回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第27讲 回归诊断实例

2020.5.27

$$D_i = \frac{1}{p} \times \frac{h_{ii}}{1-h_{ii}} r_i^2$$

影响度量表达为r和h的函数（续上节课）

Cook距离定义：$D_i = \dfrac{\| X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}^{(-i)} \|^2}{p\hat{\sigma}^2} = \dfrac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)})^{\mathsf{T}} X^{\mathsf{T}} X (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)})}{p\hat{\sigma}^2}$

*Cook*距离理解为$\hat{\boldsymbol{\beta}}$与$\hat{\boldsymbol{\beta}}^{(-i)}$的标准化平均距离：

$\hat{\boldsymbol{\beta}} \sim (\boldsymbol{\beta}, \sigma^2 (X^{\mathsf{T}} X)^{-1})$, 标准化 $(X^{\mathsf{T}} X)^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / \sigma \sim (0, I_p)$,

平均长度平方：$\| (X^{\mathsf{T}} X)^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / \sigma \|^2 / p = \dfrac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathsf{T}} (X^{\mathsf{T}} X)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p\sigma^2}$ （正态假设下 $\sim \chi^2_p$）

plug-in $\hat{\sigma}^2$： $\| (X^{\mathsf{T}} X)^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / \hat{\sigma} \|^2 / p = \dfrac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathsf{T}} (X^{\mathsf{T}} X)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p\hat{\sigma}^2}$ （正态假设下 $\sim F_{p,n-p}$）

取$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(-i)}$,即为Cook距离.

命题1: （1）$D_i = \dfrac{1}{p} \times \dfrac{h_{ii}}{1-h_{ii}} r_i^2$, 其中 $r_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ 为标准化残差。

（2）$\mathrm{DFFITS}_i = r_i^* \sqrt{\dfrac{h_{ii}}{1-h_{ii}}}$, 其中 $r_i^* = \dfrac{e_i}{\hat{\sigma}^{(-i)}\sqrt{1-h_{ii}}}$ 学生化残差.

引理1：设$A_{p\times p}$对称，$\mathbf{x},\mathbf{y}$为$p\times 1$向量，则$(A+\mathbf{x}\mathbf{y}^{\mathsf{T}})^{-1}=A^{-1}-\dfrac{A^{-1}\mathbf{x}\mathbf{y}^{\mathsf{T}}A^{-1}}{1+\mathbf{x}^{\mathsf{T}}A^{-1}\mathbf{y}}$,

引理2：假设线性模型中，对任何$i=1,\ldots,n$，$X_{(-i)}{}^{\mathsf{T}}X_{(-i)}$可逆，则

(1) $\hat{\boldsymbol{\beta}}^{(-i)}=\hat{\boldsymbol{\beta}}-\dfrac{(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i e_i}{1-h_{ii}}$，$e_i^{(-i)}\doteq y_i-\mathbf{x}_i\hat{\boldsymbol{\beta}}^{(-i)}=\dfrac{e_i}{1-h_{ii}}$

(2) 记$RSS^{(-i)}=\|\mathbf{y}_{(-i)}-X_{(-i)}\hat{\boldsymbol{\beta}}^{(-i)}\|^2$为删除数据点$i$后的残差平方和，则

$$RSS^{(-i)}=RSS-\frac{e_i^2}{1-h_{ii}},\ \hat{\sigma}^{(-i)2}=\hat{\sigma}^2+\frac{1}{n-p-1}\left(\hat{\sigma}^2-\frac{e_i^2}{1-h_{ii}}\right).$$

注：引理2说明，基于所有数据的回归结果我们就可以计算$\hat{\boldsymbol{\beta}}^{(-i)},\hat{\sigma}^{(-i)}$,进而直接计算DFBETAS, DFFITS$_i$和D$_i$等统计量，而不必删除一行数据后重新回归。

证明引理2：由于 $X=\begin{pmatrix}\mathbf{x}_1^{\mathsf{T}}\\ \dots\\ \mathbf{x}_n^{\mathsf{T}}\end{pmatrix}$，有 $X^{\mathsf{T}}X=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}$，$X^{\mathsf{T}}\mathbf{y}=\sum_{i=1}^{n}\mathbf{x}_i y_i$

$$\Rightarrow X^{\mathsf{T}}X=X_{(-i)}{}^{\mathsf{T}}X_{(-i)}+\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}},\quad X^{\mathsf{T}}\mathbf{y}=\sum_{j=1}^{n}\mathbf{x}_j y_j=X_{(-i)}{}^{\mathsf{T}}\mathbf{y}_{(-i)}+\mathbf{x}_i y_i$$

$$(1)\ \hat{\boldsymbol{\beta}}^{(-i)}=\left(X_{(-i)}{}^{\mathsf{T}}X_{(-i)}\right)^{-1}X_{(-i)}{}^{\mathsf{T}}\mathbf{y}_{(-i)}=\left(X^{\mathsf{T}}X-\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)^{-1}\left(X^{\mathsf{T}}\mathbf{y}-\mathbf{x}_i y_i\right)$$

由引理1

$$=\left((X^{\mathsf{T}}X)^{-1}+\frac{(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}(X'X)^{-1}}{1-\mathbf{x}_i^{\mathsf{T}}(X'X)^{-1}\mathbf{x}_i}\right)\left(X^{\mathsf{T}}\mathbf{y}-\mathbf{x}_i y_i\right)$$

注意 $h_{ii}=\mathbf{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i$

$$=\hat{\boldsymbol{\beta}}-(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i y_i+\frac{(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}-(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i y_i}{1-h_{ii}}$$

$$=\hat{\boldsymbol{\beta}}-(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i y_i+\frac{(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\hat{\beta}-h_{ii}(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i y_i}{1-h_{ii}}$$

$$=\hat{\boldsymbol{\beta}}-\frac{(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i\left(y_i-\mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}\right)}{1-h_{ii}}=\hat{\boldsymbol{\beta}}-\frac{(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i e_i}{1-h_{ii}}$$

(2)只需证明 $RSS^{(-i)} = RSS - \dfrac{e_i^2}{1-h_{ii}}$

$$RSS^{(-i)} = \| \mathbf{y}_{(-i)} - X_{(-i)}\hat{\boldsymbol{\beta}}^{(-i)} \|^2 \overset{由(1)}{=} \left\| \mathbf{y}_{(-i)} - X_{(-i)}\hat{\boldsymbol{\beta}} + \frac{1}{1-h_{ii}} X_{(-i)}\left[(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i e_i\right] \right\|^2$$

$\mathbf{z}$

$$= \| \mathbf{y}_{(-i)} - X_{(-i)}\hat{\boldsymbol{\beta}} \|^2 + \mathbf{z}^{\mathsf{T}}\mathbf{z} + 2\mathbf{z}^{\mathsf{T}}\left(\mathbf{y}_{(-i)} - X_{(-i)}\hat{\boldsymbol{\beta}}\right)$$

显然, $\| \mathbf{y}_{(-i)} - X_{(-i)}\hat{\boldsymbol{\beta}} \|^2 = \sum_{k\neq i} e_k^2 = RSS - e_i^2$

可以验证：$\mathbf{z}^{\mathsf{T}}\mathbf{z} = \dfrac{e_i^2 h_{ii}}{1-h_{ii}}$, $\quad 2\mathbf{z}^{\mathsf{T}}\left(\mathbf{y}_{(-i)} - X_{(-i)}\hat{\boldsymbol{\beta}}\right) = -2\dfrac{e_i^2 h_{ii}}{1-h_{ii}}$

命题1的证明：由引理2，$X\hat{\boldsymbol{\beta}} - X\hat{\boldsymbol{\beta}}^{(-i)} = \dfrac{X(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i e_i}{1-h_{ii}}$

所以 $D_i = \dfrac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)})^{\mathsf{T}} X^{\mathsf{T}} X (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(-i)})}{p\hat{\sigma}^2} = \dfrac{1}{p} \times \dfrac{h_{ii}}{1-h_{ii}} r_i^2$

（2）由 $\left(\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(-i)}\right)_i = \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}} - \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{(-i)} = \dfrac{\mathbf{x}_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\mathbf{x}_i e_i}{1-h_{ii}} = \dfrac{h_{ii}e_i}{1-h_{ii}}$

$$\Rightarrow DFFITS_i = \frac{\left(\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(-i)}\right)_i}{\hat{\sigma}^{(-i)}\sqrt{h_{ii}}} = r_i^* \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

注：引理2和命题1表明DFBETAS, DFFITS, D度量都与杠杆和标准化残差有关，当这些量较大时，蕴含了h_{ii}较大（x_i高杠杆）或 $|r_i|$较大（y_i异常）。

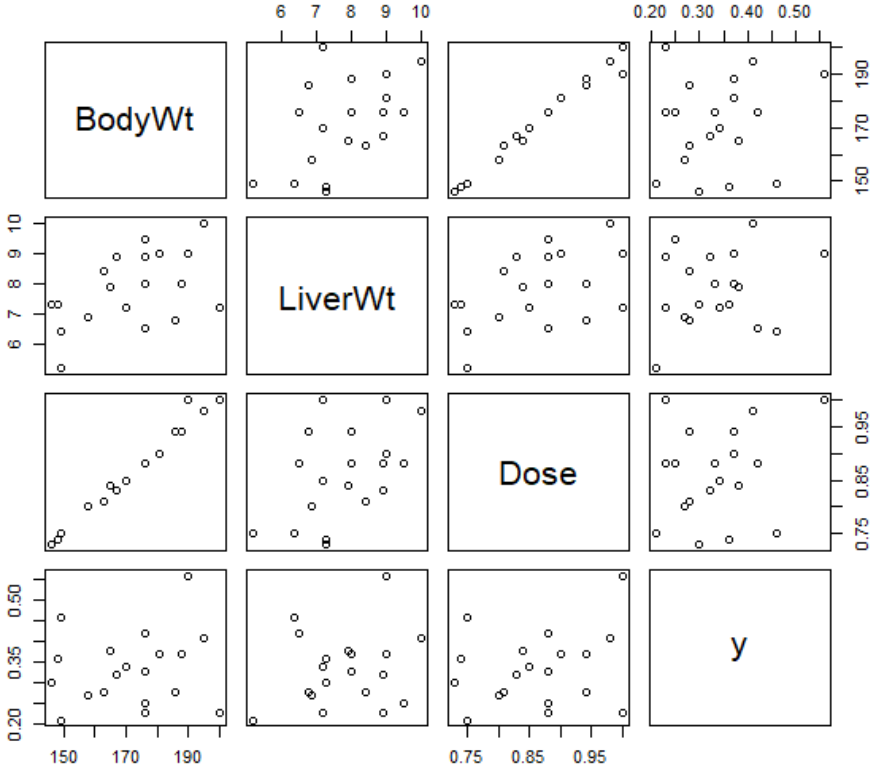
回归诊断实例分析

回归分析的回归诊断目的在于发现线性模型拟合的问题，以改进模型拟合效果。大致（但不限于）步骤如下：

- 了解问题背景和数据来源。
- 探索数据分析：变量两两散点图(plot), 直方图，盒形图，相关系数…
- 建立初始回归模型，进行回归诊断：
  - 残差分析： 发现非线性、异方差 → 数据变换 (Box-Cox)。
  - 影响分析： 发现高影响点 → 删除可以删除的高影响点。
- 再诊断，查看残差 VS 各个自变量的残差图。
  - 如果还有非线性：尝试非单调变换（或非线性模型）
  - 如果还有异方差：如果是调查数据，应用WLS； 如果不是，发现异方差结构特点，并应用IRLS.

