高级统计方法 第3次作业:

序号：1 姓名：杨程瑜 学号：20202241107 班级：网2001

**第三章**

**概念**

1. 问题（略）

这些p值基于的前提是，原假设（零假设）为这三个变量在其中另两个变量存在时，变量本身对预测值（销售量没有影响）。p值越小越有理由拒绝原假设，p值越大越没有充足理由拒绝原假设。从表3.4观察发现TV、radio的p值都很小，说明其对销售量没有影响的原假设是错误地，而newspaper的p值很大，故其对销售量没有影响的原假设是正确的。

3.问题（略）

（a）问题（略）

根据题意可以写出拟合后的函数：

Y = 50 + 20 \* X1 + 0.07 \* X2 + 35 \* X3 + 0.01 \* X4 - 10 \* X5

=50 + 20 \* X1 + 0.07 \* X2 + 35 \* X3 + 0.01 \*（X1 \* X2）-10 \*（X1 \* X3）

当预测的是男性时，其X3=0，则函数变为：

Y1 = 50 + 20 \* X1 + 0.07 \* X2 + 0.01 \*（X1 \* X2）

当预测的是女性时，其X3=1，则函数变为：

Y2 = 85 + 10 \* X1 + 0.07 \* X2 + 0.01 \*（X1 \* X2）

能够发现当IQ（X2）和GPA（X1）都确定时，GPA（X1）足够高的情况下，男性平均收入高于女性（两式只含X1的项中Y1的项系数更大，而另一含X1X2的项其系数相同）。

而其他选项都不对。

故iii项是正确的。

（b）问题（略）

代入预测函数即可

Y = 50 + 20 \* 4 + 0.07 \* 110 + 35 + 0.01 (4 \* 110) - 10 \* 4

= 137.1

（c）问题（略）

假。 交互项系数小不能直接说明二者不存在关系，必须计算回归系数的 p 值来确定交互项是否具有统计显著性。

4.问题（略）

（a）问题（略）

多项式回归的训练残差平方和比线性回归的低。

因为多项式回归可以更紧密、更灵活的拟合训练用的数据。

（b）问题（略）

测试集上结果相反，多项式回归比线性回归具有更高的测试残差平方和。

因为多项式回归在训练集上容易过度拟合，在测试集上比线性回归具有更多的误差。

（c）问题（略）

多项式回归的训练残差平方和比线性回归的低。

因为多项式回归可以更紧密、更灵活的拟合训练用的数据，不论真实函数关系是否是线性还是非线性，更灵活的模型更容易取得好的训练集拟合效果。

（d）问题（略）

根据已知条件无法判断。

如果真实函数关系更接近于线性，则可能线性回归的残差平方和更低；如果真实函数关系更接近于多项式（比如接近三次性），则可能多项式回归的残差平方和更低。

原因是偏差和方差的均衡关系：当接近多项式型时，线性模型的偏差大，方差小；多项式模型的偏差小，方差大。无法判断两个模型偏差和方差的综合大小情况。

**应用**

8.问题（略）

(a)

auto = read.csv("data/Auto.csv")

# 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))

summary(auto)

attach(auto)

lm.fit = lm(mpg~horsepower, data=auto)

summary(lm.fit)

# i

# ii

# iii

cov(auto)

cor(auto)

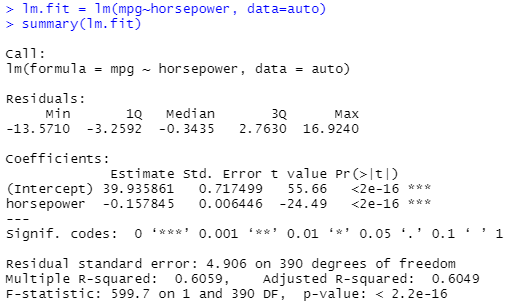
# horsepower和mpg有关系，协方差为-233.8580，相关性为-0.7784，是负相关

