高级统计方法 第N次作业:

序号： 03 姓名： 潘晨楷 学号：20212241116 班级：软2107

**作业评阅：**

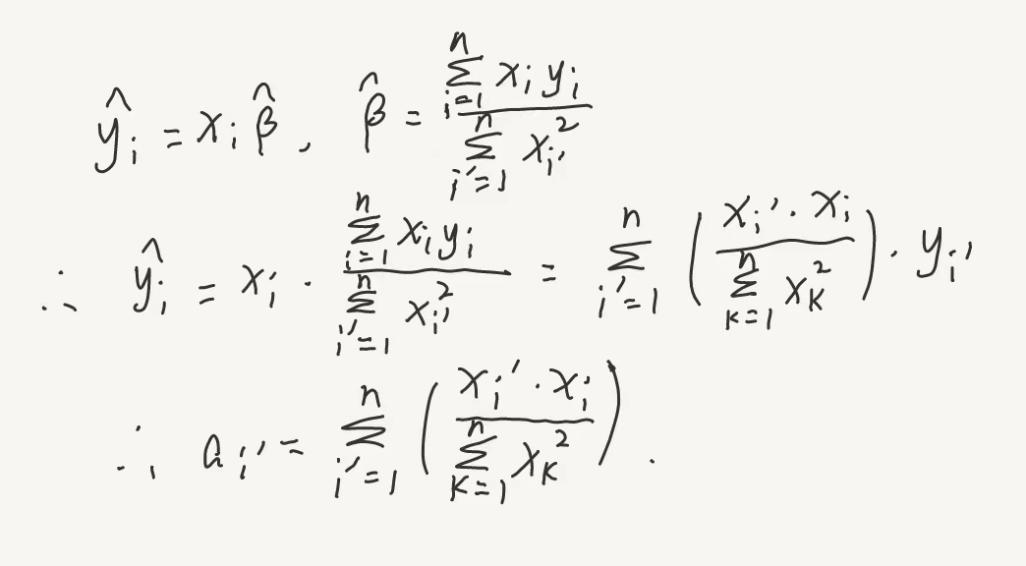
**概念**

2.问题（略）

KNN分类器和KNN回归方法在公式上是相互联系的。但是，KNN分类器的最终结果是Y的分类输出(定性)，KNN回归的输出是f(X)的定量值。

、、、、、、

1. 问题（略）

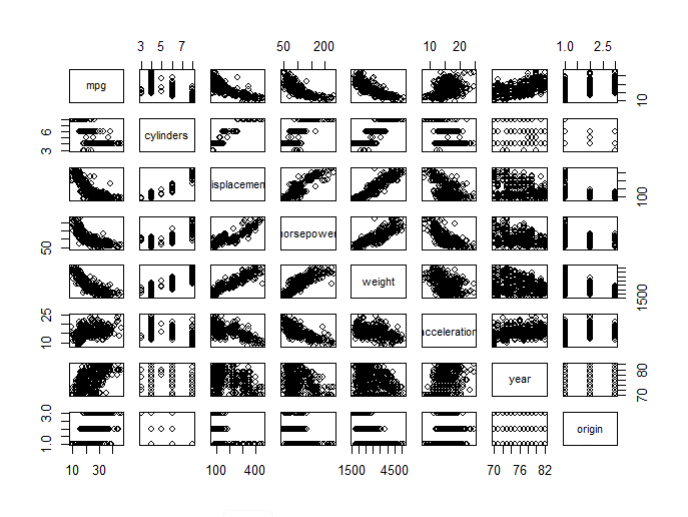


**应用**

9.问题（略）

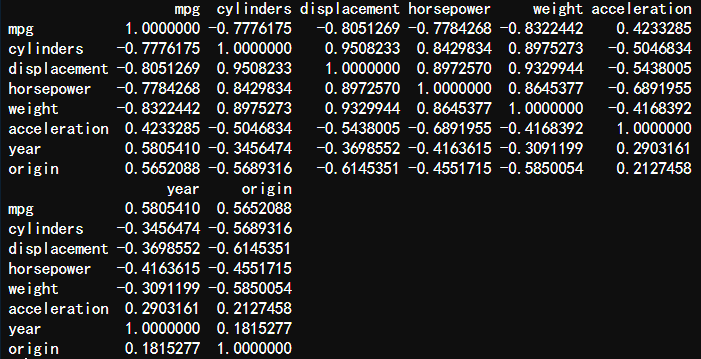
（a）问题（略）

pairs(subset(Auto,select=-name))