实例1：白鼠试验（高影响点）

一项试验希望研究动物肝脏对某种药物的吸收能力。为此随机选取19只小白鼠，口服该药物，剂量大小由体重决定 (大约为 40mg/kg 乘以体重)。一段时间后，测出肝中与所含药物重量，除以服用的剂量，得到肝吸收百分比 y。理论上，y与体重,肝重, 剂量应该没有关系。数据集： rat (alr3)

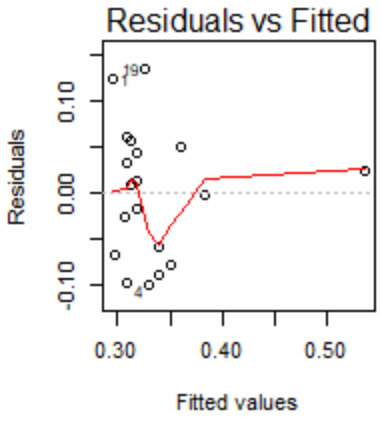


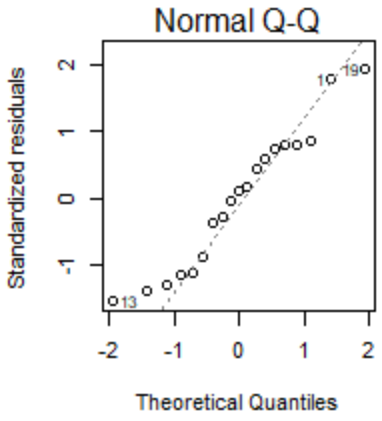
ID	BodyWt	LiverWt	Dose	y
1	176	6.5	0.88	0.42
2	176	9.5	0.88	0.25
3	190	9	1	0.56
4	176	8.9	0.88	0.23
5	200	7.2	1	0.23
6	167	8.9	0.83	0.32
7	188	8	0.94	0.37
8	195	10	0.98	0.41
9	176	8	0.88	0.33
10	165	7.9	0.84	0.38
11	158	6.9	0.8	0.27
12	148	7.3	0.74	0.36
13	149	5.2	0.75	0.21
14	163	8.4	0.81	0.28
15	170	7.2	0.85	0.34
16	186	6.8	0.94	0.28
17	146	7.3	0.73	0.3
18	181	9	0.9	0.37
19	149	6.4	0.75	0.46

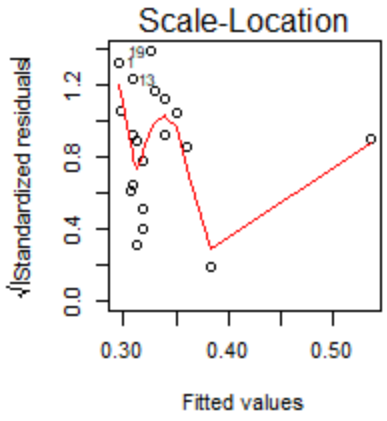
但回归分析结果表明BodyWt和Dose都显著:

lm(y ~ ., data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.266	0.195	1.4	0.19
BodyWt	-0.021	0.008	-2.7	0.02 *
LiverWt	0.014	0.017	0.8	0.42
Dose	4.178	1.523	2.7	0.02 *





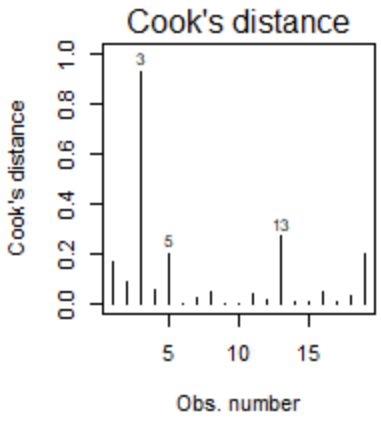


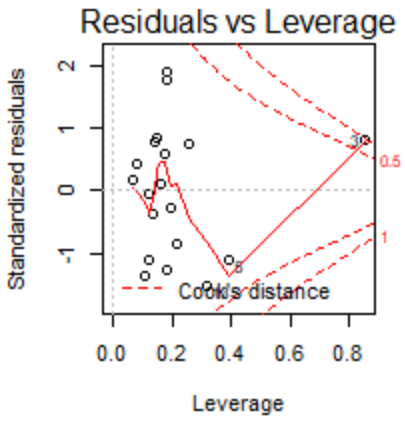
第3个小白鼠的

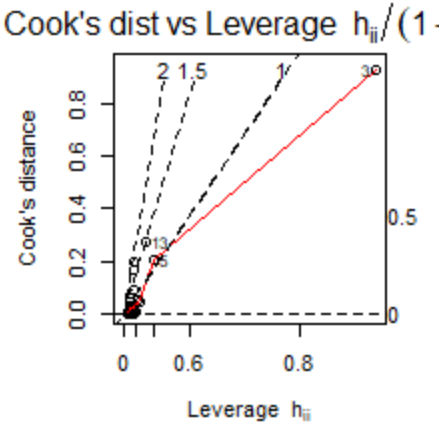
h =0.85,

D =0.93,

是高影响点。

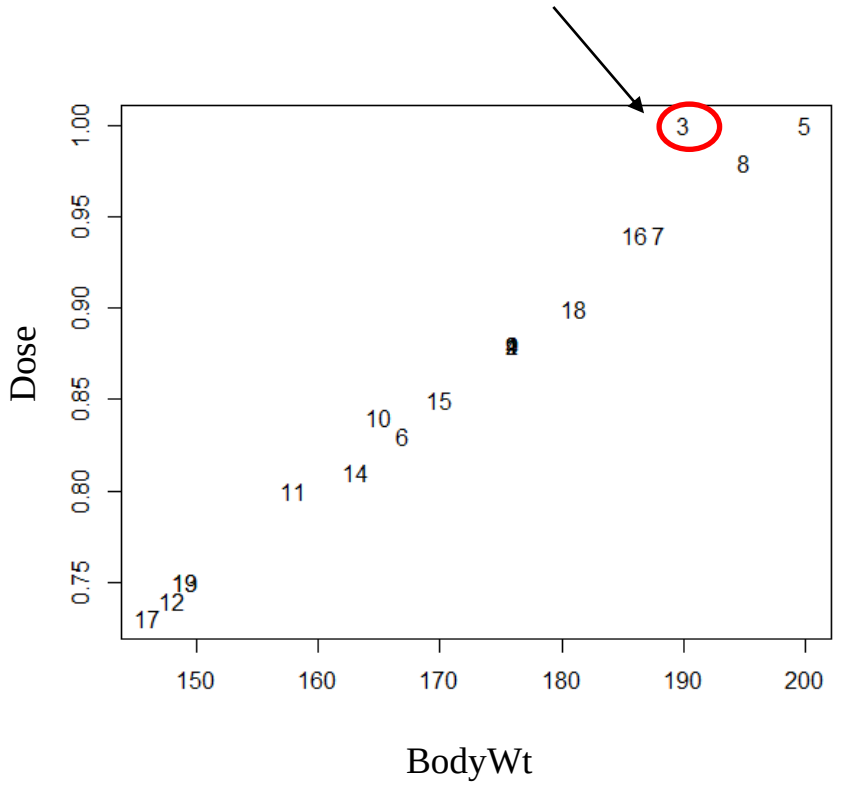






删除高影响点（慎重）

第3只白鼠的体重不是最大，
但其剂量最大（1），有误。



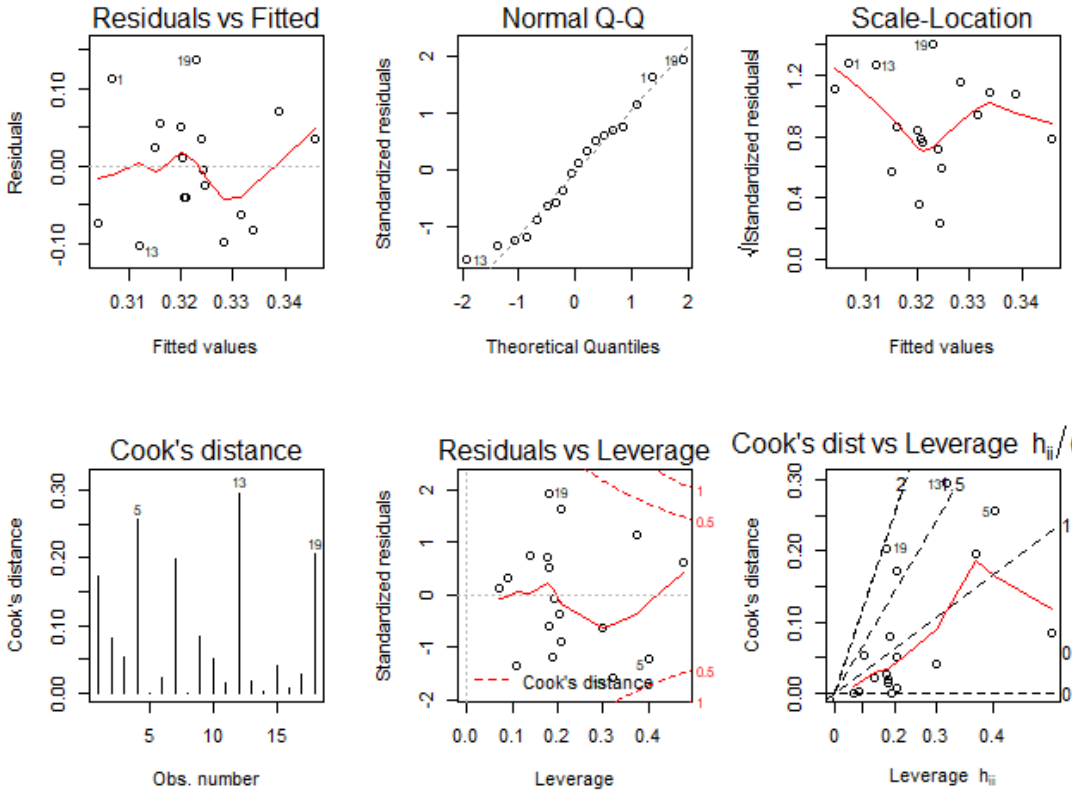
	BodyWt	LiverWt	Dose	y
1	176	6.5	0.88	0.42
2	176	9.5	0.88	0.25
3	**190**	**9**	**1**	**0.56**
4	176	8.9	0.88	0.23
5	**200**	**7.2**	**1**	**0.23**
6	167	8.9	0.83	0.32
7	188	8	0.94	0.37
8	**195**	**10**	**0.98**	**0.41**
9	176	8	0.88	0.33
10	165	7.9	0.84	0.38
11	158	6.9	0.8	0.27
12	148	7.3	0.74	0.36
13	149	5.2	0.75	0.21
14	163	8.4	0.81	0.28
15	170	7.2	0.85	0.34
16	186	6.8	0.94	0.28
17	146	7.3	0.73	0.3
18	181	9	0.9	0.37
19	149	6.4	0.75	0.46

删除第三行数据重新分析，所有自变量不再显著，这符合实际情况

lm(formula = y ~ ., data = rat[-3,])

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.311	0.205	1.52	0.15
BodyWt	-0.008	0.019	-0.42	0.68
LiverWt	0.009	0.019	0.48	0.64
Dose	1.485	3.713	0.40	0.70



另一种策略（高度相关的自变量）：

因为BodyWt, Dose几乎完全成正比，模型中可去掉BodyWt 或Dose之一。

lm(y ~ . - Dose, data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.17836	0.22778	0.8	0.4
BodyWt	0.00035	0.00151	0.2	0.8
LiverWt	0.01233	0.02041	0.6	0.6

lm(y ~ . - BodyWt, data = rat)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1174	0.2191	0.5	0.6
LiverWt	0.0087	0.0201	0.4	0.7
Dose	0.1736	0.2863	0.6	0.6

