高级统计方法 第4次作业:

序号：27 姓名：王琪瑞 学号：20202241014 班级：软2002

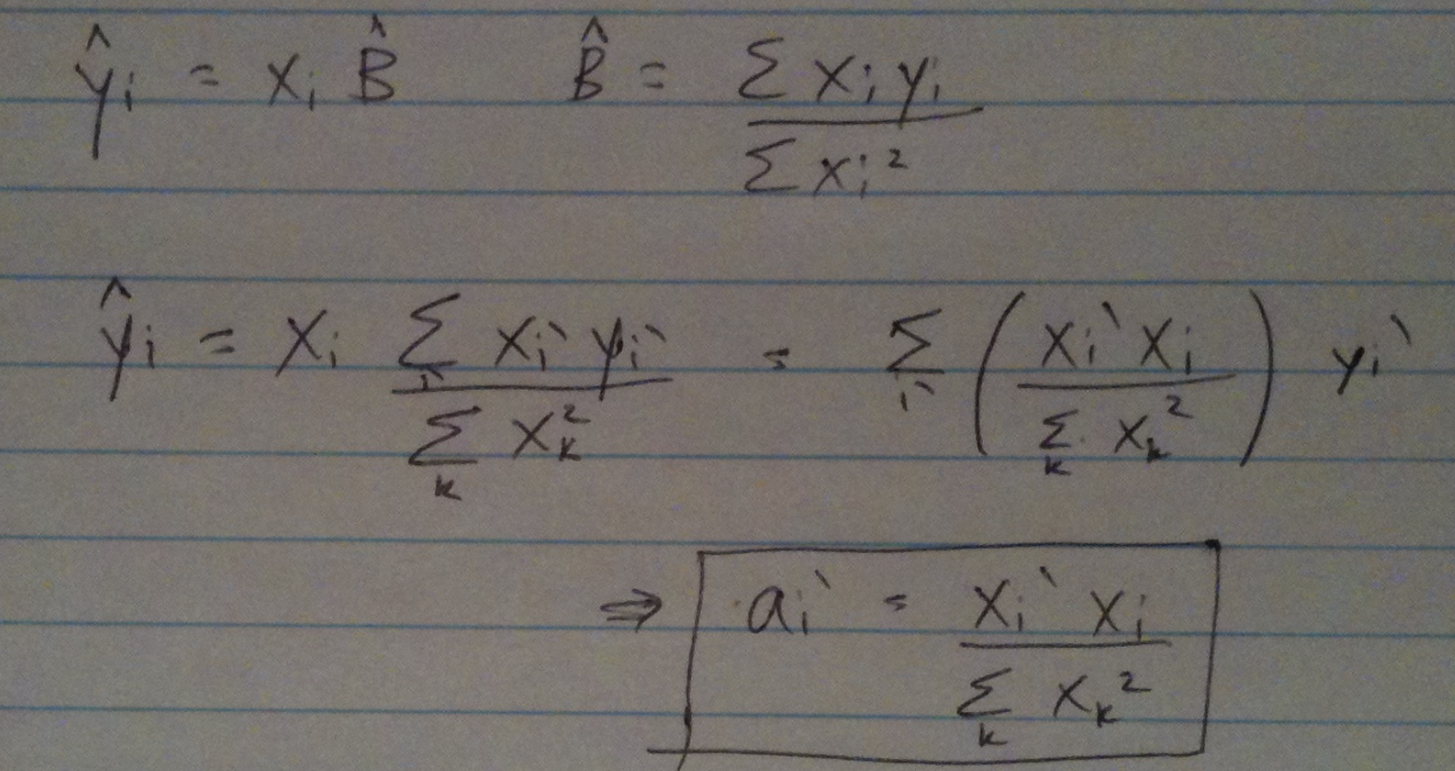
**概念**

2.问题（略）

答：KNN分类器和KNN回归方法在公式上密切相关。然而，KNN分类器的最终结果是分类输出的Y(定性)，而KNN回归的输出预测的是f(X)的定量值。

5.问题（略）

答：



**应用**

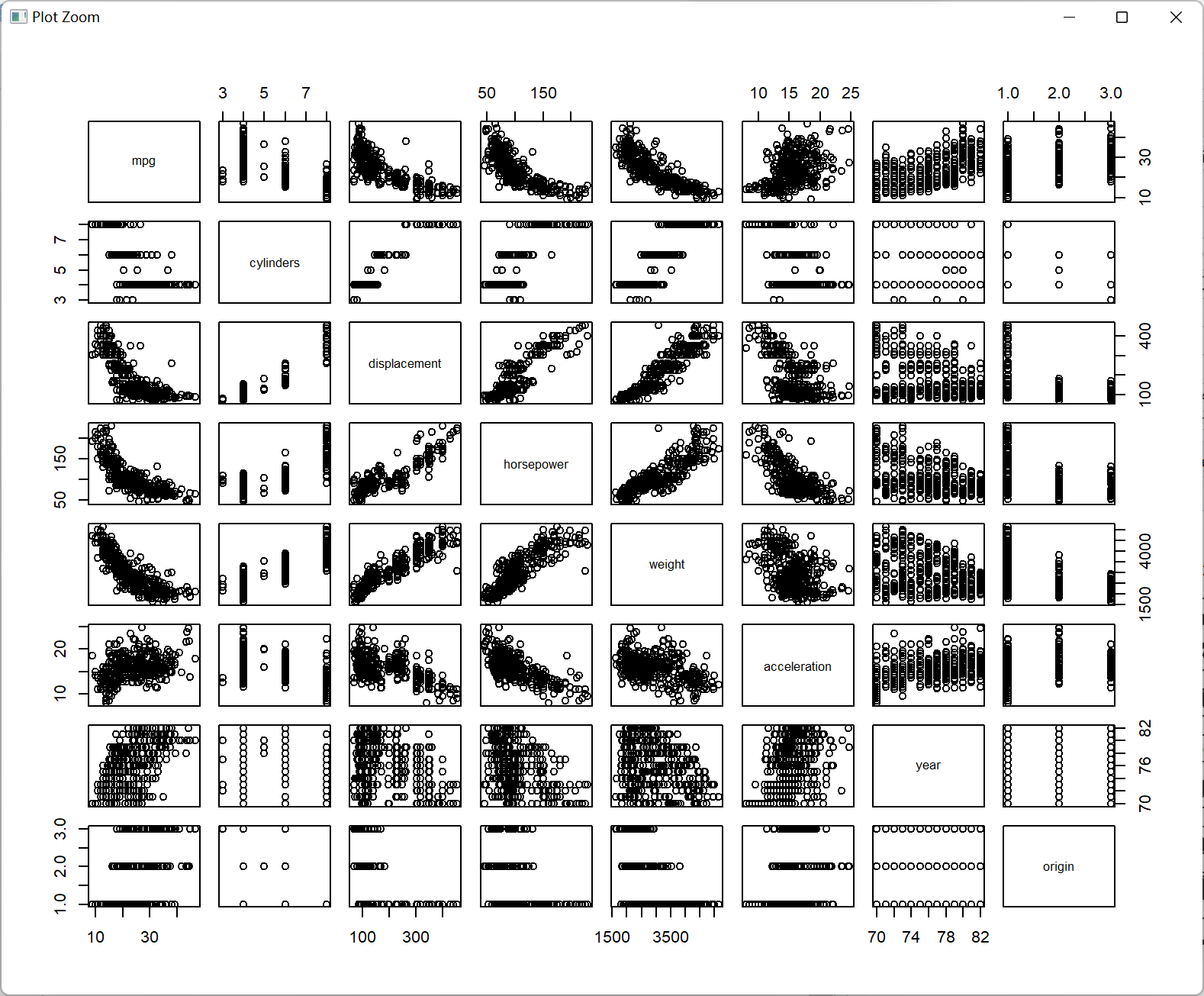
**9.问题（略）**

（a）问题（略）

脚本：

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

截图：

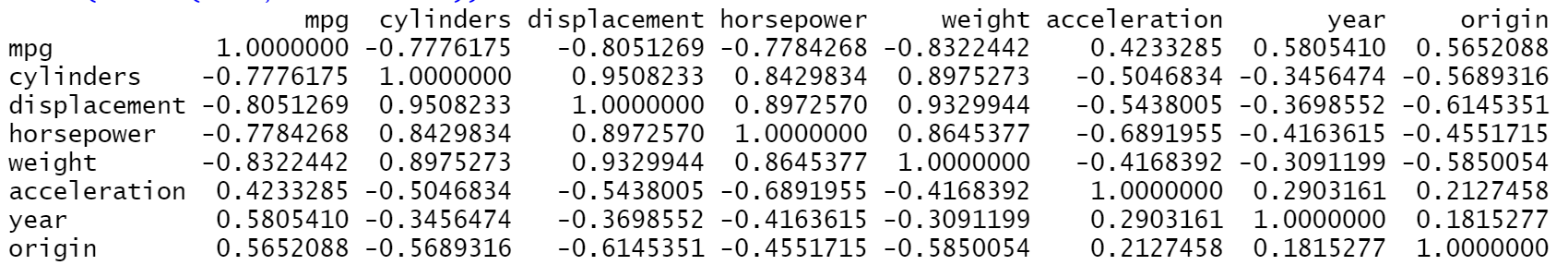


（b）问题（略）

脚本：

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

截图：



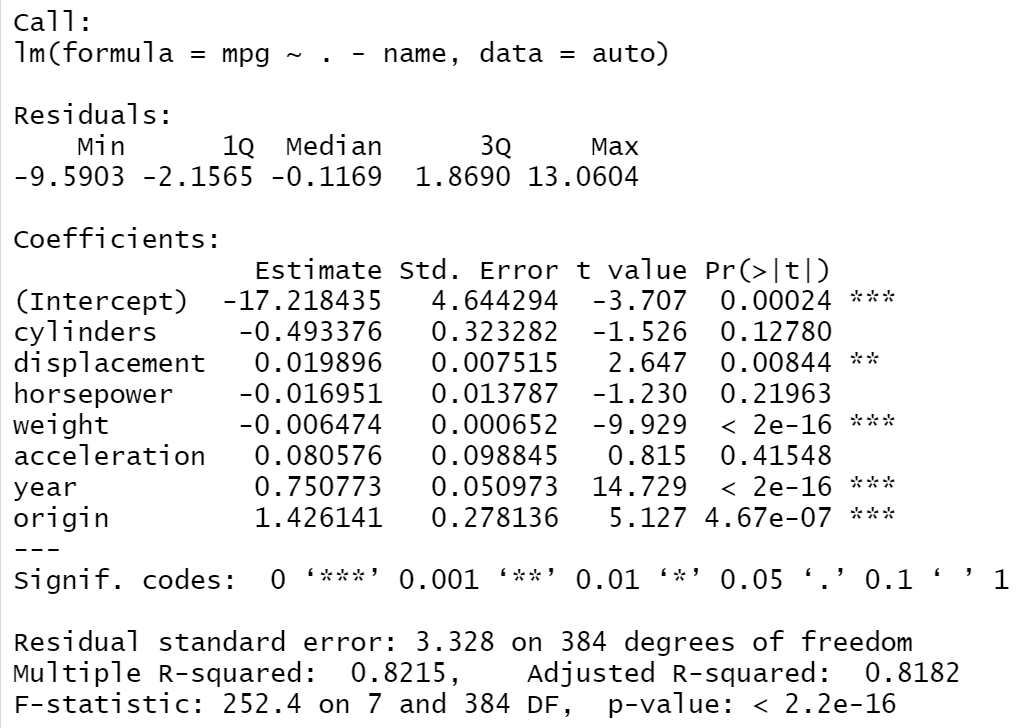
（c）问题（略）

脚本：

lm.fit=lm(mpg~.-name,data=auto)

summary(lm.fit)

截图：



i．

有关系。通过检验，把预测变量的系数均为0做零假设，检验结果F统计量远大于1，且P值接近0，拒绝零假设，即认为预测变量和响应变量之间有关系。

ii．

查看每个变量的P值，可以得出displacement, weight, year, origin与响应变量在统计上存在显著关系；而cylinders, horsepower, acceleration不存在显著关系。

iii．

年份的系数是0.7508，说明在其他因素不变的情况下，每经过一年，mpg平均增加约0.75，将近每年增加1个单位的油耗。

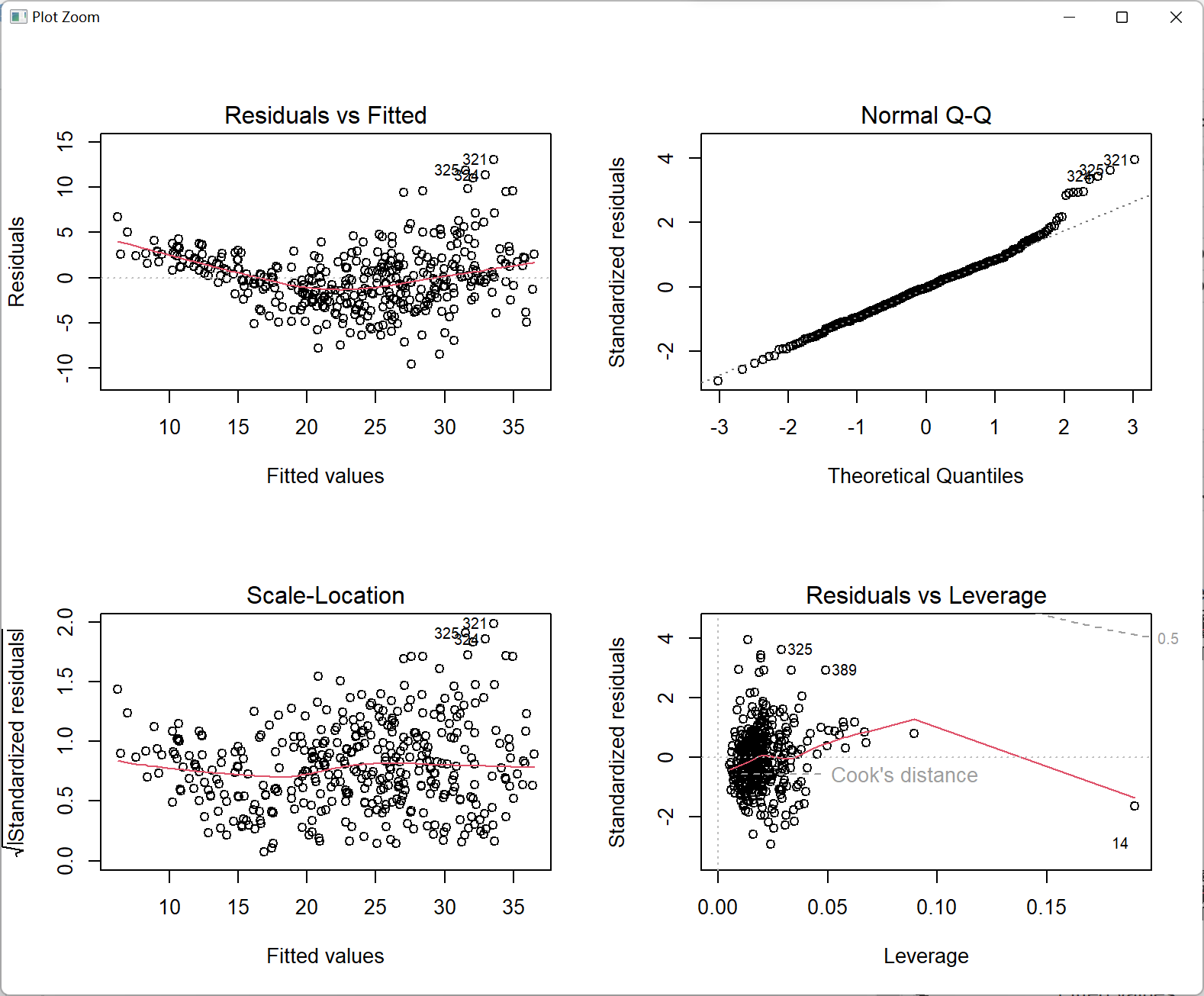
（d）问题（略）

脚本：

par(mfrow=c(2,2))

plot(lm.fit)