（b）问题（略）

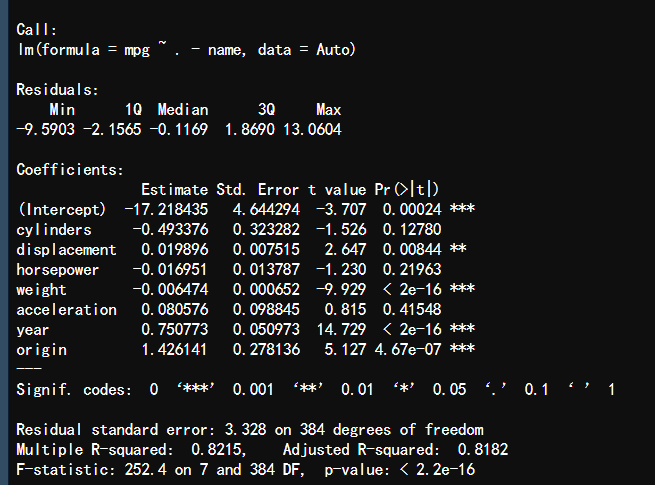
cor(subset(Auto, select=-name))



（c）问题（略）

lm.fit1 = lm(mpg~.-name, data=Auto)

summary(lm.fit1)

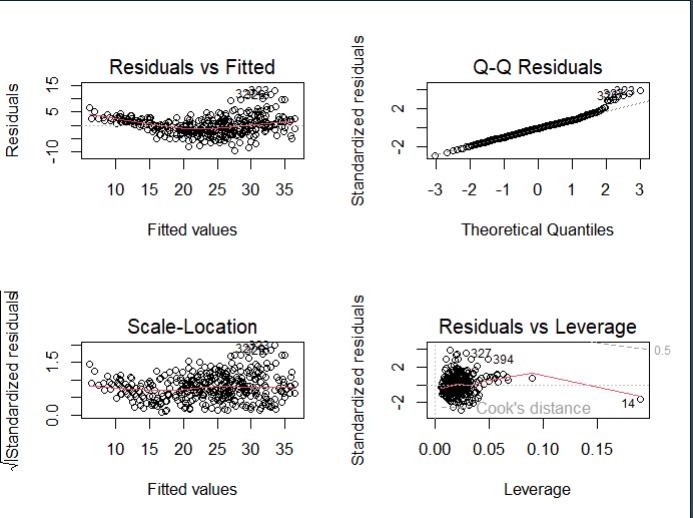


1. 通过检验是否所有回归系数为零的零假设，预测因子和响应之间存在关系。F统计量离1很远(p值很小)，表明有证据反对原假设。所以预测变量和响应变量之间有关系。
2. 查看与每个预测器的t统计量相关的p值，我们可以看到displacement、weight、 year和origin在统计上有显著的关系，而cylinders、horsepower和acceleration则没有。
3. Year的回归系数为0.7508，表明每一年，mpg的系数增加。换句话说，汽车的燃油效率每年提高近1英里/加仑。

（d）问题（略）

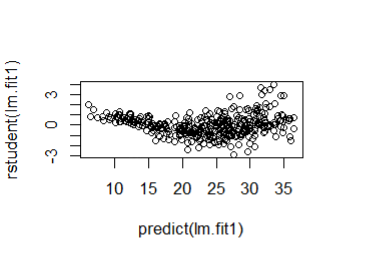
par(mfrow=c(2,2))

plot(lm.fit1)



1. 由拟合残差图可以看出，残差并不处于一个稳定的状态，模型欠缺一定的准确性。
2. 从Normal Q-Q图可以得到残差基本符合正态分布。
3. 从拟合值-方差图可以得到方差处于一个基本确定的状态。
4. 从杠杆图可以看到14号点被识别出了异常高杠杆。

plot(predict(lm.fit1), rstudent(lm.fit1))