变量不再显著，且第3只白鼠不再是高影响点。对与第3只小白鼠，(BodyWt,Dose)异常，但BodyWt不太异常，Dose 也不太异常。

实例2：杂志广告收入（BC变换）

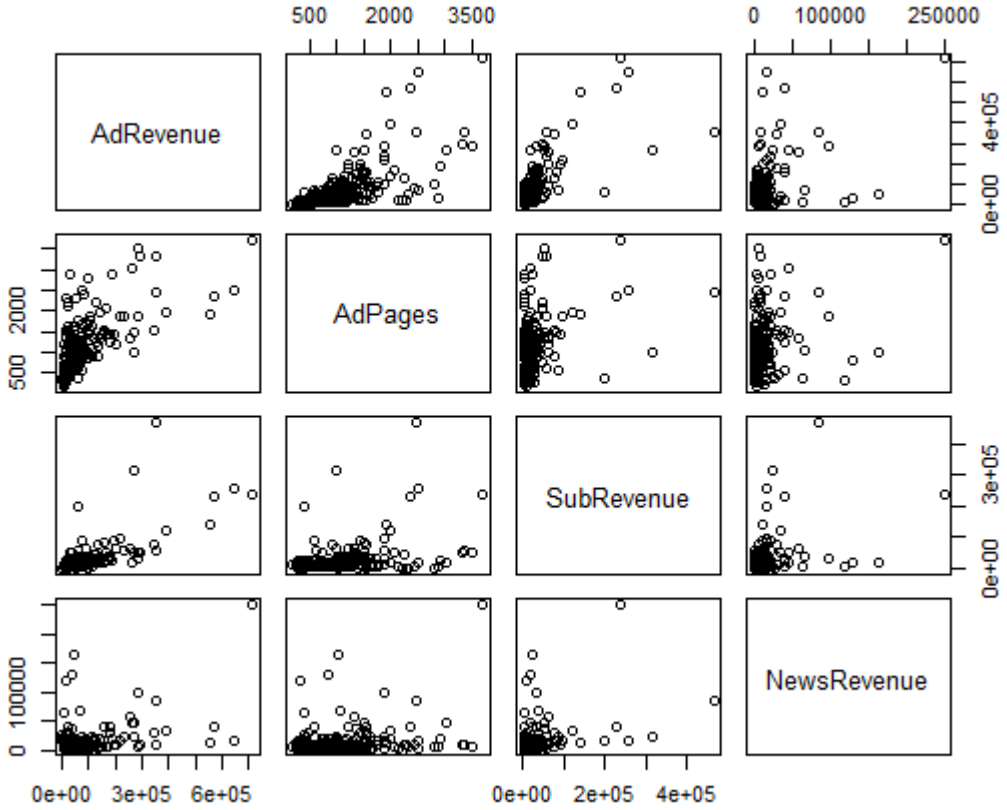
数据magazines (http://staff.ustc.edu.cn/~ynyang/lm2020/lab/magazines.xls) 统计了2003年204 种美国杂志广告收入情况。关心的问题是付费广告页数与广告收入的关系。4个变量如下：

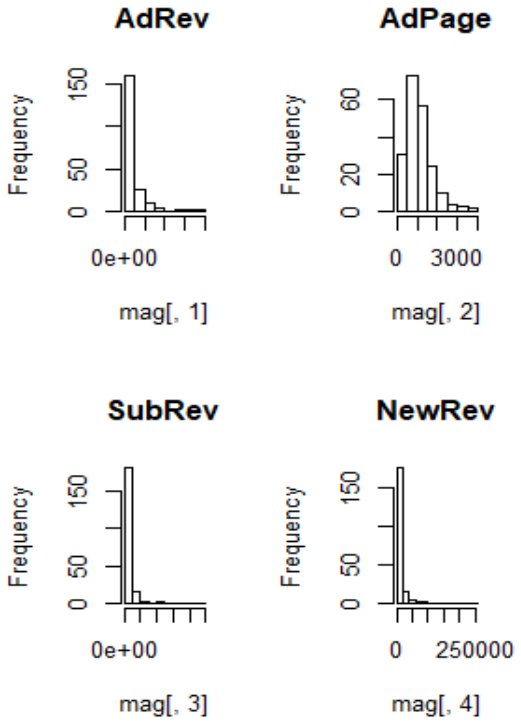
变量	解释
AdRevenue	广告收入（1000$)
AdPages	付费广告页数
SubRevenue	订阅收入
NewsRevenue	零售收入（newsstand sales）

Magazine	AdRevenue	AdPages	SubRevenue	NewsRevenue
Weekly World News	2280	300	854	16568
National Examiner	3382	380	968	27215
J-14	4218	250	2206	12453
Soap Opera Weekly	4622	439	5555	24282
Easyriders	5121	523.69	4155	9929
Official Xbox Magazine	5838	541.66	4311	10320
Weight Watchers	6986	287.27	9202	4048
Globe	7634	380	2180	63771

探索数据分析(plot,corrplot,boxplot, hist etc)

plot(mag)； pairs(mag)



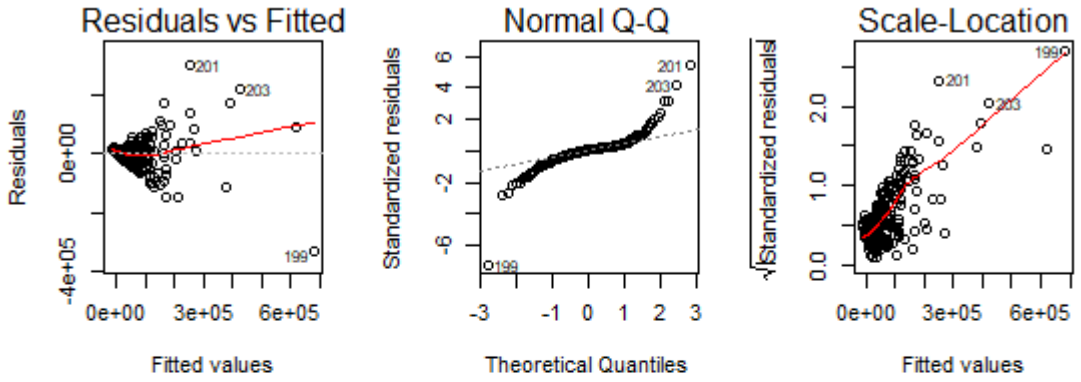


AdRevenu： 与AdPages ， SubRevenue呈正的相关关系；
SubRevenue、NewsAvenue 与AdPages关系不大。
各个变量都是偏态分布的(skewed), 这提示应该做对数或幂次小于1的幂次变换。

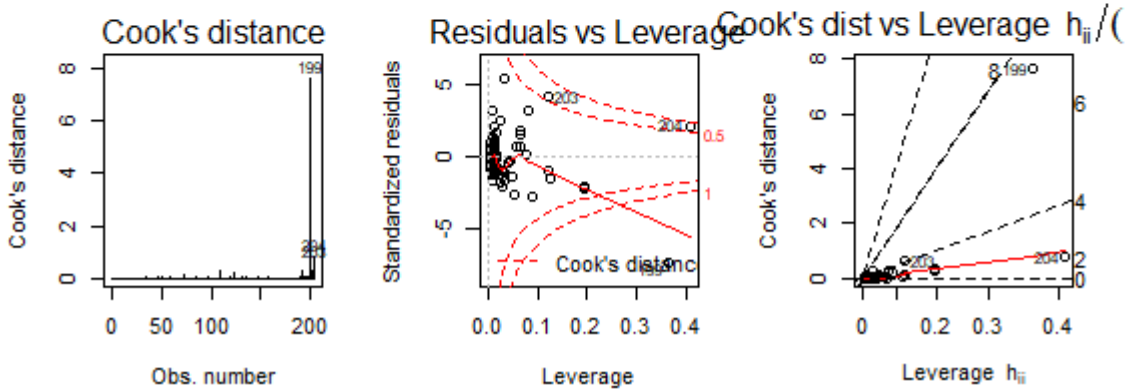
fit0：拟合初始模型

```
fit0 = lm(AdRevenue~. , data=magazines)
plot(fit0,1:6 )
```

$$\text{AdRevenue} = \beta_0 + \beta_1 \times \text{Adpages} + \beta_2 \times \text{SubRevenues} + \beta_3 \times \text{NewsRevenue} + \varepsilon,$$



残差图（第一个图）表明方差不齐；
QQ图说明残差分布非正态。
第三个图表明很强的趋势

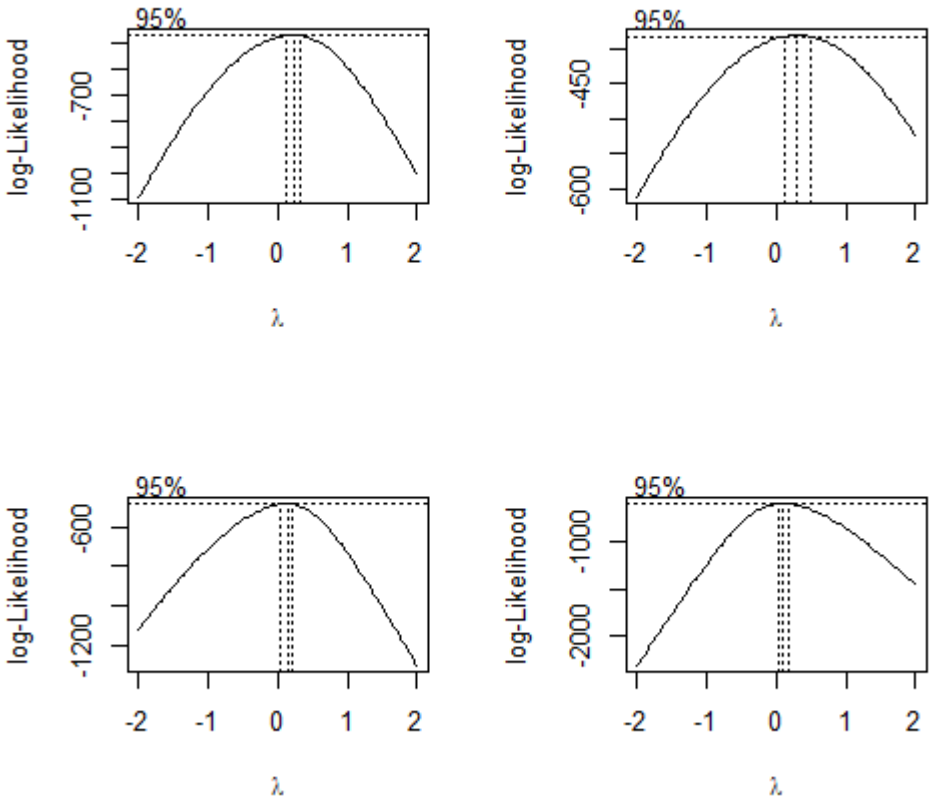


有高影响点，$D_{199}=8$，
该杂志为 TV guide, 与其它杂志明显不同，可以考虑删除。但删除之后发现192 (Reader's Digest)又变成高影响($D_{192}=4$)