截图：

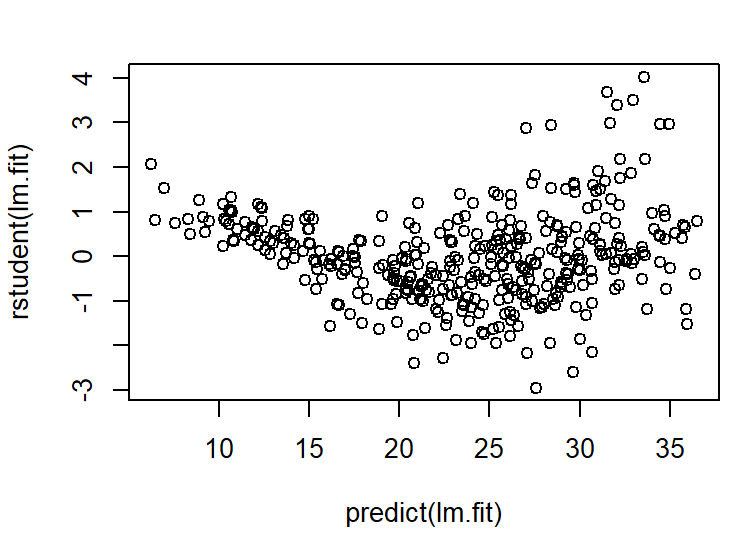


1. 由拟合残差图可以看出，残差并不处于一个稳定的状态，模型欠缺一定的准确性。
2. 从Normal Q-Q图可以得到残差基本符合正态分布。
3. 从拟合值-方差图可以得到方差处于一个基本确定的状态。
4. 从杠杆图可以看到14号点被识别出了异常高杠杆。

脚本：

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

截图：



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：

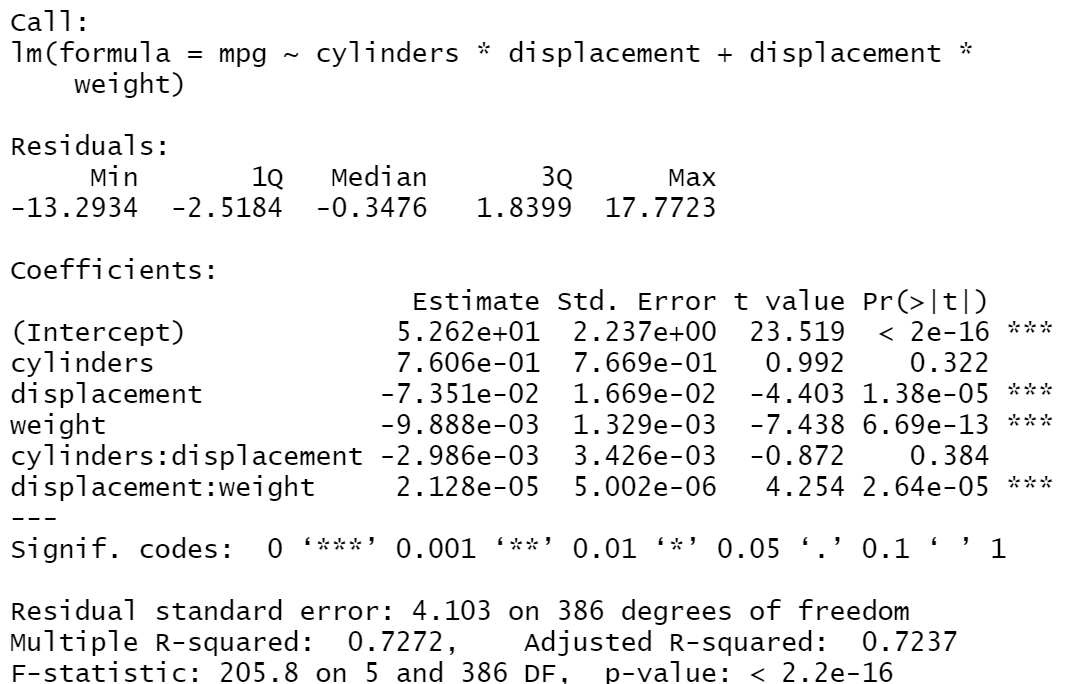
cylinders，displacement和displacement，weight

脚本：

lm.fit = lm(mpg~cylinders\*displacement+displacement\*weight)

summary(lm.fit)

截图：



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）

对此前回归中不显著的变量做变换。

脚本：

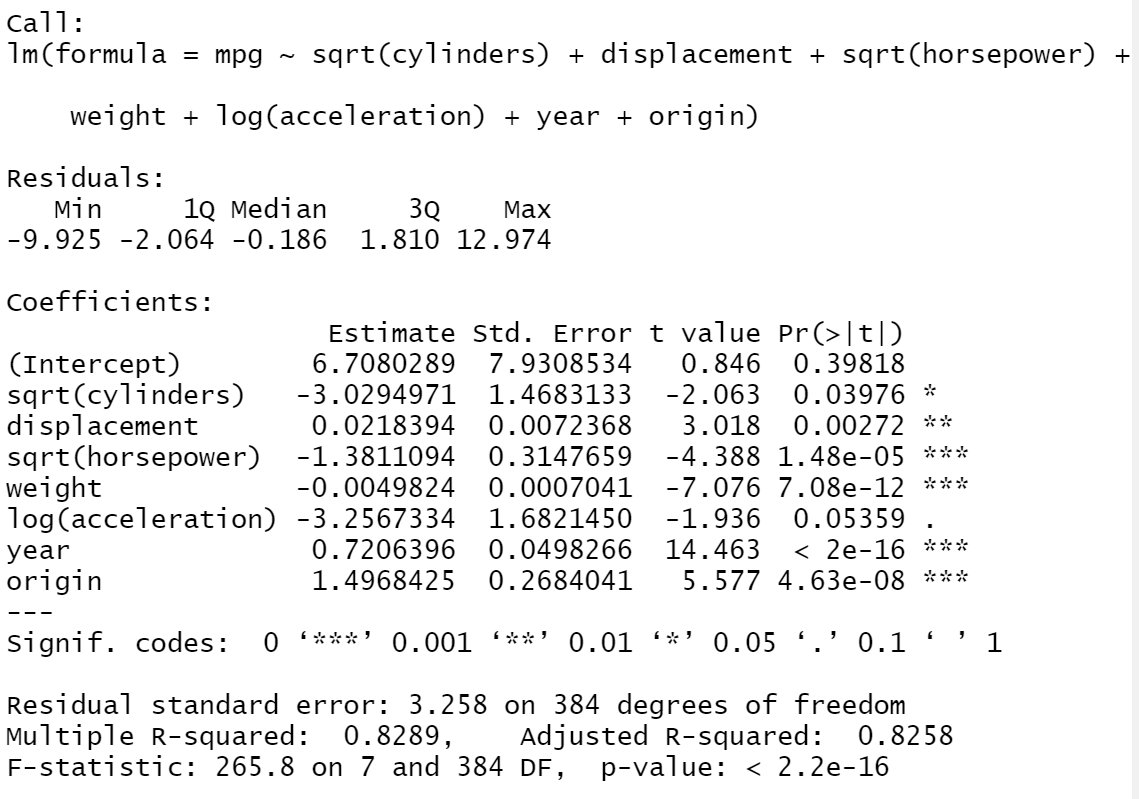
lm.fit=lm(mpg~sqrt(cylinders)+displacement+sqrt(horsepower)+weight+log(acceleration)+year+origin)

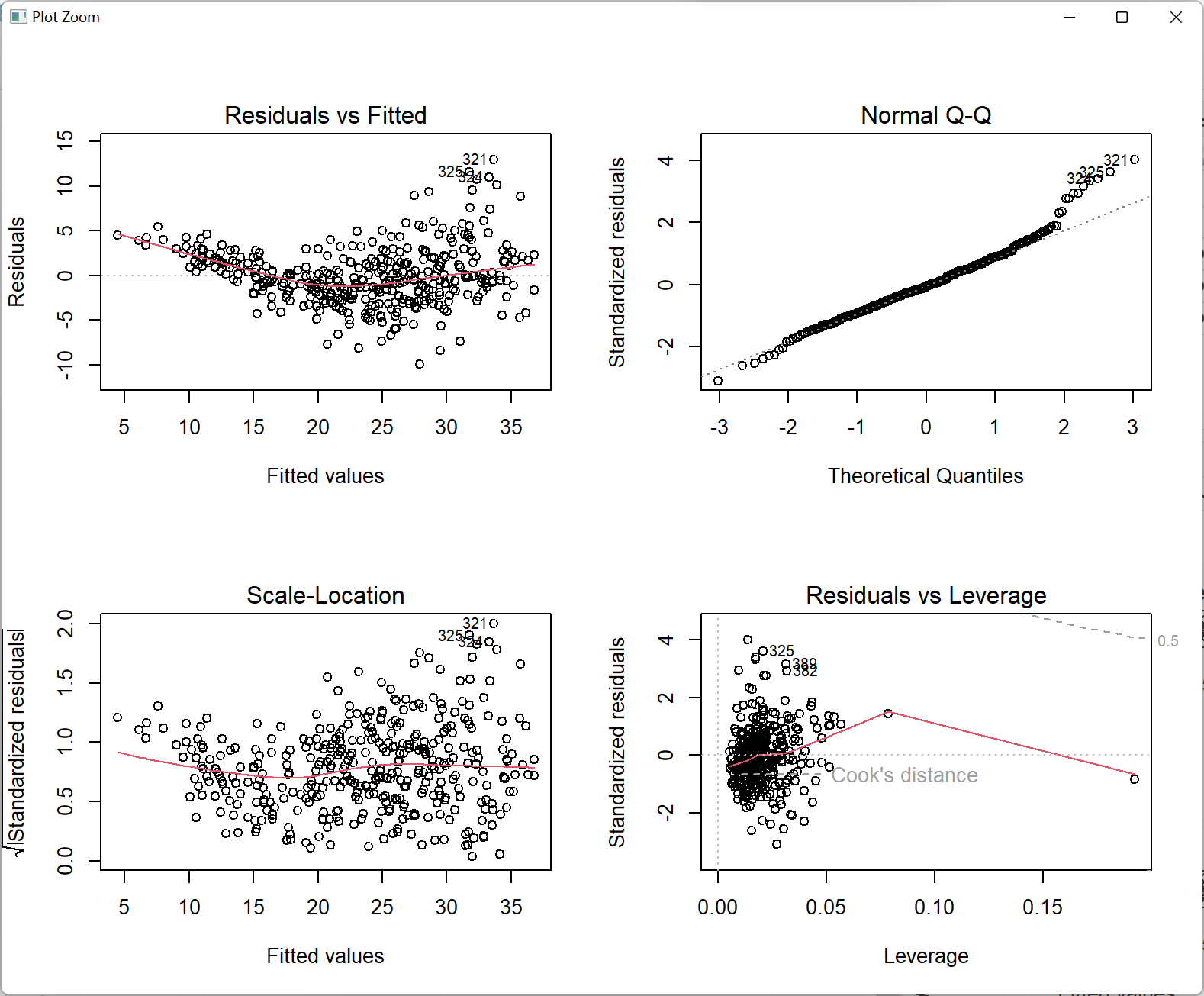
summary(lm.fit)

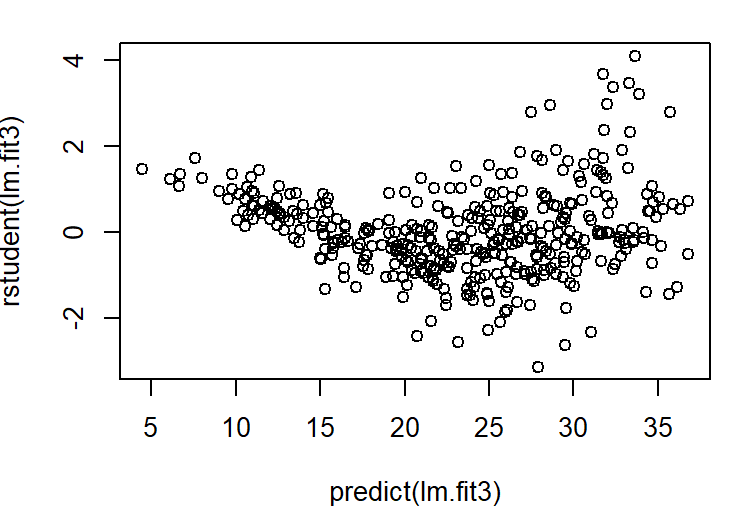
plot(lm.fit)

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

截图：







从结果看，各个变量的p值均很小，即在统计上有显著关系。

残差相对于此前更加稳定，方差依旧处于一个稳定的范围内，但仍有一部分残差较大，并且还是存在有异常高杠杆作用的点。

**10.问题（略）**

（a）问题（略）

脚本：

library(ISLR2)