从图中可以看出有一些异常大的离群点。

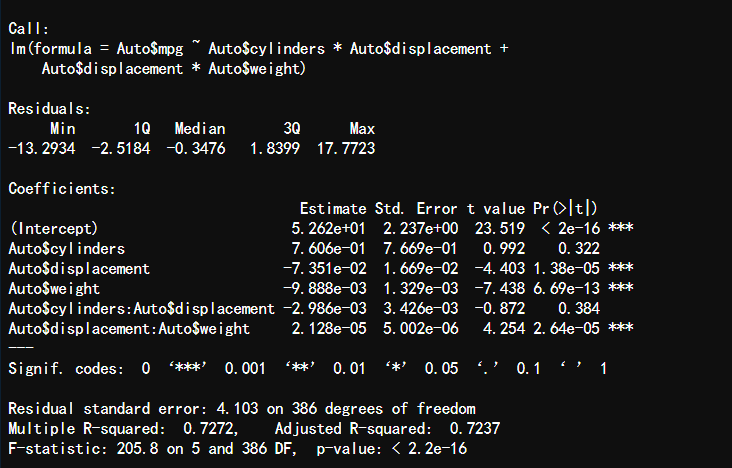
（e）问题（略）

从相关系数矩阵中，找到两个相关系数大于0.9的变量对：

cylinders，displacement和displacement，weight

lm.fit9=lm(Auto$mpg~Auto$cylinders\*Auto$displacement+Auto$displacement\*Auto$weight)

summary(lm.fit9)

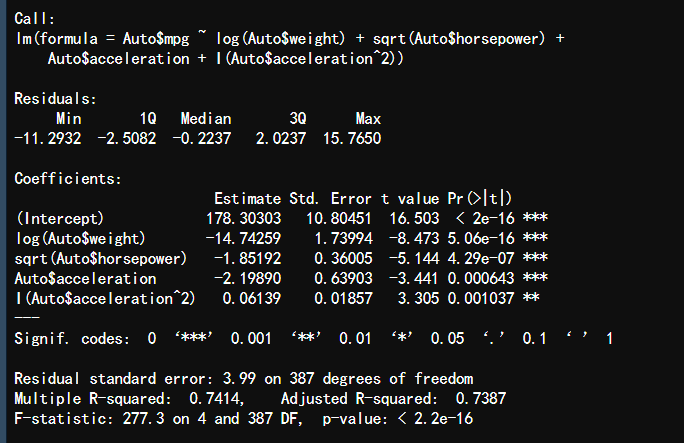


从结果的p值看，displacement和weight的交互作用在统计上显著，而 cylinders和displacement的交互作用则不显著。

（f）问题（略）

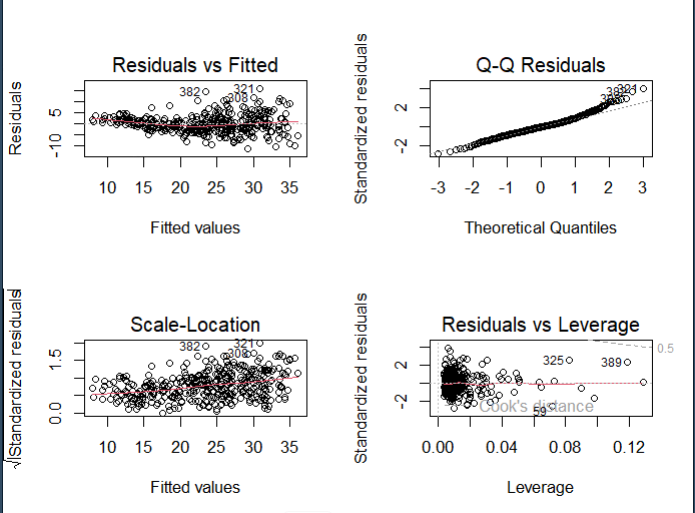
>lm.fit10=lm(Auto$mpg~log(Auto$weight)+sqrt(Auto$horsepower)+Auto$acceleration+I(Auto$acceleration^2))

> summary(lm.fit10)

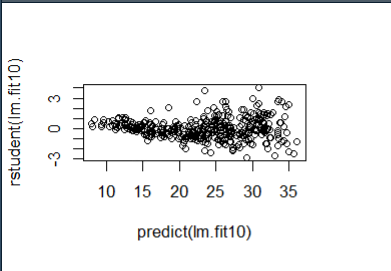


> par(mfrow=c(2,2))

> plot(lm.fit10)



plot(predict(lm.fit10), rstudent(lm.fit10))



显然，从p值来看，log(weight)、sqrt(horsepower)和acceleration^2都具有某种统计显著性。残差图比所有线性回归项的图具有更少的可识别模式。学生化残差显示潜在的异常值(>3)。杠杆图显示了三个以上的高杠杆点。