Box-Cox变换

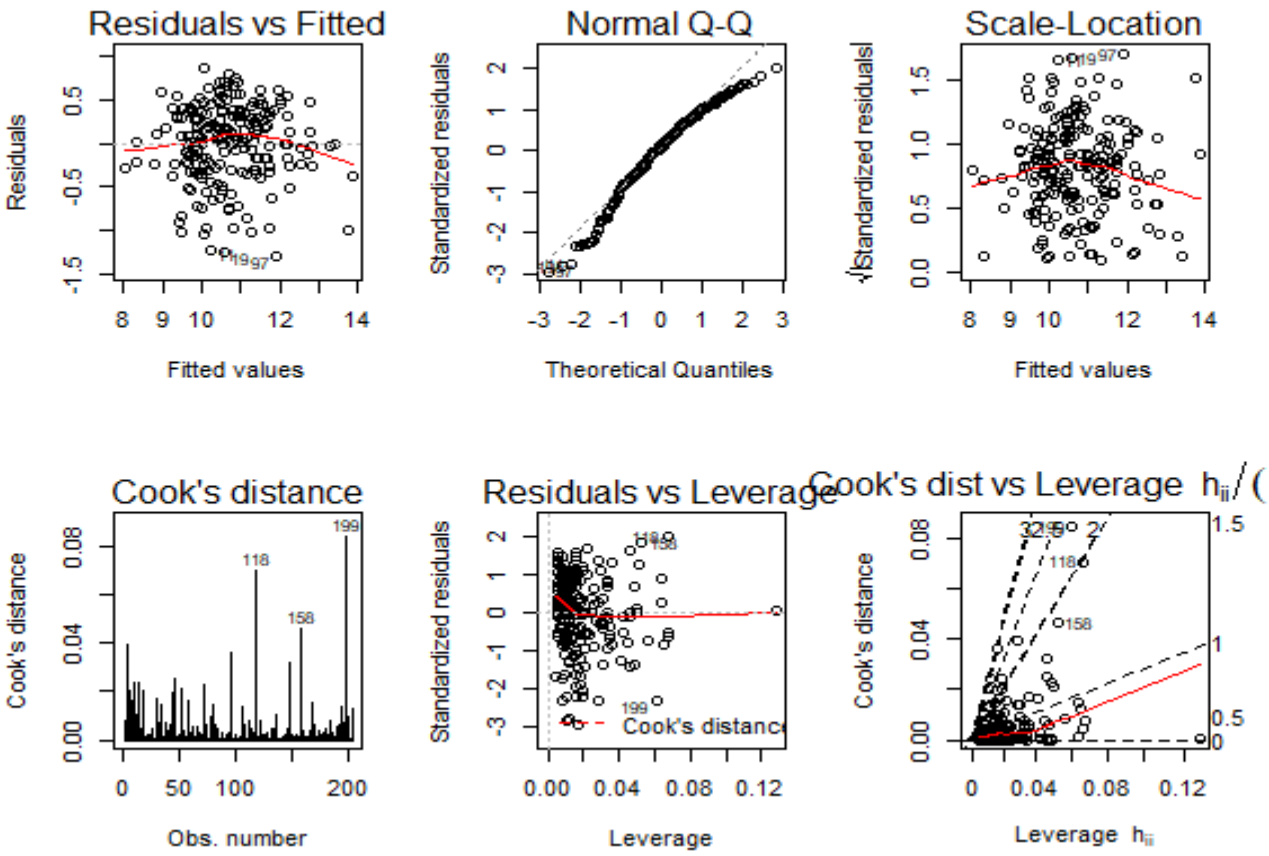
或许高影响点与数据的偏态分布有关，我们暂不删除TV Guide，首先做Box-Cox变换，再检查高影响点。我们发现4个变量都应该做对数变换。

```
par(mfrow=c(2,2))
boxcox(AdRevenue~., data=mag )
boxcox( AdPages~., data=mag  )
boxcox(SubRevenue~., data=mag )
boxcox(NewsRevenue~., data=mag )
```



fit1：拟合变换模型

$$\log(\text{AdRevenue}) = \beta_0 + \beta_1 \times \log(\text{Adpages}) + \beta_2 \times \log(\text{SubRevenues}) + \beta_3 \times \log(\text{NewsRevenue}) + \varepsilon,$$



Cook距离D没有异常，

第50个杂志(Builder)的杠杆值0.12，平均来看杠杆值约为p/n=4/204=0.02，所以Builder的自变量高度异常。其零售收入很少。

但Builder杂志的D值不大（D=1e-5），该点不是高影响的。

```
> summary(fit1)

Call:
lm(formula = AdRevenue ~ ., data = log(mag))

Coefficients:
                Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)     -2.02894    0.41407  -4.900   1.98e-06 ***
AdPages          1.02918    0.05564  18.497    < 2e-16 ***
SubRevenue       0.55849    0.03159  17.677    < 2e-16 ***
NewsRevenue  0.04109    0.02414   1.702     0.0903 .

Residual standard error: 0.4483 on 200 degrees of freedom
Multiple R-squared:  0.8326,    Adjusted R-squared:  0.8301
F-statistic: 331.6 on 3 and 200 DF,  p-value: < 2.2e-16
```

NewsRevenue不显著，SubRevenue的系数0.55,

与0.50差异不显著:

$(0.55-0.50)/0.03159=1.583,\ p=0.114,$

拟合得到的回归方程为：

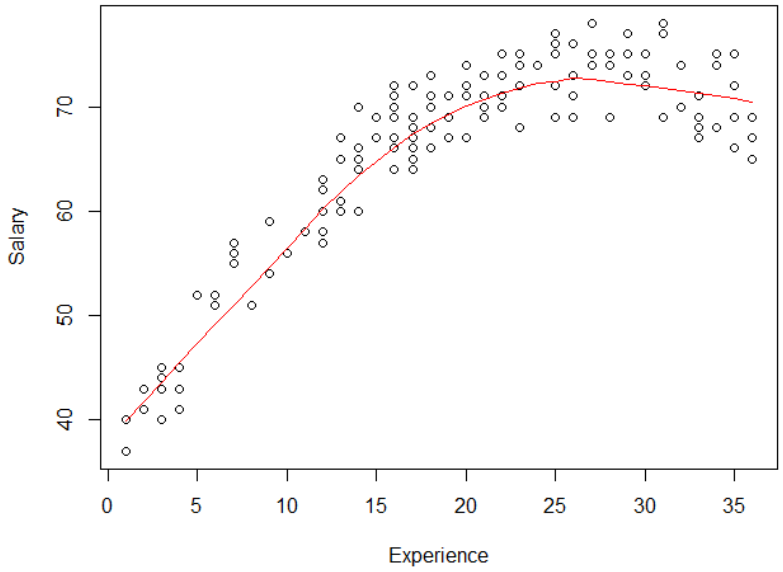
$$\log(\text{AdRevenue})=-2+1\times\log(\text{Adpages})+0.5\times\log(\text{SubRevenues})$$

$$\text{AdRevenue}=0.135\times\text{Adpages}\times\sqrt{\text{SubRevenues}}$$

实例3：工资与工龄（多项式拟合/非单调变换）

数据集se (http://staff.ustc.edu.cn/~ynyang/lm2020/lab/se.xls) 是调查了134个职员(包括会计、工程师、系统管理员等）工资与工作经验数据。

变量	解释
Salary	年工资（1000$）
Experience	工龄（年）



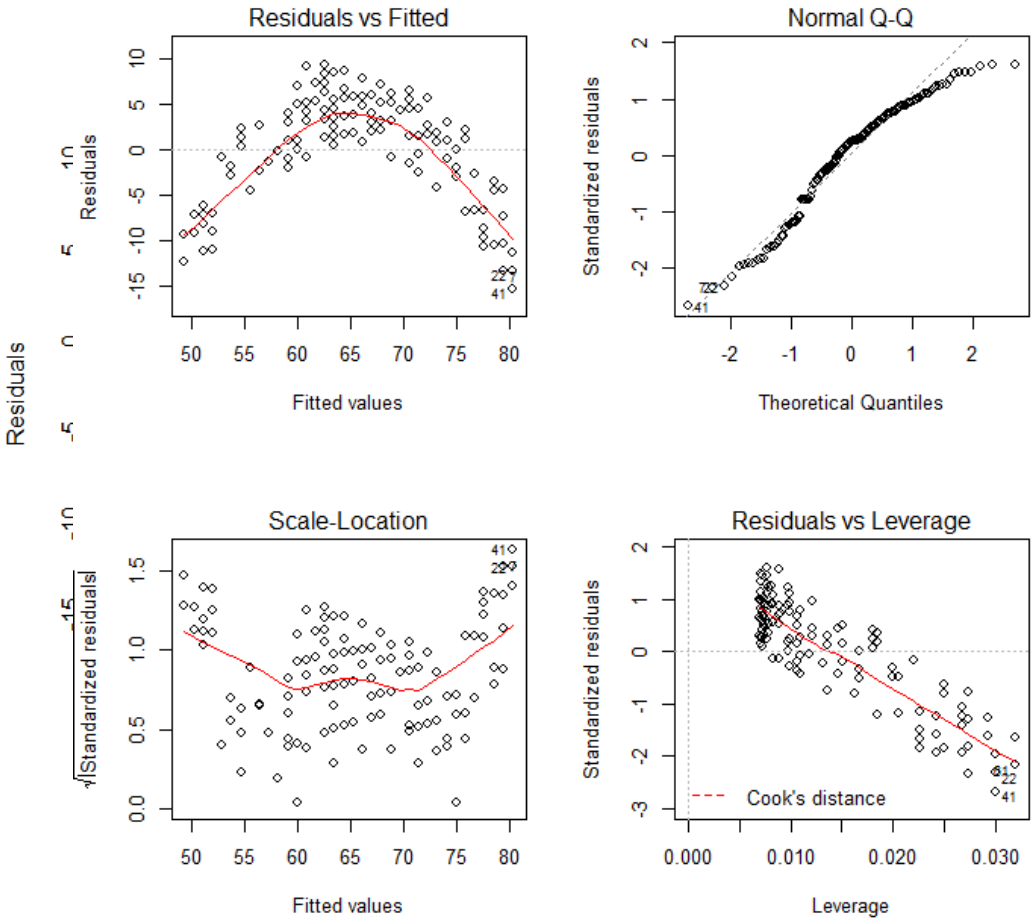

```
> plot( se )
> lines(lowess(se，  f=2/3), col = 2) # lowess方法
```

Salary	Experience
71	26
69	19
73	22
69	17
65	13
75	25
66	35
66	16
67	16
69	16
76	26
72	16
69	25
45	4
72	17
62	12
74	23

拟合简单线性模型

```
> a=lm(Salary~Experience, data=se)
> plot(a )
```

$$\text{Salary} = \beta_0 + \beta_1 \times \text{Experience} + \varepsilon,$$

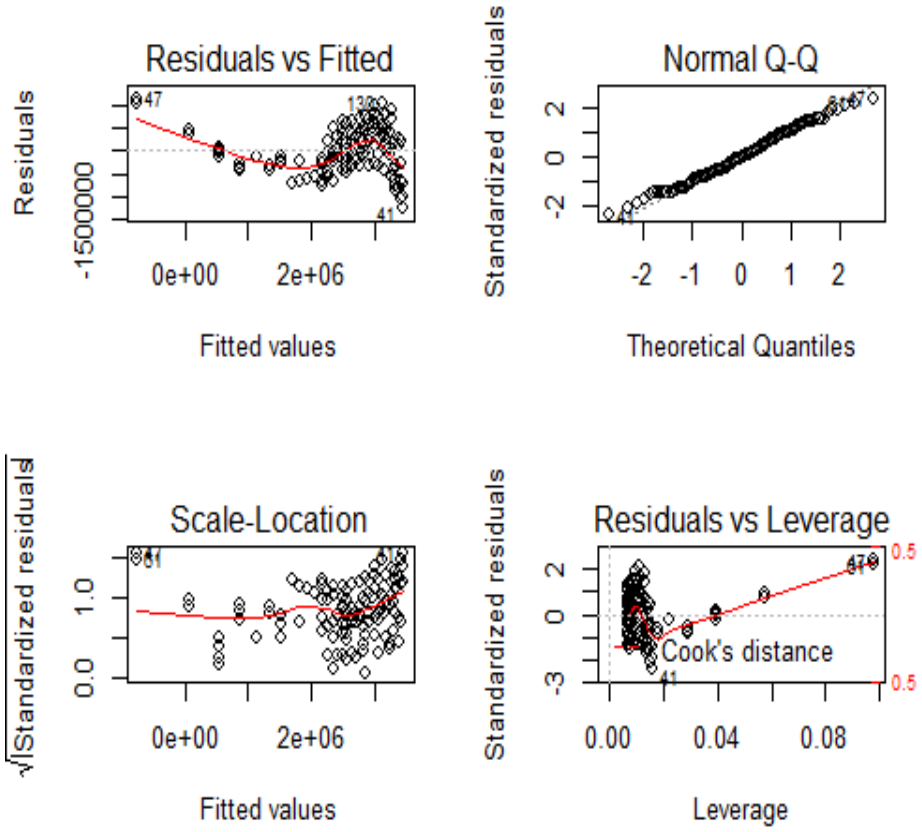


残差有非线性趋势

应用Box-Cox变换（Salary和Experience分别作立方和对数变换）后依旧有非线性现象。

$$\text{Salary}^3 = \beta_0 + \beta_1 \times \log(\text{Experience}) + \varepsilon,$$

