> h.1 <- read.csv("C:/Users/Linna\_hu/Desktop/group project/variables.csv")

> View(h.1)

> budget = h.1[,1]

> company = h.1[,2]

> quarter=h.1[,3]

> revenue=h.1[,4]

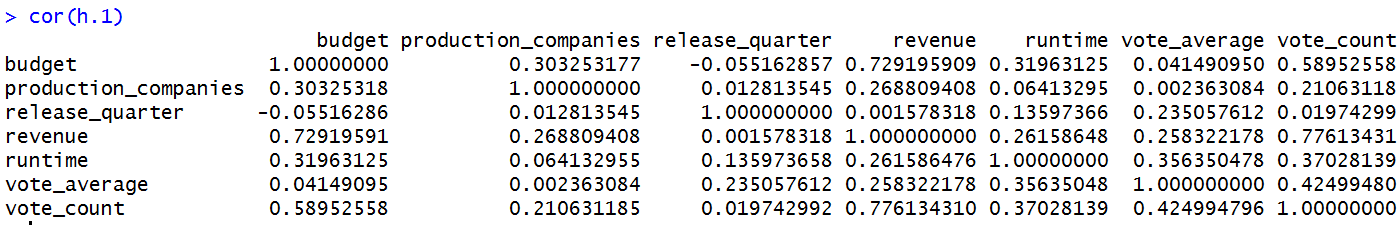
> runtime=h.1[,5]

> vote\_average=h.1[,6]

> vote\_count=h.1[,7]

> x3 =factor(quarter)