但是，从上面的图中观察到两个问题:1)残差与拟合图表明模型中的异方差(非恒定方差超过平均值)。2) Q-Q图表明残差存在一定的异常。

因此，需要对我们的模型进行更好的转换。从9（a）的相关矩阵。displacement，horsepower和weight与我们的响应mpg表现出类似的非线性模式。这种非线性模式非常接近于对数形式。因此，在下一次尝试中，我们使用log(mpg)作为响应变量。

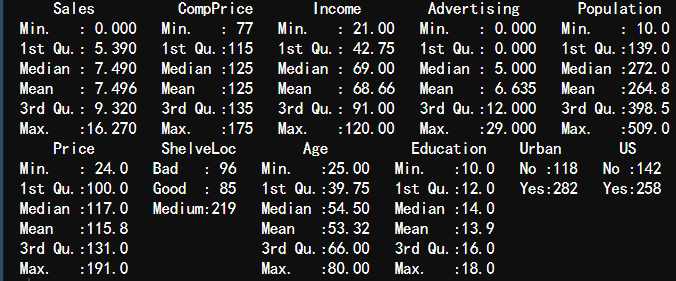
结果表明，mpg的对数变换得到了更好的模型拟合(更好的R^2，残差正态性)。

1. 问题（略）
2. 问题（略）

install.packages("ISLR")

library(ISLR)

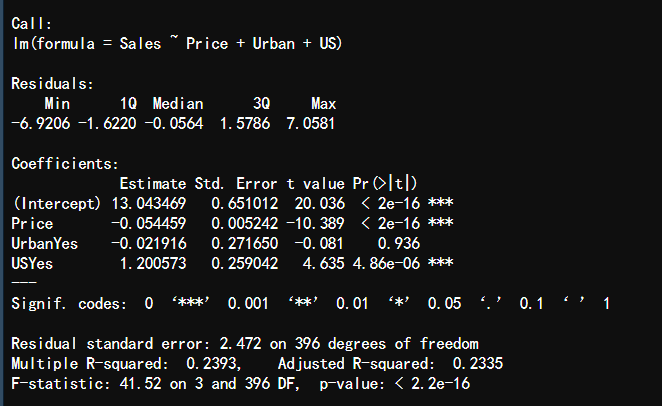
summary(Carseats)



attach(Carseats)

lm.fit = lm(Sales~Price+Urban+US)

summary(lm.fit)



1. 问题（略）

Price：

其p值很接近0，说明销售额和价格之间存在统计上的显著关系，该系数为-0.054459，说明存在的是负关系，随着价格的增加，销售额减少。

UrbanYes：

其有着高p值，说明商店位置在不在城市与销售额没有统计上的 显著关系。

USYes：

其p值很接近0，说明商店在不在美国与销售额存在统计上的显著关系，其系数是1.2，说明当商店在美国时，销售额平均增加1200。

1. 问题（略）

Sales = 13.04 - 0.05446 Price - 0.022 UrbanYes + 1.20 USYes

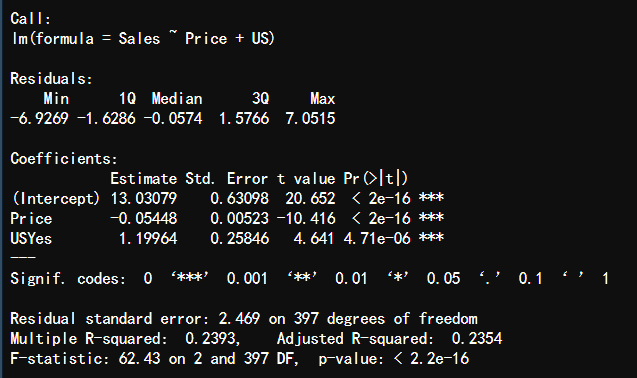
1. 问题（略）

Price和USYes可以拒绝零假设，因为其p值很接近0。

1. 问题（略）

lm.fit2 = lm(Sales ~ Price + US)

summary(lm.fit2)

