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

第28讲 回归诊断实例（续）

2020.5.29

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)

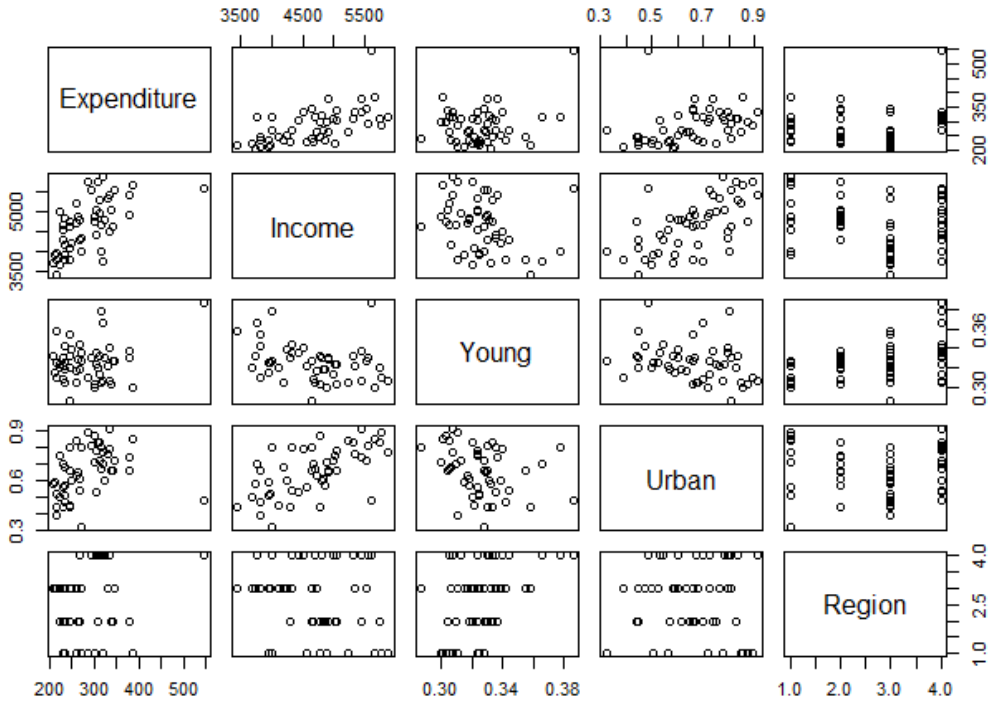
变量	解释
Expenditure	各州年度人均教育花费
Income	各州人均收入
Young	18岁以下人口比例
Urban	城市人口比例
Region	地区(1:东北, 2:中部和北部, 3:南部, 4:西部)

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

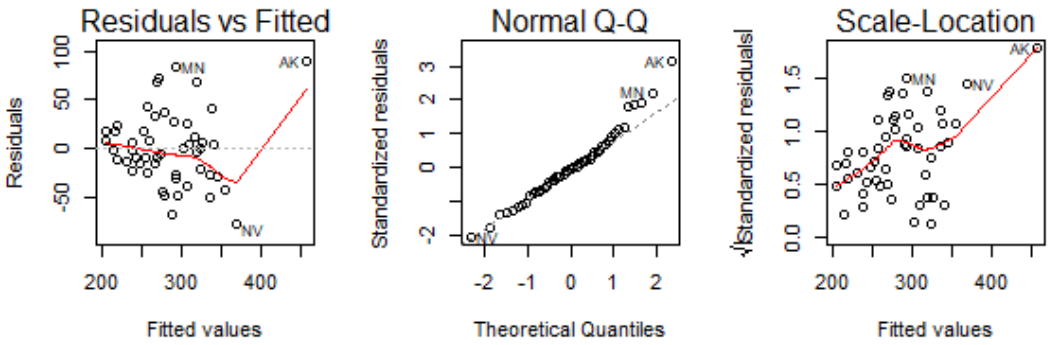


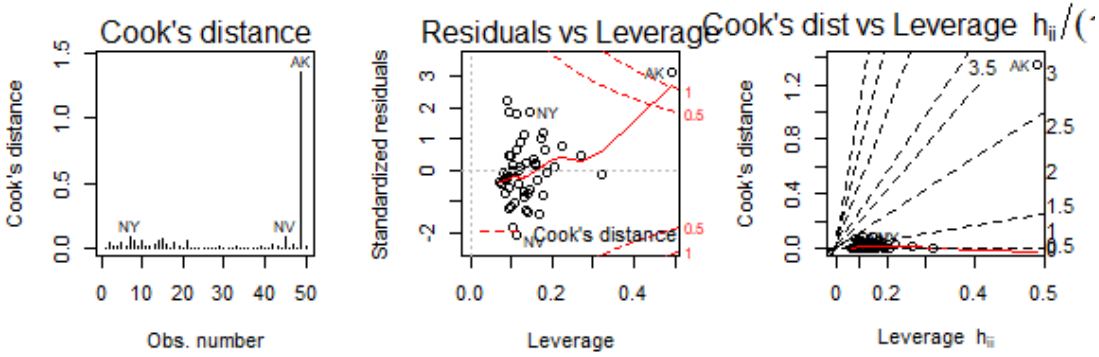
Expenditure ： 与Income，Urban呈线性关系；在各个地区大致持平；
Income: 与Urban呈线性关系； 与Young相关性不大；各地区差异不明显
Young:与Urban有某种负相关； 各地区差异较大，西部最多，东部最少。

fit0：拟合初始模型

```
fit0 = lm(Expenditure~. , data=edu)
plot(fit0,1:6 )
```

$$\text{Expenditure}_i = \beta_0 + \beta_1 \times \text{Income}_i + \beta_2 \times \text{Urban}_i + \beta_3 \times \text{Young}_i + \sum_{k=2}^{4} \alpha_k I_{(\text{Region}_i = k)} + \varepsilon_i, i = 1,...,50$$





