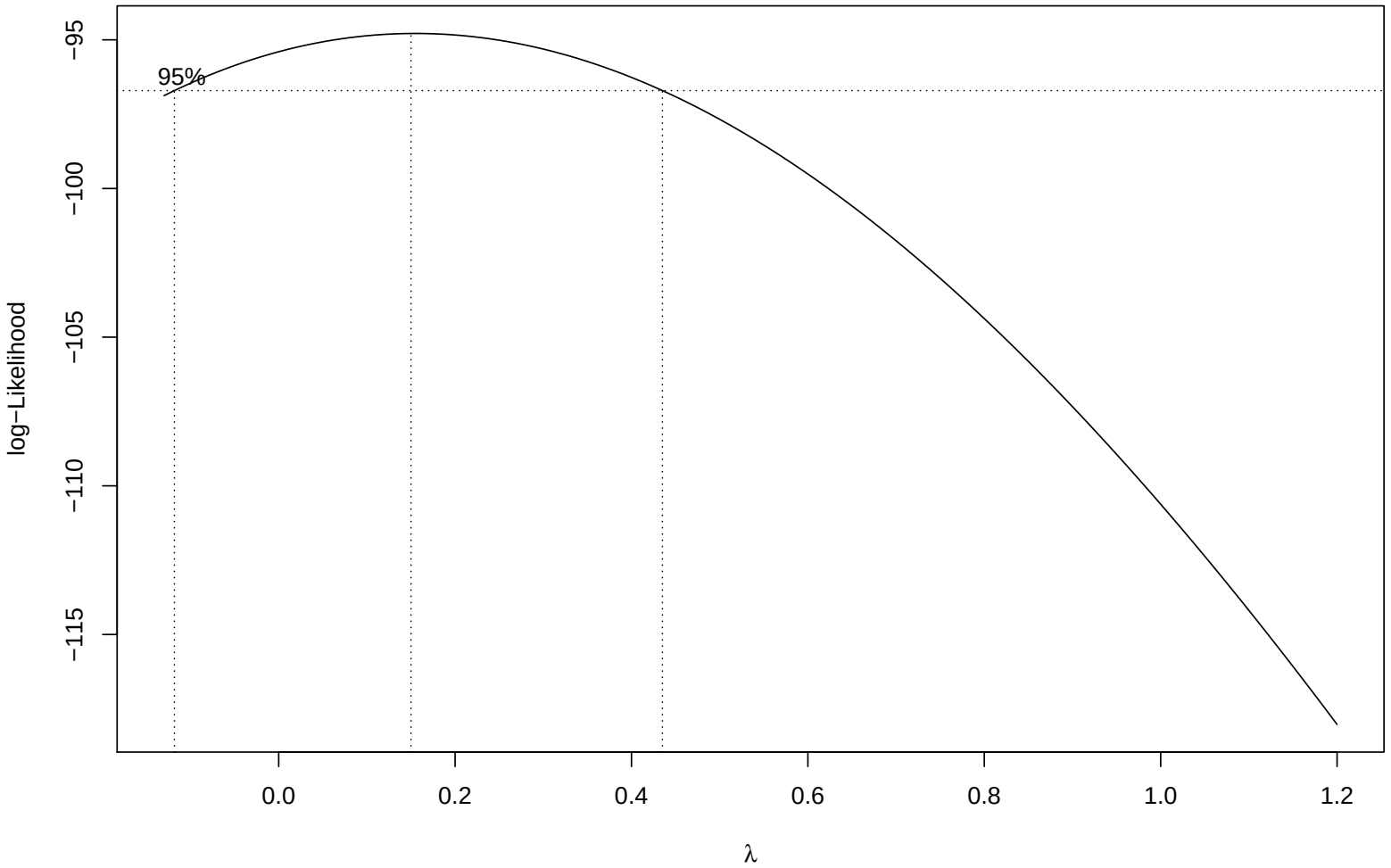
回归结果

```
##
## Call:
## lm(formula = sales ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -262.57 -123.80  -28.58   97.11  419.31
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   64.190     39.683   1.618    0.111
## X              6.975      1.045   6.672 7.43e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 158.1 on 63 degrees of freedom
## Multiple R-squared:  0.414,  Adjusted R-squared:  0.4047
## F-statistic: 44.52 on 1 and 63 DF,  p-value: 7.429e-09
```

