高级统计方法 第4次作业:

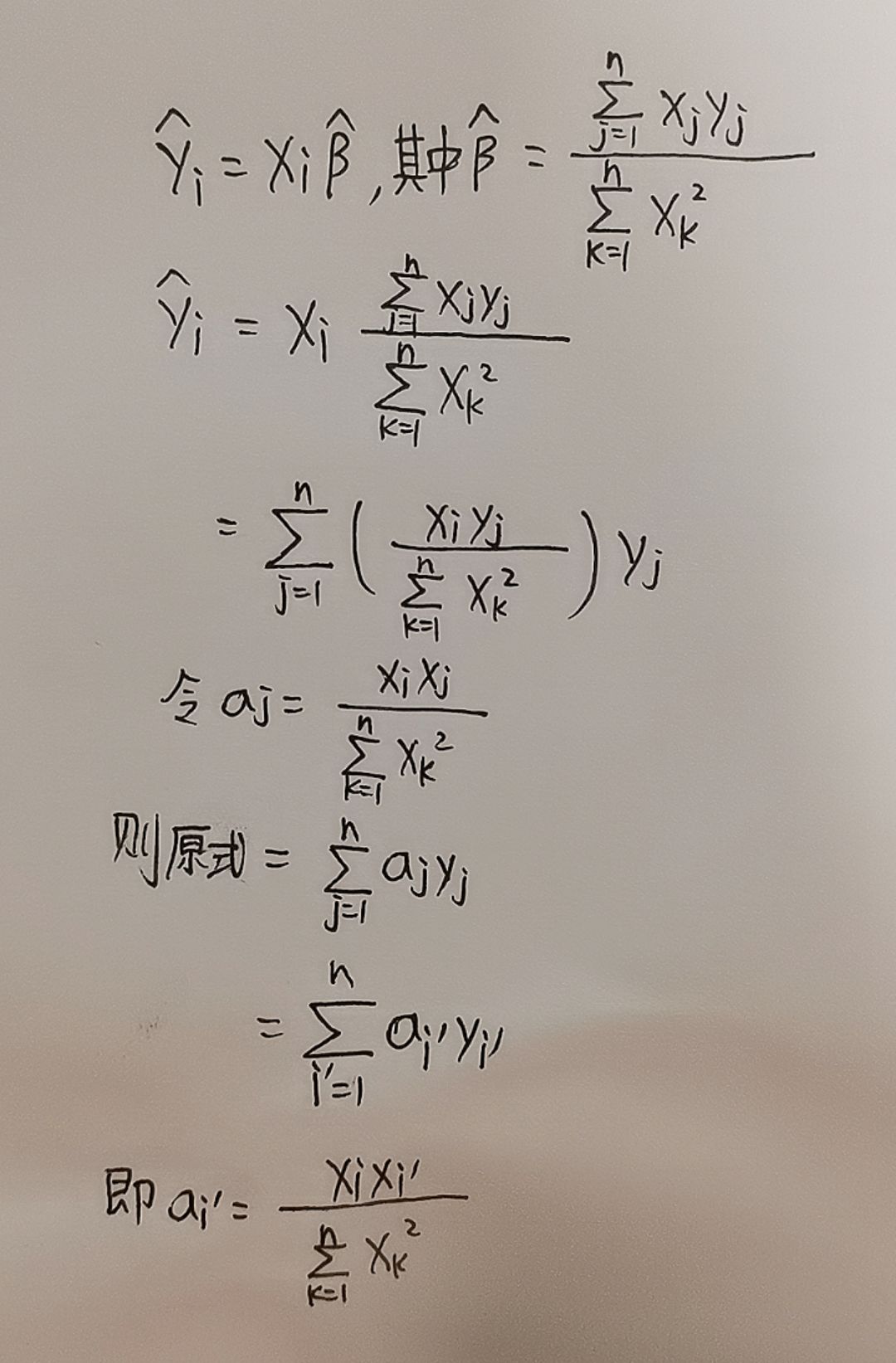
序号：1 姓名：杨程瑜 学号：20202241107 班级：网2001

**概念**

2.问题（略）

算法思想或者说公式相同，都是从要预测的点的最近K个近邻点中，根据这些点的特征来预测。而KNN分类器根据最近的K个点中类别数占最多的类别，来给预测点进行预测；而KNN回归方法根据最近的K个点的平均值来给预测点进行预测。一个是定性的输出，一个是定量值的输出。