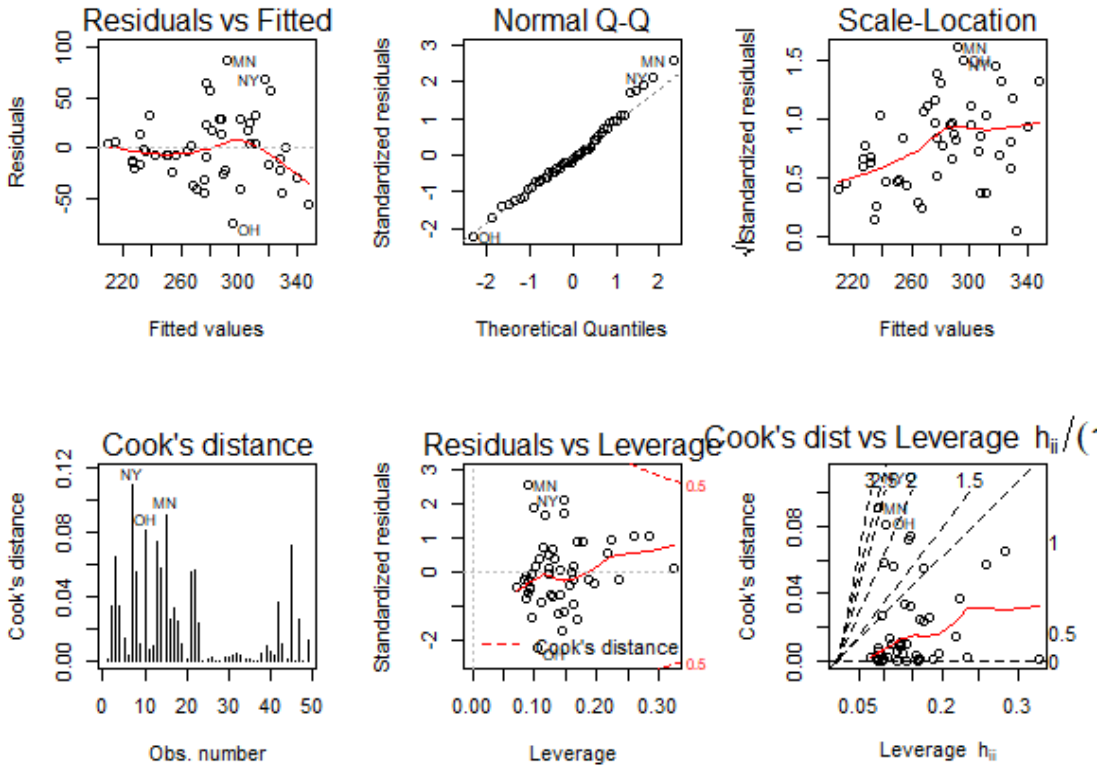
残差图表明方差不齐，
AK是高影响点：D=1.3,， h=0.5

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：

$$\text{Expenditure}_i = \beta_0 + \beta_1 \times \text{Income}_i + \beta_2 \times \text{Urban}_i + \beta_3 \times \text{Young}_i + \sum_{k=2}^{4} \alpha_k I_{(\text{Region}_i = k)} + \varepsilon_i,\ i \neq 49$$



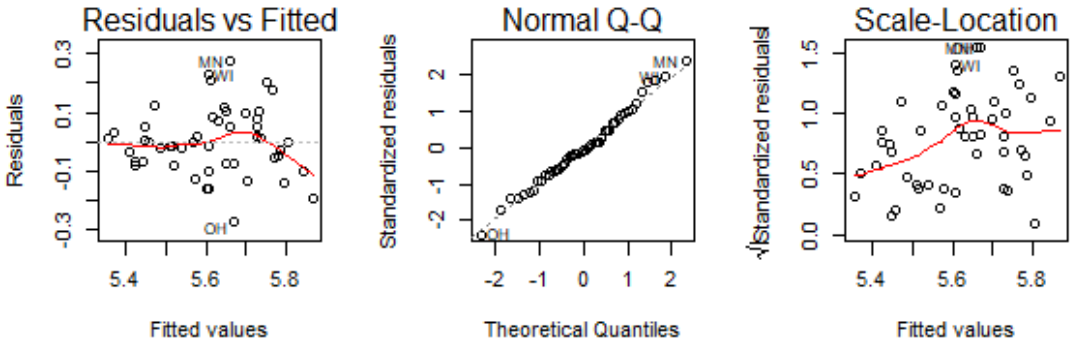
不再有高影响点，
残差图表明方差不齐。

fit2: 数据变换后重新拟合

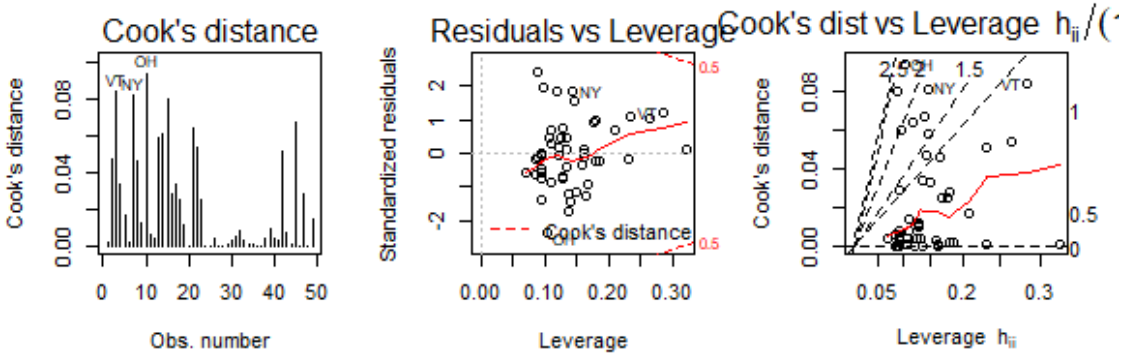
对Expenditure、income做对数变换（boxcox）

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

$$\log(\text{Expenditure}_i) = \beta_0 + \beta_1 \times \log(\text{Income}_i) + \beta_2 \times \text{Urban}_i + \beta_3 \times \text{Young}_i + \sum_{k=2}^{4} \alpha_k I_{(\text{Region}_i = k)} + \varepsilon_i \,,\ i \neq 49(AK)$$



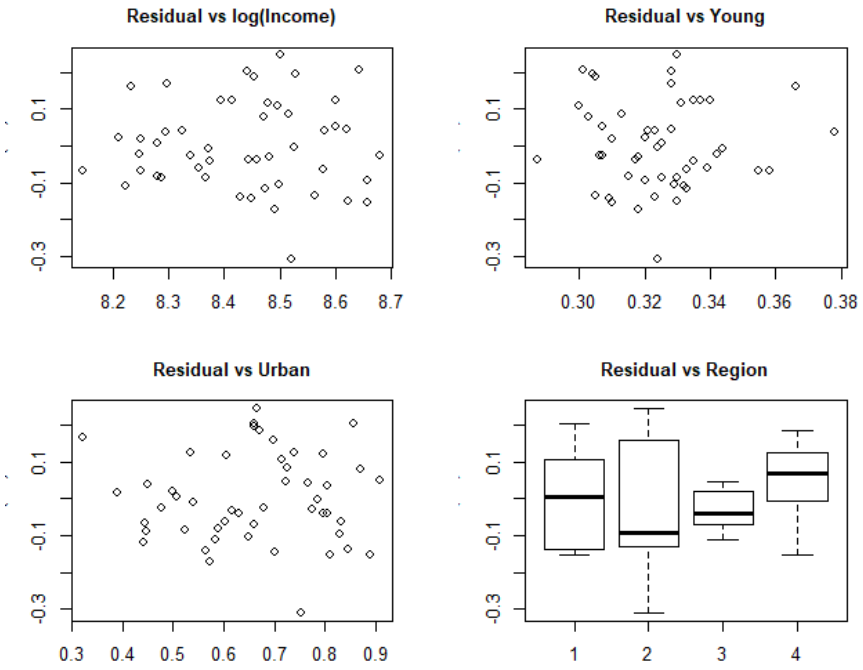
与fit2结果类似，方差仍不齐。具体是哪个自变量导致的方差不齐？为此我们可以画出残差 vs 各个自变量的残差图。