Box-Cox 变换是单调变换。非单调的非线性无法用Box-Cox变换消除。



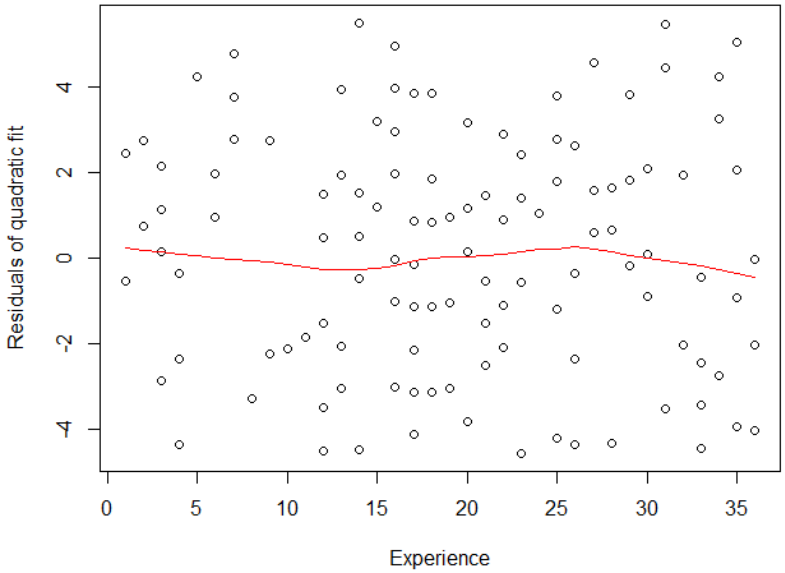
多项式拟合

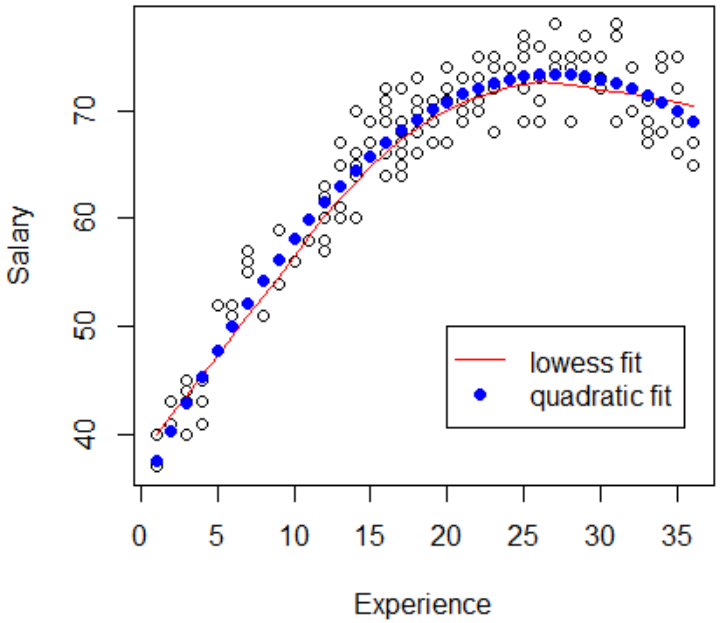
```
> lm(Salary~Experience+ I(Experience^2), data=se)
```

$$\text{Salary} = a + b \times \text{Experience} + c \times \text{Experience}^2 + \varepsilon$$

虽然这不是严格意义上的线性模型，但如果令$\text{Esq} = \text{Experience}^2$,

则该模型是两个自变量的线性模型：$\text{Salary} = a + b \times \text{Experience} + c \times \text{Esq} + \varepsilon$





实例4：调查汇总数据（异方差）

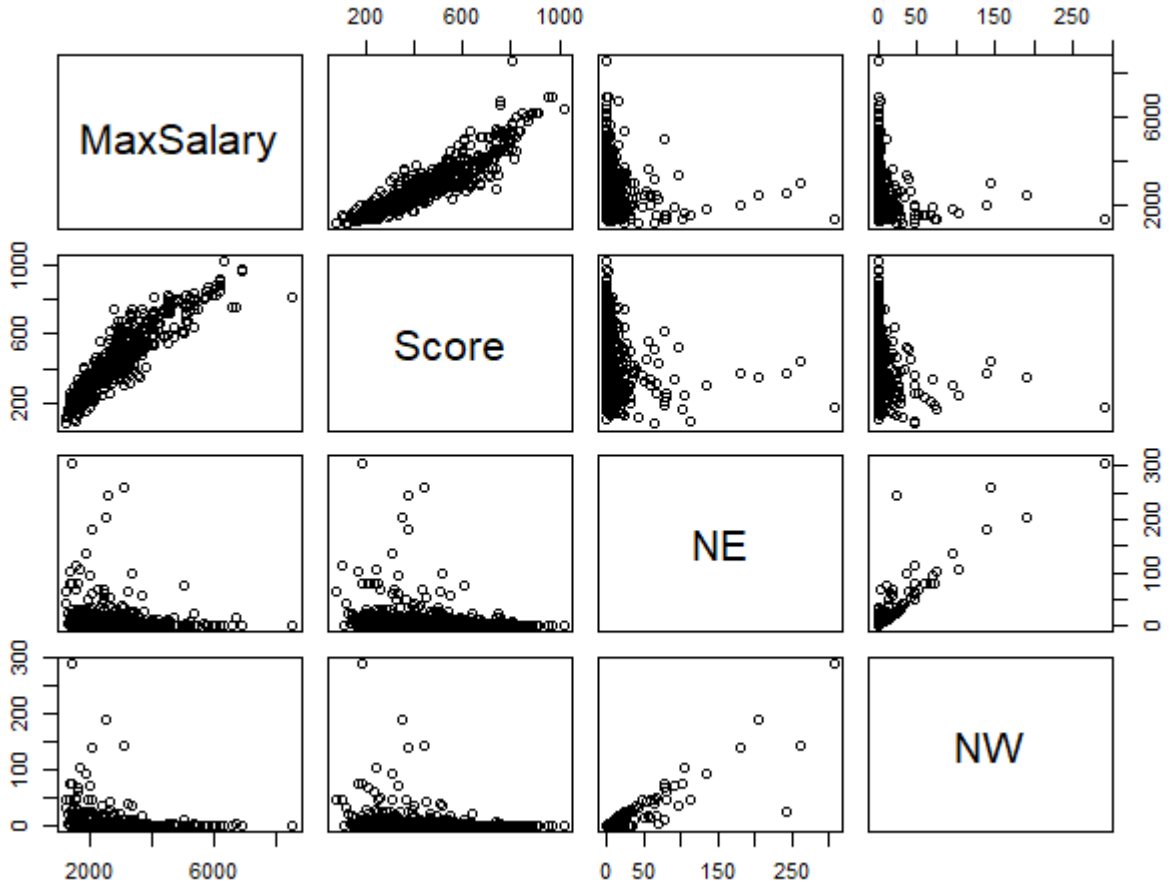
数据集 salarygov (alr3) 汇总了美国政府某部门495种职位的信息，把包括每种职位的最高工资、每种职位的人数、女性人数和职位难度系数(Score)。目的是研究工资与职位难度的关系，特别地我们关心女性占主导的职位的工资情况是否偏低。变量具体描述如下：

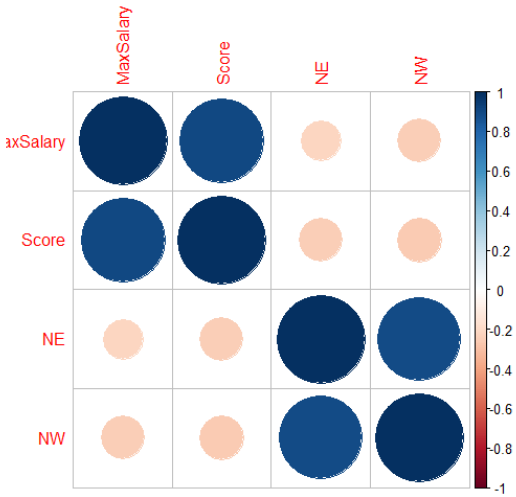
变量	描述
MaxSalary	职位最高工资
Score	职位难度系数（82-1017）
NE	该职位的雇员总数(number of employees)
NW	该职位的女性人数(number of women)

JobClass	NW	NE	Score	MaxSalary
Account_clerk	52	68	258	1549
Account_clerk_Intermediate	26	29	269	1712
Account_clerk_Principal	10	13	321	2182
Account_clerk_Senior	16	24	273	1982
Accountant	1	12	352	2555
Accountant_Chief	0	5	709	4060
Accountant_Principal	0	4	505	3424
Accountant_Senior	2	18	404	3031

探索数据分析(plot,corrplot,boxplot, hist etc)

pairs(gov)
corrplot(cor(gov))

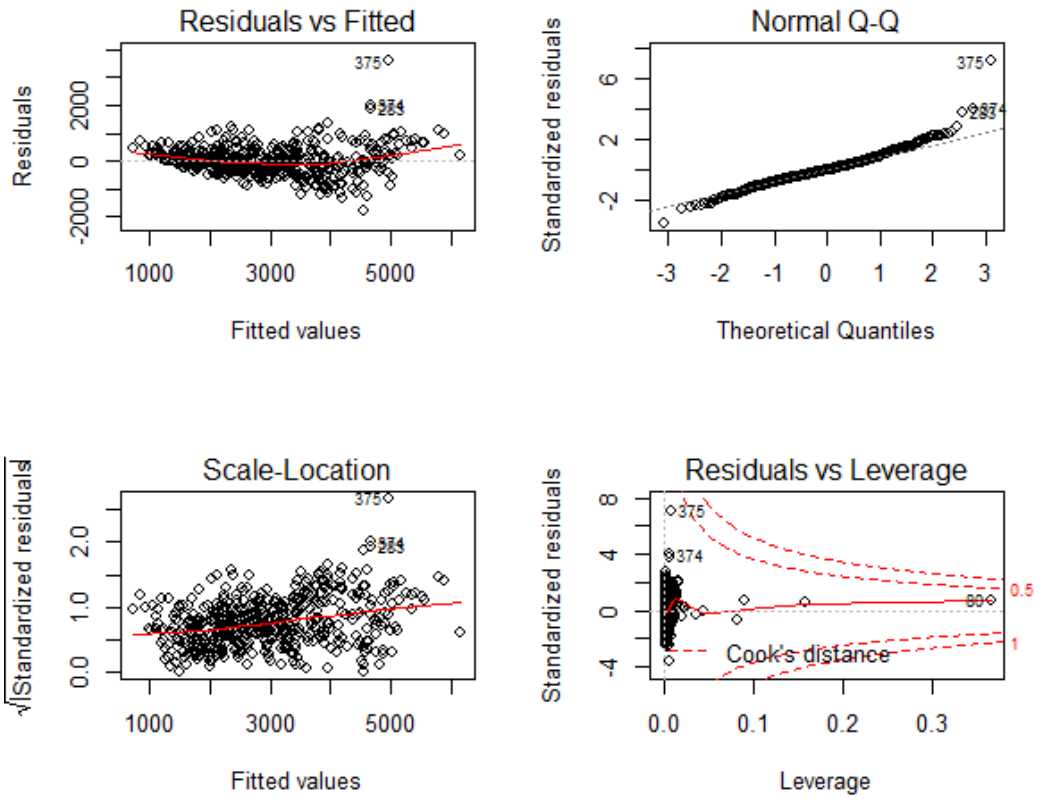




fit0：方差齐性模型

极大统计量的方差与样本量的关系不确定，通常不是O(1/n)的形式。

$$\text{MaxSalary}_i = \beta_0 + \beta_1 \times \text{Score}_i + \beta_2 \times \text{NW}_i + \beta_2 \times \text{NE}_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$$