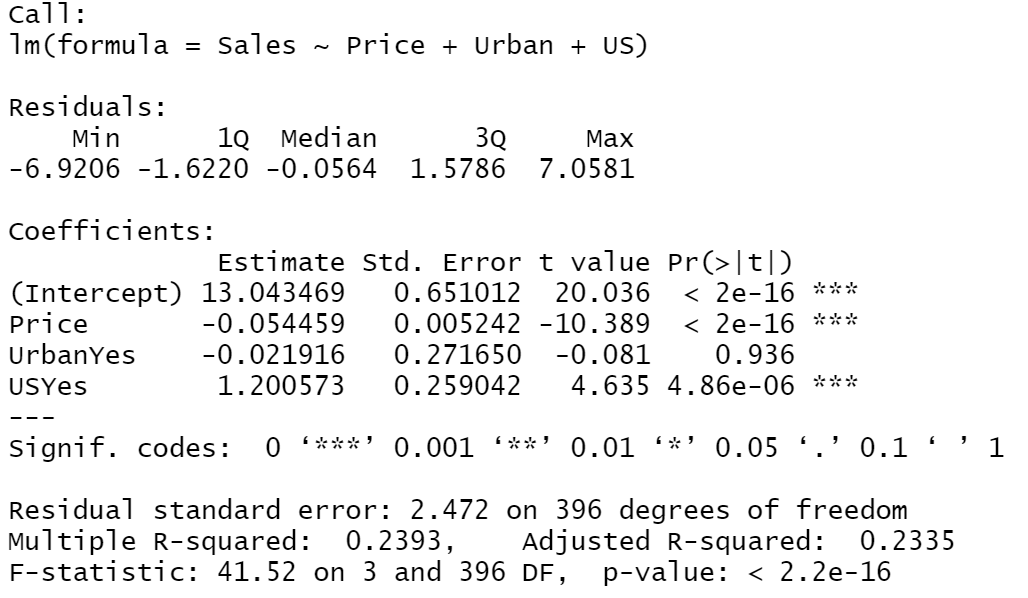
summary(Carseats)

attach(Carseats)

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

summary(lm.fit)

截图：



（b）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。

（c）问题（略）

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

（d）问题（略）

Price和USYes可以拒绝零假设，因为其p值很接近0。

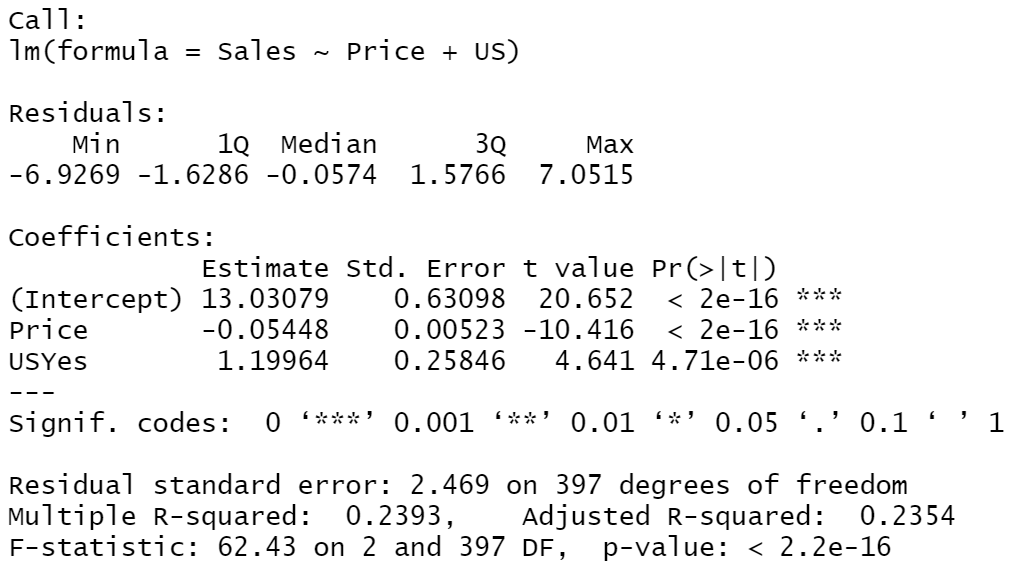
（e）问题（略）

脚本：

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

summary(lm.fit)

截图：



（f）问题（略）

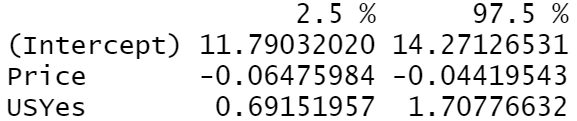
（e）中模型的RSE和R方均比（a）中略小一点，拟合效果差距不大，（e）中的模型对数据的拟合会比（a）略好一点。

（g）问题（略）

脚本：

confint(lm.fit2)

截图：

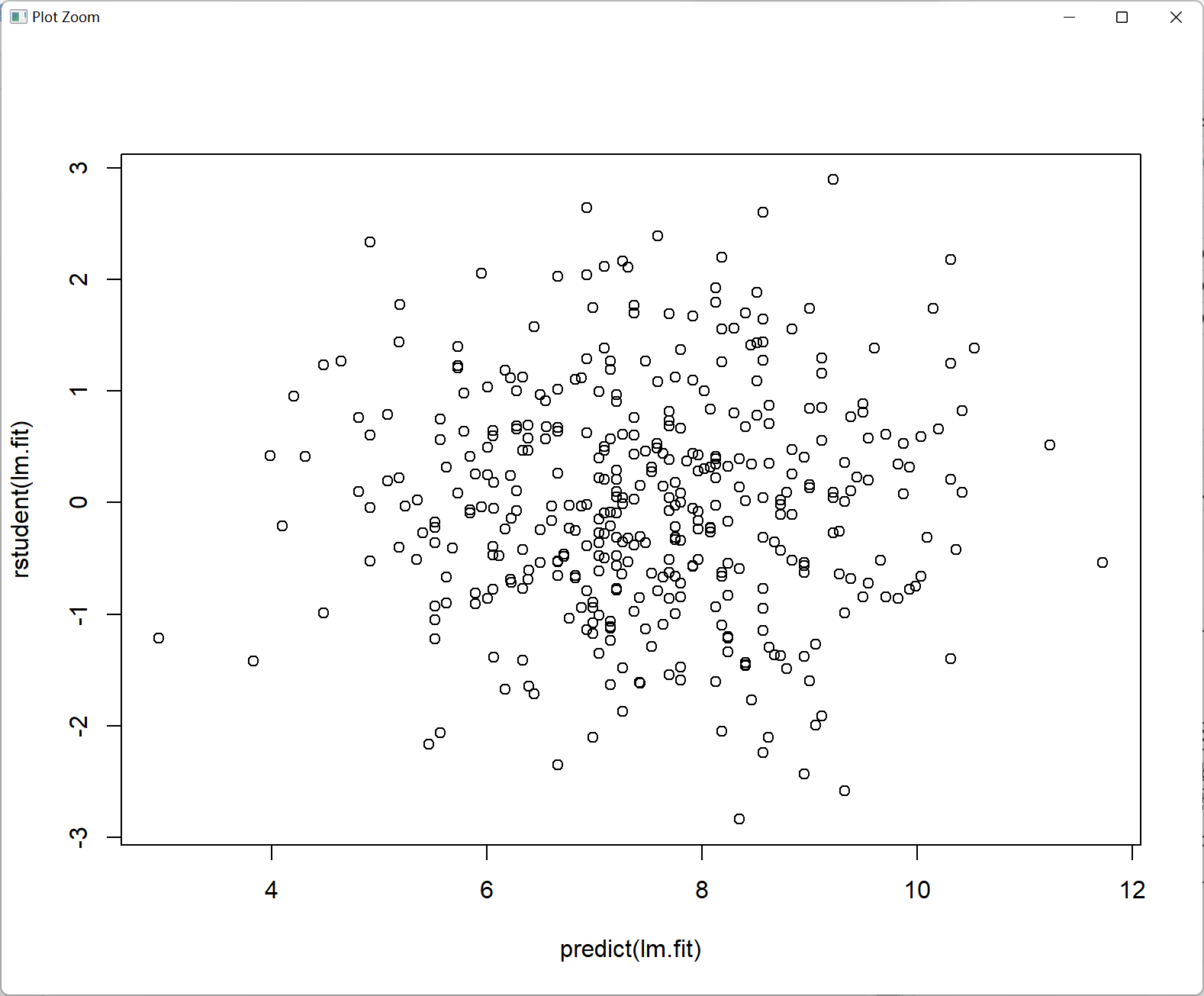


（h）问题（略）

脚本：

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

截图：



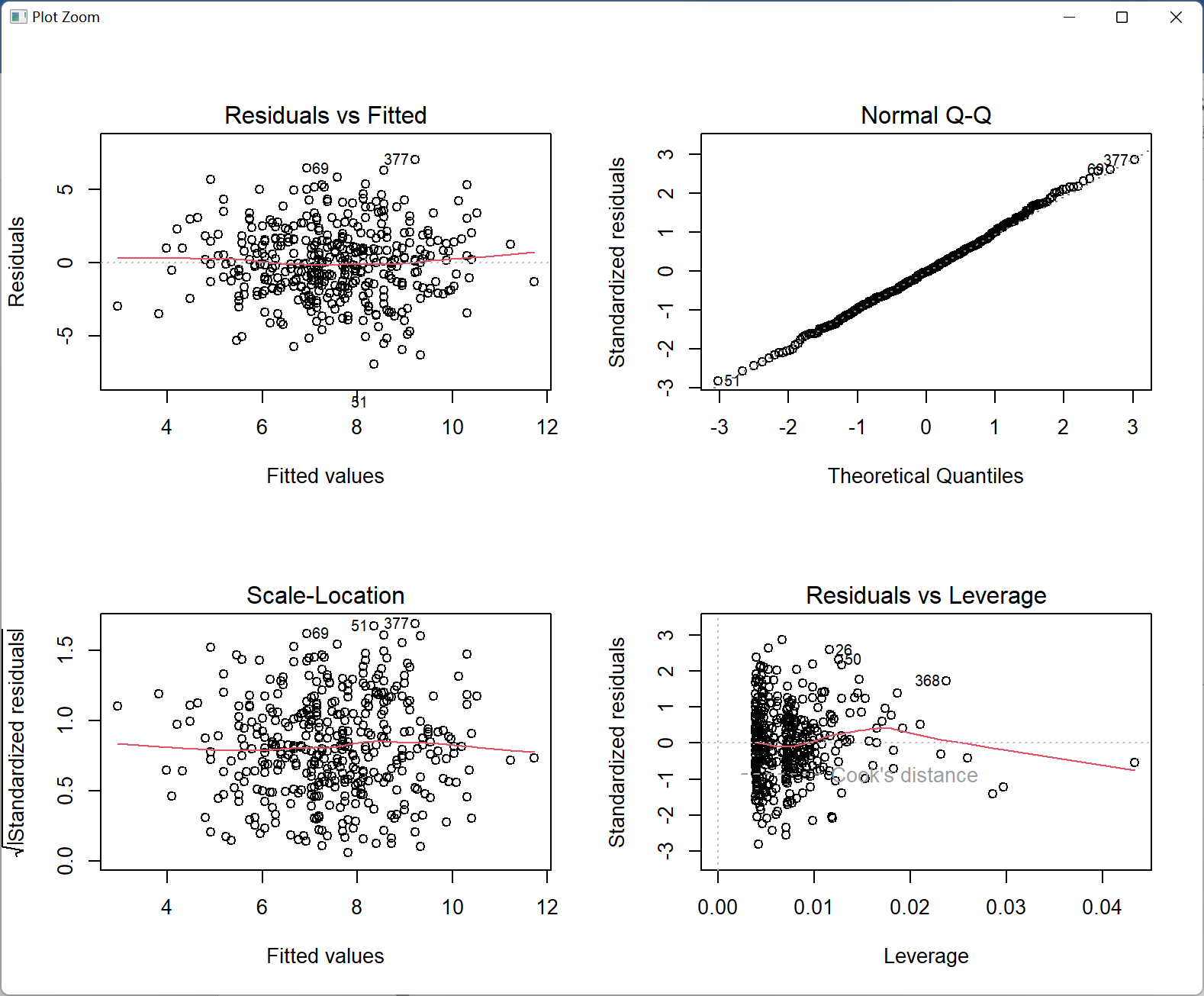
由图可知所有的残差都在(-3，3)的区间内，没有异常的离群点。

脚本：

par(mfrow=c(2,2))

plot(lm.fit)

截图：



有小部分散点具有高杠杆率。

**11.问题（略）**

（a）问题（略）

脚本：

set.seed(1)

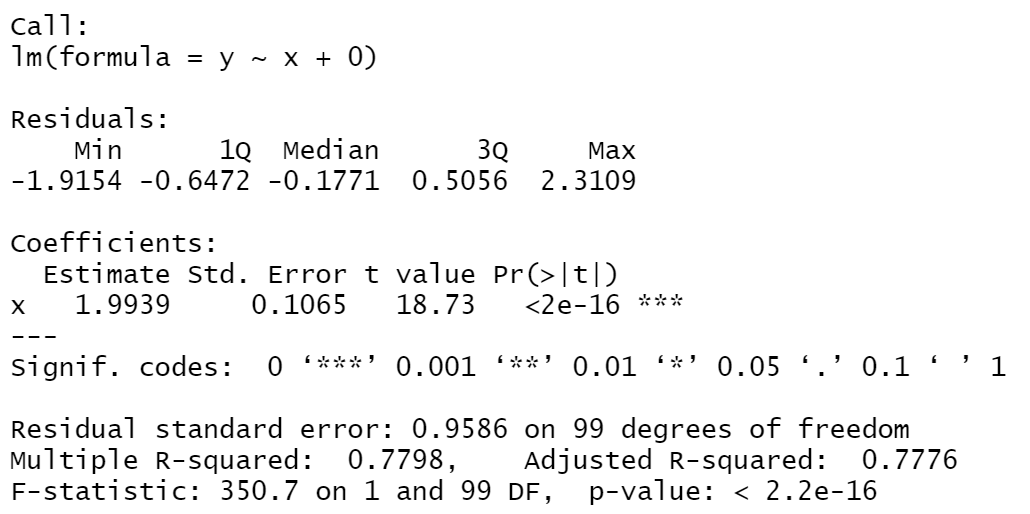
x=rnorm(100)

y=2\*x+rnorm(100)

lm.fit = lm(y~x+0)

summary(lm.fit)

截图：



估计系数：1.9939 标准差：0.1065 t统计量：18.73 p值：<2e-16

其p值很接近0，因此拒绝原假设。

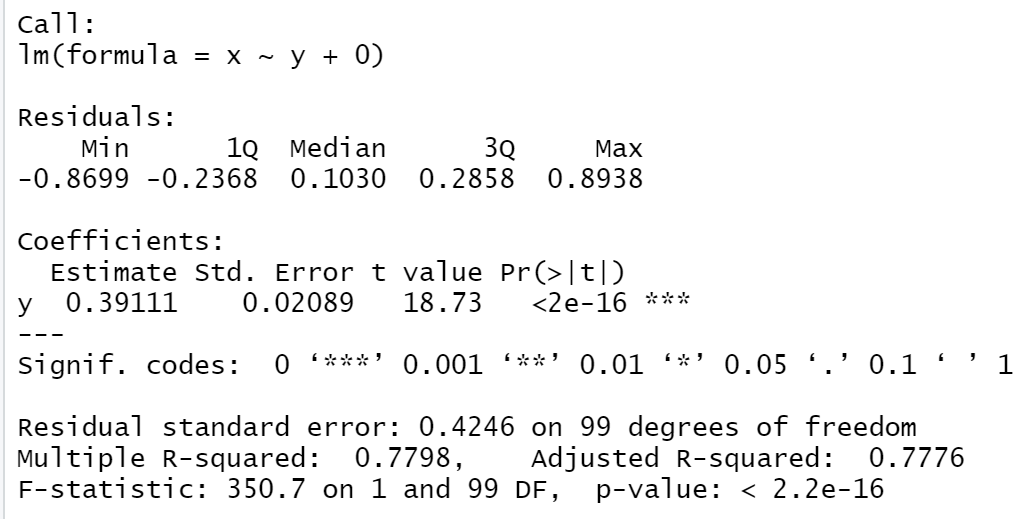
（b）问题（略）

脚本：

lm.fit = lm(x~y+0)

summary(lm.fit)

截图：



估计系数：0.39111 标准差：0.02089 t统计量：18.73 p值：<2e-16

其p值很接近0，因此拒绝原假设。

（c）问题（略）

（a），（b）都是对同一根线的拟合结果。二者有着相同的t统计量值和对应的p值，R方值，F统计量值和对应的p值。二者系数相乘也接近1.