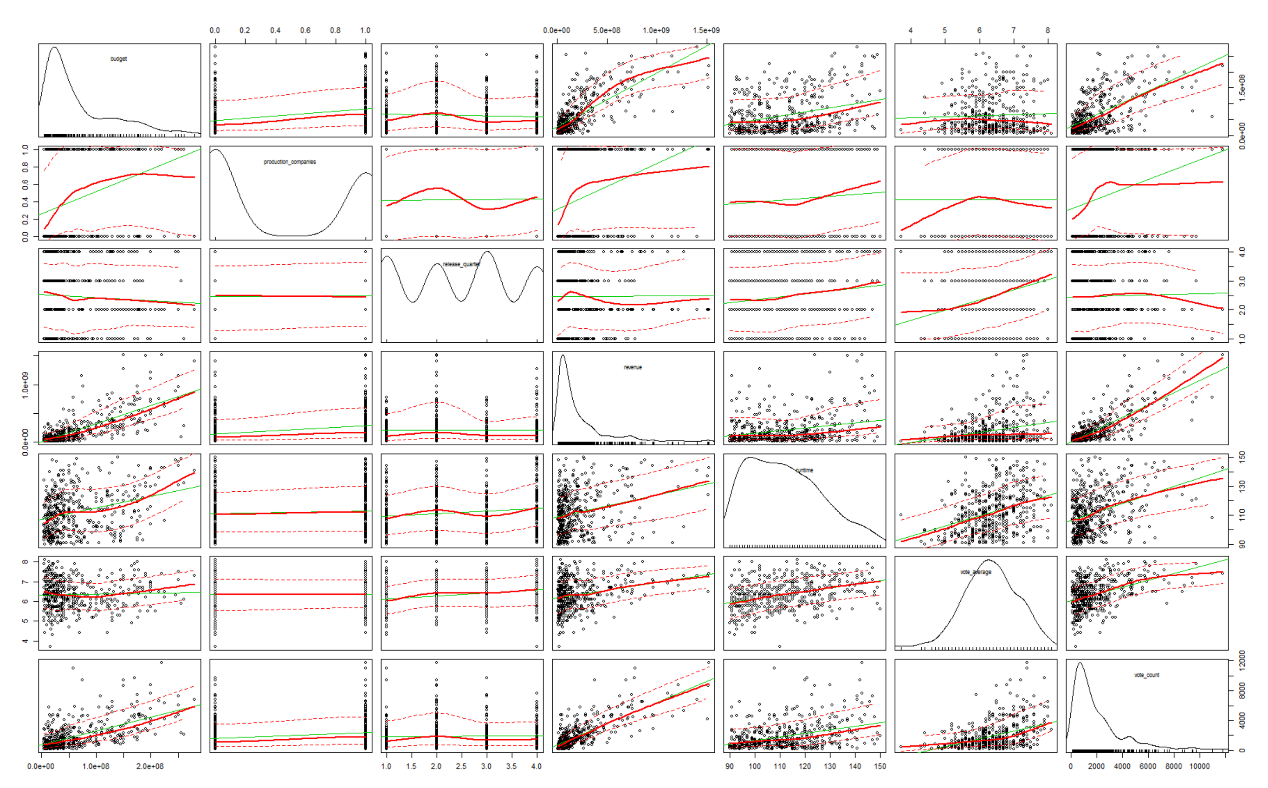
变量选择：



将关系可视化：

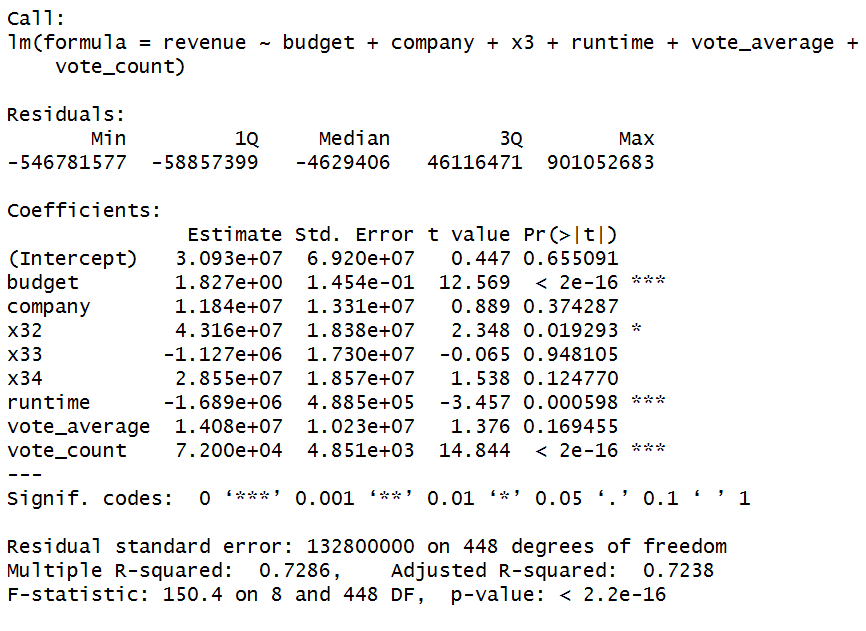
> library(car)

> scatterplotMatrix(h.1)



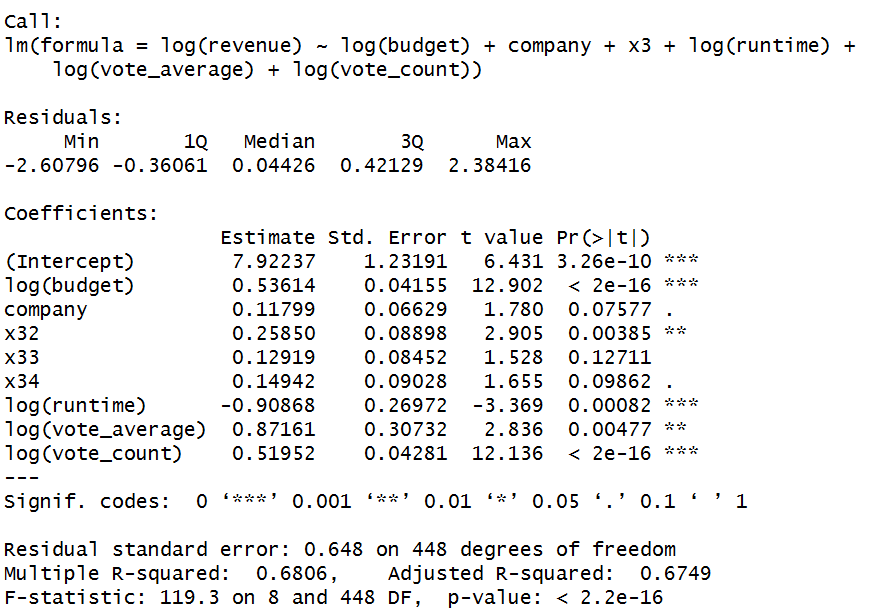
> lm.test <- lm(revenue~ budget + company +x3 +runtime + vote\_average +vote\_count)