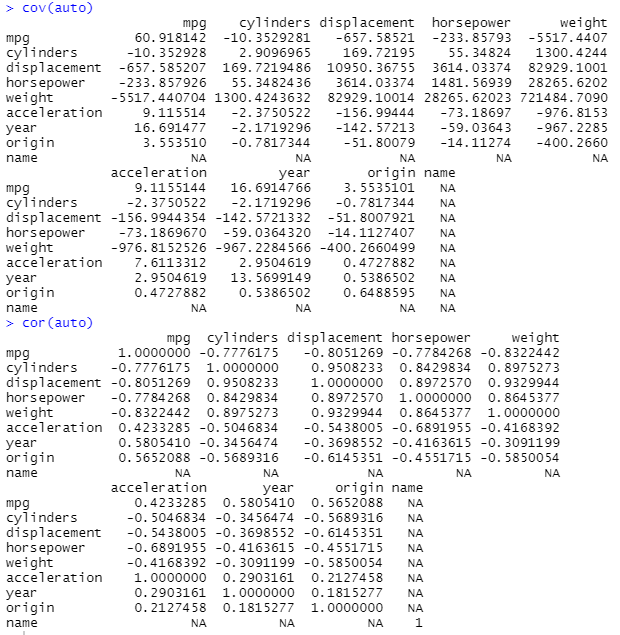
# iv

# 由于值为字符型，需要转型为数字

predict(lm.fit, data.frame(horsepower = 98))

# 是24.46708



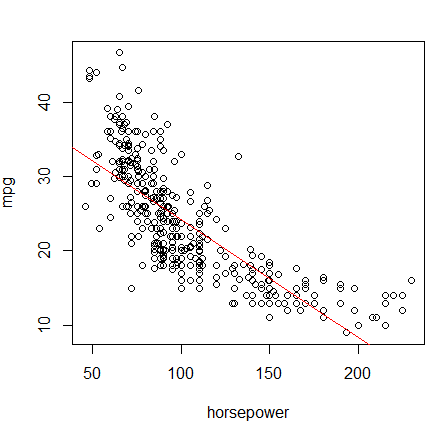




# (b)

plot(horsepower, mpg)

abline(lm.fit, col="red")



# (c)

par(mfrow=c(2,2))

plot(lm.fit)

# 通过b中的散点图能够发现，红色拟合直线与散点图并不紧密贴合

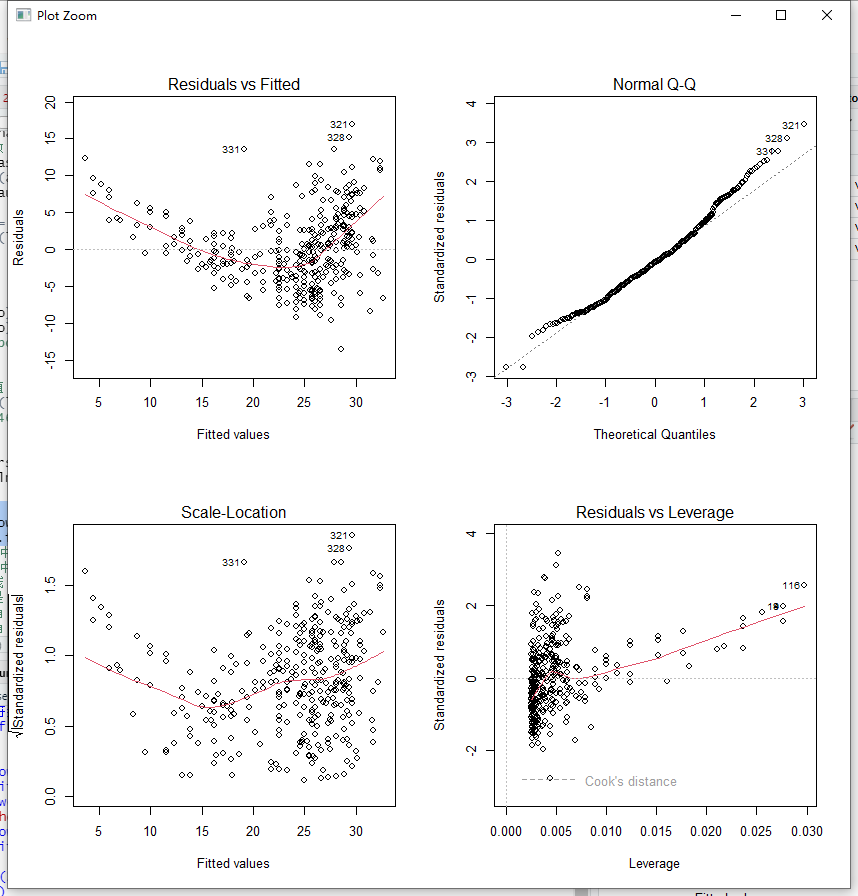
# 通过c中左上角残差图与拟合图，发现残差值与拟合图存在曲线贴合关系

# 如果线性相关则残差值应该和拟合没有任何系统关联，故线性模型不能很好拟合

# QQ图是判断正态分布的，数据基本吻合正态分布

# 左下角是方差分布，先下降后上升，基本符合

# 右下角是极端点分布，发现没有特别极端的点（都在零点几）



**第二章**

**应用**

9.问题（略）

library(ISLR)