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

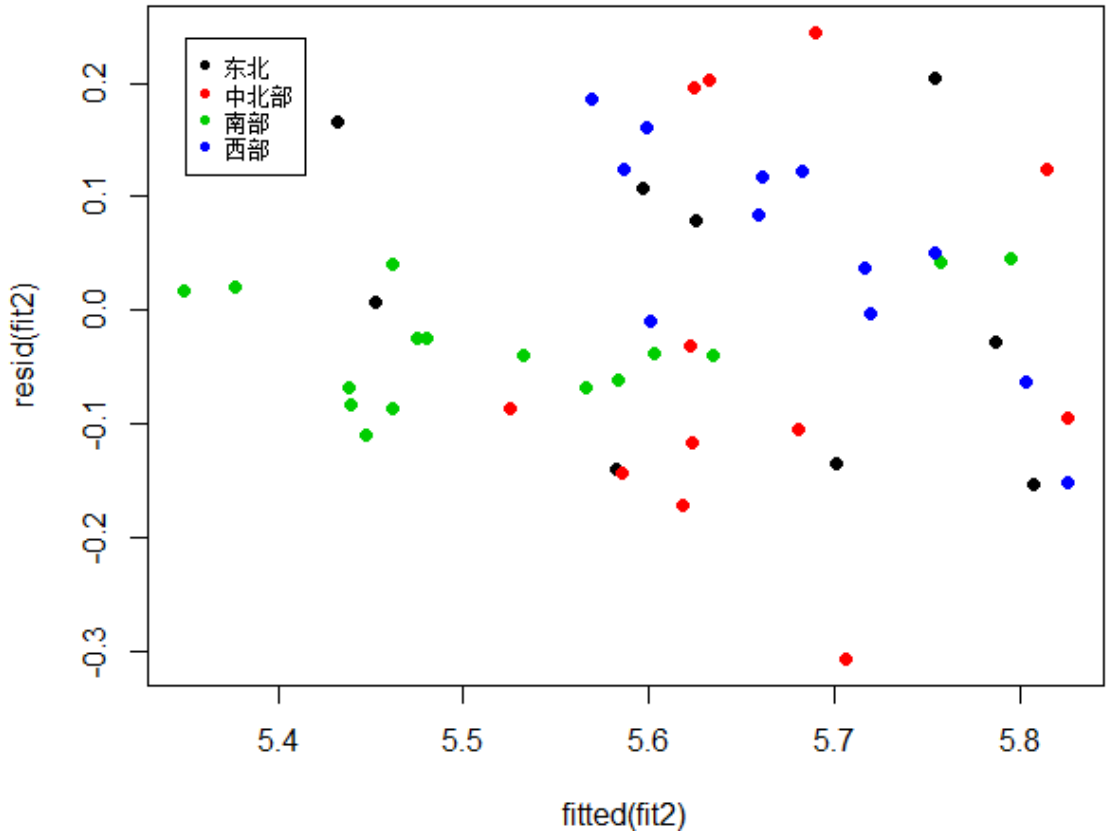
残差对每个自变量的残差图有助于发现非线性或异方差具体与哪个自变量有关。



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随拟合值增大而增大的现象 。

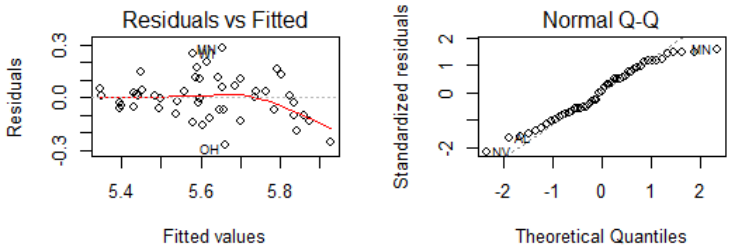



```
par(mfrow=c(1,2))
plot(fitted(fit2), resid(fit2), col=as.numeric(edu[-49,5]),pch=16)
legend(5.37, -0.13,pch=16, legend=c("东北","中北部", "南部","西部"),col=1:4,cex=0.75,)
```

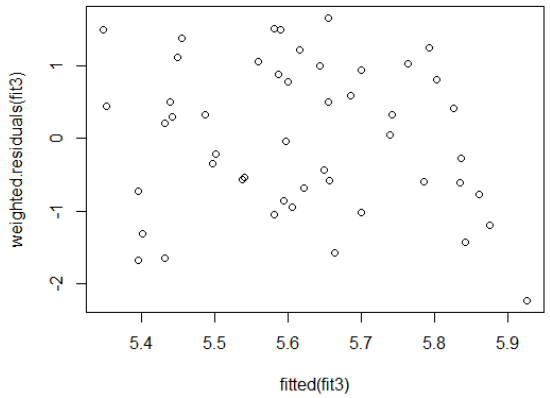
fit3: 异方差模型，IRLS方法求解

$$\log(\text{Expenditure}_i) = \beta_0 + \beta_1 \times \log(\text{Income}_i) + \beta_2 \times \text{Urban}_i + \beta_3 \times \text{Young}_i + \sum_{k=2}^{4} \alpha_k I_{(\text{Region}_i = k)} + \varepsilon_i,$$

$i \neq 49(AK)$. 假设4个区的误差方差不同,分别为$\sigma_1^2,...,\sigma_4^2$ (未知).



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：



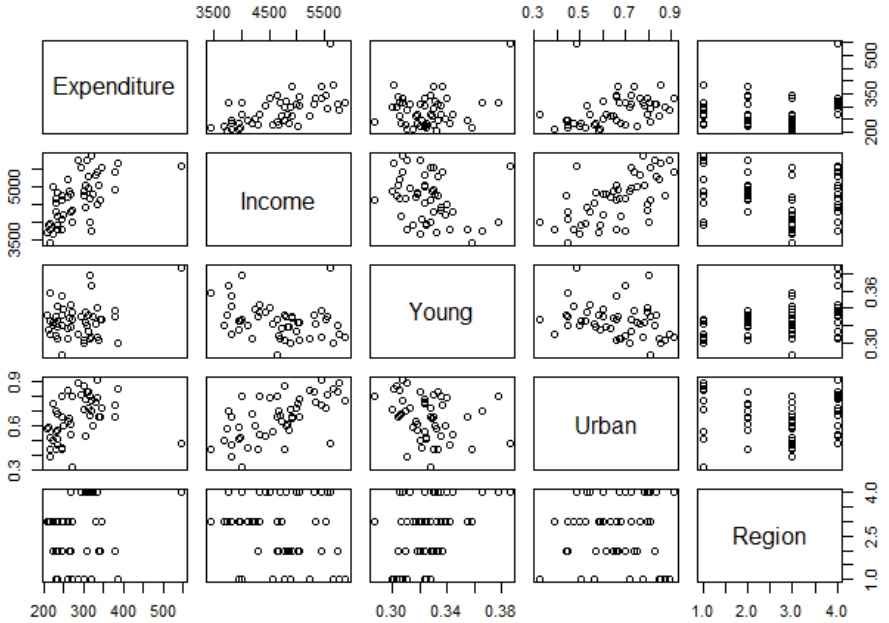

```
fit.ini = fit= lm(log(Expenditure)~., data=edu[-49,] )
repeat{
                beta=coef(fit)
                res=resid(fit)
                sigmasq1 = sum(res[1:9]^2)/(9)
                sigmasq2 = sum(res[10:21]^2)/(12)
                sigmasq3 = sum(res[22:37]^2)/(16)
                sigmasq4 = 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
```