> summary(lm.test)

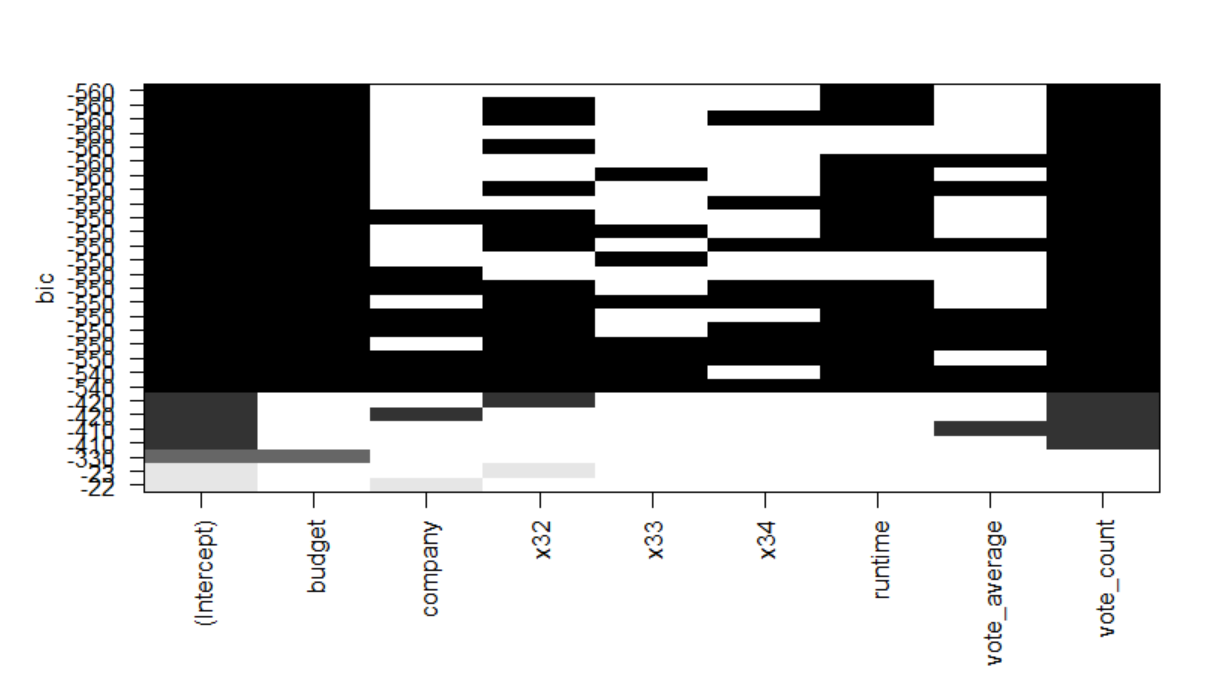


> lm.test <- lm(log(revenue)~ log(budget) + company +x3 +log(runtime) + log(vote\_average) +log(vote\_count))

> summary(lm.test)