(a)

# name和origin是定性的，其他是定量的（查看数据集每一列的实际意义）

sapply(Auto, is.factor)

# 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

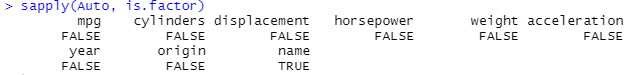
Auto$horsepower[which(Auto$horsepower=="?")]=NA

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

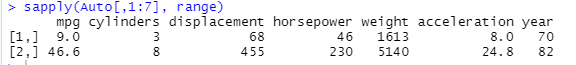
Auto = as.data.frame(lapply(Auto, as.numeric))



(b)

# 使用 sapply 批处理函数对矩阵的所有单元使用 range 函数处理

sapply(Auto[,1:7], range)

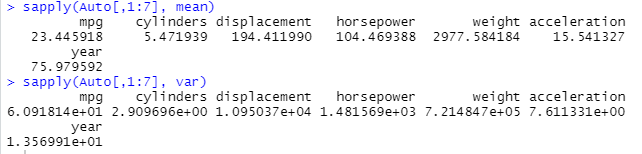


(c)

# 同上使用 sapply -> mean, sd

sapply(Auto[,1:7], mean)

sapply(Auto[,1:7], var)



(d)

# 剔除数据使用负号索引

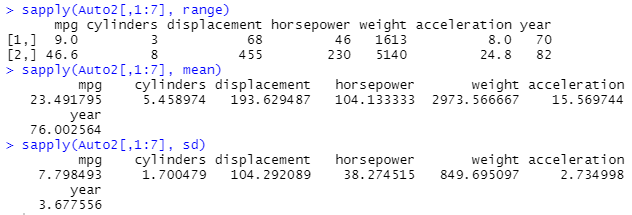
Auto2 = Auto[-10,]

Auto2 = Auto2[-85,]

sapply(Auto2[,1:7], range)

sapply(Auto2[,1:7], mean)

sapply(Auto2[,1:7], var)



(e)

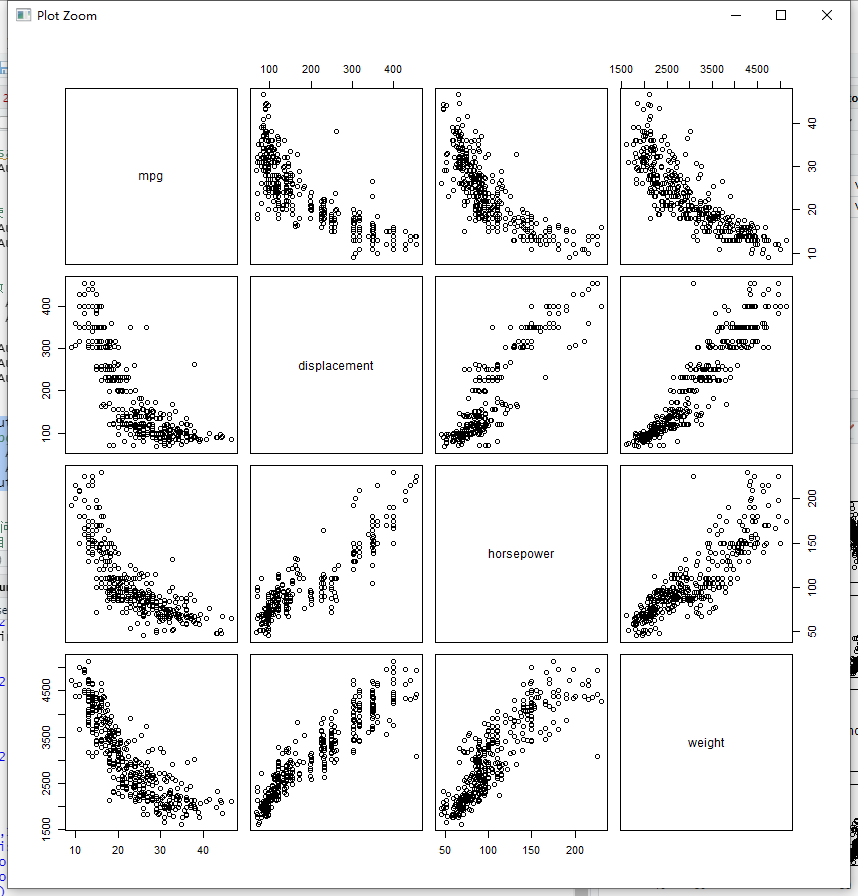
pairs(Auto[,1:7])