1. 问题（略）



**应用**

1. 问题（略）

(a)问题（略）

library(ISLR)

# 发现数据中有缺值的数据单元(表示为？)，先替换？为NA，再使用na.omit进行处理

Auto$horsepower[which(Auto$horsepower=="?")]=NA

Auto$mpg[which(Auto$mpg=="?")]=NA

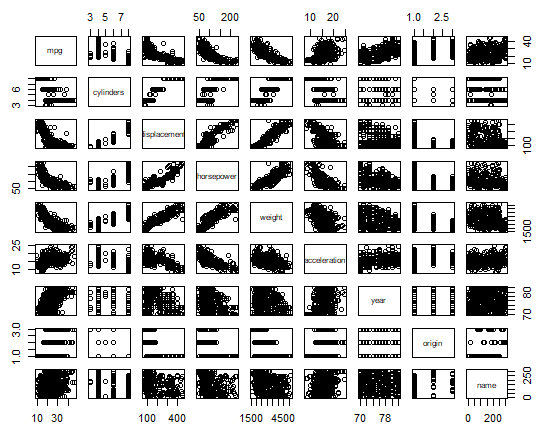
Auto = na.omit(Auto)

# 发现数据都为字符型，直接操作是无法计算的，遂改为数字型

Auto = as.data.frame(lapply(Auto, as.numeric))

attach(Auto)

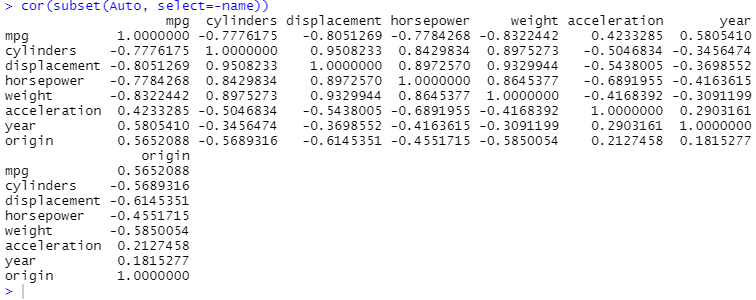
pairs(Auto)



(b)问题（略）

# 使用subset取子集，使用select参数加负号去除name

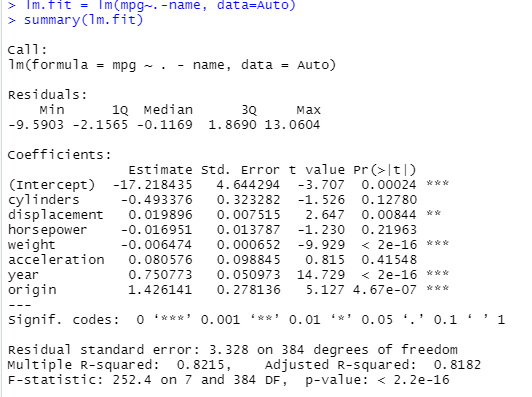
cor(subset(Auto, select=-name))



(c)问题（略）

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

summary(lm.fit)



i问题（略）

有关系，基于零假设：预测变量和响应变量之间所有回归系数均为0

而其F统计量为252.4，远大于1，p值远小于0.05

ii问题（略）

displacement,weight,year,origin存在显著关系

cylinders,horsepower,acceleration不存在显著关系

iii问题（略）

year的系数约为0.758，代表year每增长一年，mpg增长大约0.758个单位

即汽车每年的燃油效率提高约0.758mpg/年

(d)问题（略）

par(mfrow=c(2,2))

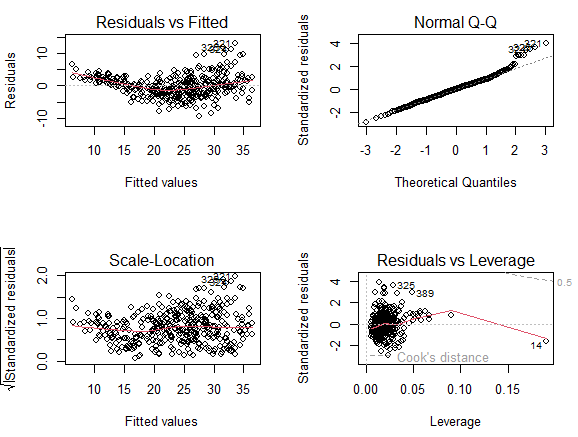
plot(lm.fit)