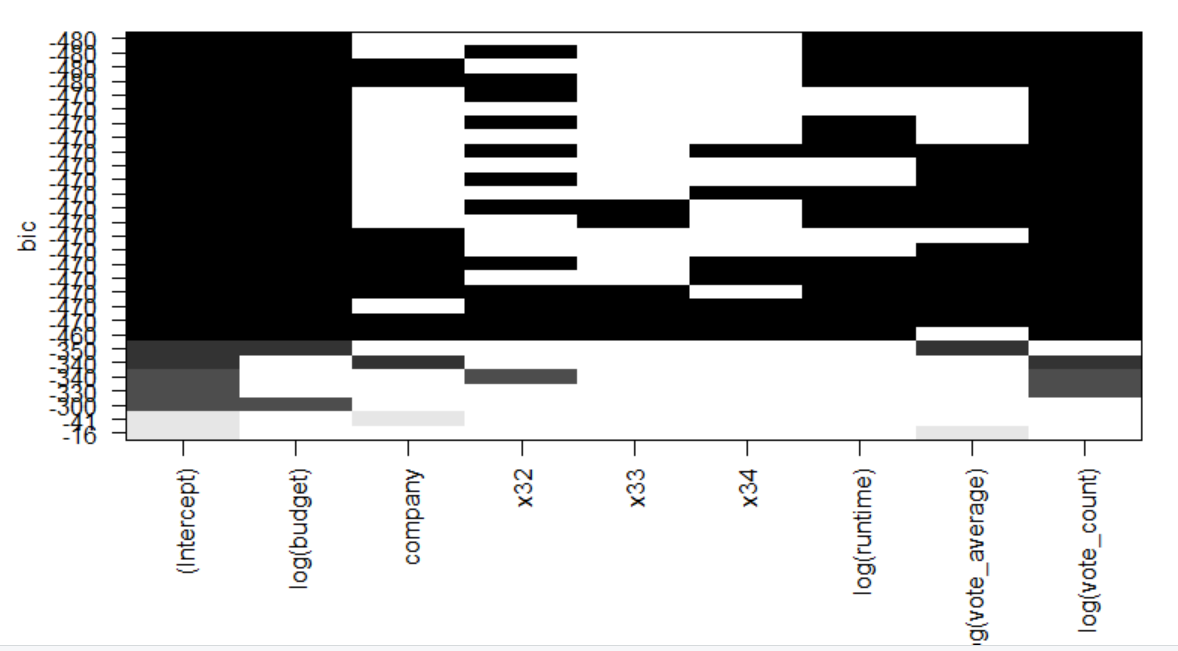






> leaps<-regsubsets(log(revenue)~log(budget) + company +x3 + log(runtime) + log(vote\_average) + log(vote\_count),data =h.1,nbest=4)

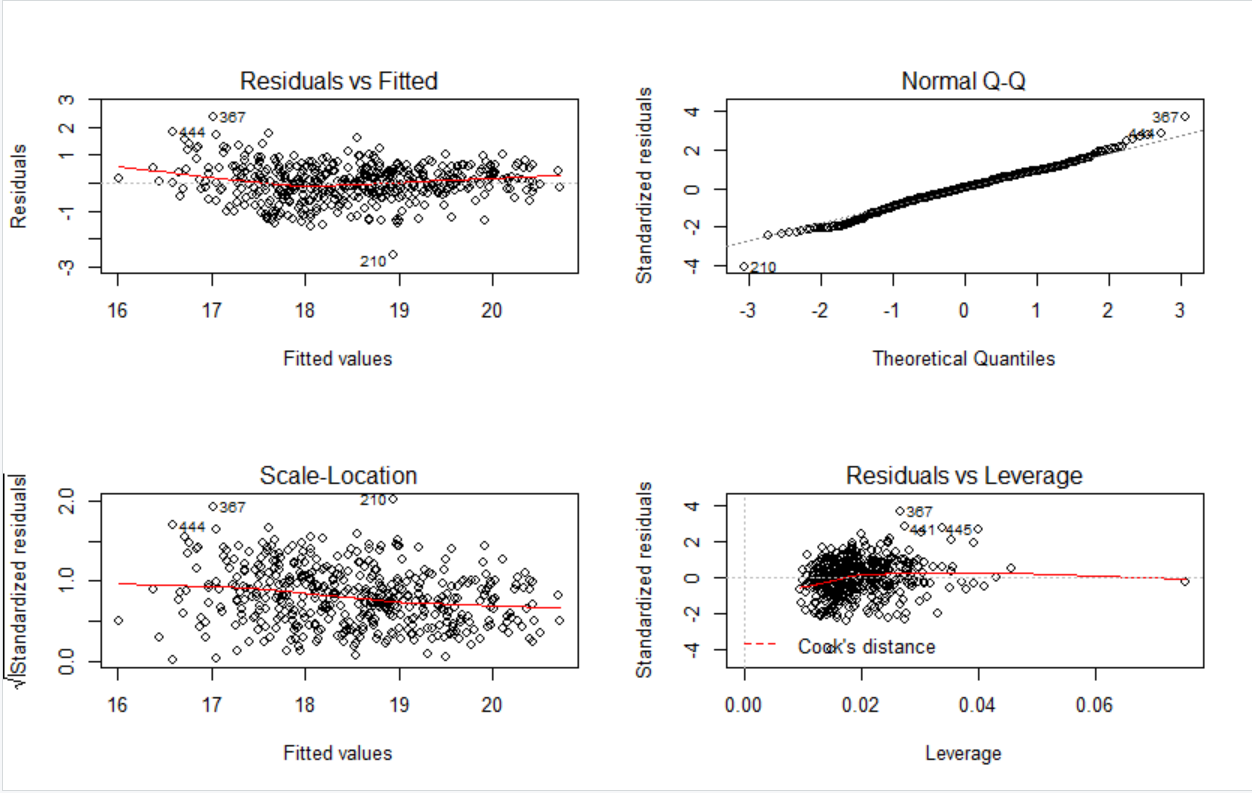
> plot(leaps)