残差有非线性趋势和方差随拟合值增大的趋势。能否应用WLS方法？

这是抽样调查汇总数据，响应变量是一种职位中若干人的最大值，但最大值的方差形式未知（不与1/NE成正比），所以无法应用WLS（见下页说明）。

异方差模型

$$\text{MaxSalary}_i = \beta_0 + \beta_1 \times \text{Score}_i + \beta_2 \times \text{NW}_i + \beta_2 \times \text{NE}_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2 / w_i)$$

w未知：响应MaxSalary是一类岗位中NE个职员的最大值，即极大统计量，其渐近方差 σ^2/w中的w 未知（w不等于NE）。

靠中间的分位数的渐近方差$\propto 1/n$,但极大或极小统计量的渐近方差不是$O(1/n)$.

n个随机变量$x_1,...,x_n$ iid ~ $f(x)$, 次序统计量$x_{(1)} \le x_{(2)} \le ... \le x_{(n)}, 0 < q < 1,$

$x_{(nq)}$为$x_1,...,x_n$的q分位数 ,ξ_q为f的分位点，则当$n \to \infty$,近似地有

$$x_{(nq)} \sim N\left(\xi_q, \frac{q(1-q)}{n\left[f(\xi_q)\right]^2}\right), \quad \text{var}(x_{(nq)}) \propto 1/n$$

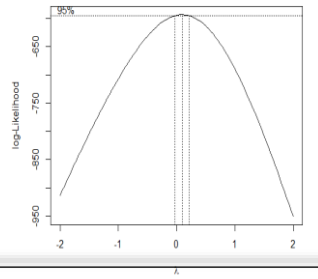
上述结论对$q = 0,1$不成立，即最大值$x_{(n)} = \max(x_1,...,x_n)$的分布未必是正态，

且其渐近方差与总体f有关，且不是O$(1/n)$的阶，比如

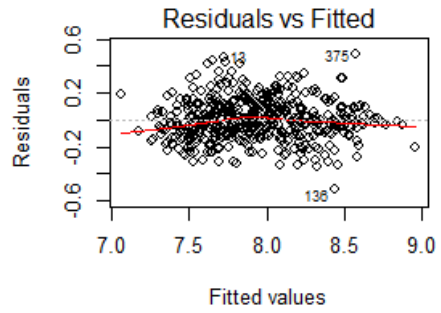
- f 正态：$\text{var}(x_{(n)}) \propto 1/\log(n)$
- f 均匀：$\text{var}(x_{(n)}) \propto 1/n^2$

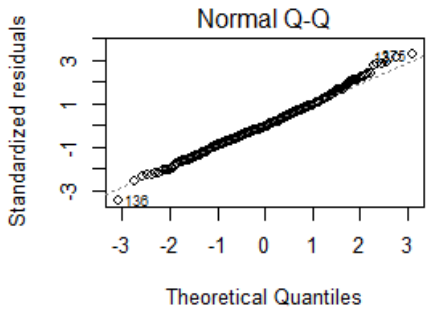
fit1： Box-Cox变换

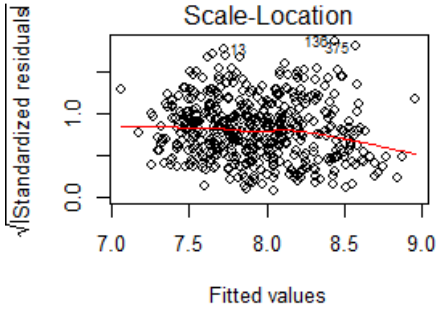
```
Boxcox(MaxSalary~.,data=gov)
fit1 = lm(MaxSalarye~. , data=gov)
```

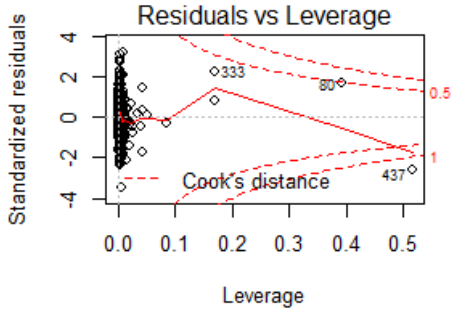


$$\log(\text{MaxSalary})_i = \beta_0 + \beta_1 \times \text{Score}_i + \beta_2 \times \text{NW}_i + \beta_2 \times \text{NE}_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$$









响应变量做对数变化之后，fit0中的异方差和非线性都消失了。

R^2=0.8472，NW的系数小于0

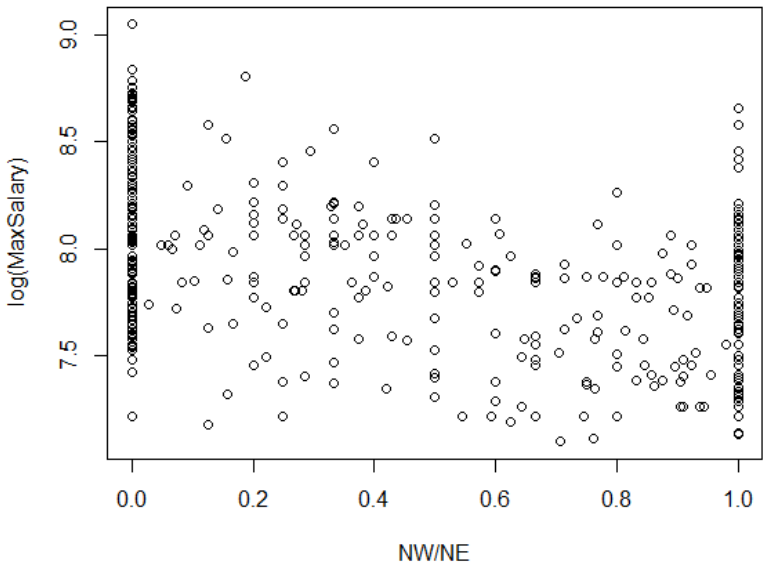
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.05509	1.97e-02	358.64	0.00e+00
Score	0.00187	3.76e-05	49.70	1.05e-193
NW	-0.00299	6.99e-04	-4.28	2.27e-05
NE	0.00174	5.12e-04	3.41	7.10e-04

fit2：简化模型

fit0 = lm(log(MaxSalary)~Score+I(NW/NE))

为了考察女性占优的职位情况，定义女性比率：$r_i = \mathrm{NW}_i / \mathrm{NE}_i$

$$\log(\mathrm{MaxSalary})_i = \beta_0 + \beta_1 \times \mathrm{Score}_i + \beta_2 \times r_i + \varepsilon_i, \varepsilon_i \sim (0, \sigma^2)$$



Call:
lm(formula = log(MaxSalary) ~ Score + I(NW/NE), data = gov)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.162e+00	2.177e-02	328.955	<2e-16 ***
Score	1.767e-03	3.709e-05	47.644	<2e-16 ***
I(NW/NE)	-1.452e-01	1.651e-02	-8.796	<2e-16 ***

Residual standard error: 0.1419 on 492 degrees of freedom
Multiple R-squared: 0.8628, Adjusted R-squared: 0.8623
F-statistic: 1548 on 2 and 492 DF, p-value: < 2.2e-16

实例5：教育花费（异方差，IRLS）

问题背景：数据集edu 是1975年美国50个州的青少年教育费用数据, 变量如下表所示，关心的问题是人均教育花费与其它变量的关系. (http://staff.ustc.edu.cn/~ynyang/lm2020/lab/edu.xls)

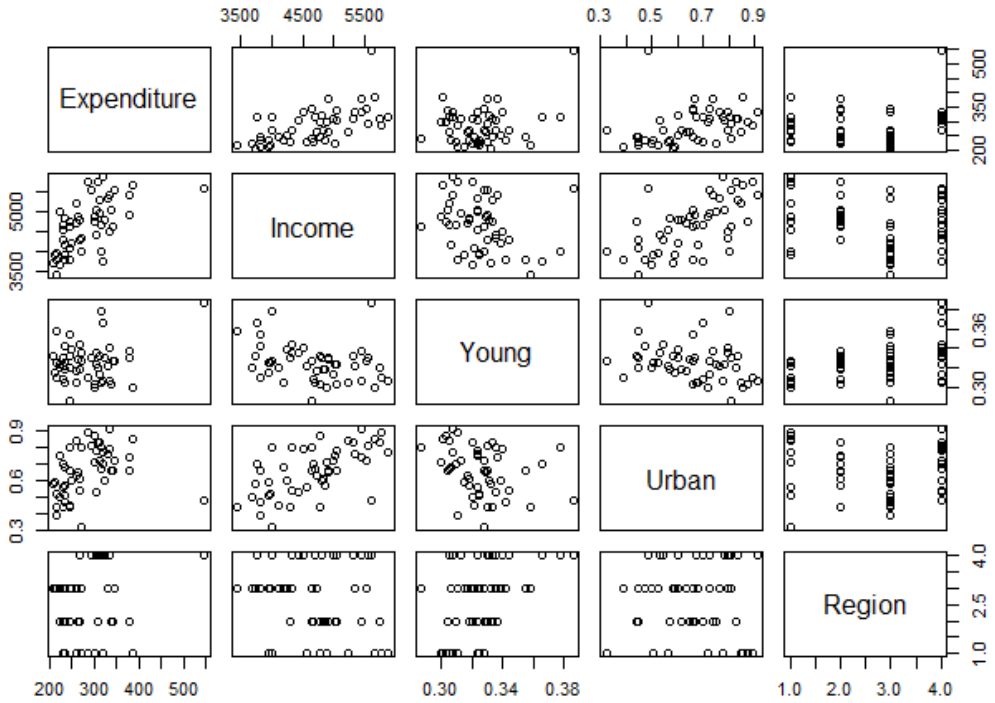
变量	解释
Expenditure	各州年度人均教育花费
Income	各州人均收入
Young	18岁以下人口比例
Urban	城市人口比例
Region	地区(1:东北, 2:中部和北部, 3:南部, 4:西部)

state	Expenditure	Income	Young	Urban	Region
ME	235	3944	0.325	0.508	1
NH	231	4578	0.323	0.564	1
VT	270	4011	0.328	0.322	1
MA	261	5233	0.305	0.846	1
RI	300	4780	0.303	0.871	1
CT	317	5889	0.307	0.774	1
NY	387	5663	0.301	0.856	1
NJ	285	5759	0.31	0.889	1
PA	300	4894	0.3	0.715	1
OH	221	5012	0.324	0.753	2
IN	264	4908	0.329	0.649	2
IL	308	5753	0.32	0.83	2
MI	379	5439	0.337	0.738	2
WI	342	4634	0.328	0.659	2
MN	378	4921	0.33	0.664	2
IA	232	4869	0.318	0.572	2
MO	231	4672	0.309	0.701	2
ND	246	4782	0.333	0.443	2
SD	230	4296	0.33	0.446	2
NE	268	4827	0.318	0.615	2
KS	337	5057	0.304	0.661	2
DE	344	5540	0.328	0.722	3
MD	330	5331	0.323	0.766	3
VA	261	4715	0.317	0.631	3
WV	214	3828	0.31	0.39	3
NC	245	4120	0.321	0.45	3
SC	233	3817	0.342	0.476	3
GA	250	4243	0.339	0.603	3
FL	243	4647	0.287	0.805	3
KY	216	3967	0.325	0.523	3
TN	212	3946	0.315	0.588	3
AL	208	3724	0.332	0.584	3
MS	215	3448	0.358	0.445	3
AR	221	3680	0.32	0.5	3
LA	244	3825	0.355	0.661	3
OK	234	4189	0.306	0.68	3
TX	269	4336	0.335	0.797	3
MT	302	4418	0.335	0.534	4
ID	268	4323	0.344	0.541	4
WY	323	4813	0.331	0.605	4
CO	304	5046	0.324	0.785	4
NM	317	3764	0.366	0.698	4
AZ	332	4504	0.34	0.796	4
UT	315	4005	0.378	0.804	4
NV	291	5560	0.33	0.809	4
WA	312	4989	0.313	0.726	4
OR	316	4697	0.305	0.671	4
CA	332	5438	0.307	0.909	4
AK	546	5613	0.386	0.484	4
HI	311	5309	0.333	0.831	4