残差


Standardized Residuals
residual values
2
1
0
−1
100
200
300
400
500
Fitted values

boxcox 结果

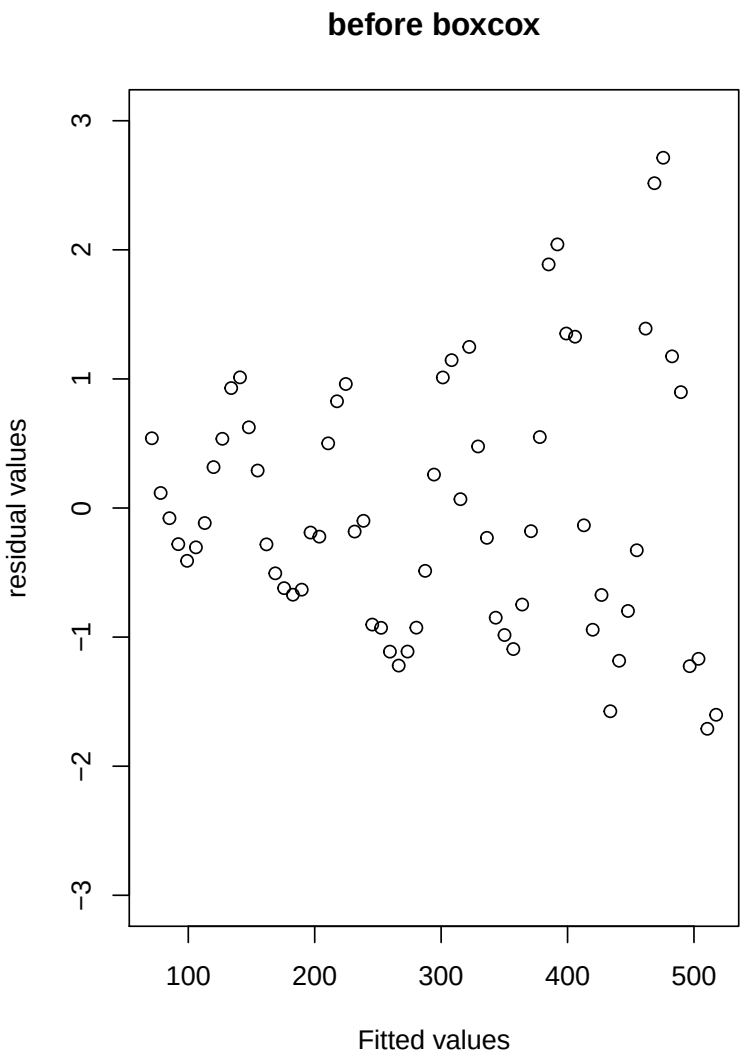


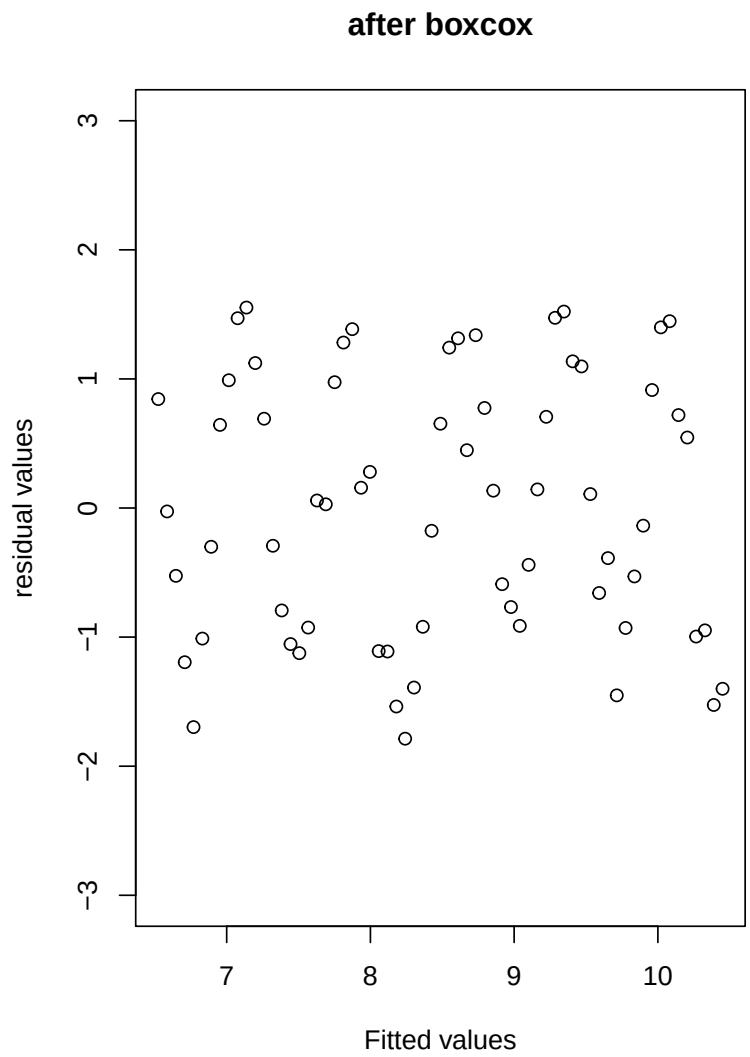

```
## lambda =  0.15
```

修正模型

```
##
## Call:
## lm(formula = (sales^lambda - 1)/lambda ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16849 -1.11715  0.03469  1.09524  1.86477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.462756   0.307030  21.049  < 2e-16 ***
## X           0.061346   0.008088   7.585  1.9e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.223 on 63 degrees of freedom
## Multiple R-squared:  0.4773, Adjusted R-squared:  0.469
## F-statistic: 57.53 on 1 and 63 DF,  p-value: 1.902e-10
```

修正前后的残差对比


before boxcox
residual values
Fitted values
−3
−2
−1
0
1
2
3
100
200
300
400
500


after boxcox
residual values
Fitted values
−3
−2
−1
0
1
2
3
7
8
9
10

离群值 (outlier)

产生离群值的原因：

1. 主观原因：收集和记录数据时出现错误
2. 客观原因：重尾分布（比如，t 分布）和混合分布

离群值的简单判断：

1. 数据散点图
2. 学生化残差图，如果$|t_i| > 3$ (或者 2.5,2)，则对应的数据判定为离群值。
3. 离群值的统计检验方法，M-估计 (Maximum likelihood type estimators)
4. Cook 距离：
   `https://en.wikipedia.org/wiki/Cook%27s_distance`