（d）问题（略）

脚本：

(sqrt(length(x)-1) \* sum(x\*y)) / (sqrt(sum(x\*x) \* sum(y\*y) - (sum(x\*y))^2)

与上述所求t统计量值相等。

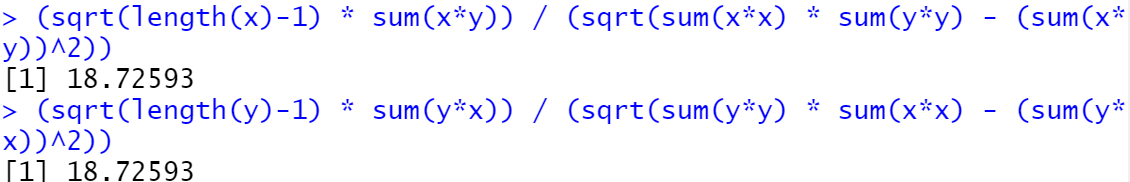
（e）问题（略）

将上述（d）命令中的x和y互换位置，发现没有影响，即：

(sqrt(length(y)-1) \* sum(y\*x)) / (sqrt(sum(y\*y) \* sum(x\*x) - (sum(y\*x))^2))

而其中length(x)=length(y)，所以t(y,x)=t(x,y)。

运行结果也是如此：



（f）问题（略）

脚本：

lm.fit = lm(y~x)

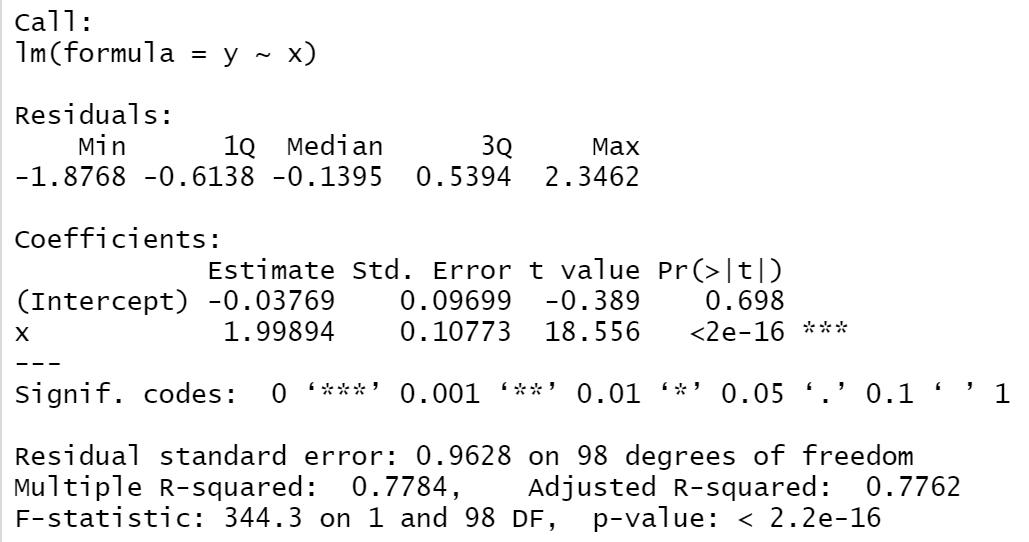
lm.fit2 = lm(x~y)

summary(lm.fit)

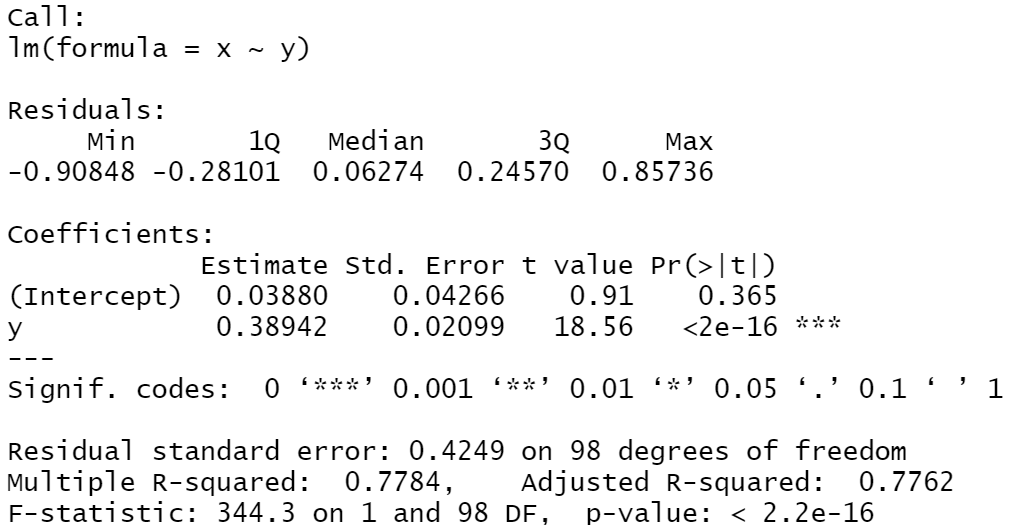
summary(lm.fit2)

截图：

y对x：



x对y：



可以看出二者的t统计量相同，均为18.56。

**13.问题（略）**

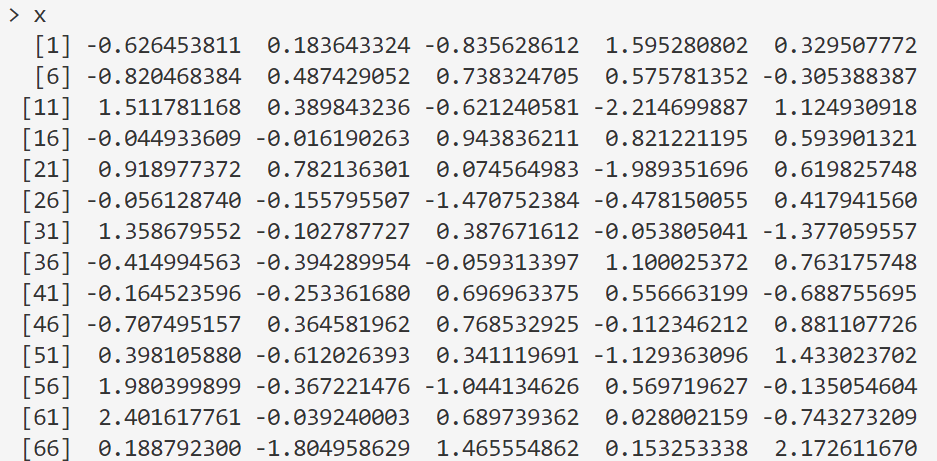
（a）问题（略）

脚本：

set.seed(1)

x = rnorm(100,0,1)

截图：

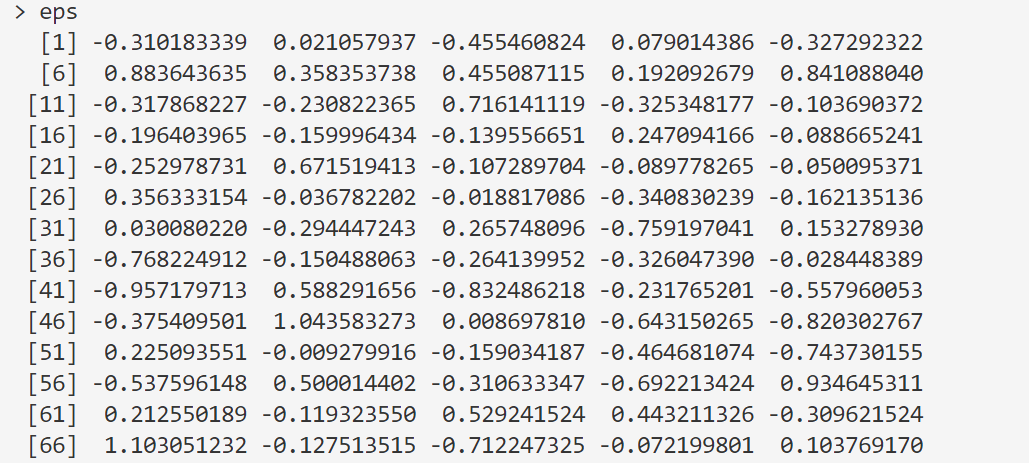


（b）问题（略）

脚本：

eps = rnorm(100, 0, sqrt(0.25))

截图：



（c）问题（略）

脚本：

y = -1 + 0.5\*x + eps

截图：



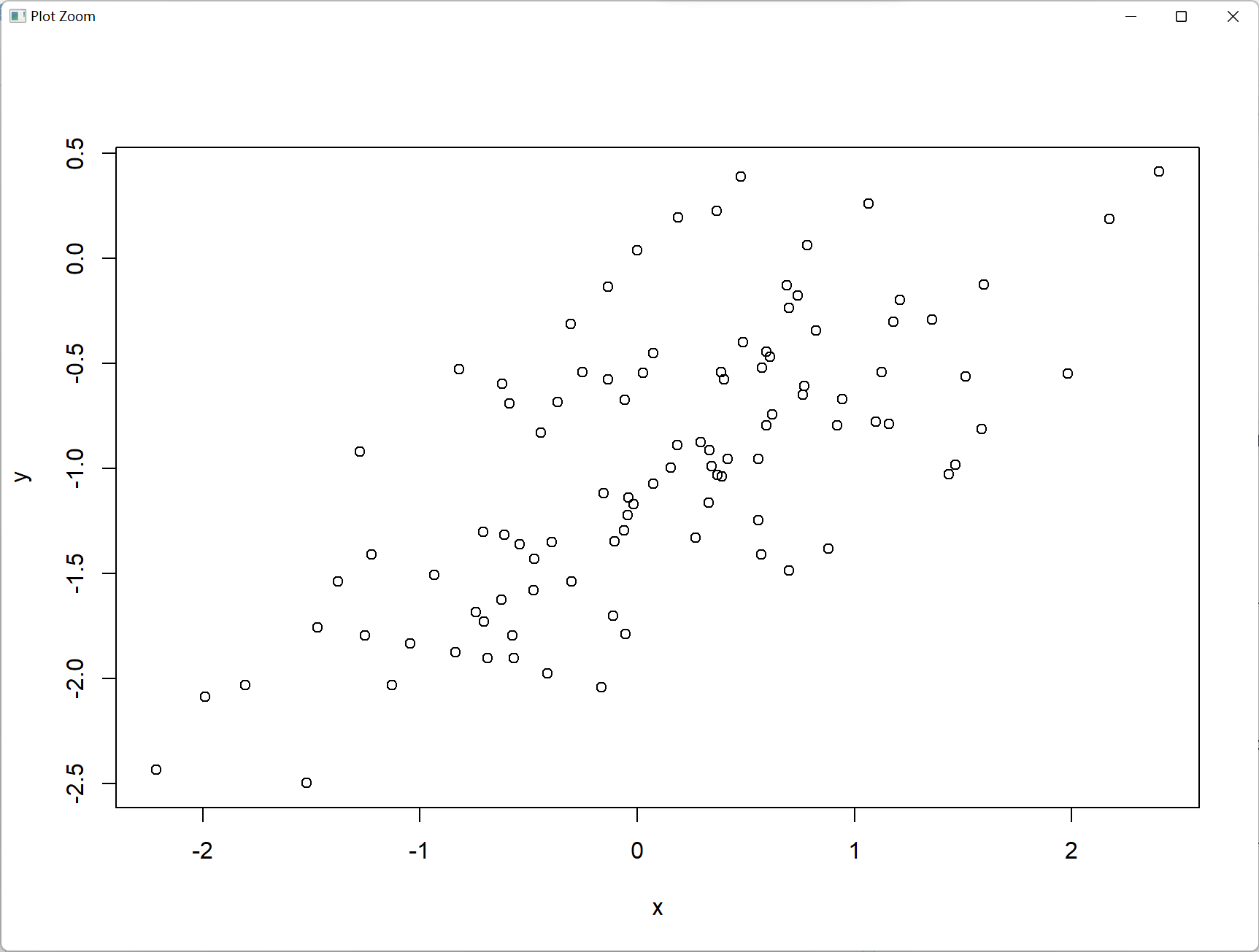
y的长度是100，β0是-1，β1是0.5。

（d）问题（略）

脚本：

plot(x, y)