$$\sigma_1^2 = 0.018,\ \sigma_2^2 = 0.028,\ \sigma_3^2 = 0.001,\ \ \sigma_4^2 = 0.013$$

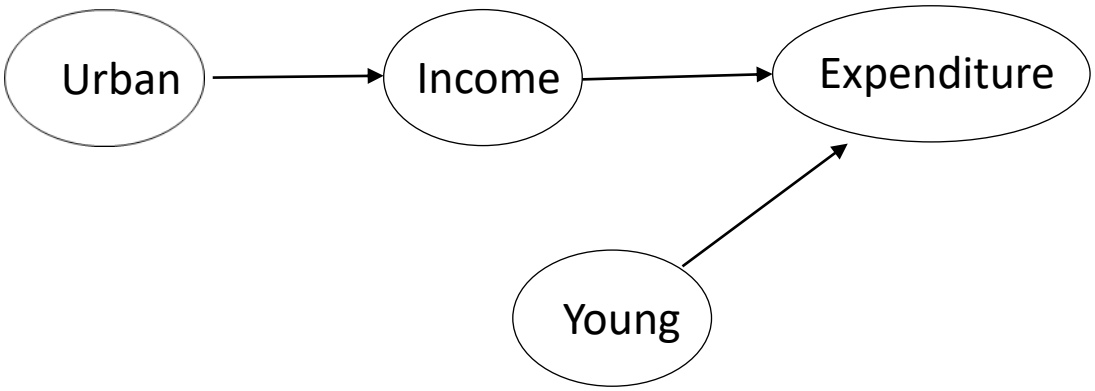
```
Call:
lm(formula = log(Expenditure) ~ log(Income) + Young + Urban +
    Region, data = edu[-49, ], weights = w)

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）：



实例6：休斯顿房价（提高拟合效果）

问题背景：休士顿纪事报（Husont Chronicle) 2007年4月15号刊登了2006年Houston1922个房区（subdevision）的房产价格信息统计表。每个房区的房价以售出的房子价格的中位数代表。变量描述如下：

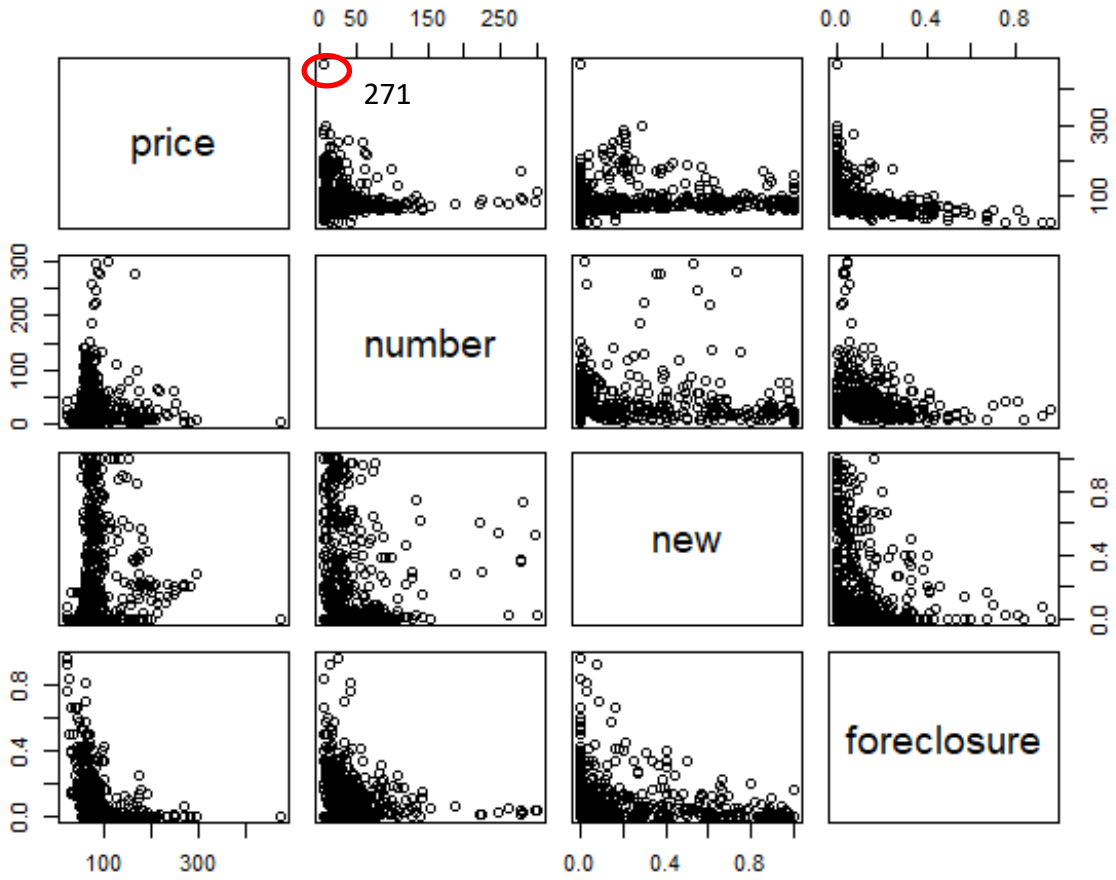
变量	解释
price	中位数价格（每平方英尺）
Number	2006年售出的房子数目
new	售出房子中新房（建于2005-2006）的比例
foreclosure	拍卖售出房子的比例

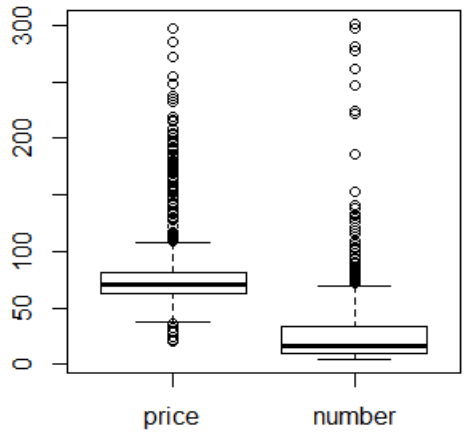
当购房者因偿还房贷违约而被取消赎回房产抵押的权利时，银行作为借贷方成为该房产的所有 人并以低于市场价的方式拍卖房产，变量 foreclosure 即代表的是这种拍卖所占的百分比。我们感兴趣的是变量new, foreclosure 与房价的关系，以及如何利用这种关系预测房价。

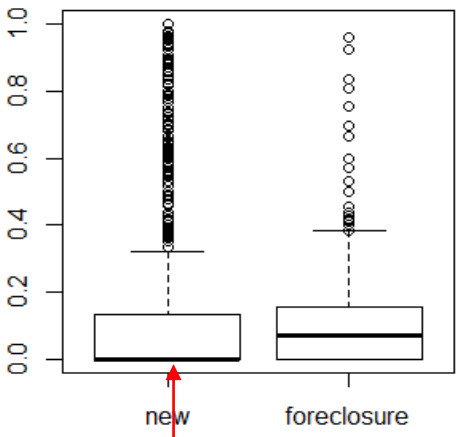
部分数据(961x4): http://staff.ustc.edu.cn/~ynyang/lm2020/lab/houston-train.xls

探索数据分析