Cook 距离

Cook 距离定义为：

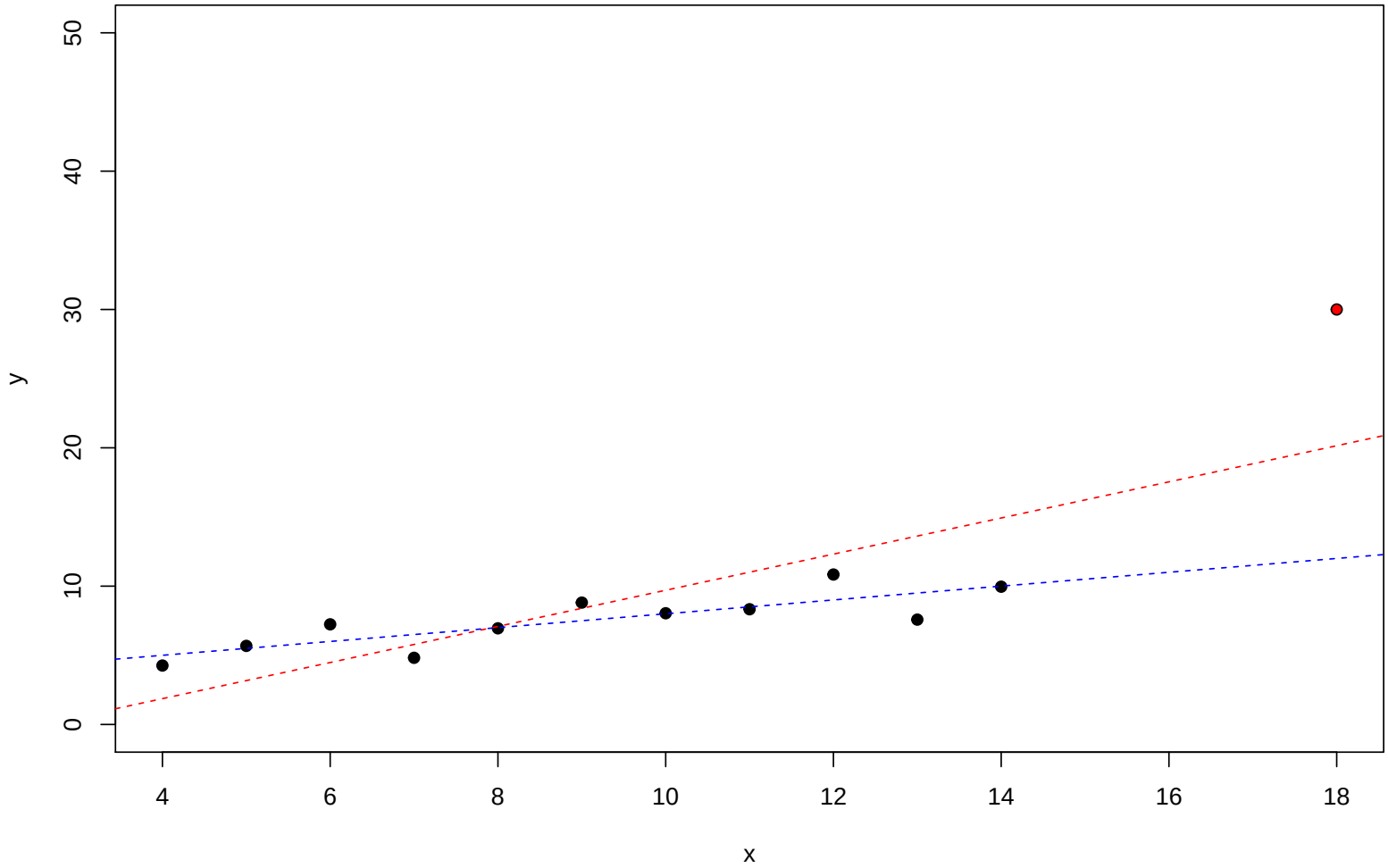
$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^\top X^\top X (\hat{\beta} - \hat{\beta}_{(i)})}{p\hat{\sigma}^2}, i = 1, \ldots, n,$$

- 其中$\hat{\beta}_{(i)}$ 为剔除第i 个数据得到β 的最小二乘估计
- 在 R 中用命令 `cooks.distance()`可以得到 Cook 统计量的值