探索数据分析(plot,corrplot,boxplot, hist etc)

```
plot(edu)
corrplot(cor(edu))
```

1:东北, 2:中部和北部, 3:南部, 4:西部

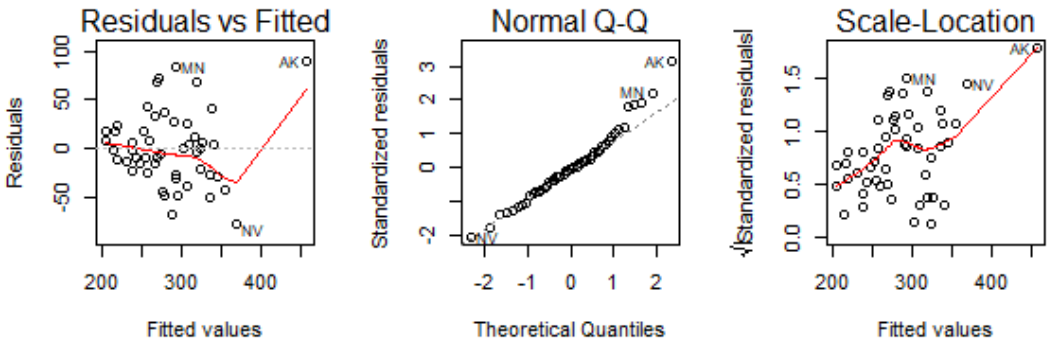


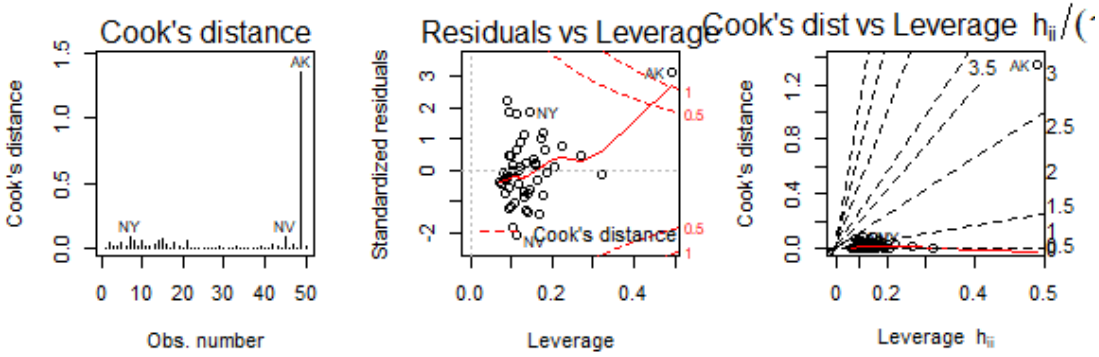
Expenditure ： 与Income，Urban呈线性关系；在各个地区大致持平；
Income: 与Urban呈线性关系； 与Young相关性不大；各地区差异不明显
Young:与Urban有某种负相关； 各地区差异较大，西部最多，东部最少。

fit0：拟合初始模型

```
fit0 = lm(Expenditure~. , data=edu[-49,])
Plot(fit0,1:6 )
```

$$\text{Expenditure}_i = \beta_0 + \beta_1 \times \text{Income}_i + \beta_2 \times \text{Urban}_i + \sum_{k=2}^{4} \alpha_k I_{(\text{Region}_i = k)} + \varepsilon_i, i = 1,\ldots,50$$





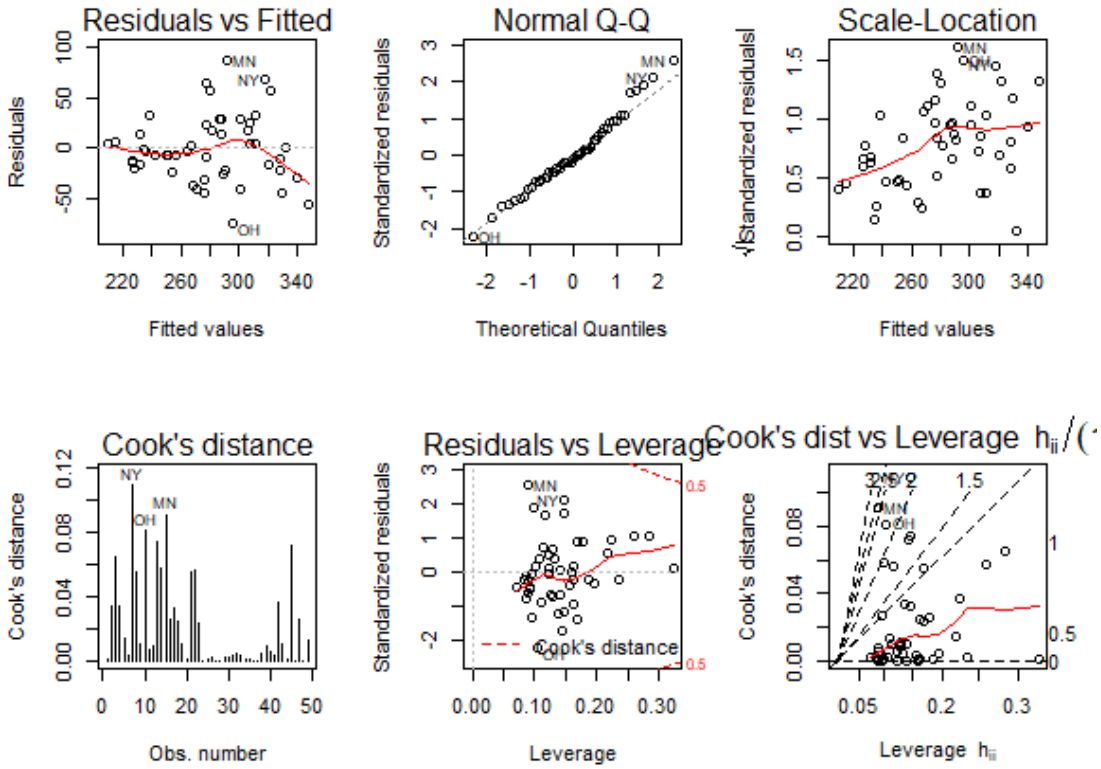
残差图表明方差不齐，
AK是高影响点：D=1.3，h=0.5

fit1：删除高影响点后重新拟合

```
fit1 = lm(Expenditure~. , data=edu[-49,])
plot(fit1,1:6 )
```

AK(Alaska) 高影响，地理位置特殊，可以删除。重新拟合：

$$\text{Expenditure}_i = \beta_0 + \beta_1 \times \text{Income}_i + \beta_2 \times \text{Urban}_i + \sum_{k=2}^{4} \alpha_k I_{(\text{Region}_i = k)} + \varepsilon_i,\ i \neq 49(AK)$$



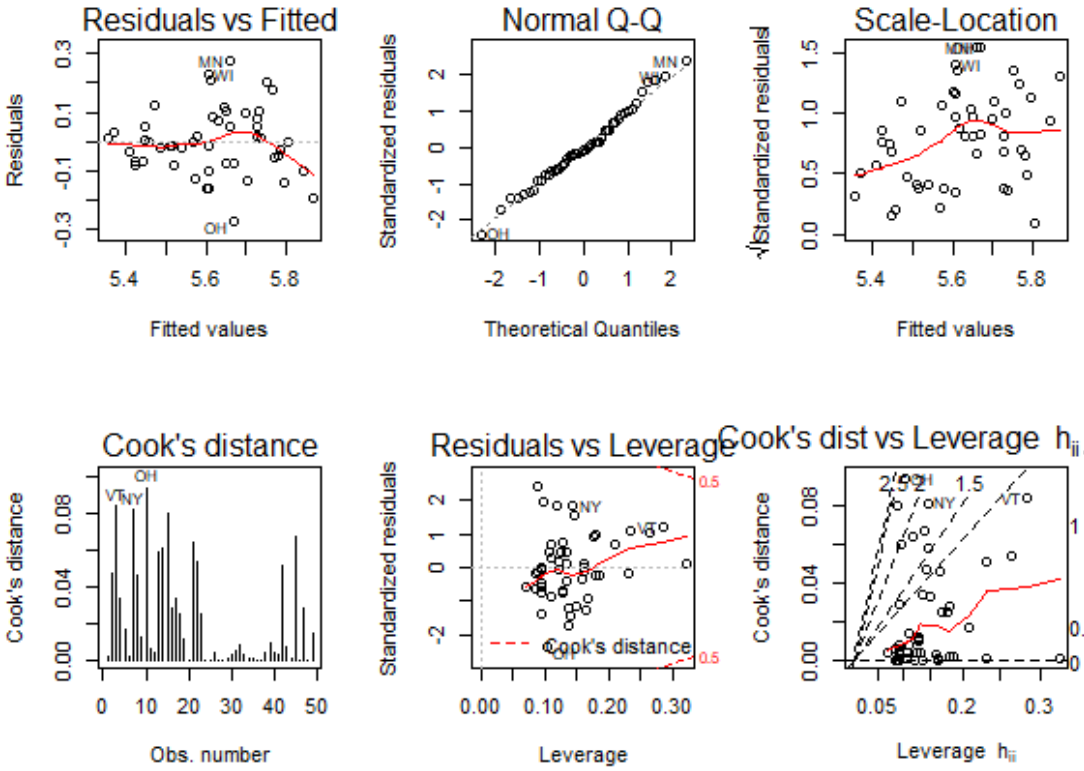
不再有高影响点，
残差图表明方差不齐。

fit2: 数据变换后重新拟合

对Expenditure、income做对数变换（boxcox）

```
library(MASS)
boxcox(Expenditure~. , data=edu[-49,])
boxcox(Income~., data=edu[-49,])
fit2 = lm(log(Expenditure)~. , data=edu[-49,])
```

$$\log(\text{Expenditure}_i) = \beta_0 + \beta_1 \times \log(\text{Income}_i) + \beta_2 \times \text{Urban}_i + \sum_{k=2}^{4} \alpha_k I_{(\text{Region}_i = k)} + \varepsilon_i \,,\ i \neq 49(AK)$$