```
houston= read.table("http://staff.ustc.edu.cn/~ynyang/lm2020/lab/houston-train.xls",head=T)
pairs(houston)
houston=houston[-271, ]  ##删除271
```

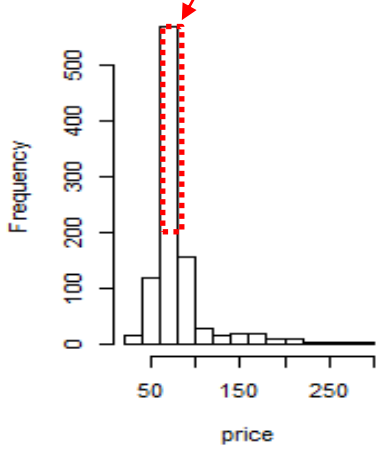


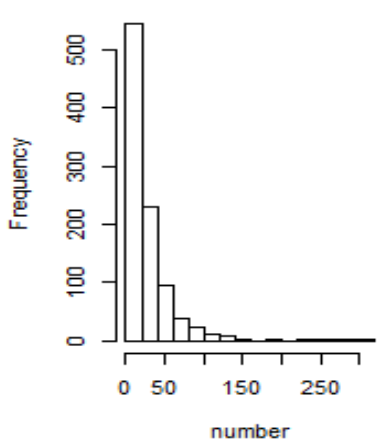


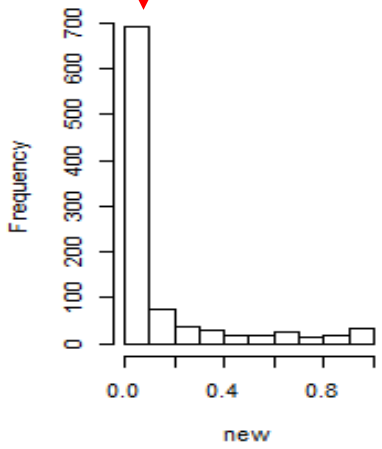


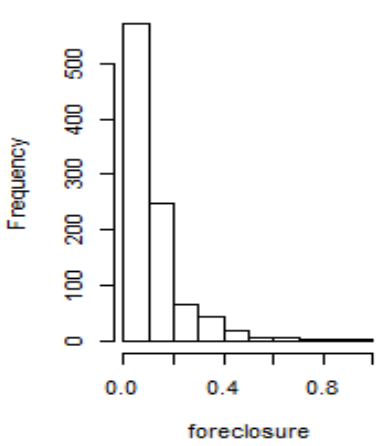
price分布不均衡，中间部分频率过高，BC变换不能改善这种情况。

new极不均衡， new=0的房区占比60%。







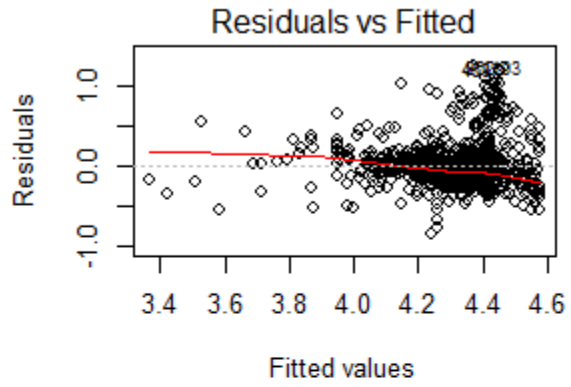


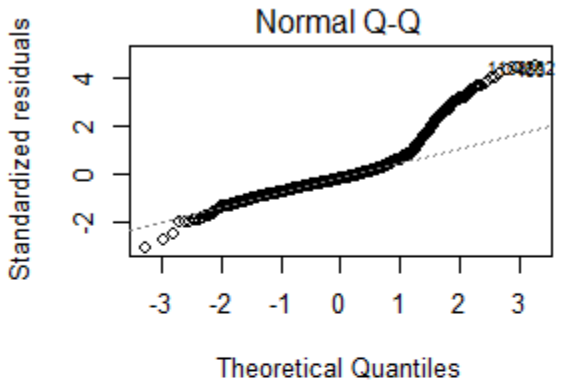
price频率最高的那部分（60<price<80）大概有560个房区，其中360个是没有新房子出售的房区（new=0），其余200个房区有新房出售。即红色方框部分代表没有新房出售的房区。

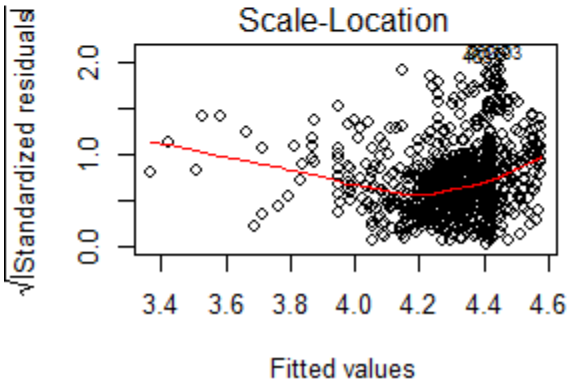
fit1: 拟合模型

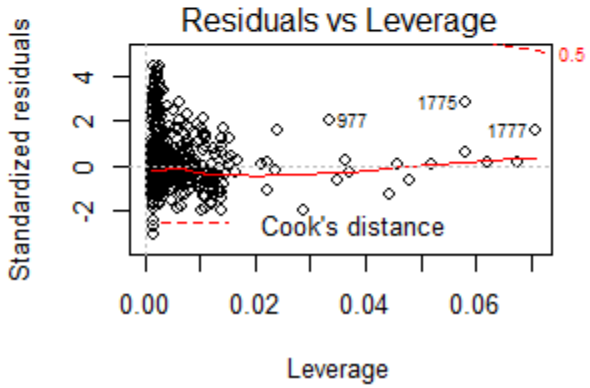
lm(log(price) ~new+foreclosure,data=houston)

$$\log(\text{price}) = \beta_0 + \beta_1 \times \text{new} + \beta_2 \times \text{foreclosure} + \varepsilon, \quad \varepsilon \sim (0, \sigma^2) \text{或} \varepsilon \sim (0, \sigma^2 / \text{number})$$