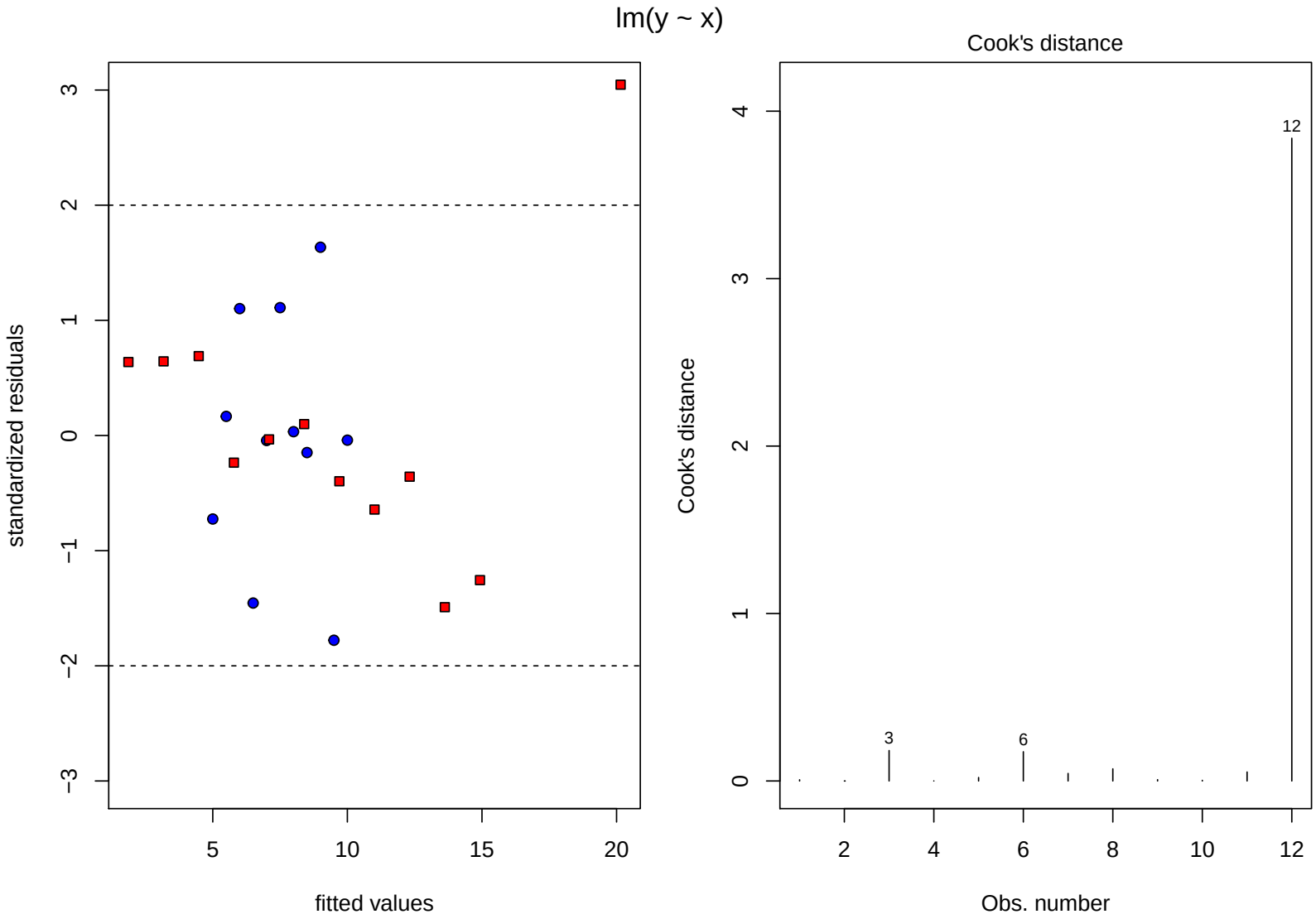
如果发现特别大的D_i 一定要特别注意：

- 检查原始数据是否有误，如有，改正后重新计算；否则，剔除对应的数据
- 如果没有足够理由剔除影响大的数据，就应该采取收集更多的数据或者采用更加稳健的方法以降低强影响数据对估计和推断的影响，从而得到比较稳定的回归方程。

异常值



残差图


lm(y ~ x)
standardized residuals
fitted values
Cook's distance
Cook's distance
Obs. number
12
3
6