回归诊断：

二次拟合诊断图：

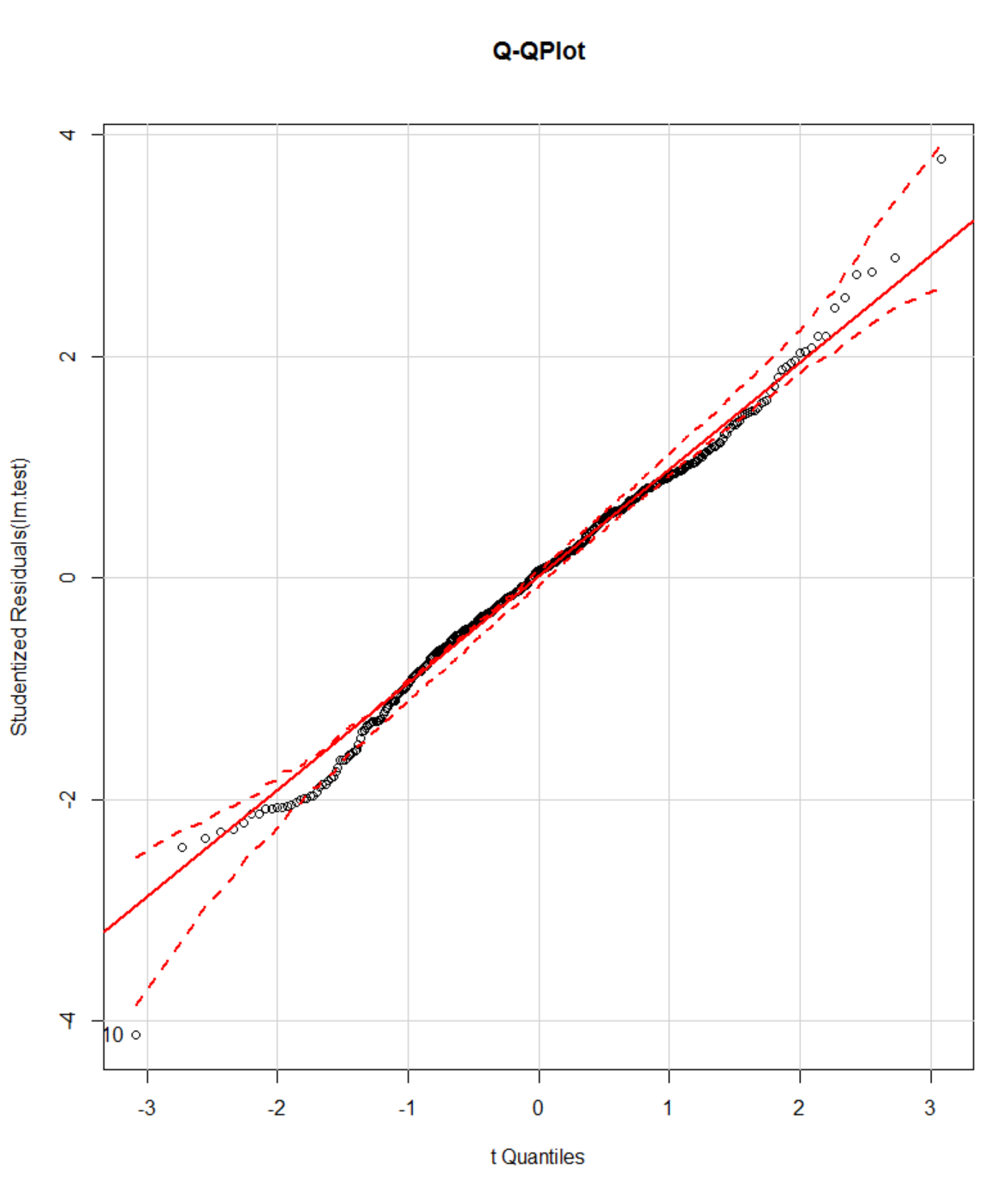
> par(mfrow=c(2,2))