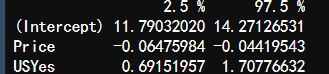


1. 问题（略）

（e）中模型的RSE和R^2均比（a）中略小一点，拟合效果差距不大，（e）中的模型对数据的拟合会比（a）略好一点。

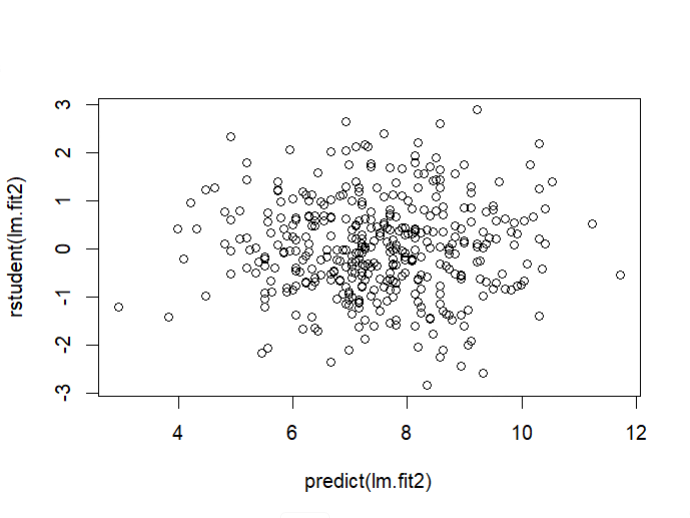
1. 问题（略）

confint(lm.fit2)



1. 问题（略）

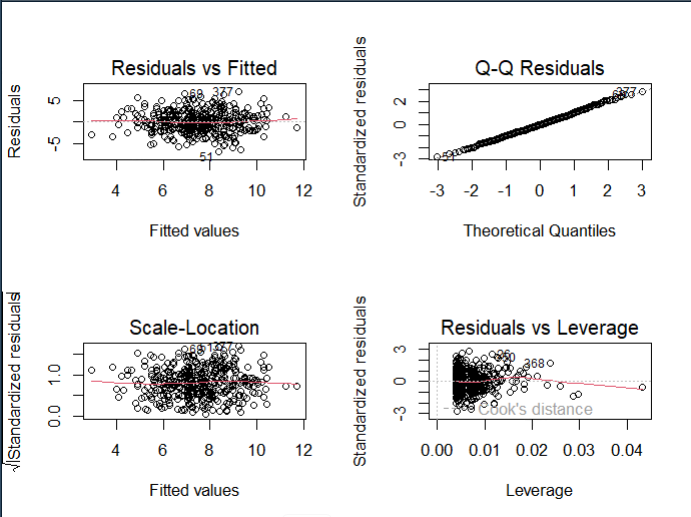
plot(predict(lm.fit2), rstudent(lm.fit2))



由图可知所有的残差都在(-3，3)的区间内，没有异常的离群点。

par(mfrow=c(2,2))

plot(lm.fit2)



在杠杆统计图上，有一些观测值大大超过(p+1)/n(0.0076)，这表明相应的点具有高杠杆。

1. 问题（略）
2. 问题（略）

set.seed(1)

x1 = runif(100)

x2 = 0.5 \* x1 + rnorm(100)/10

y = 2 + 2\*x1 + 0.3\*x2 + rnorm(100)

函数形式是：y = 2 + 2\*x1 + 0.3\*x2 +eps

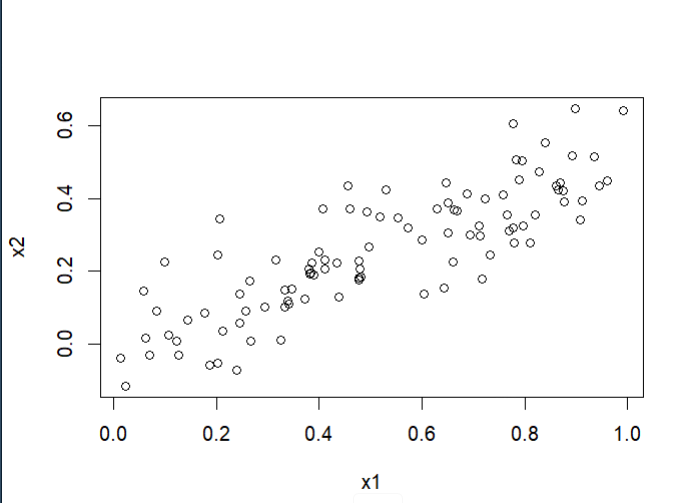
回归系数：β0=2，β1=2，β2=0.3

1. 问题（略）

cor(x1, x2)