变量选择

- 完全子集法
- 向前回归法（每步只能增加）
- 向后回归法（每步只能剔除）
- 逐步回归法（可剔除也可增加）

比较准则

- AIC 准则 (Akaike information criterion)：由日本统计学家 Hirotugu Akaike 提出并命名

$$AIC(A) = \ln Q(A) + \frac{2}{n}\#(A)$$

- BIC 准则 (Bayesian information criterion)：由 Gideon E. Schwarz (1978, Ann. Stat.) 提出

$$BIC(A) = \ln Q(A) + \frac{\log n}{n}\#(A)$$

其中$A \subset \{1, \ldots, p-1\}$，$Q(A)$ 为选择子集A 中变量跑回归的残差平方和，$\#(A)$ 表示A 中元素的个数。**AIC/BIC** 越小越好。

案例

某种水泥在凝固时放出的热量Y 与水泥中四种化学成分X_1, X_2, X_3, X_4 有关，现测得 13 组数据，如表所示。希望从中选出主要的变量，建立Y 关于它们的线性回归方程。

x1	x2	x3	x4	y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

所有变量参与回归

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = cement)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1750 -1.6709  0.2508  1.3783  3.9254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.4054    70.0710   0.891   0.3991
## x1            1.5511     0.7448   2.083   0.0708 .
## x2            0.5102     0.7238   0.705   0.5009
## x3            0.1019     0.7547   0.135   0.8959
## x4           -0.1441     0.7091  -0.203   0.8441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.446 on 8 degrees of freedom
## Multiple R-squared:  0.9824, Adjusted R-squared:  0.9736
## F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

R 语言关键命令

```
step(object, scope, scale = 0,
     direction = c("both", "backward", "forward"),
     trace = 1, keep = NULL, steps = 1000, k = 2, ...)
```

- object: 初始回归模型，比如 `object = lm(y=1,data=cement)`
- scope: 为逐步回归搜索区域，比如 `scope = ~x1+x2+x3+x4`
- trace: 表示是否保留逐步回归过程
- direction: 搜索方向
- k: 为 AIC 惩罚因子，即$AIC(A) = \ln Q(A) + \frac{k}{n}\#(A)$，默认$k = 2$，当$k = \log n$ 是为 BIC 准则。

逐步回归过程

```
## Start:  AIC=71.44
## y ~ 1
##
##        Df Sum of Sq     RSS    AIC
## + x4    1   1831.90  883.87 58.852
## + x2    1   1809.43  906.34 59.178
## + x1    1   1450.08 1265.69 63.519
## + x3    1    776.36 1939.40 69.067
## <none>              2715.76 71.444
##
## Step:  AIC=58.85
## y ~ x4
##
##        Df Sum of Sq     RSS    AIC
## + x1    1    809.10   74.76 28.742
## + x3    1    708.13  175.74 39.853
## <none>               883.87 58.852
## + x2    1     14.99  868.88 60.629
## - x4    1   1831.90 2715.76 71.444
##
## Step:  AIC=28.74
## y ~ x4 + x1
##
##        Df Sum of Sq     RSS    AIC
## + x2    1     26.79   47.97 24.974
## + x3    1     23.93   50.84 25.728
## <none>               74.76 28.742
## - x1    1    809.10  883.87 58.852
## - x4    1   1190.92 1265.69 63.519
##
## Step:  AIC=24.97
## y ~ x4 + x1 + x2
##
```