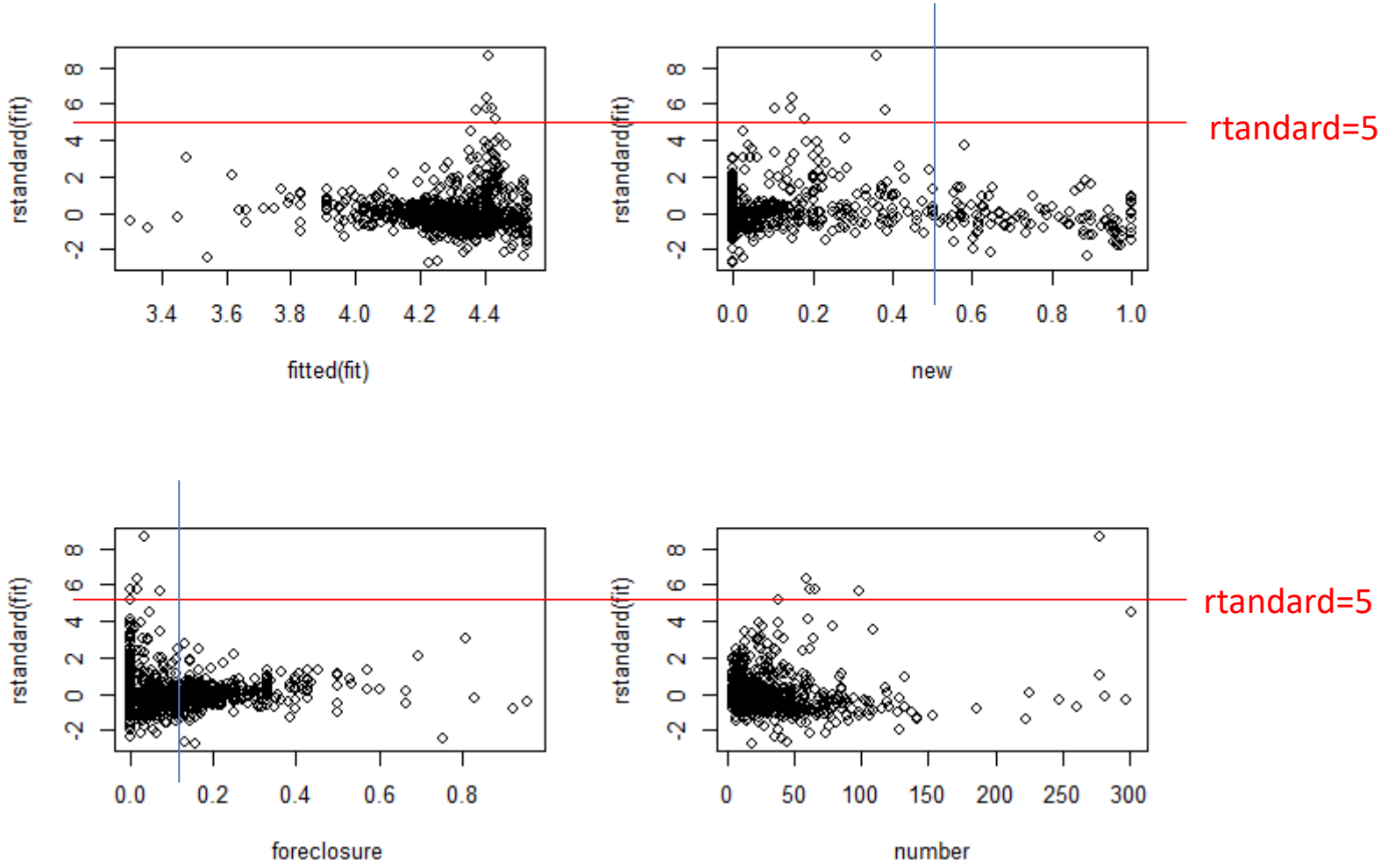








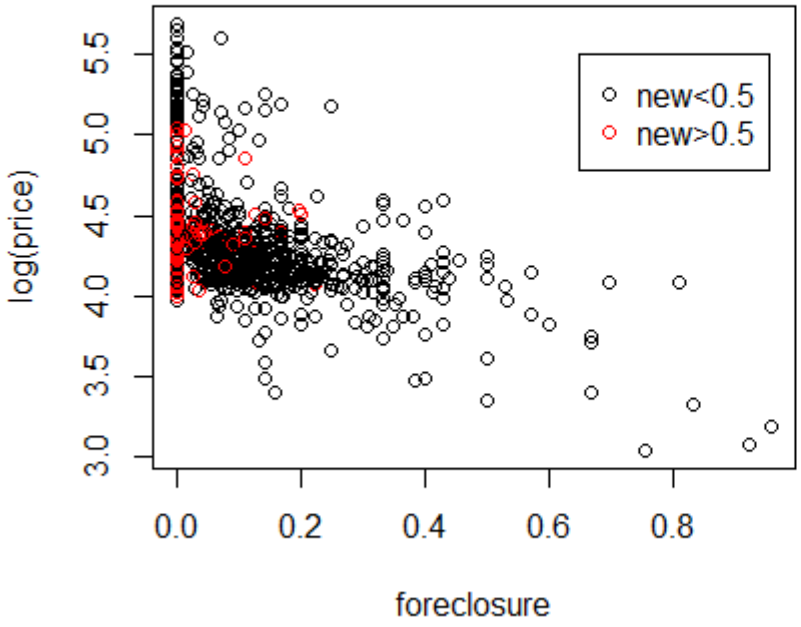
残差图：标准化残差VS每个自变量（加权残差图类似)

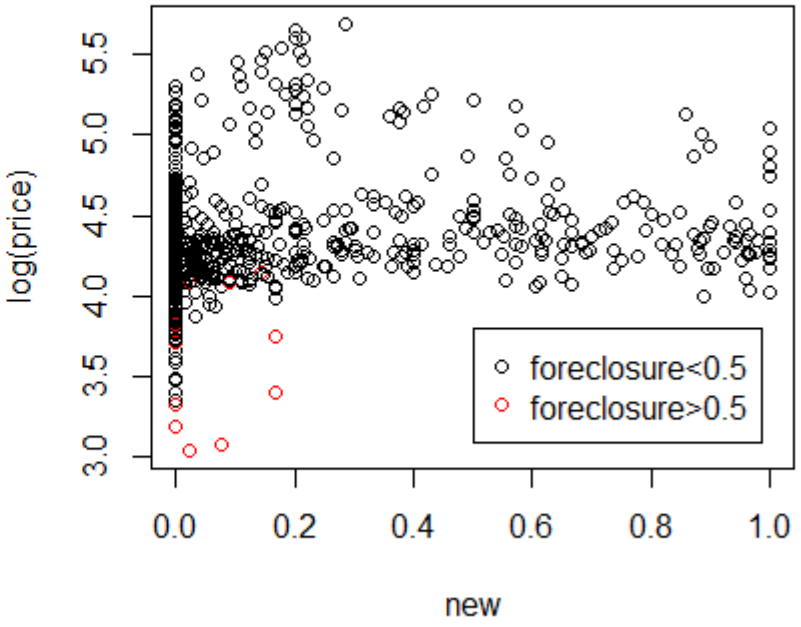


Outlier（rtandard>5) 集中在foreclosure ~0-0.1 附近,和new<0.5(即这些outlier都是旧房子）。图3说明误差方差随foreclosure增大而减小。

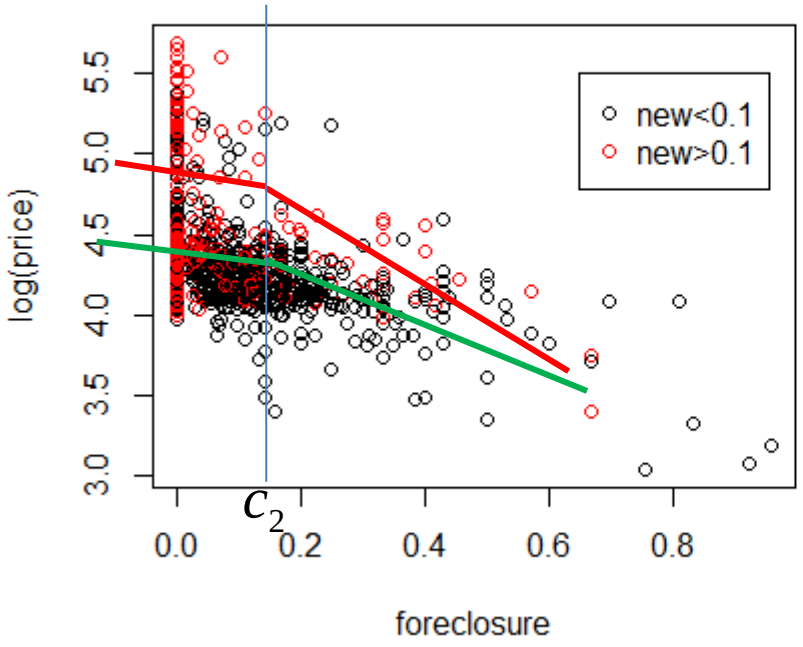
再看散点图

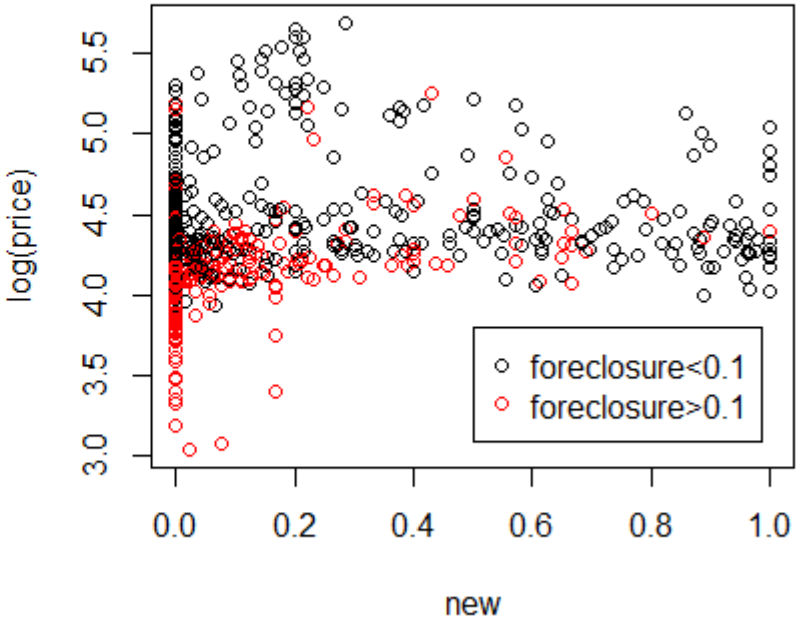
```
plot(log(price)~foreclosure, data=houston,col=1+(new>0.1 ))
plot(log(price)~new, data=houston,col=1+(foreclosure>0.15 ))
```





- 特别高的高房价（log(price)>5.5,price>245) 房子集中于fore~0和new<0.5;
- price 与foreclosure负相关，与new关系不明显
- 左图红点：中位数房子是新的 （new>0.5), 它们房价适中，拍卖率较低
- 右图红点：中位数房子是拍卖的(foreclosure>0.5)，它们房价很低，在老房区中（新房率<20%）。





- 老房区（new<0.1) 的价格较低，拍卖率低的小区 (foreclosure<0.1)房价较高。
- 左图：log(price)与foreclosure的关系在new较大与较小时不同，在foreclosure较大与较小时也不同。
- 右图： foreclosure较小时price较大。

fit2: 拟合交互作用模型（非线性