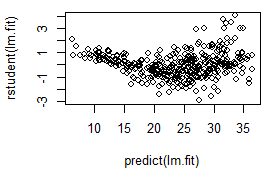
# 分析残差图，发现有比较明显的曲线（先下降后上升）

# 存在较大的离群点、高杠杆点14（约0.2处，而大部分点不超过0.05）

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

# 画出学生化残差图，发现其中有大于3的点(高杠杆点)，可能存在异常





(e)问题（略）

lm.fit1 = lm(mpg~cylinders\*displacement+displacement\*weight)

summary(lm.fit1)

# 存在统计显著的交互作用

# 观察发现cylinders:displacement交互项的p值为0.384>0.05，不够显著

# 而displacement:weight的p值远小于0.05，足够显著

(f)问题（略）

lm.fit2 = lm(mpg~log(weight)+log(horsepower)+log(acceleration))

summary(lm.fit2)

lm.fit2 = lm(mpg~sqrt(weight)+sqrt(horsepower)+sqrt(acceleration))

summary(lm.fit2)

lm.fit2 = lm(mpg~I(weight^2)+I(horsepower^2)+I(acceleration^2))

summary(lm.fit2)

# 通过尝试发现如下变量显著性水准比较高

lm.fit3 = lm(mpg~log(horsepower)+sqrt(horsepower)+log(weight)+I(acceleration^2))

summary(lm.fit3)

# 画图进行分析

par(mfrow=c(2,2))

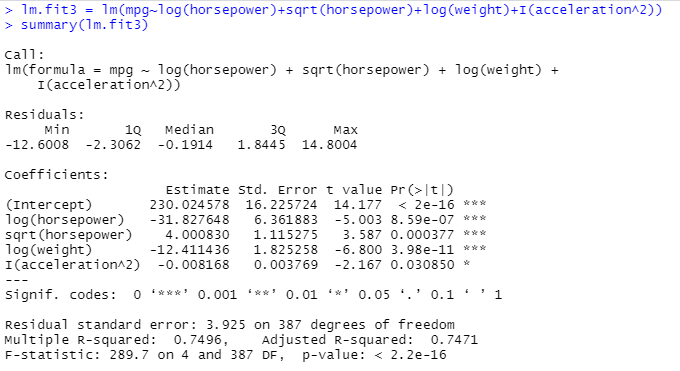
plot(lm.fit2)

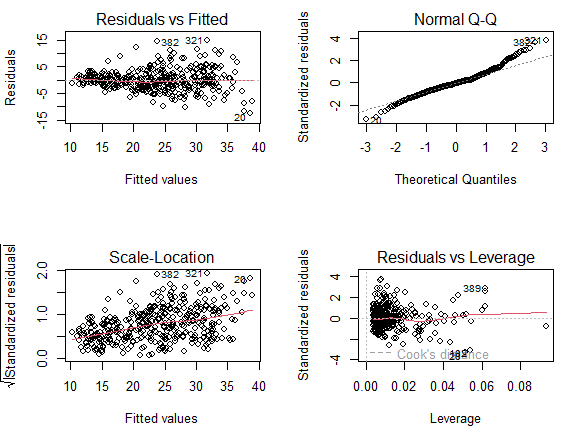
# 残差图曲线趋势不明显，拟合效果比较好

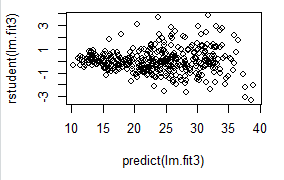
# 存在离群点

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

# 画出学生化残差图，发现其中有大于3的点，可能存在异常







10.问题（略）

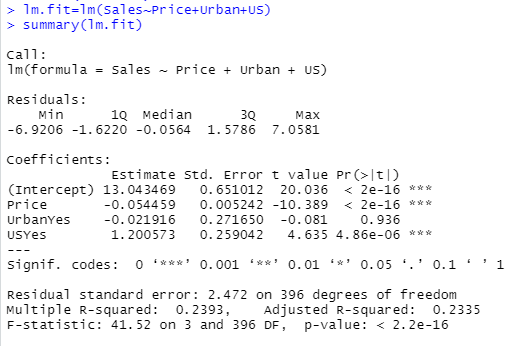
(a)问题（略）

library(ISLR)

attach(Carseats)