截图：



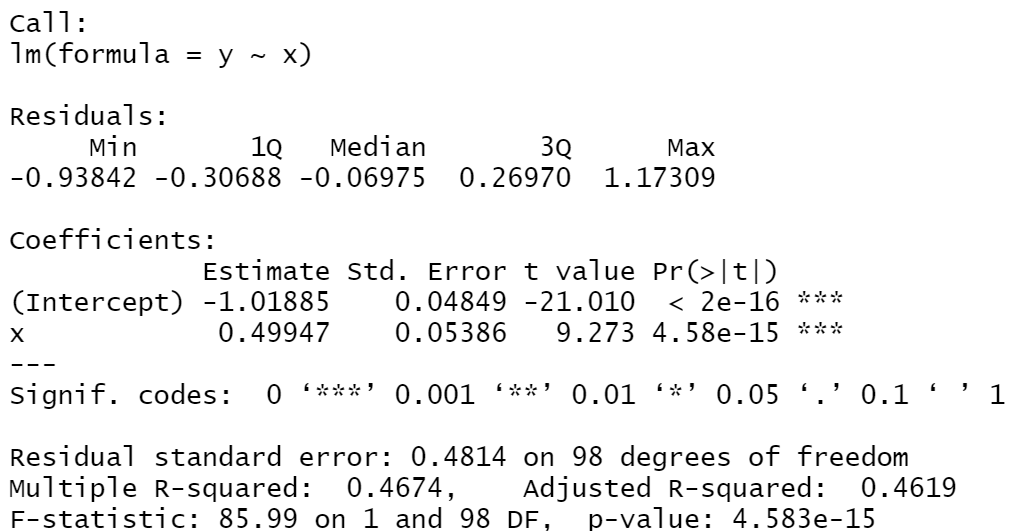
x和y有着较良好的线性关系，且为正相关。

（e）问题（略）

脚本：  
lm.fit = lm(y~x)

summary(lm.fit)

截图：



拟合系数为-1.01885和0.49947，并且F统计量远远大于1且p值很接近0，拟合效果具有显著性。

（f）问题（略）

脚本：

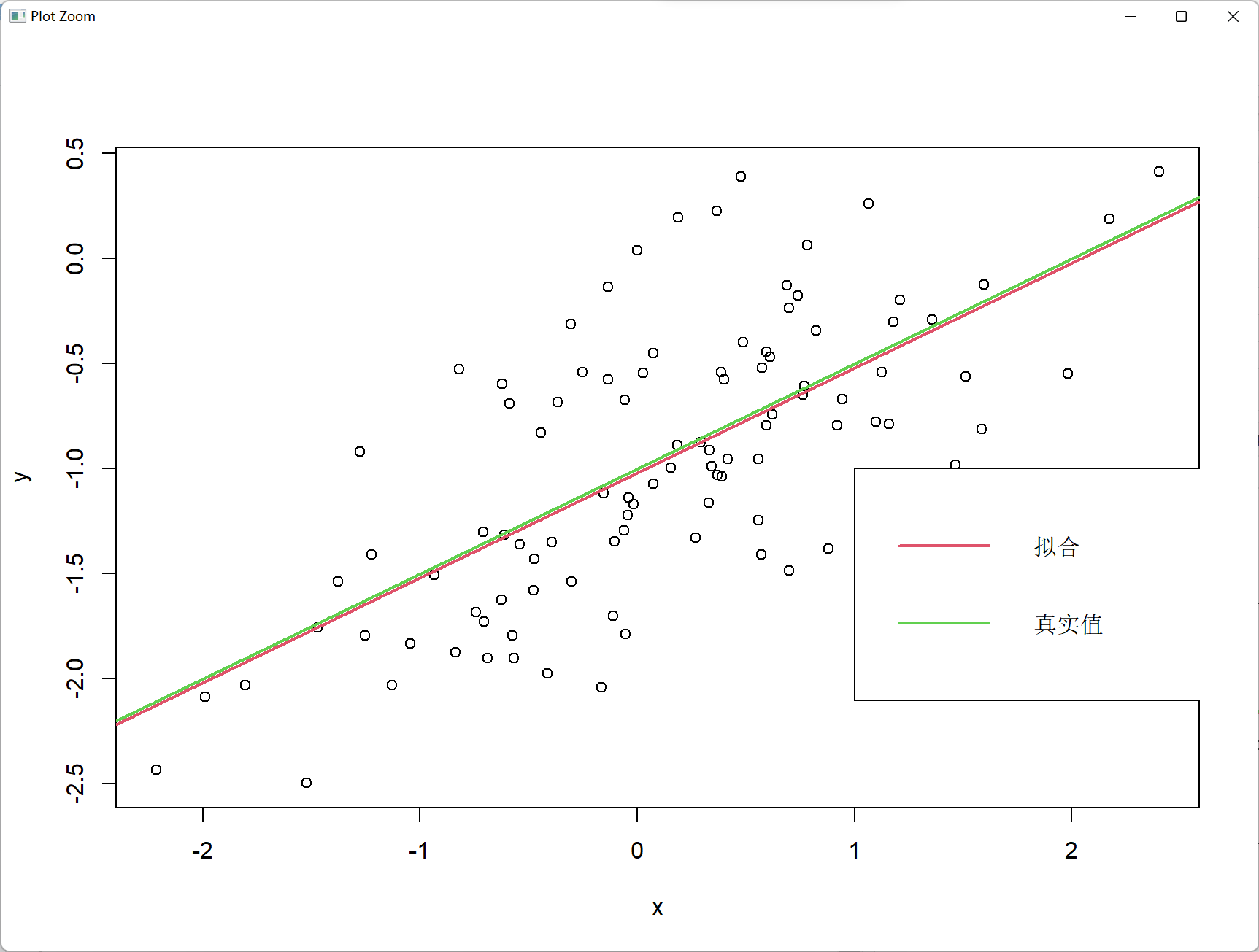
plot(x, y)

abline(lm.fit, lwd=2, col=2)

abline(-1, 0.5, lwd=2, col=3)

legend(-1, legend = c("拟合", "真实值"), col=2:3, lwd=2)

截图：



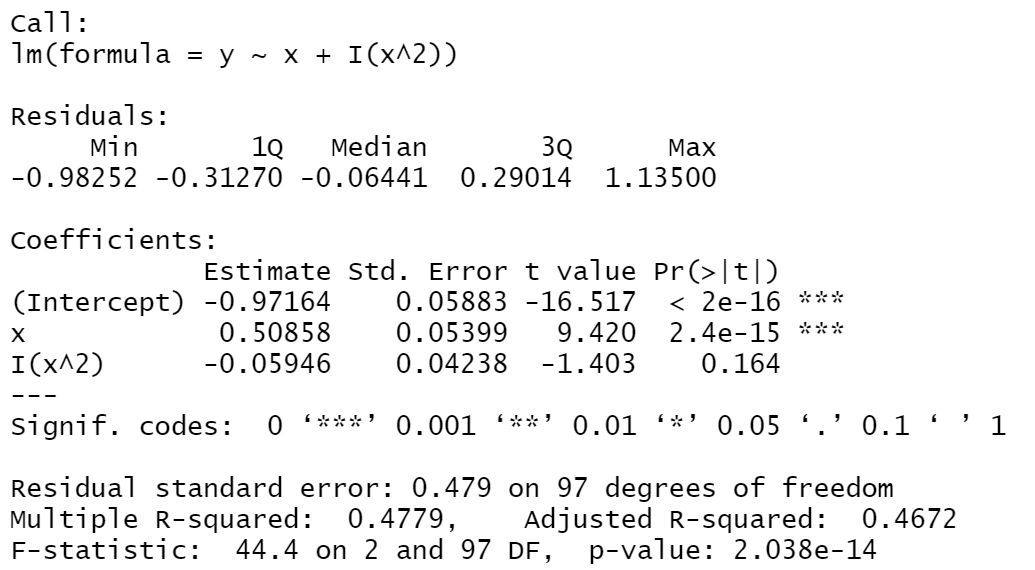
（g）问题（略）

脚本：

lm.fit2 = lm(y~x+I(x^2))

summary(lm.fit2)

截图：



拟合效果有所提升，RSE有所降低，R方有所提升。但x^2的p值说明其在统计上的显著性不好。

（h）问题（略）

脚本：

set.seed(1)

eps1 = rnorm(100, 0, 0.125)

x1 = rnorm(100)

y1 = -1 + 0.5\*x1 + eps1

plot(x1, y1)

lm.fit1 = lm(y1~x1)

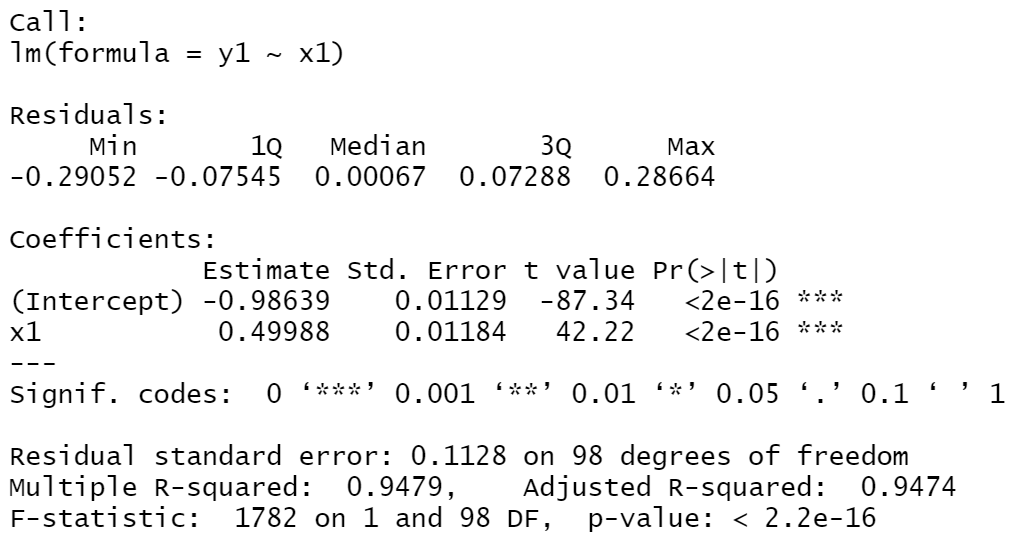
summary(lm.fit1)

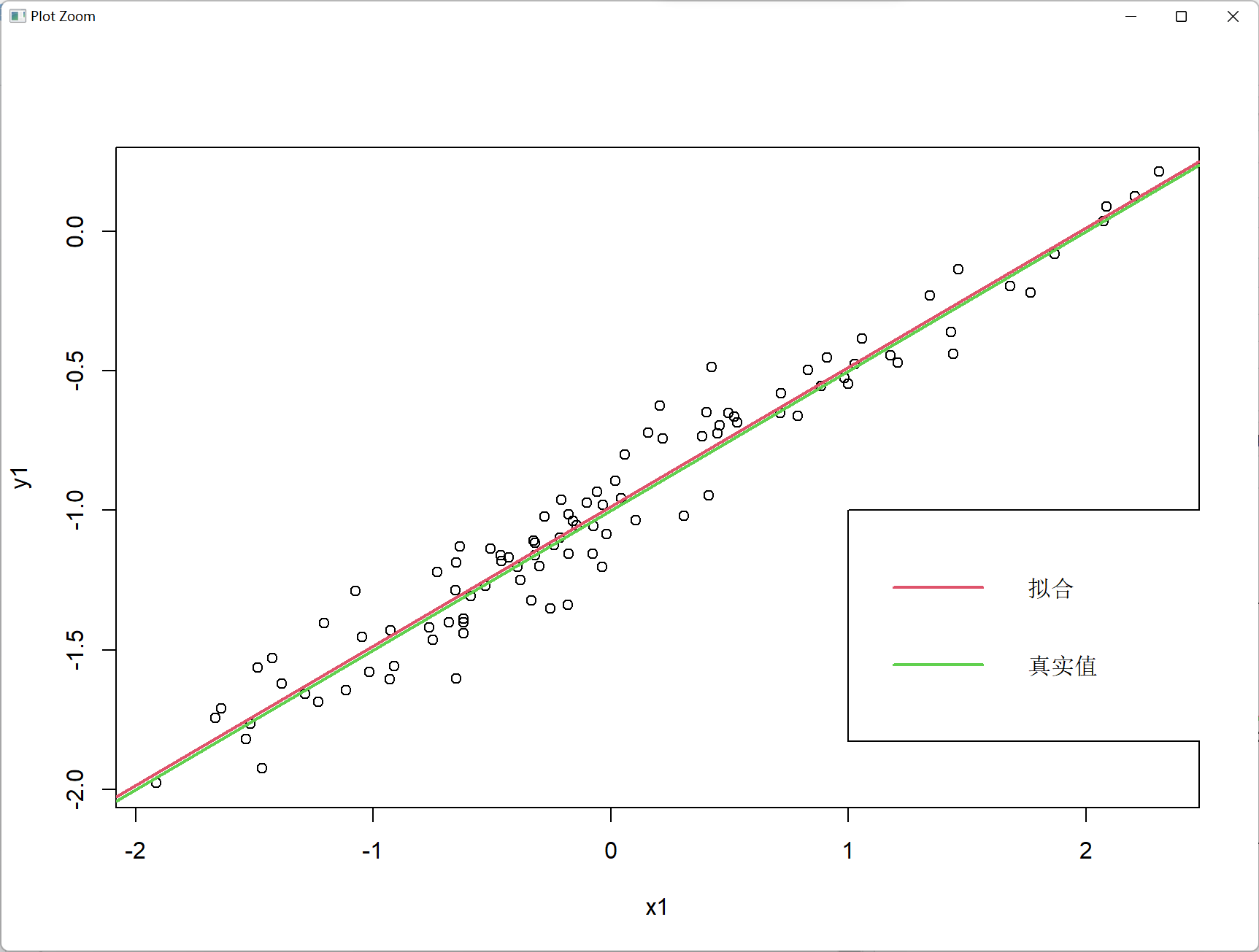
abline(lm.fit1, lwd=2, col=2)

abline(-1, 0.5, lwd=2, col=3)

legend(-1, legend = c("拟合", "真实值"), col=2:3, lwd=2)

截图：





从RSE显著减小和R方显著增大来看，成功减少了数据里的噪声。

（i）问题（略）

脚本：

set.seed(1)

eps2 = rnorm(100, 0, 1)

x2 = rnorm(100)

y2 = -1 + 0.5\*x2 + eps2

plot(x2, y2)

lm.fit2 = lm(y2~x2)