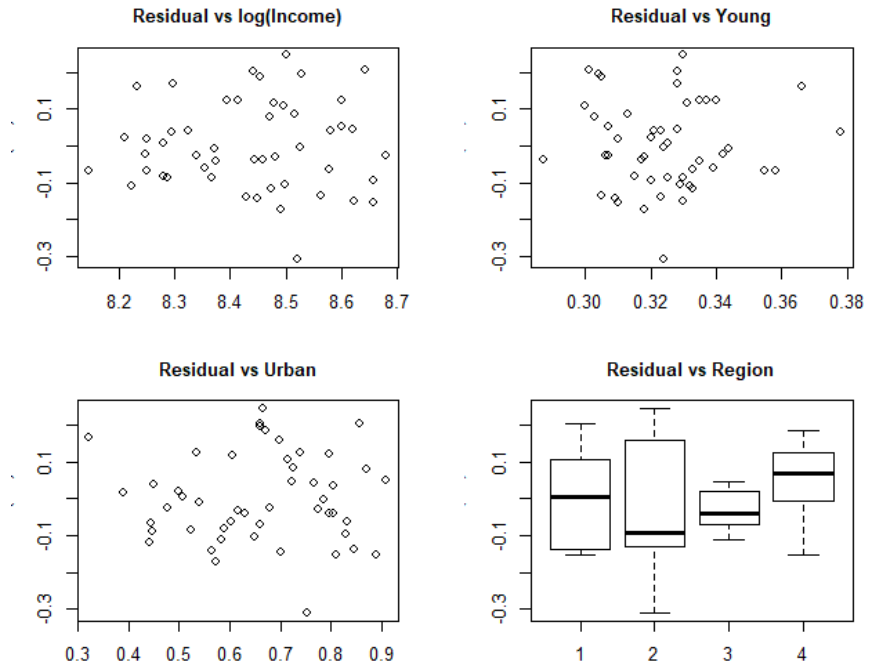


与fit2结果类似，方差仍不齐。具体是哪个自变量导致的方差不齐？为此我们可以画出残差 vs 各个自变量的残差图。

标准的残差图（(残差 vs 拟合值) 中，拟合值作为所有自变量的代表（自变量的最优组合）并不能完全代表各个自变量。下面我们考察残差 vs 各个自变量

残差图：残差 vs 自变量

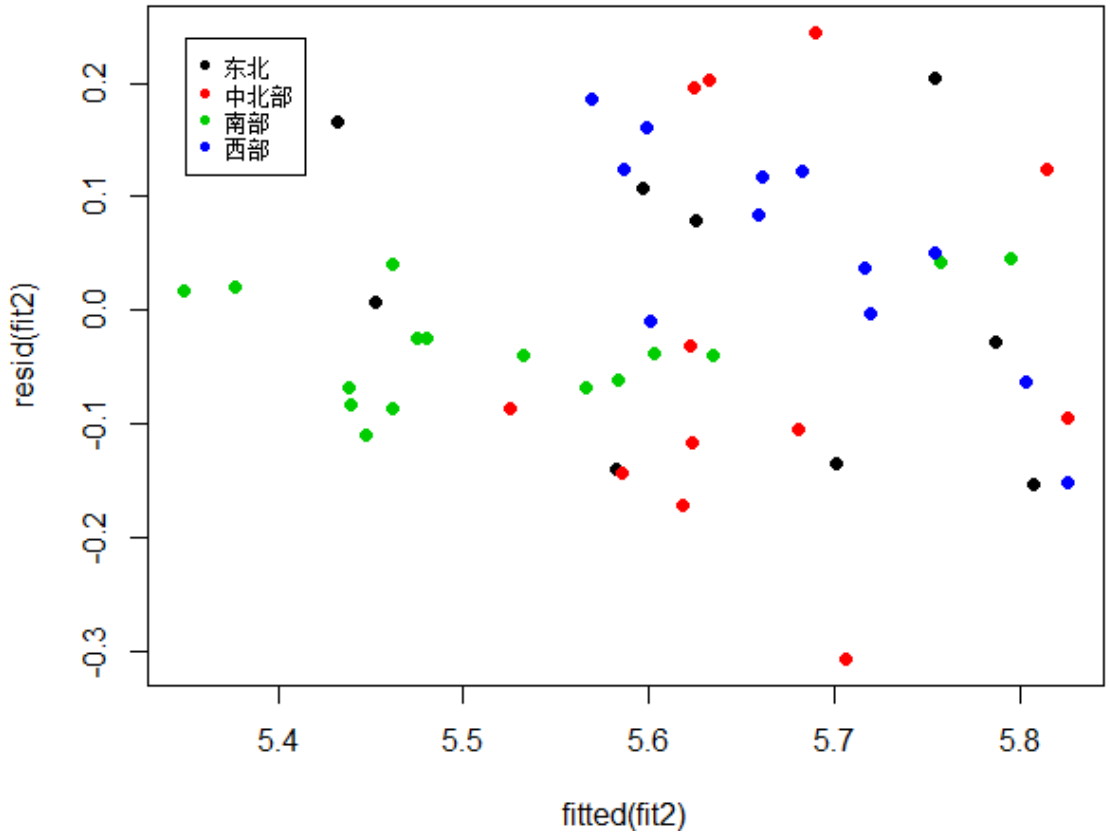
残差对每个自变量的残差图有助于发现非线性或异方差具体与哪个自变量有关。



前3个自变量-残差图无异方差现象，最后一个残差图表明4个地区之间方差变化较大。

事实上，Expenditure是人均花费，应该假设异方差模型，以每个州1975年的人口数作为权重，应用WLS方法。我们没有人口数据， 但Region某种意义上代表了人口差异。

在标准的残差图（残差 vs 拟合值）中标记4个region，可以看到每个区内方差基本是常数，南部 （Region=3)方差很小，而中北部方差最大同时拟合值也较大，在导致了残差图上方差岁拟合值增大而增大的现象 。

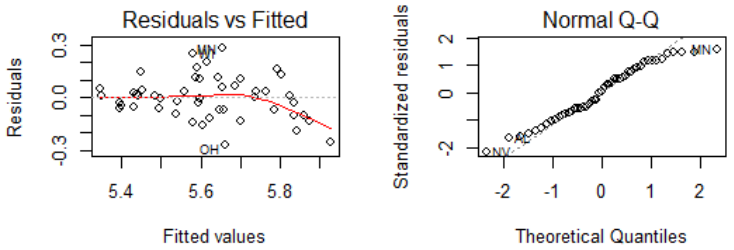



```
par(mfrow=c(1,2))
plot(fitted(fit2), resid(fit2), col=as.numeric(edu[-49,5]),pch=16)
legend(5.37, -0.13,pch=16, legend=c("东北","中北部", "南部","西部"),col=1:4,cex=0.75,)
```

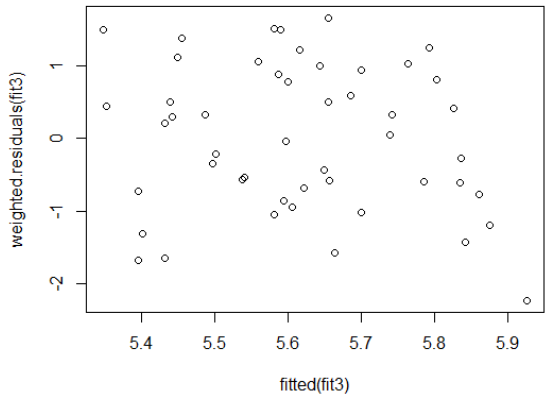
fit3: 异方差模型，IRLS方法求解

$$\log(\text{Expenditure}_i) = \beta_0 + \beta_1 \times \log(\text{Income}_i) + \beta_2 \times \text{Urban}_i + \sum_{k=2}^{4} \alpha_k I_{(\text{Region}_i = k)} + \varepsilon_i \,,\ i \neq 49(AK),$$

假设4个区的误差方差不同,分别为$\sigma_1^2,...,\sigma_4^2$



WLS并不能消除异方差，而是将异方差现象作为权重，残差图几乎与fit2相同。加权残差图方差为常数：



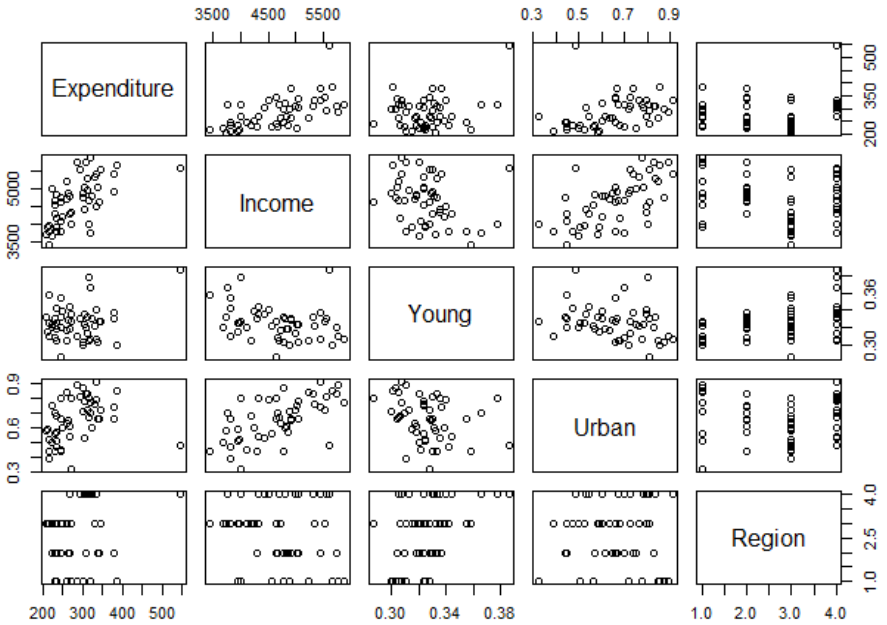

```
fit.ini = fit= lm(log(Expenditure)~., data=edu[-49,] )
repeat{
                beta=coef(fit)
                res=resid(fit)
                sigmasq1 = sum(res[1:9]^2)/(9)
                sigmasq2 = sum(res[10:21]^2)/(12)
                sigmasq3 = sum(res[22:37]^2)/(16)
                sigmasq4 = sum(res[38:49]^2)/(12)
                sigma2=c(rep(sigmasq1,9),
                rep(sigmasq2,12),rep(sigmasq3,16),rep(sigmasq4,12))
                w=1/sigma2
                fit = lm(log(Expenditure)~log(Income)+Young+Urban.,
data=edu[-49,], weight=w)
                beta.new=coef(fit)
                beta.new
                delta=sum(abs(beta.new-beta))
                print(delta)
                if (delta<1e-10) break
                beta=beta.new
        }
fit3=fit # final fit
unique(sigma2) ## sigma2 for the 4 regions:  0.018 0.028 0.001 0.013
```

$$\sigma_1^2 = 0.017,\ \sigma_2^2 = 0.03,\ \sigma_3^2 = 0.001,\ \ \sigma_4^2 = 0.017$$

```
Call:
lm(formula = log(Expenditure) ~ log(Income) + Young + Urban +
    Region, data = edu[-49, ], weights = w)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.84887    0.85433  -5.676 1.16e-06 ***
log(Income)  1.12248    0.09759  11.502 1.47e-14 ***
Young        3.05170    0.52806   5.779 8.25e-07 ***
Urban        0.00663    0.09209   0.072    0.943
Region2     -0.04600    0.06872  -0.669    0.507
Region3     -0.01197    0.04729  -0.253    0.801
Region4      0.06839    0.06082   1.124    0.267
---
Residual standard error: 1.028 on 42 degrees of freedom
Multiple R-squared:  0.8766,    Adjusted R-squared:  0.8589
F-statistic: 49.71 on 6 and 42 DF,  p-value: < 2.2e-16
```

散点图表明Expenditure 与Urban 是显著正相关的，但在控制其它变量之后，特别是控制了Income之后，它们不再相关。



大致上， 我们可以猜想如下因果路径图（因为是观察研究，该因果路径的正确性存疑）：