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

summary(lm.fit)



(b)问题（略）

# Price：有显著性，Sales和Price呈负相关，其系数为-0.054

# UrbanYes：p值远大于0.05，没有显著性

# USYes：有显著性，Sales和是否在美国存在正相关，如果商店在美国，则销售额将增加约 1201 件

(c)问题（略）

# 将系数代入式子即可

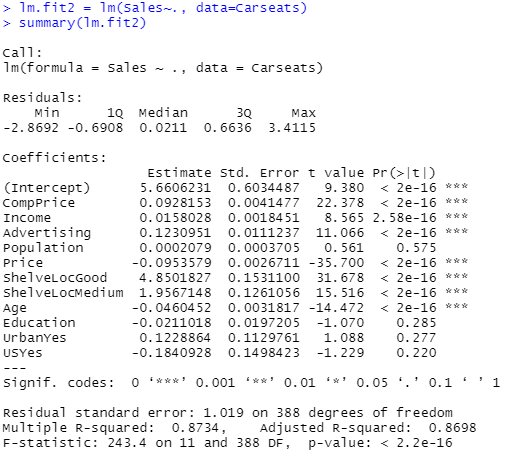
# Sales = 13.04 + -0.05 Price + -0.02 UrbanYes + 1.20 USYes

(d)问题（略）

lm.fit2 = lm(Sales~., data=Carseats)

summary(lm.fit2)

# 可以看到ComPrice,Income,Advertising,Price,ShelveLocGood,ShelveLocMedium,Age的p值显著，可以拒绝零假设

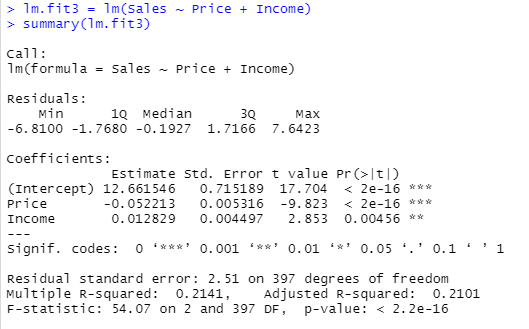


(e)问题（略）

lm.fit3 = lm(Sales ~ Price + Income)

summary(lm.fit3)

# Sales与Price和Income的拟合



(f)问题（略）

# 观察summary函数显示的RSE与R-squared

# (a) RSE=2.472 R-squared=0.2393

# (e) RSE=2.51 R-squared=0.2141

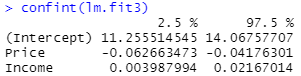
# 拟合情况比较类似，都只能解释约20%的方差，但是a中模型R-squared大些，拟合的效果稍微更好

(g)问题（略）

confint(lm.fit3)

# Price的95%置信区间为[-0.0627, -0.0417]

# Income的95%置信区间为[0.0040, 0.0217]



(h)问题（略）

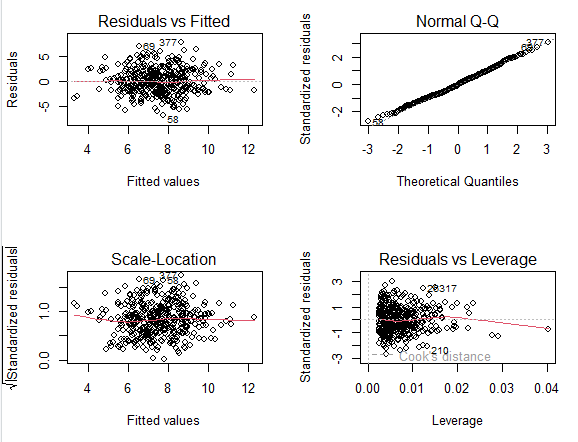
par(mfrow=c(2,2))