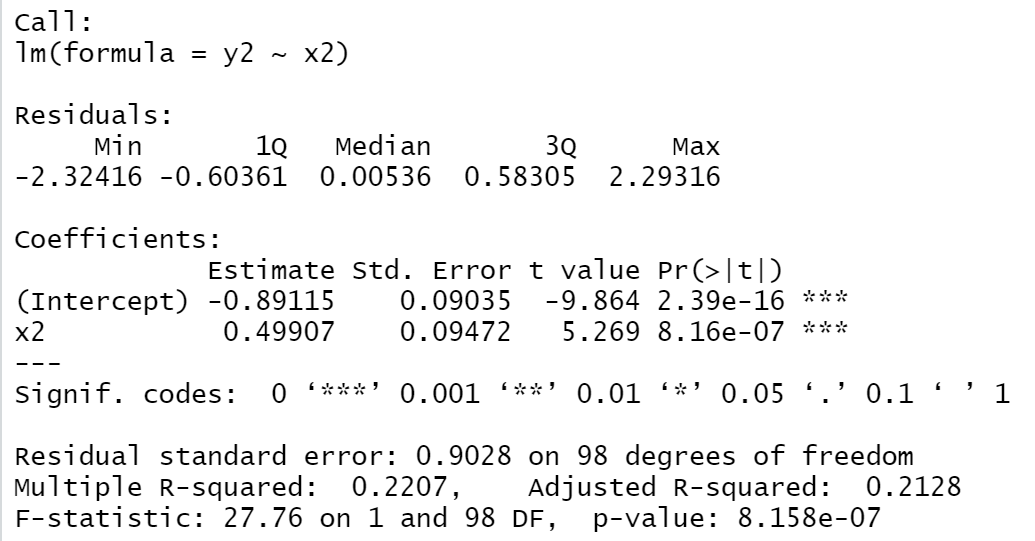
summary(lm.fit2)

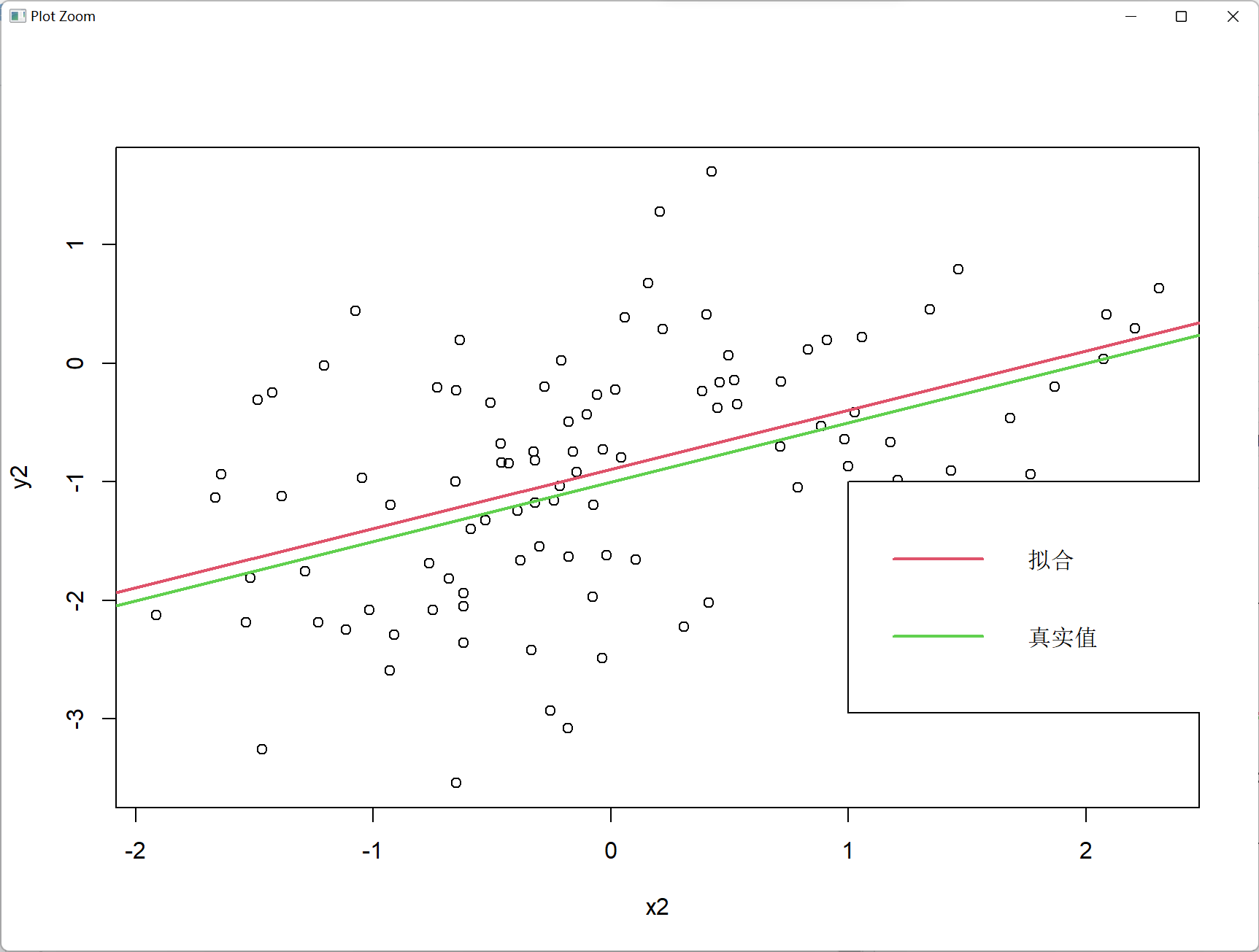
abline(lm.fit2, lwd=2, col=2)

abline(-1, 0.5, lwd=2, col=3)

legend(-1, legend = c("拟合", "真实值"), col=2:3, lwd=2)

截图：





从RSE显著增大和R方显著减小来看，成功增大了数据里的噪声。

（j）问题（略）

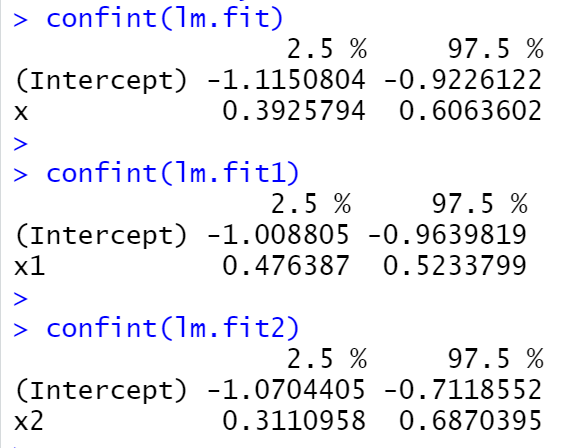
脚本：

confint(lm.fit)

confint(lm.fit1)

confint(lm.fit2)

截图：



所有的区间似乎都以大约0.5为中心，第二次拟合的区间比第一次拟合的区间窄，最后一次拟合的区间比第一次拟合的区间宽。

**14.问题（略）**

（a）问题（略）

脚本：

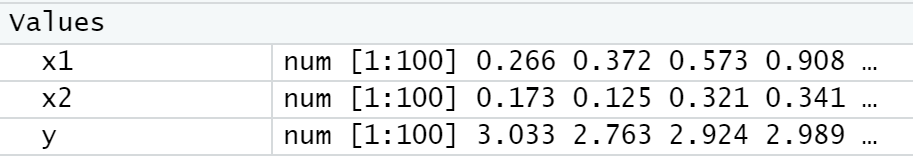
set.seed(1)

x1 = runif(100)

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

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

截图：



函数形式是：y = 2 + 2\*x1 + 0.3\*x2 +eps

回归系数：β0=2，β1=2，β2=0.3

（b）问题（略）

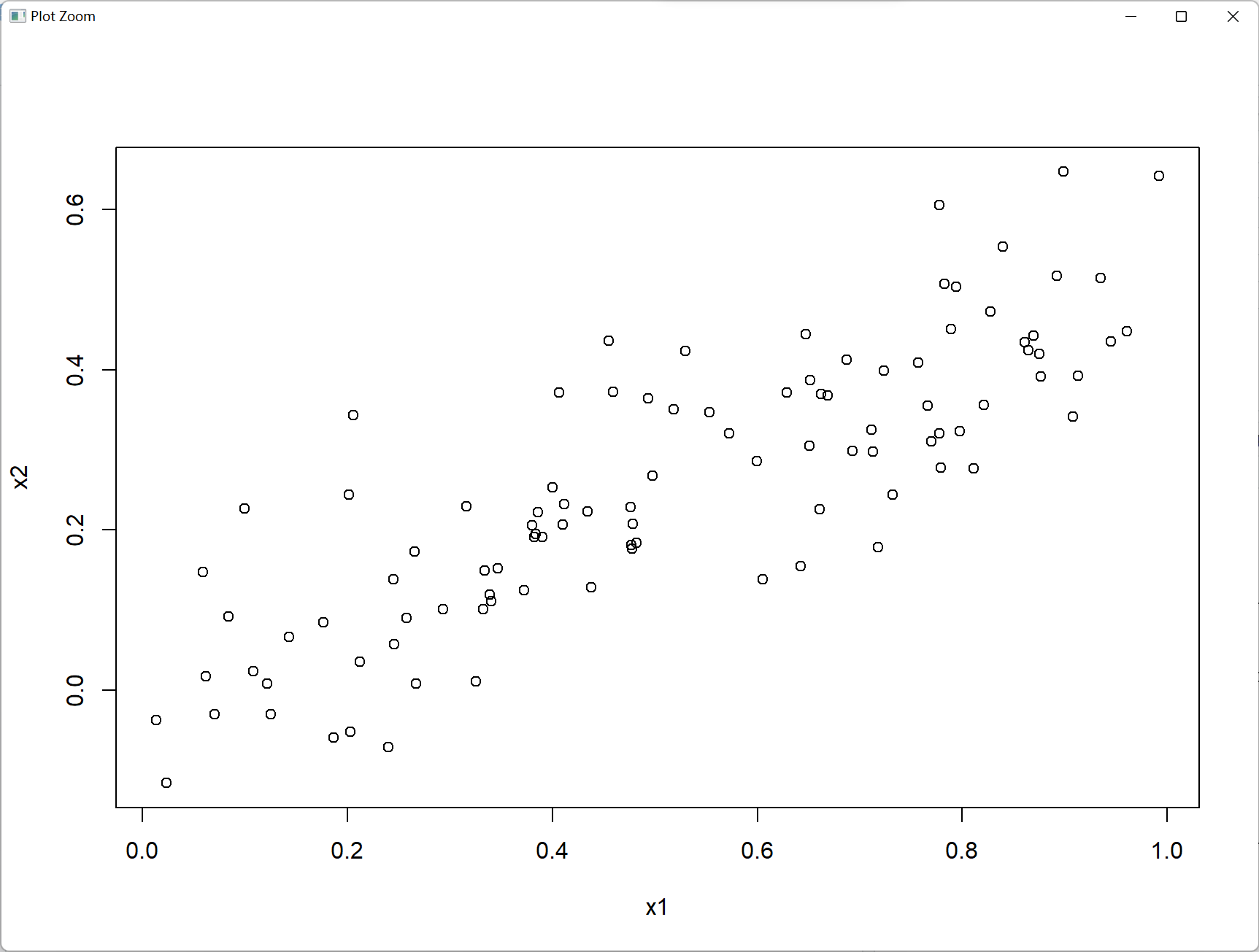
脚本：

cor(x1, x2)

plot(x1, x2)

截图：



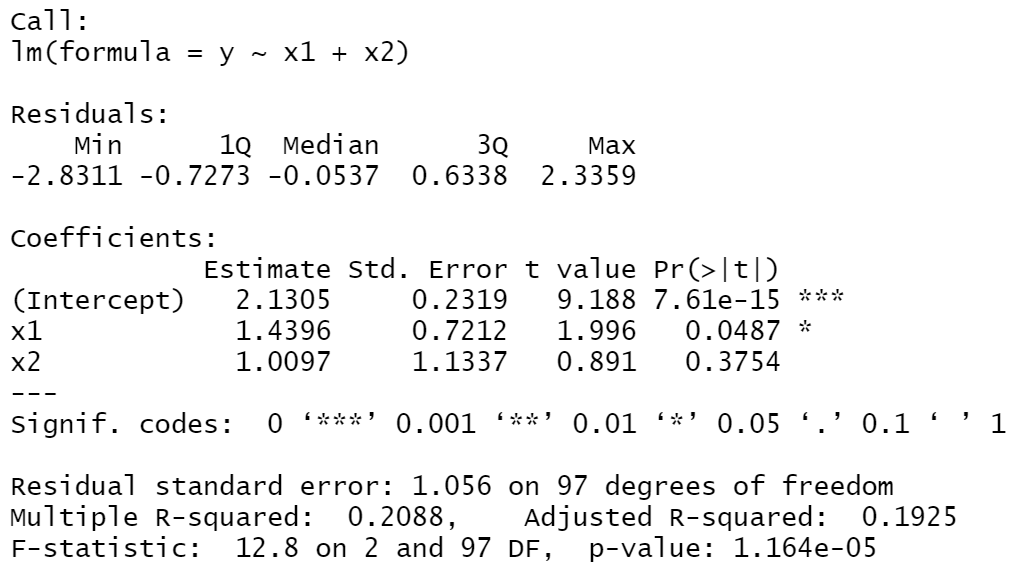


（c）问题（略）

脚本：

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

summary(lm.fit)

截图：  


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。

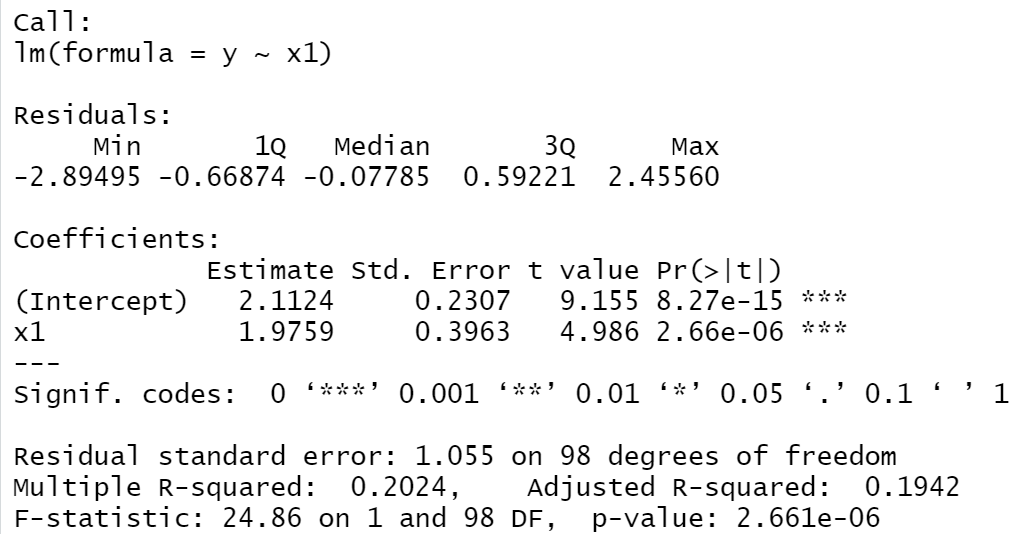
（d）问题（略）

脚本：

lm.fit = lm(y~x1)

summary(lm.fit)