>plot(lm.test)

标记异常值：

> qqPlot(lm.test,labels=row.names(h.1),id.method="identify",simulate=TRUE,main="Q-QPlot")

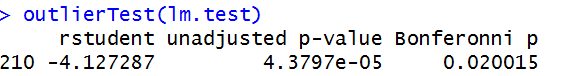
[1] "210"



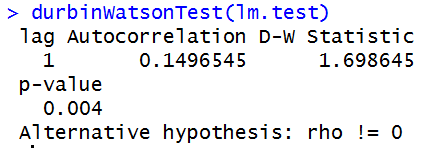
> outlierTest(lm.test)

rstudent unadjusted p-value Bonferonni p

210 -4.127287 4.3797e-05 0.020015



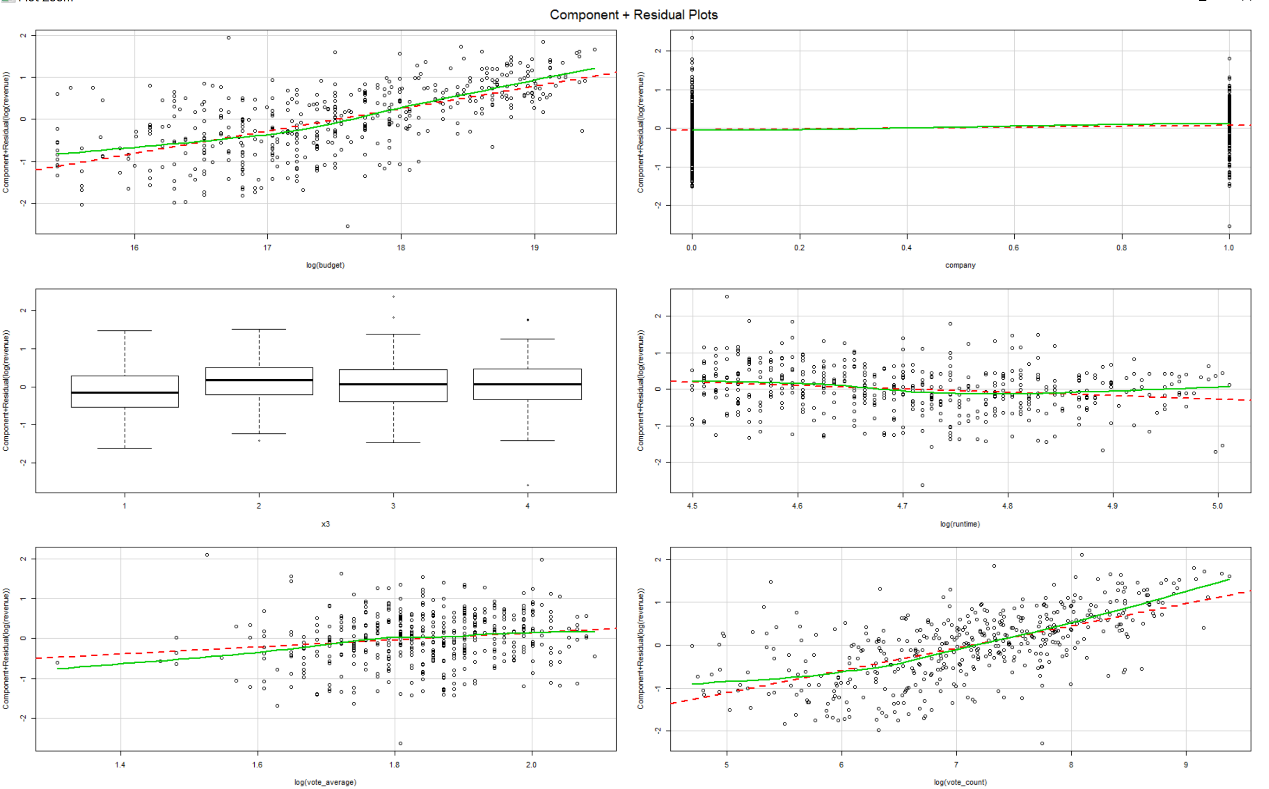
2、误差的独立性



P值不显著说明无自相关性，误差之间独立

3、线性

> crPlots(lm.test)