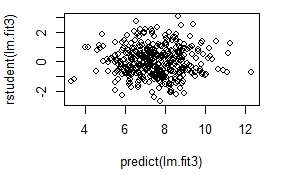
plot(lm.fit3)

# 显然存在离群点（0.04处）

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

# 大部分点比较集中，高杠杆点不明显





11.问题（略）

(a)问题（略）

set.seed(1)

x=rnorm(100)

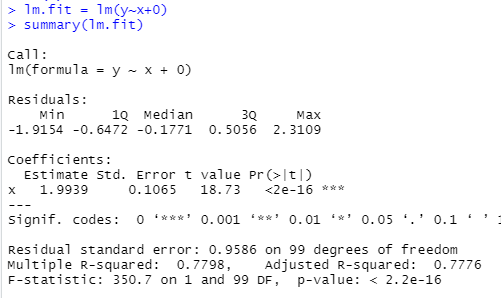
y=2\*x+rnorm(100)

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

summary(lm.fit)

# x的估计系数为1.9939,标准差为0.1065,t统计量为18.73

# p值远小于0.05,表示有显著性



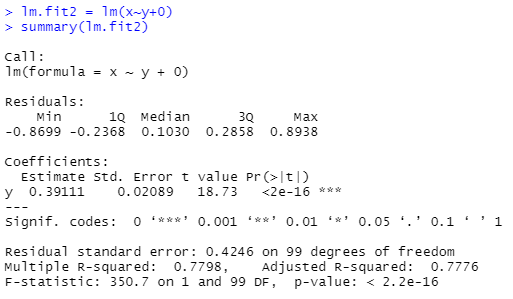
(b)问题（略）

lm.fit2 = lm(x~y+0)

summary(lm.fit2)

# y的估计系数为0.3911,标准差为0.0209,t统计量为18.73

# p值远小于0.05,表示有显著性



(c)问题（略）

# 发现两者t统计量是相同的

# 实际上两者是对同一条直线的拟合，互为反函数

(d)问题（略）

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



(e)问题（略）

# 从(d)中计算得出t=18.7259,而从x对y与y对x的回归中都能够得出t统计量为18.73

(f)问题（略）

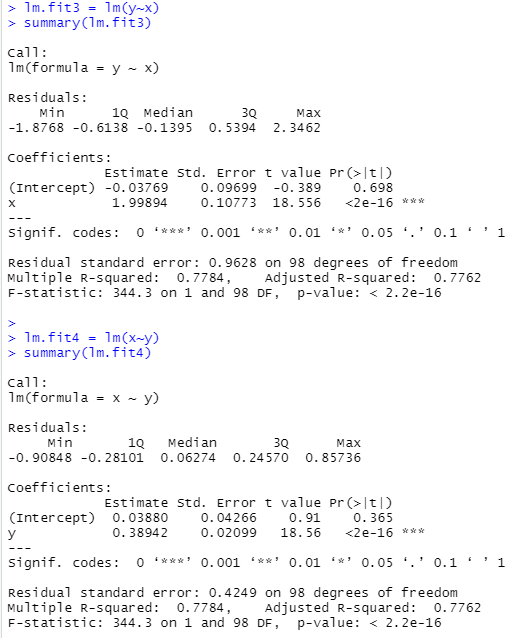
lm.fit3 = lm(y~x)

summary(lm.fit3)

lm.fit4 = lm(x~y)

summary(lm.fit4)

# t统计量都是18.56



13.问题（略）

(a)问题（略）

rm(list = ls())

set.seed(1)

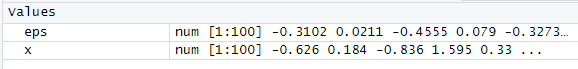
x=rnorm(100)