# 其中mpg和displacement、horsepower、weight，两两之间互有关系

Auto3 = Auto[,1:5]

Auto3 = Auto3[,-2]

pairs(Auto3)



(f)

# 从(e)问中能够得到结论：displacement、horsepower、weight都可以预测mpg

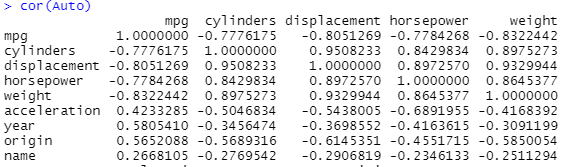
# 再从相关性的角度来说明

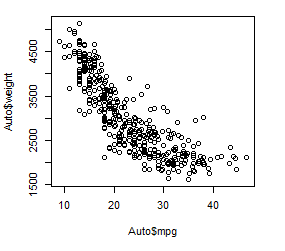
cor(Auto)

# 发现weight和mpg的相关系数为-0.8322，显然有明显的负相关，可以用来预测mpg

plot(Auto$mpg, Auto$weight)

# 散点图大概能够看出是非线性的负相关关系





1. 问题（略）

(a)

library(MASS)

print(Boston)

Boston

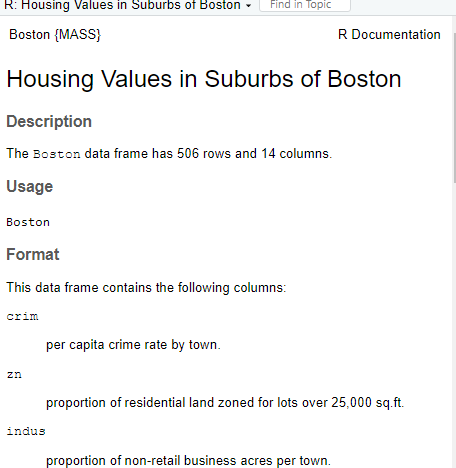
# 数据说明信息

print(?Boston)

# 506行，14列

dim(Boston)



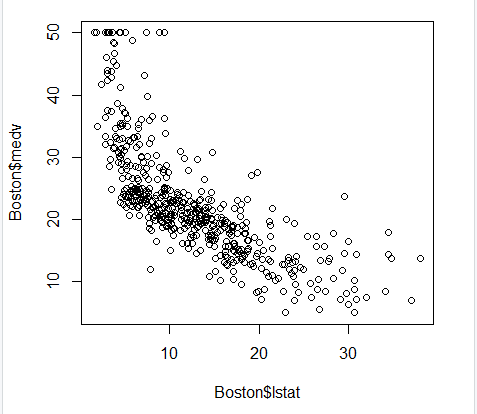


(b)

# 低地位人口的占比与自住房价值的中值之间的关系

# 随着低地位人口占比增加，自住房价值中值呈下降趋势

plot(Boston$lstat, Boston$medv)



(c)

# 住宅越老，犯罪率越高

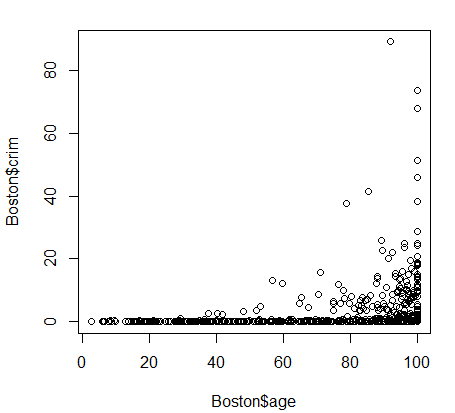
plot(Boston$age, Boston$crim)