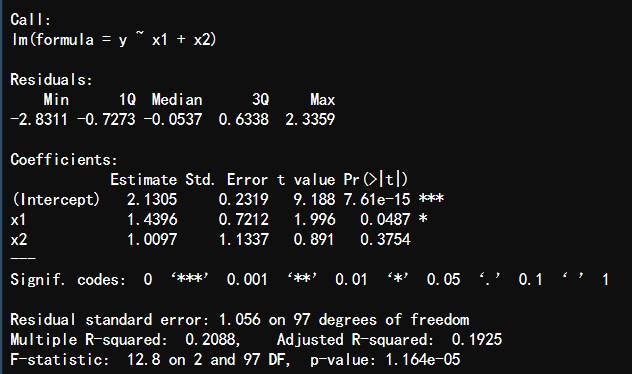
plot(x1, x2)



1. 问题（略）

lm.fit = lm(y~x1+x2)

summary(lm.fit)



β0=2.1305，β1=1.4396，β2=1.0097。

β0接近真实系数，β1和β2不太接近。

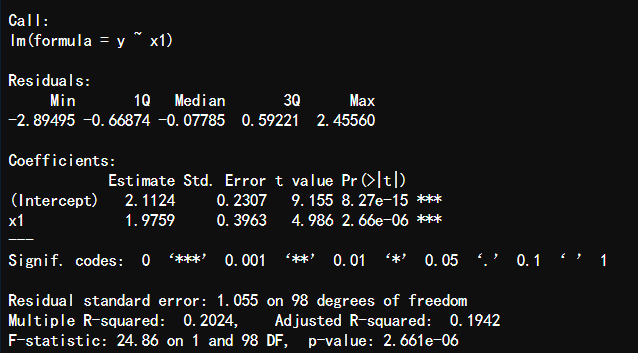
能拒绝β1=0的假设，因为其p值小于0.05。

不能拒绝β2=0的假设，因为其p值很大。

1. 问题（略）

lm.fit = lm(y~x1)

summary(lm.fit)

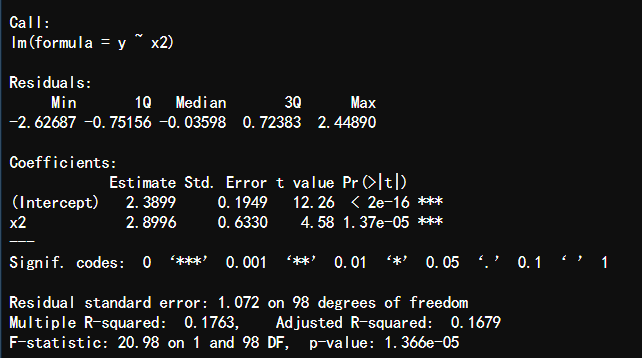


由于p值很接近0，能拒绝零假设。

1. 问题（略）

lm.fit = lm(y~x2)

summary(lm.fit)



由于p值很接近0，能拒绝零假设。

1. 问题（略）

不矛盾。x1和x2具有共线性，在回归的时候很难同时区分回归效果。而分别做回归的时候，具有明显的线性关系能够得到良好的回归效果。

1. 问题（略）

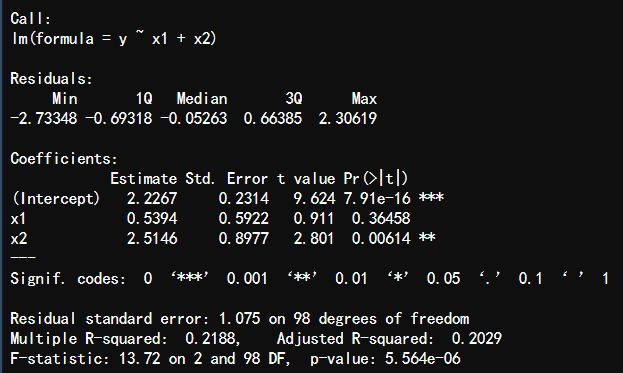
x1 = c(x1, 0.1)

x2 = c(x2, 0.8)

y = c(y, 6)

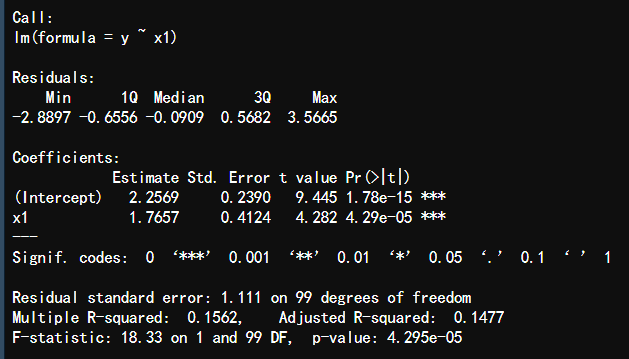
lm.fit1 = lm(y~x1+x2)

summary(lm.fit1)