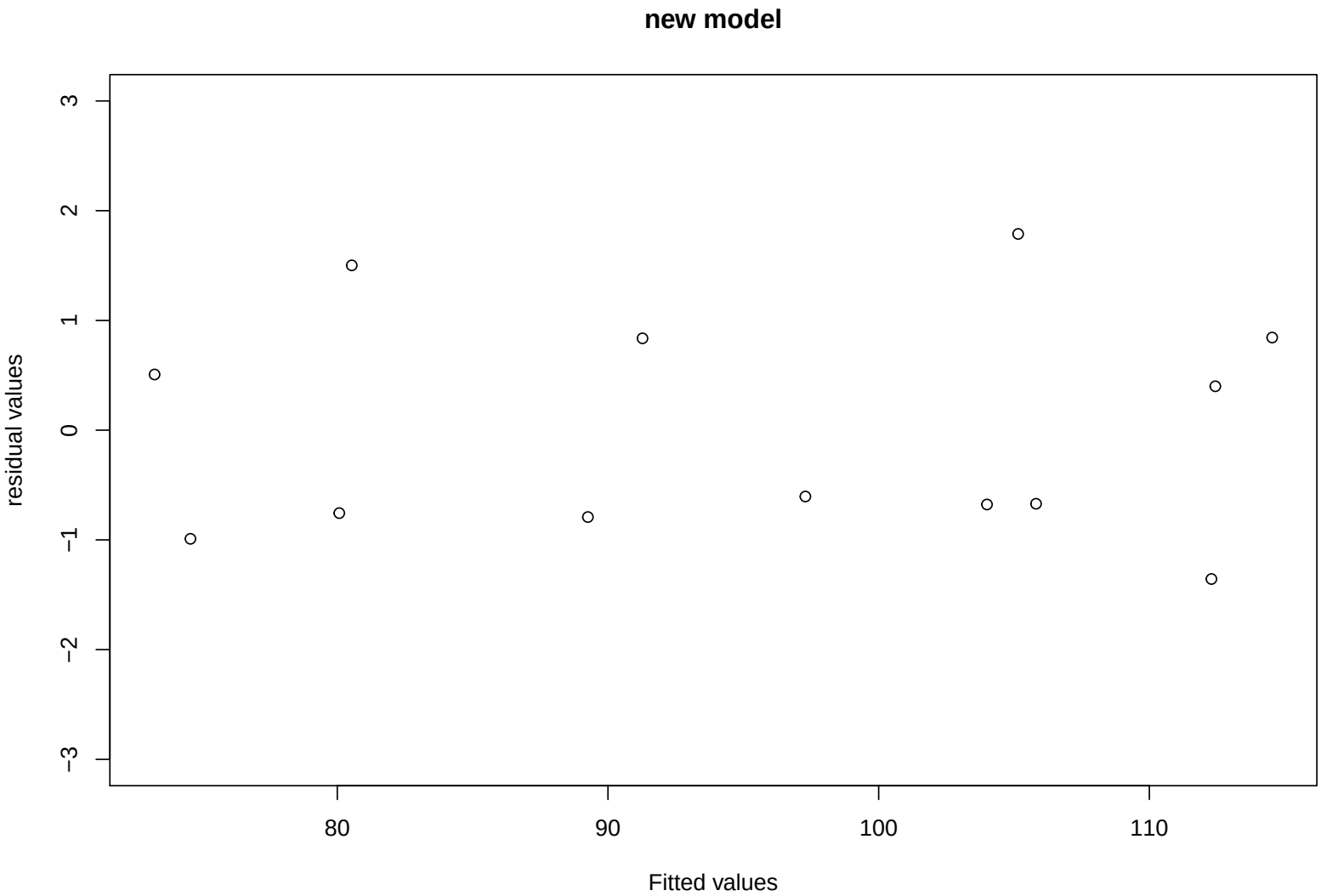
逐步回归结果（剔除第 3 个变量）

```
##
## Call:
## lm(formula = y ~ x4 + x1 + x2, data = cement)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0919 -1.8016  0.2562  1.2818  3.8982
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.6483    14.1424   5.066 0.000675 ***
## x4           -0.2365     0.1733  -1.365 0.205395
## x1            1.4519     0.1170  12.410 5.78e-07 ***
## x2            0.4161     0.1856   2.242 0.051687 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.309 on 9 degrees of freedom
## Multiple R-squared:  0.9823, Adjusted R-squared:  0.9764
## F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

模型改进：剔除第 4 个变量

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = cement)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.893 -1.574 -1.302  1.363  4.048
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
## x1           1.46831    0.12130   12.11 2.69e-07 ***
## x2           0.66225    0.04585   14.44 5.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.406 on 10 degrees of freedom
## Multiple R-squared:  0.9787, Adjusted R-squared:  0.9744
## F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

改进模型的残差图