```
lm(log(price) ~ new+foreclosure +forclosure*I(new<0.1)
+ foreclosure*I(forclosure<0.15) ,data=houston)
```

$$\log(\text{price}) = \beta_0 + \beta_1 \times \text{new} + \beta_2 \times \text{foreclosure} + \alpha_1 1_{(\text{new}<c_1)} + \alpha_2 1_{(\text{foreclosure}<c_2)}$$
$$+ \gamma_1 \times \text{foreclosure} \times 1_{(\text{new}<c_1)} + \gamma_2 \times \text{foreclosure} \times 1_{(\text{foreclosure}<c_2)} + \varepsilon,$$

我们取$c_1 = 0.1, c_2 = 0.15$

上述模型等价于：

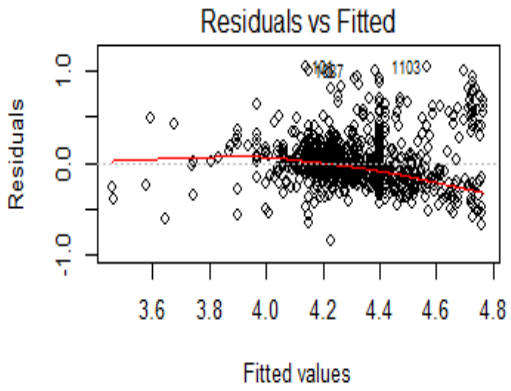
$\text{new} \geq c_1, \text{foreclosure} \geq c_2:\ \log(\text{price}) = \beta_0 + \beta_1 \times \text{new} + \beta_2 \times \text{foreclosure} + \varepsilon,$

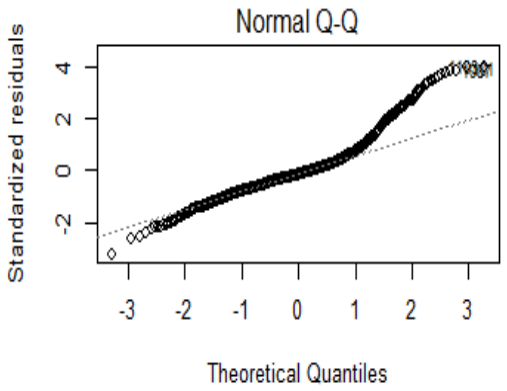
$\text{new} \geq c_1, \text{foreclosure} < c_2:\ \log(\text{price}) = \beta_0 + \alpha_2 + \beta_1 \times \text{new} + (\beta_2 + \gamma_2) \times \text{foreclosure} + \varepsilon,$

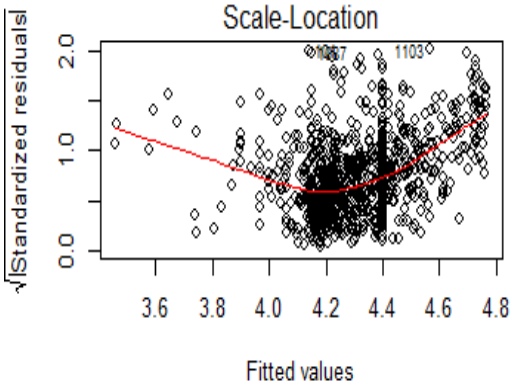
$\text{new} < c_1, \text{foreclosure} \geq c_2:\ \log(\text{price}) = \beta_0 + \alpha_1 + \beta_1 \times \text{new} + (\beta_2 + \gamma_1) \times \text{foreclosure} + \varepsilon,$

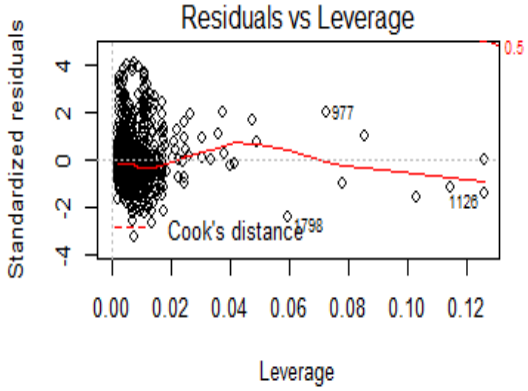
$\text{new} < c_1, \text{foreclosure} < c_2:\ \log(\text{price}) = \beta_0 + \alpha_1 + \alpha_2 + \beta_1 \times \text{new} + (\beta_2 + \gamma_1 + \gamma_2) \times \text{foreclosure} + \varepsilon,$

即在(new, foreclosure)的(高, 高), (高, 低), (低, 高), (低, 低)四种水平组合下， foreclosue 的回归系数不同, 截距项不同。


Residuals vs Fitted
Residuals
Fitted values
1103


Normal Q-Q
Standardized residuals
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Fitted values
1103


Residuals vs Leverage
Standardized residuals
Cook's distance
977
1126
Leverage