截图：



由于p值很接近0，能拒绝零假设。

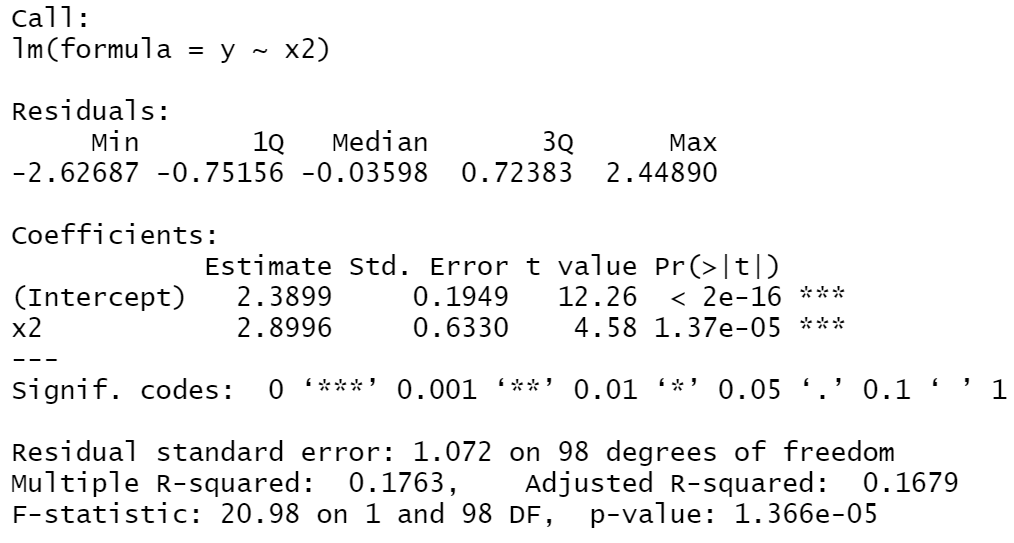
（e）问题（略）

脚本：

lm.fit = lm(y~x2)

summary(lm.fit)

截图：



由于p值很接近0，能拒绝零假设。

（f）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。

（g）问题（略）

脚本：

x1 = c(x1, 0.1)

x2 = c(x2, 0.8)

y = c(y, 6)

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

summary(lm.fit1)

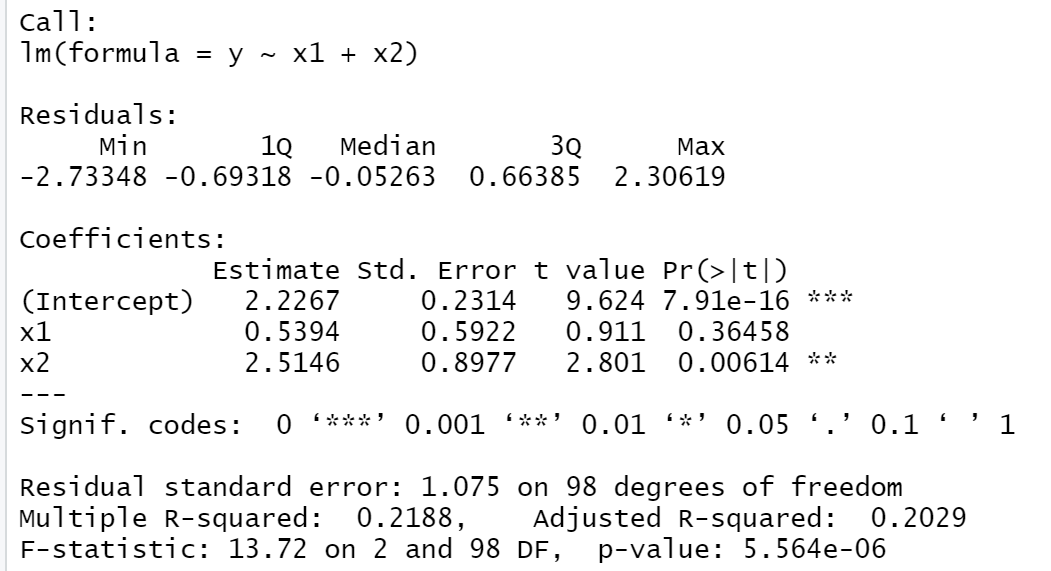
lm.fit2 = lm(y~x1)

summary(lm.fit2)

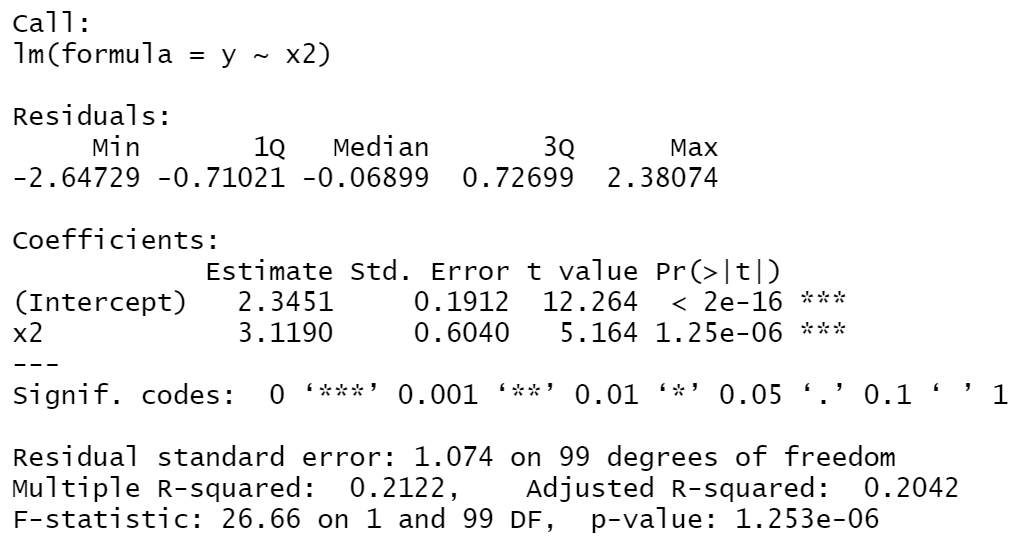
lm.fit3 = lm(y~x2)

summary(lm.fit3)

截图：







在x1和x2的回归中，p值告诉我们，x1不显著，x2显著。

脚本：

par(mfrow=c(2,2))

plot(lm.fit1)

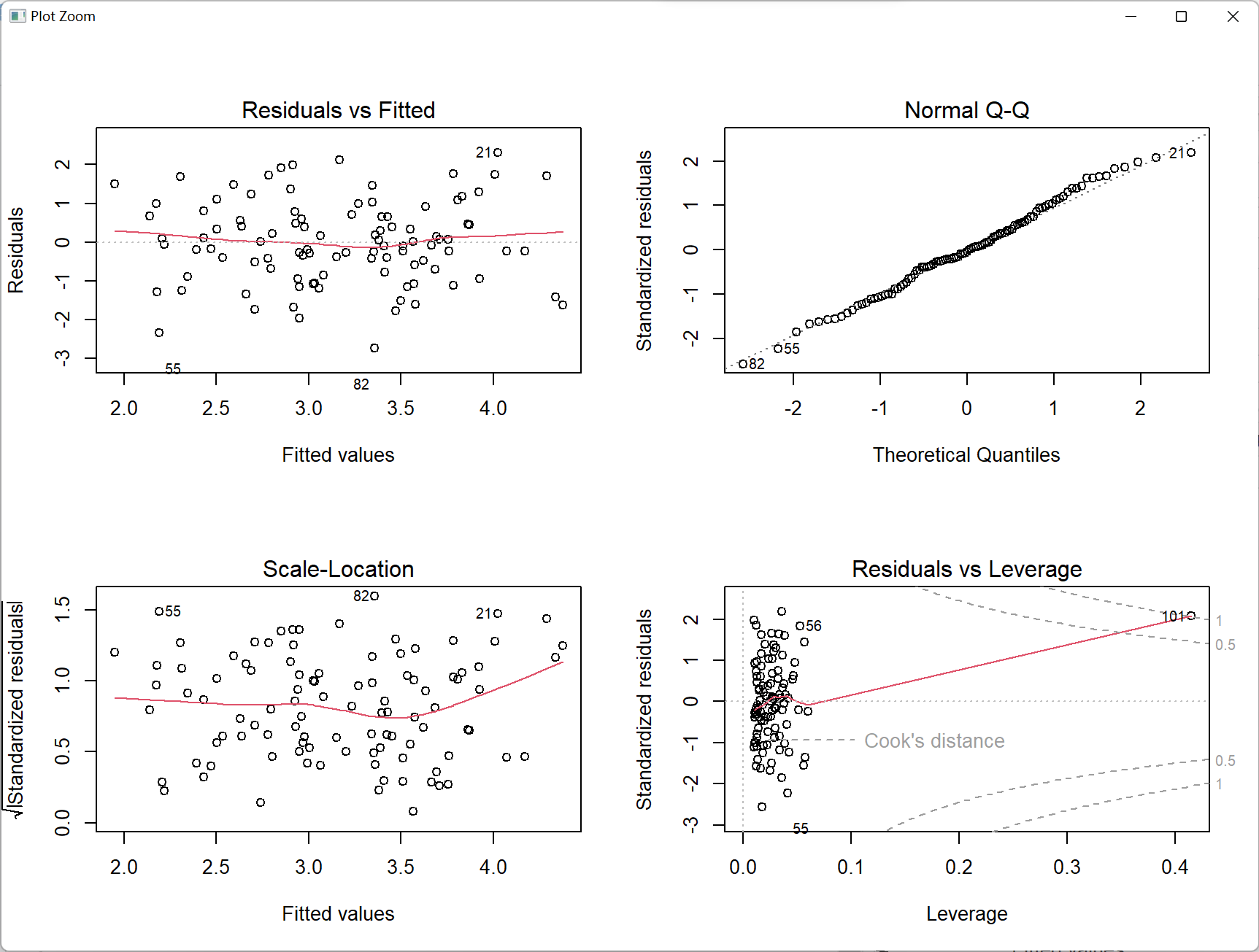
par(mfrow=c(2,2))

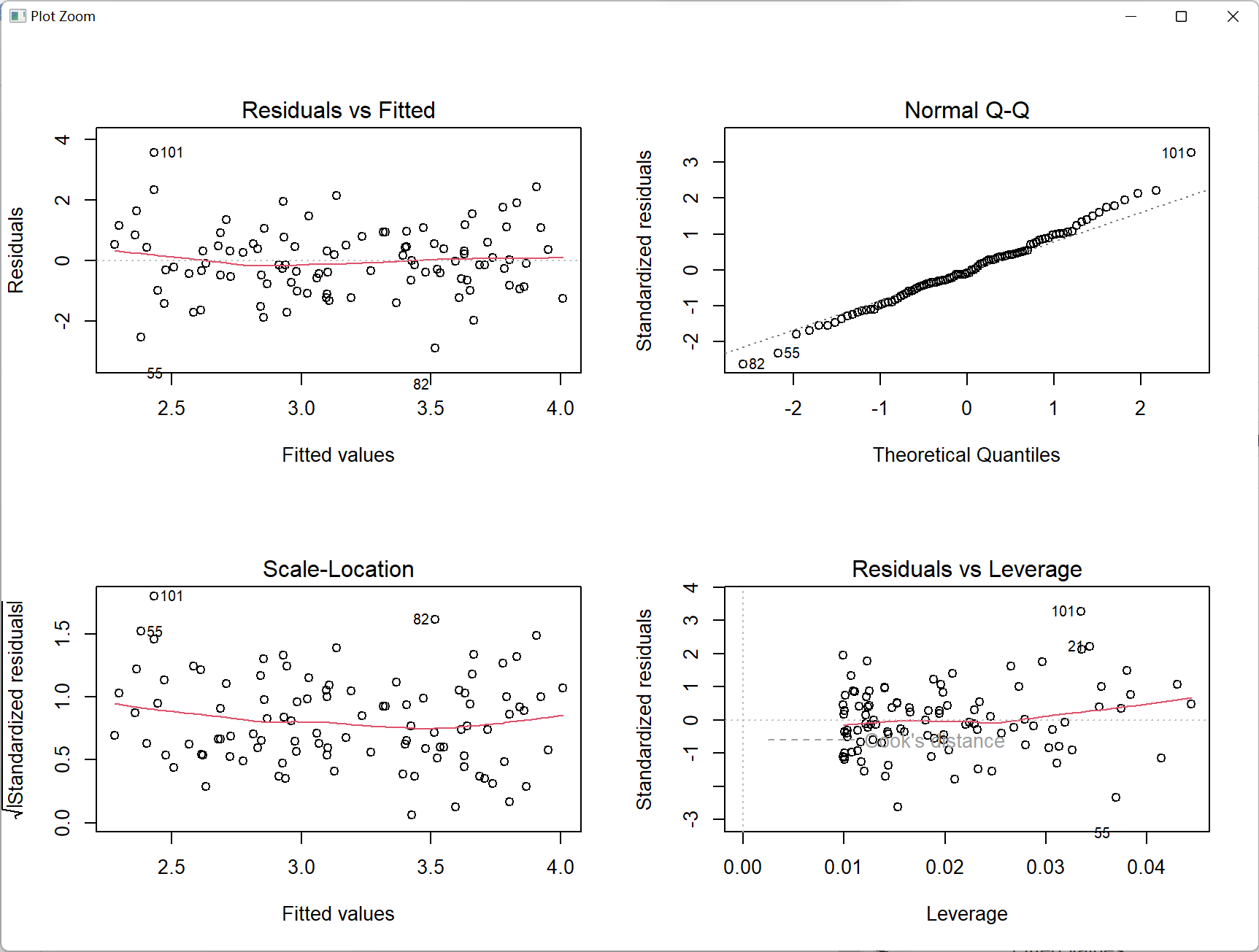
plot(lm.fit2)