new model
residual values
3
2
1
0
−1
−2
−3
80
90
100
110
Fitted values

LASSO

现代变量选择方法：LASSO (Least Absolute Shrinkage & Selection Operator) 是斯坦福大学统计系 Tibshirani 于 1996 年发表的著名论文 “Regreesion shrinkage and selection via the LASSO” (Journal of Royal Statistical Society, Seriers B, 58, 267-288) 中所提出的一种变量选择方法。

Figure 7: Rob Tibshirani

https://statweb.stanford.edu/~tibs/

$$\hat{\beta}_{LASSO} = \arg\min ||Y - X\beta||^2 \text{ subject to } \sum_{i=0}^{p-1} |\beta_i| \leq t,$$

等价于

$$\hat{\beta}_{LASSO} = \arg\min ||Y - X\beta||^2 + \lambda \sum_{i=0}^{p-1} |\beta_i|.$$

在 R 语言中可以使用 `glmnet`包来实施 LASSO 算法。

回归分析与因果分析

即使建立了回归关系式并且统计检验证明相关关系成立，也只能说明研究的变量是统计相关的，而不能就此断定变量之间有因果关系。

案例 (Ice Cream Causes Polio)：小儿麻痹症疫苗发明前，美国北卡罗来纳州卫生部研究人员通过分析冰淇淋消费量和小儿麻痹症的关系发现当冰淇淋消费量增加时，小儿麻痹疾病也增加。州卫生部发生警告反对吃冰淇淋来试图阻止这种疾病的传播。

没有观察的混杂因素 ——温度

Polio and ice cream consumption both increase in the summertime. Summer is when the polio virus thrived.

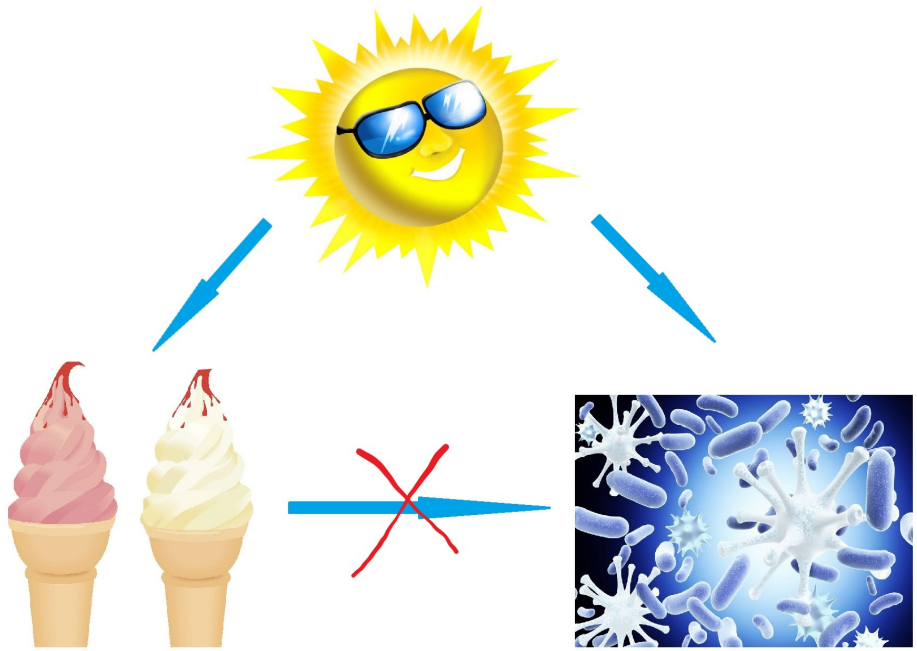


Figure 8: The danger of mixing up causality and correlation