(b)问题（略）

# rnorm函数，参数分别为向量长度、均值、标准差

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



(c)问题（略）

y = -1 + 0.5\*x + eps

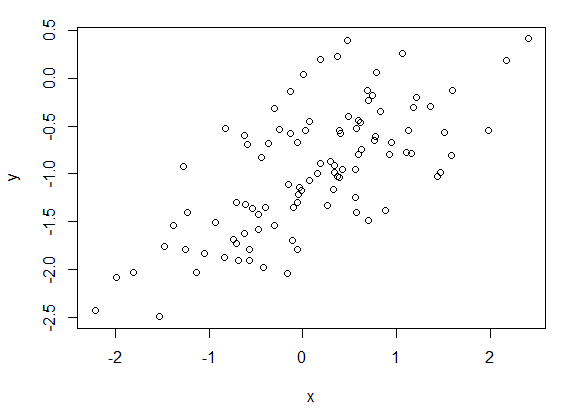
# y的长度为100,β0=-1,β1=0.5

(d)问题（略）

plot(x, y)

# 散点图总体呈上升趋势

# 即x和y大致上是正相关的



(e)问题（略）

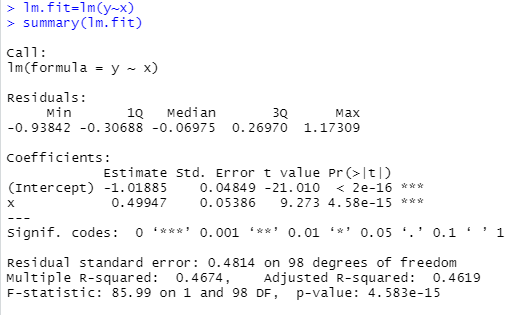
lm.fit=lm(y~x)

summary(lm.fit)

# 拟合的结果：截距为-1.0189，x的系数为0.4995，和β0、β1很相近

# F统计量为85.99，比较大

# 同时p值远小于0.05，具有显著性，结果显示拟合的效果较好



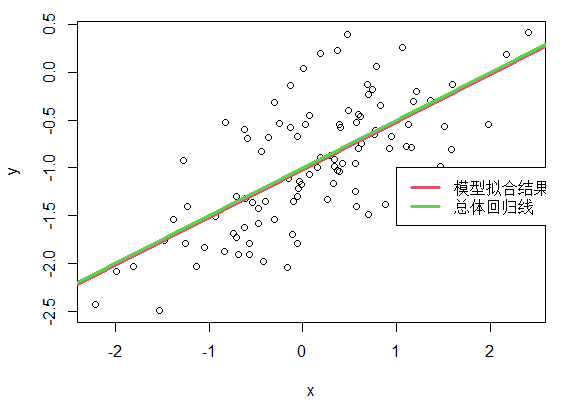
(f)问题（略）

plot(x, y)

abline(lm.fit, lwd=3, col=2)

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

legend(-1, legend = c("模型拟合结果", "总体回归线"), col=2:3, lwd=3)



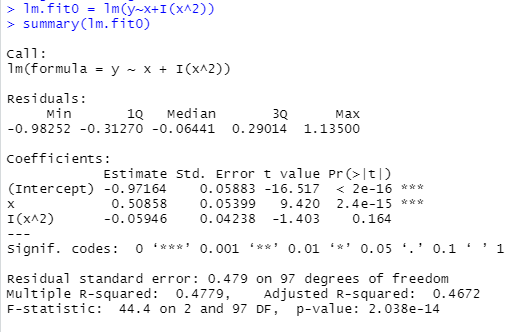
(g)问题（略）

lm.fit0 = lm(y~x+I(x^2))

summary(lm.fit0)

# 模型的RSE减少了，且R-squared增加了，表明模型拟合效果略微变好了

# 虽然二次项的t统计量的p值显示，其并没有显著性



(h)问题（略）

set.seed(1)

# 减小数据噪声，将其方差设为0.1

eps1 = rnorm(100, 0, sqrt(0.1))

x1 = rnorm(100)

y1 = -1 + 0.5\*x1 + eps1

plot(x1, y1)

lm.fit1 = lm(y1~x1)

summary(lm.fit1)

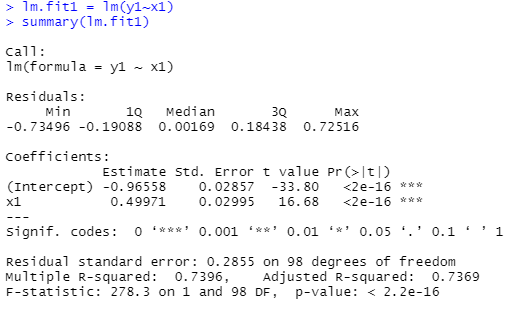
abline(lm.fit1, lwd=3, col=2)