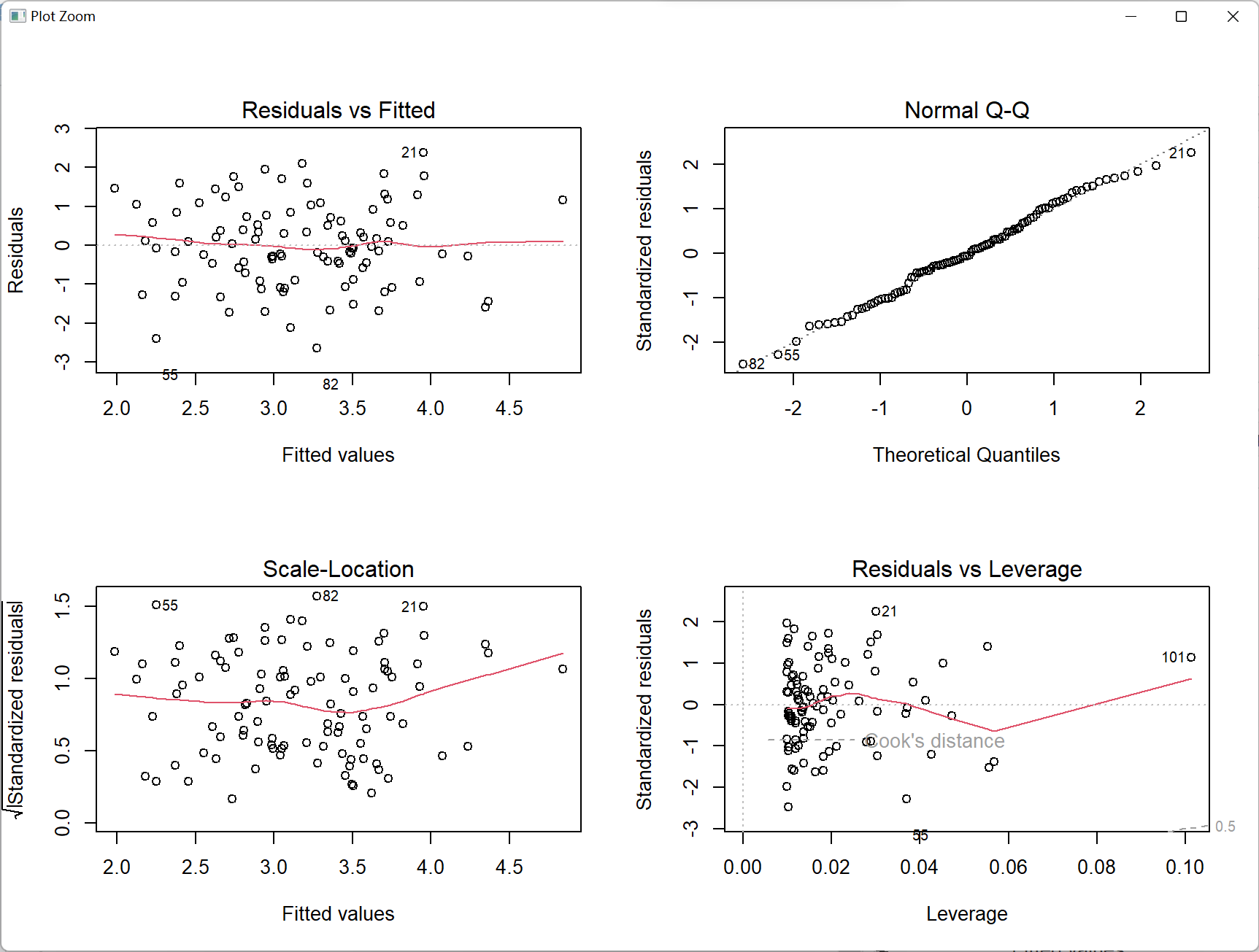
par(mfrow=c(2,2))

plot(lm.fit3)

截图：







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。

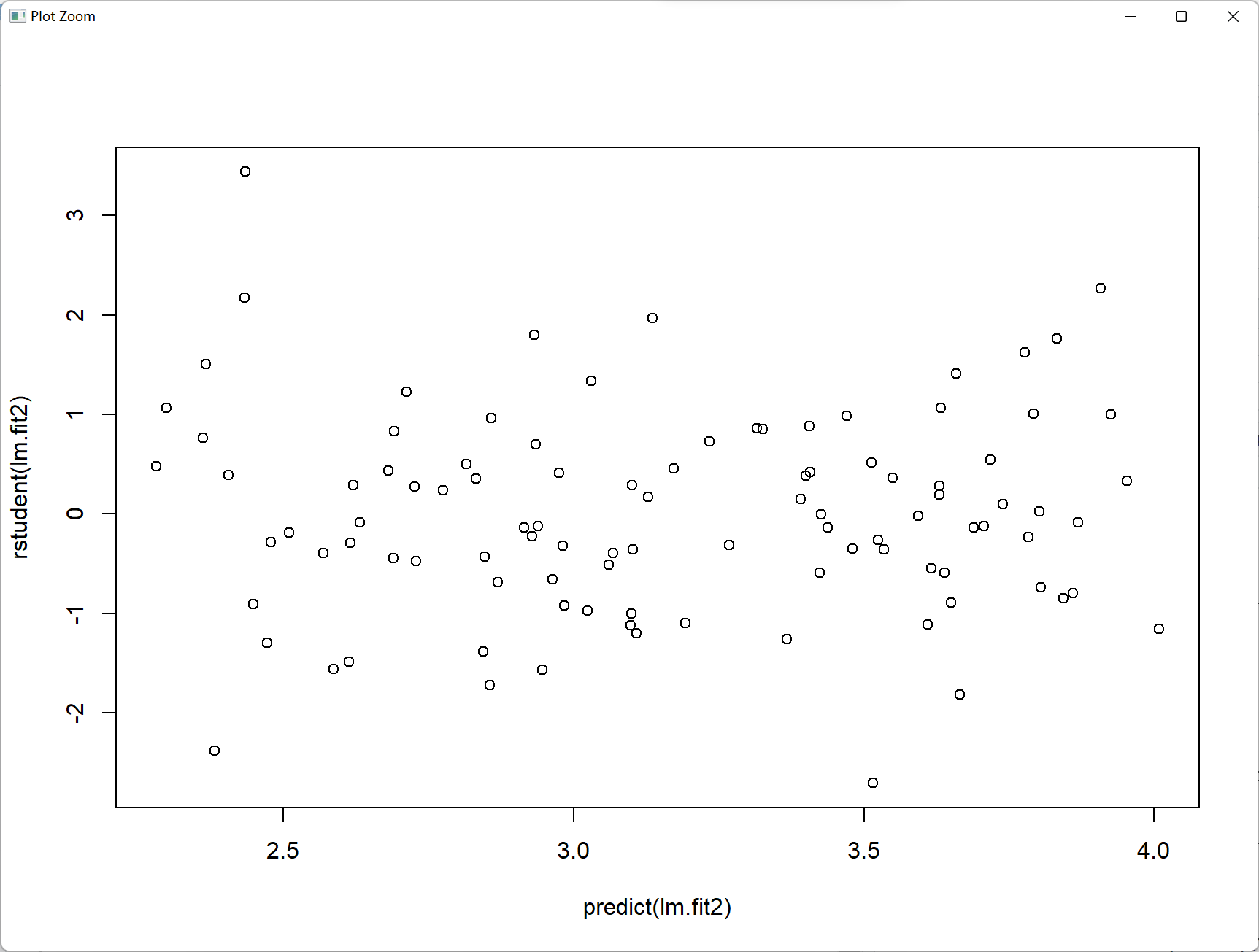
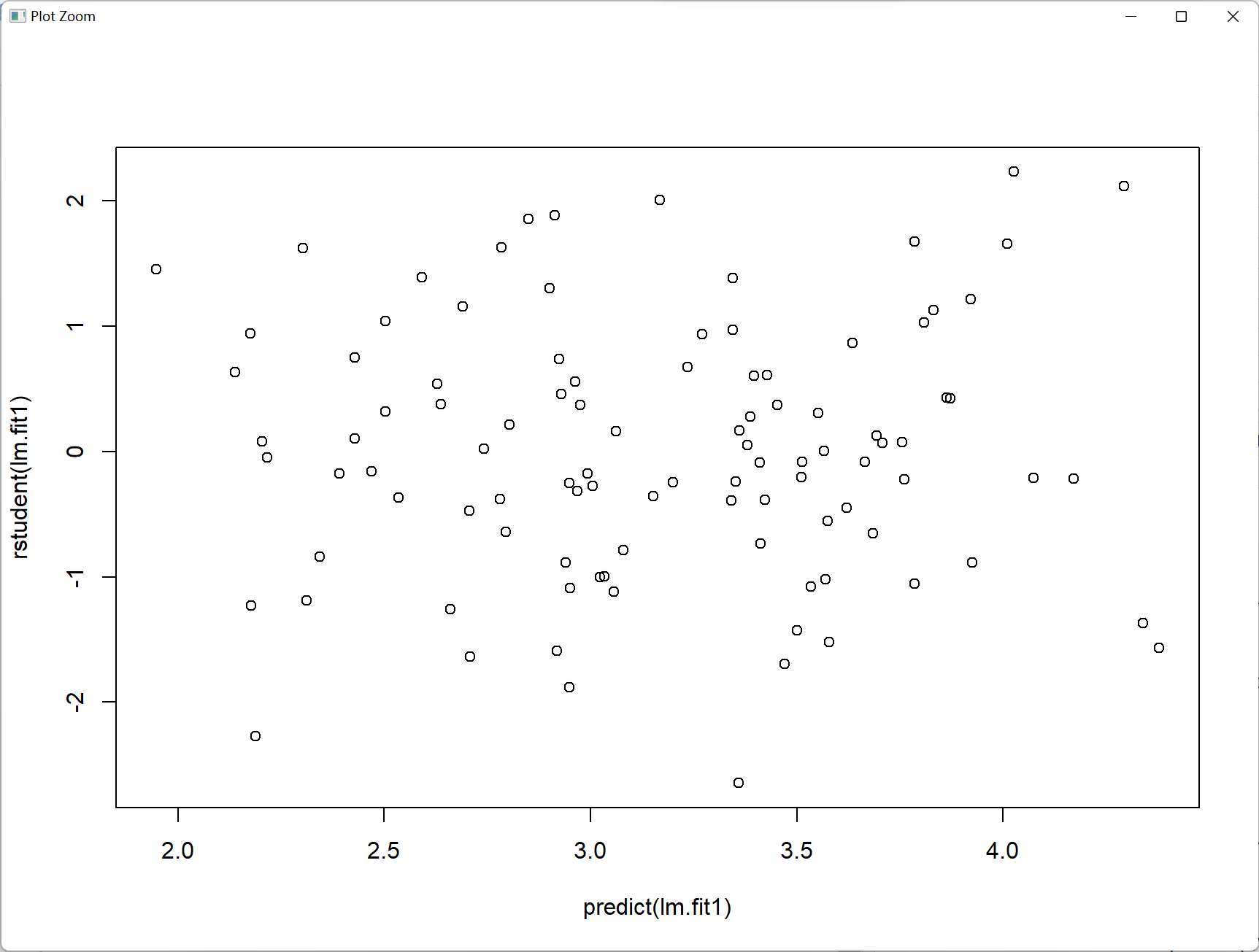
脚本：

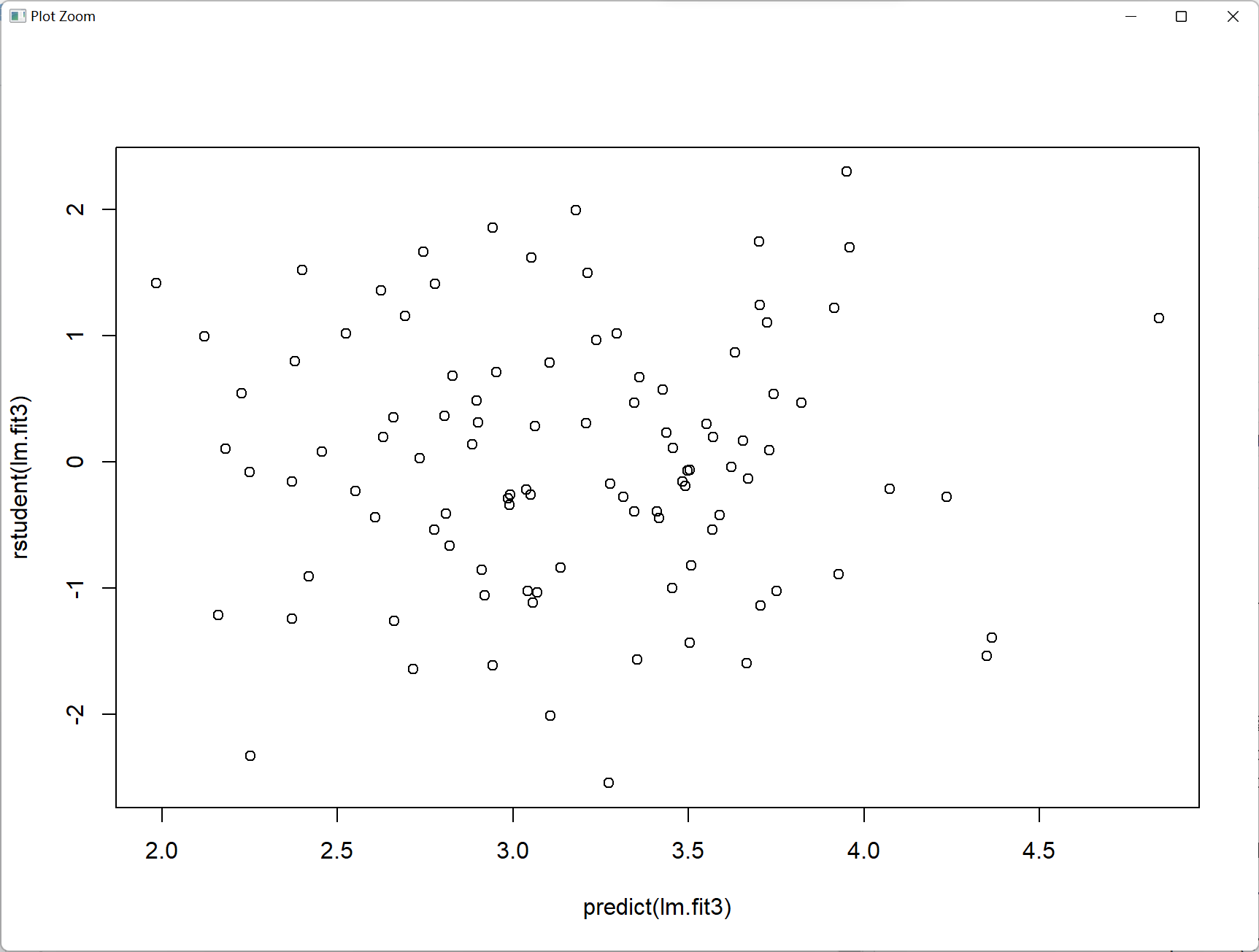
plot(predict(lm.fit1), rstudent(lm.fit1))

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

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

截图：





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。