线性相关性良好

4、同方差性

> ncvTest(lm.test)

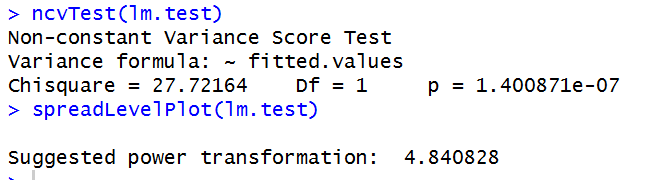
Non-constant Variance Score Test

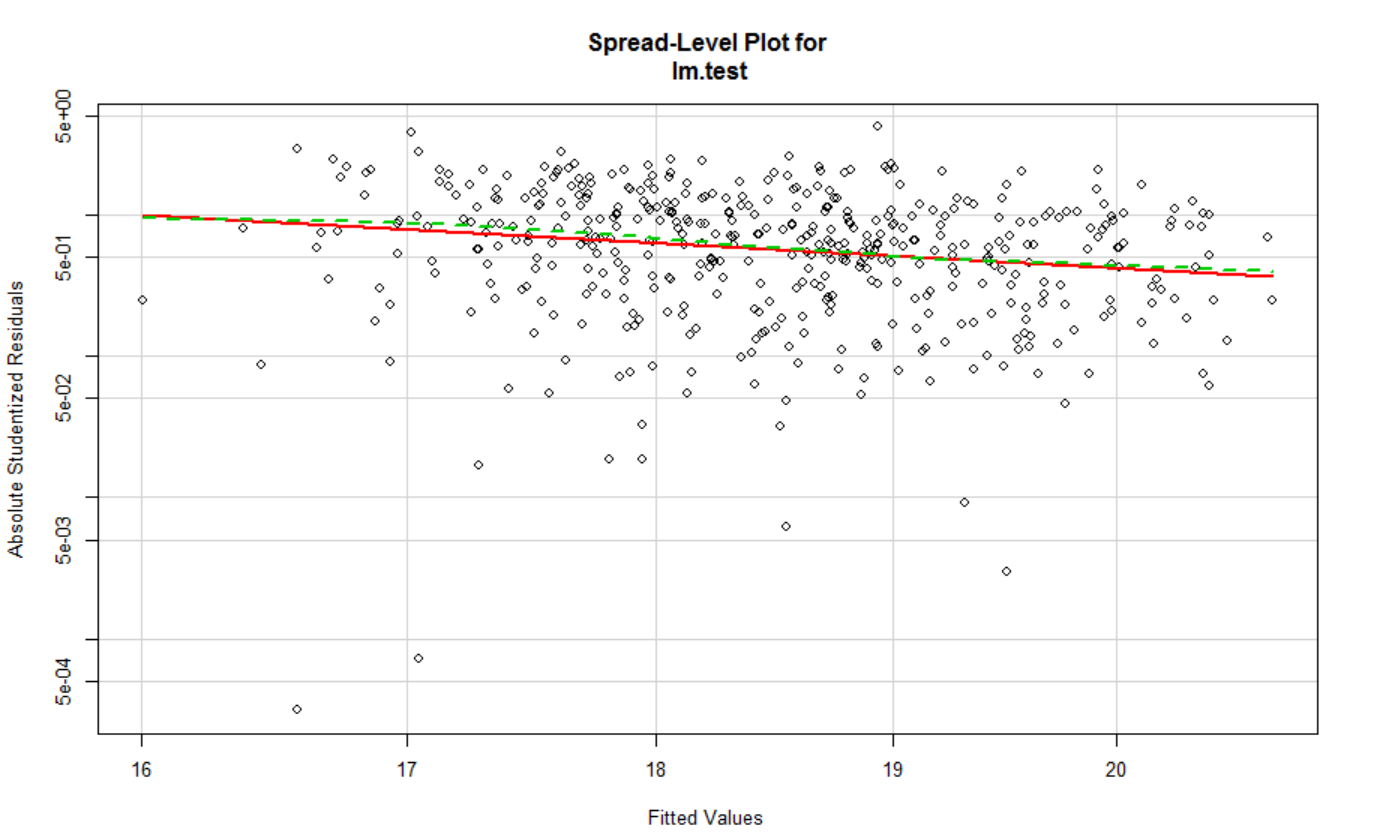
Variance formula: ~ fitted.values

Chisquare = 27.72164 Df = 1 p = 1.400871e-07

> spreadLevelPlot(lm.test)