```
fit1:
lm(formula = log(price) ~ new + foreclosure, data =houston, weights = number)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.40957    0.01312 336.177  < 2e-16 ***
new          0.12541    0.03246   3.864 0.000119 ***
foreclosure -1.15313    0.07817 -14.751  < 2e-16 ***
---
Residual standard error: 1.356 on 957 degrees of freedom
Multiple R-squared:  0.2277,    Adjusted R-squared:  0.2261
F-statistic: 141.1 on 2 and 957 DF,  p-value: < 2.2e-16
```

```
fit2：
lm(formula = log(price) ~ new + foreclosure * I(foreclosure >     0.15) + foreclosure * I(new > 0.1), data = houston)

Coefficients:
                                    Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)                          4.40061   0.01622      271.250   < 2e-16 ***
new                                 -0.38939   0.05699      -6.833     1.49e-11 ***
foreclosure                         -1.73830   0.20418       -8.514    < 2e-16 ***
I(foreclosure > 0.15)TRUE           -0.01784   0.03994      -0.447     0.655266
I(new > 0.1)TRUE                     0.40722   0.03733       10.908      < 2e-16 ***
foreclosure:I(foreclosure > 0.15)TRUE  0.77444   0.23248      3.331     0.000898 ***
foreclosure:I(new > 0.1)TRUE        -0.50509   0.16579  -    3.046      0.002379 **
---
Residual standard error: 0.2602 on 953 degrees of freedom
Multiple R-squared:  0.3334,    Adjusted R-squared:  0.3292
F-statistic: 79.43 on 6 and 953 DF,  p-value: < 2.2e-16
```

Fit2相比于fit1，R2从0.22增加到0.33，参数增加了4个。拟合效果有较大幅度改善。

预测

如果再引入一些变量，比如加入number, 加入高阶项，高阶交互项等，比如调正阈值c_1, c_2, 拟合效果会有进一步提升，但模型的预测能力不一定相应提升。所谓预测，就是对未知房价的房区，通过其已知的自变量预测其房价. 假设我们得到fit2拟合模型：

$$\log(\text{price}) = \hat{\beta}_0 + \hat{\alpha}_1 1_{(\text{new}<0.1)} + \hat{\alpha}_2 1_{(\text{fore}<0.15)} + \hat{\beta}_1 \times \text{new} + \hat{\beta}_2 \times \text{fore} + \hat{\gamma}_2 \times \text{fore} \times 1_{(\text{fore}<0.15)}$$

我们预测new=0.3, fore=0.1, number=9的房区的房价中位数，

$$\log(\text{price})_{\text{预测}} = \hat{\beta}_0 + \hat{\alpha}_2 + \hat{\beta}_1 \times \text{new} + (\hat{\beta}_2 + \hat{\gamma}_2) \times \text{fore} = \hat{\beta}_0 + \hat{\alpha}_2 + 0.3\hat{\beta}_1 + 0.1(\hat{\beta}_2 + \hat{\gamma}_2)$$

如果我们知道上述房区的房价 $\text{price}_{\text{true}}$, 那么预测误差为

$$\log(\text{price})_{\text{预测}} - \log(\text{price})_{\text{true}}$$

评估预测误差：交叉验证（cross-validation)

为了评价模型的预测效果，一般使用交叉验证方法，把数据集houston分成两部分：

- 一部分用于拟合模型（训练），称为训练数据集(training data)；
- 另一部分数据用来测试训练得到的模型，称为测试数据集。

使用训练数据得到拟合结果，预测测试集中的响应变量，得到预测误差平方。比较各种训练模型的预测误差，选取预测误差最小的作为最终预测模型。

拟合、但不要过度拟合：
后面我们将会看到，好的预测模型在训练数据上拟合效果不一定最好，但也一定不是太差。