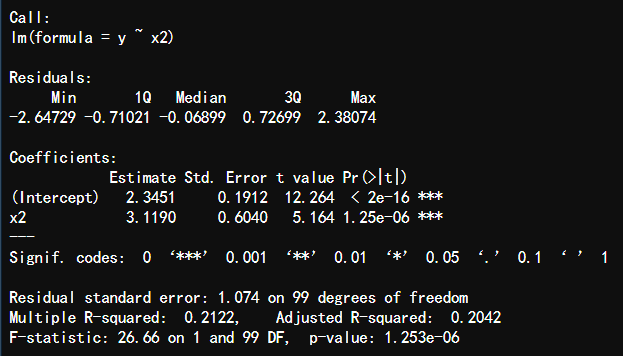
lm.fit2 = lm(y~x1)

summary(lm.fit2)



lm.fit3 = lm(y~x2)

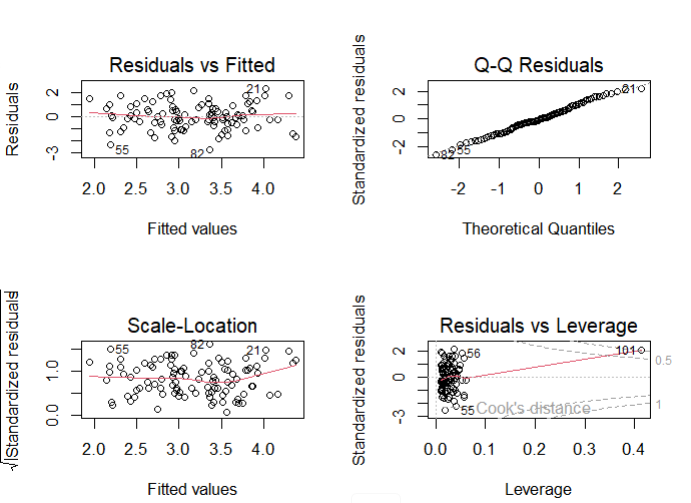
summary(lm.fit3)



在x1和x2的回归中，p值告诉我们，x1不显著，x2显著。

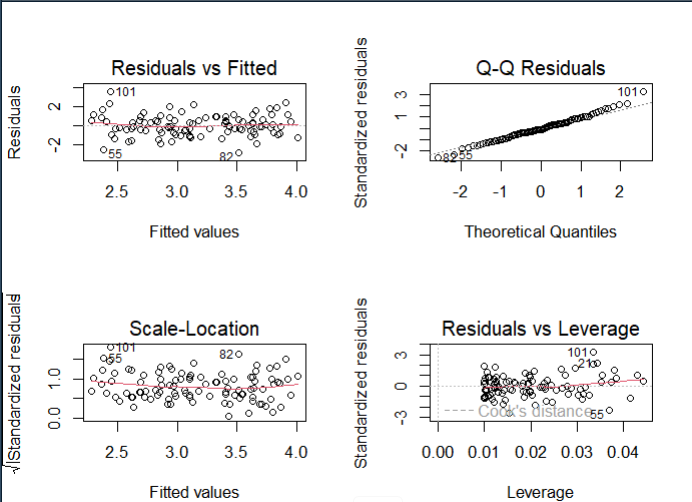
par(mfrow=c(2,2))

plot(lm.fit1)



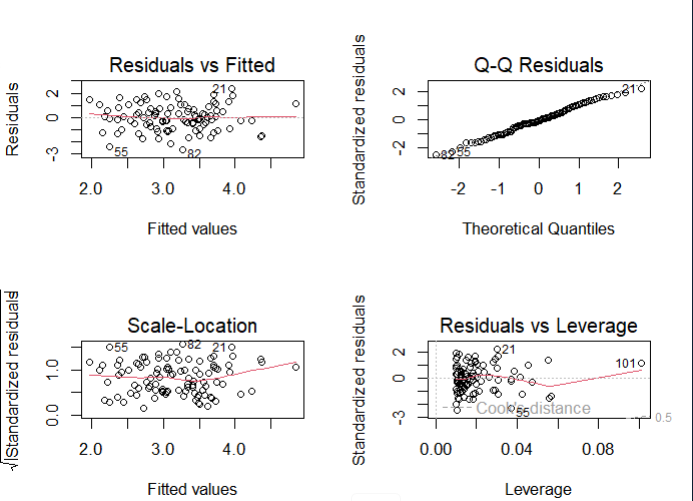
par(mfrow=c(2,2))

plot(lm.fit2)