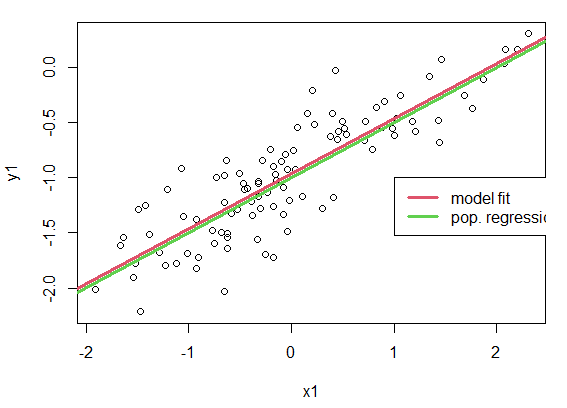
abline(-1, 0.5, lwd=3, col=3)

legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)

# 结果是显然模型的RSE减少了，且R-squared增加了

# 说明模型的拟合效果随着噪声减小而变好了





(i)问题（略）

# 增加数据的噪声，将方差设为0.5

eps2 = rnorm(100, 0, sqrt(0.5))

x2 = rnorm(100)

y2 = -1 + 0.5\*x2 + eps2

plot(x2, y2)

lm.fit2 = lm(y2~x2)

summary(lm.fit2)

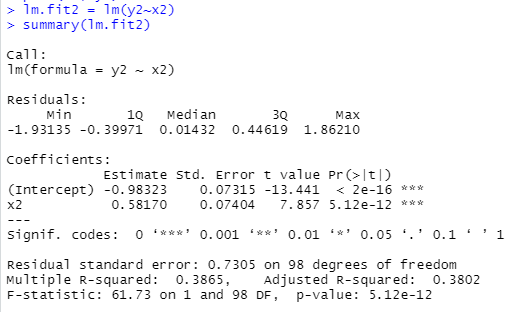
abline(lm.fit2, lwd=3, col=2)

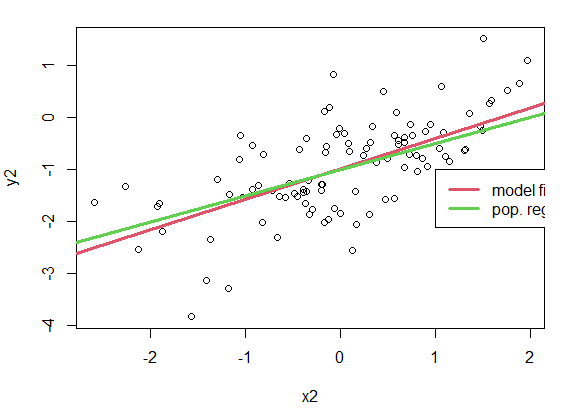
abline(-1, 0.5, lwd=3, col=3)

legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)

# 结果是显然模型的RSE增加了，且R-squared降低了

# 说明模型的拟合效果随着噪声增加而变差了





(j)问题（略）

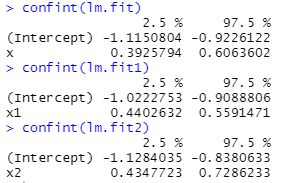
confint(lm.fit)

confint(lm.fit1)

confint(lm.fit2)

# 三个区间都以0.5为中心，宽度不同

# 其中噪声越小，范围越窄；噪声越大，范围越宽



14.问题（略）

(a)问题（略）

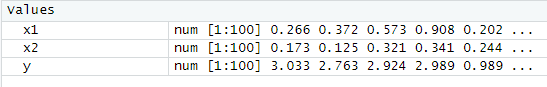
set.seed(1)

x1 = runif(100)