Suggested power transformation: 4.840828

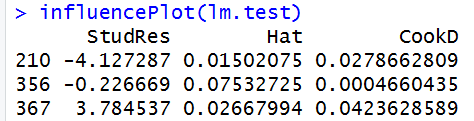


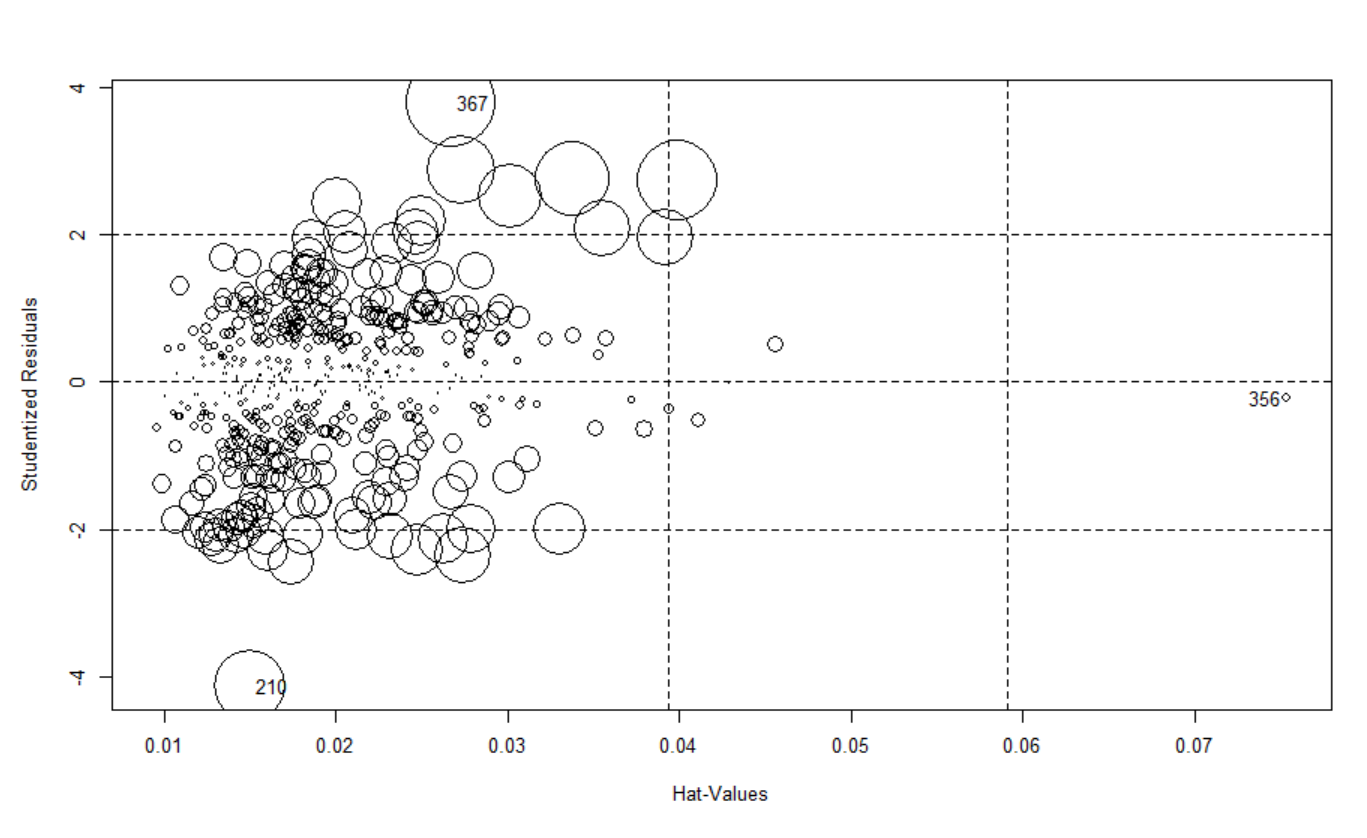


其中的点应该在水平的最佳拟合曲线周围呈现水平随机分布

5、强影响点、离群点、高杠杆值点：

> influencePlot(lm.test)





6、多重共线性multicollinearity

表明存在多重共线性关系