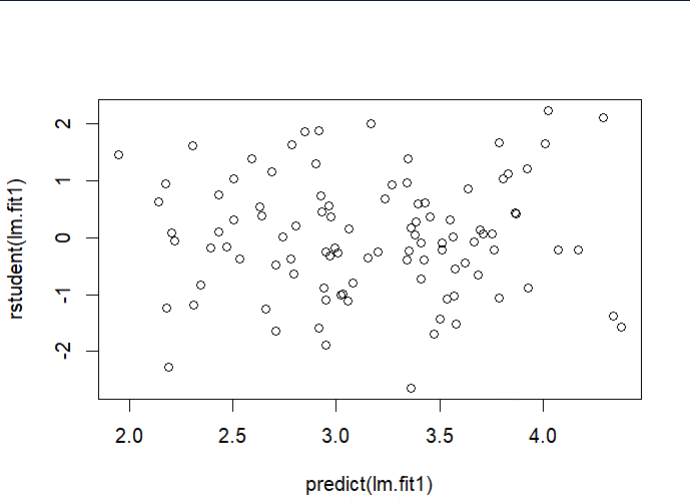
par(mfrow=c(2,2))

plot(lm.fit3)

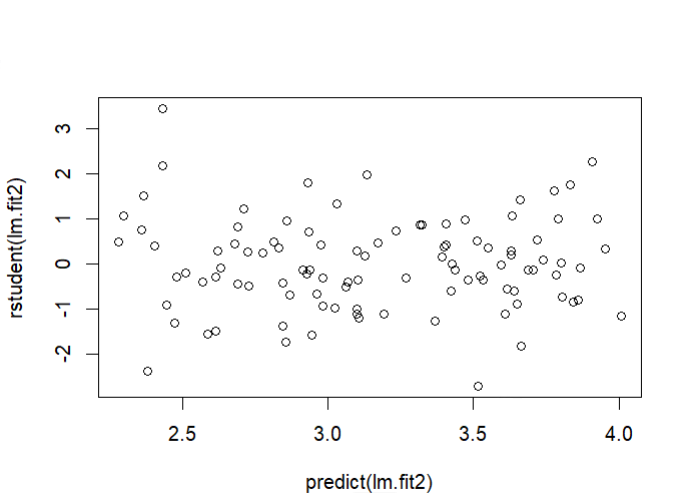


在x1和x2共同拟合和x2单独拟合的模型中，出现了high leverage作用的点。

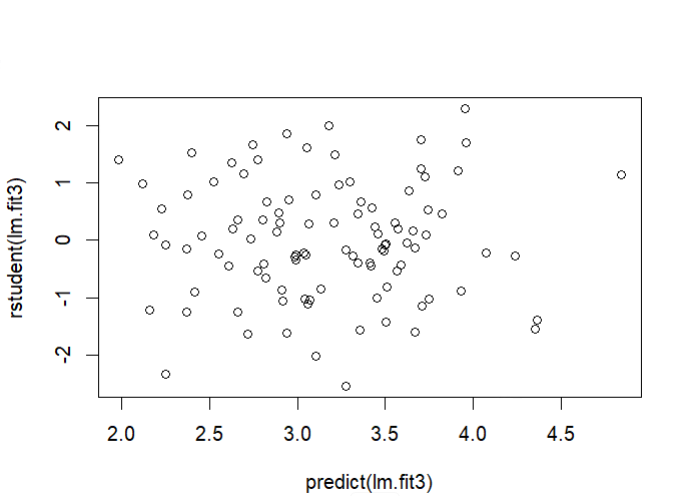
plot(predict(lm.fit1), rstudent(lm.fit1))



plot(predict(lm.fit2), rstudent(lm.fit2))



plot(predict(lm.fit3), rstudent(lm.fit3))



只有x1单独拟合的模型中发现了一个Outlier，其他数据都在[-3,3]这个区间内。