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

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

# 系数：β0=2,β1=2,β2=0.3



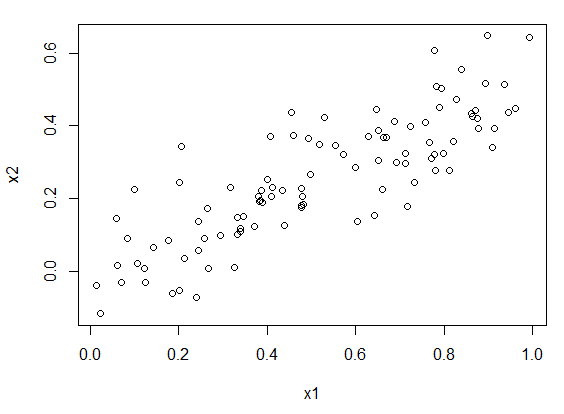
(b)问题（略）

cor(x1, x2)

# 0.8351

plot(x1, x2)





(c)问题（略）

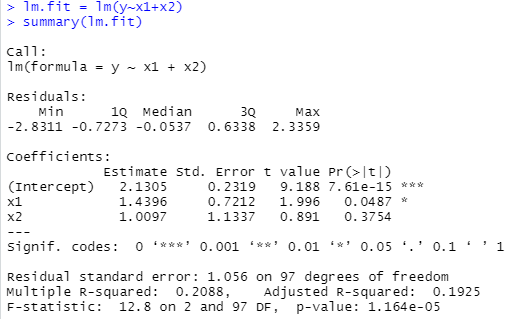
lm.fit = lm(y~x1+x2)

summary(lm.fit)

# β0=2.1305,β1=1.4396,β2=1.0097

# 可以拒绝β1的原假设，其p值小于0.05，具有显著性

# 不能拒绝β2的原假设，其p值大于0.05，不具有显著性

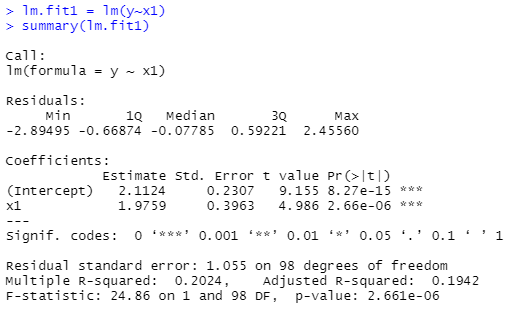


(d)问题（略）

lm.fit1 = lm(y~x1)

summary(lm.fit1)

# 可以拒绝β1的原假设，其p值小于0.05，具有显著性

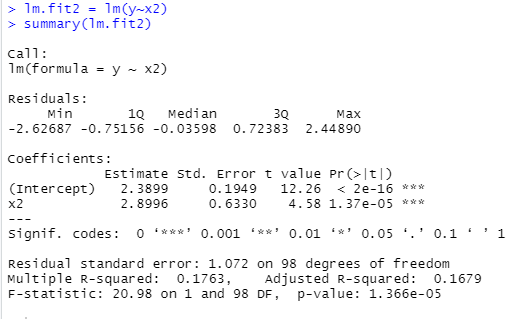


(e)问题（略）

lm.fit2 = lm(y~x2)

summary(lm.fit2)

# 可以拒绝β2的原假设，其p值小于0.05，具有显著性



(f)问题（略）

# 不矛盾，因为x2由x1线性生成，故x1与x2具有共线性

# 当单独分别和y进行回归拟合时，显著性强；当同时拟合时，很难区分他们的影响

(g)问题（略）

x1 = c(x1, 0.1)

x2 = c(x2, 0.8)

y = c(y, 6)

lm.fit3 = lm(y~x1+x2)

summary(lm.fit3)

# x1的t统计量的p值大于0.05，没有统计显著性

# x2的t统计量的p值远小于0.05，统计显著

lm.fit4 = lm(y~x1)

summary(lm.fit4)

# x1的t统计量的p值远小于0.05，统计显著

lm.fit5 = lm(y~x2)

summary(lm.fit5)

# x2的t统计量的p值远小于0.05，统计显著

par(mfrow=c(2,2))

plot(lm.fit3)

# 存在明显的离群点（101）

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

par(mfrow=c(2,2))

plot(lm.fit4)

plot(predict(lm.fit4), rstudent(lm.fit4))

# 存在一个高杠杆点（大于3）

par(mfrow=c(2,2))

plot(lm.fit5)

# 存在明显的离群点（101）

plot(predict(lm.fit5), rstudent(lm.fit5))