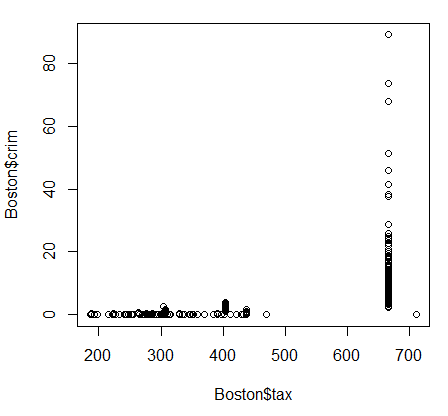
# 税率越高，犯罪率越高

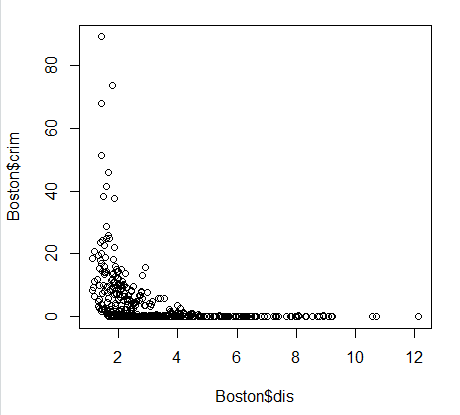
plot(Boston$tax, Boston$crim)

# 距离就业中心越近，犯罪率越高

plot(Boston$dis, Boston$crim)







(d)

par(mfrow=c(3,3))

# 大部分城市犯罪率较低（小于20占大多数）

hist(Boston$crim[Boston$crim>1], breaks=30)

hist(Boston$crim[Boston$crim>2], breaks=30)

hist(Boston$crim[Boston$crim>5], breaks=30)

# 大部分城市税率在250-400之间，还有很多一部分城市的税率在660附近(666)

hist(Boston$tax, breaks=30)

hist(Boston$tax[Boston$tax>400], breaks=30)

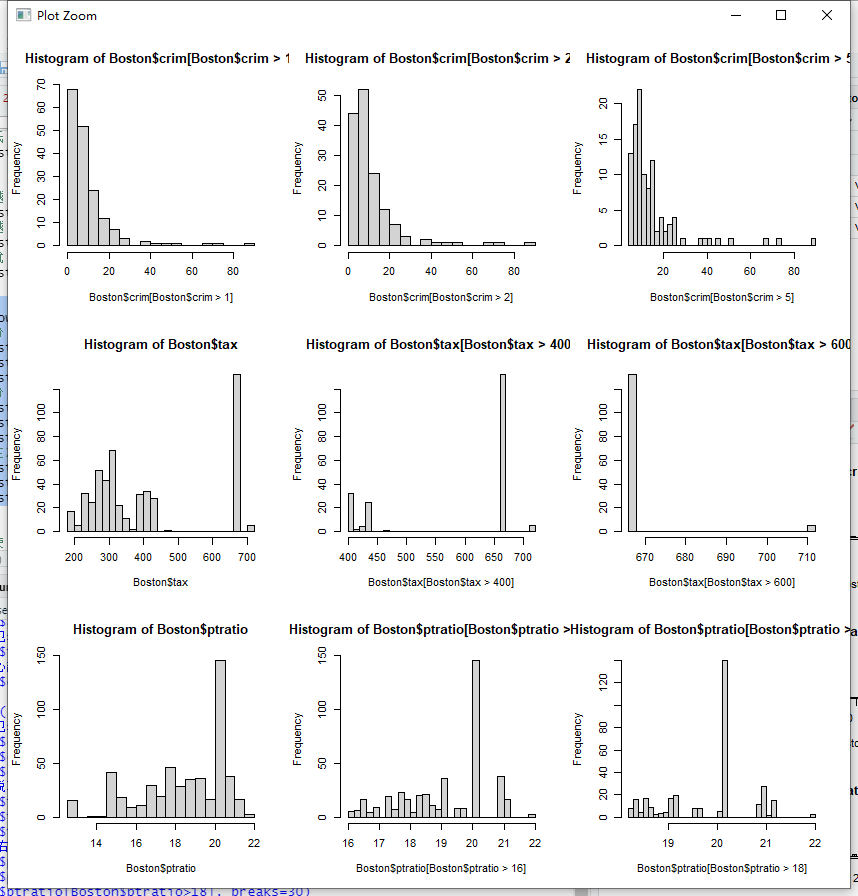
hist(Boston$tax[Boston$tax>600], breaks=30)

# 除了在20左右分布大量数据外，其他师生比率下分布比较均匀

hist(Boston$ptratio, breaks=30)

hist(Boston$ptratio[Boston$ptratio>16], breaks=30)

hist(Boston$ptratio[Boston$ptratio>18], breaks=30)



(e)

# 可以使用plyr包的count函数直接计数

library(plyr)

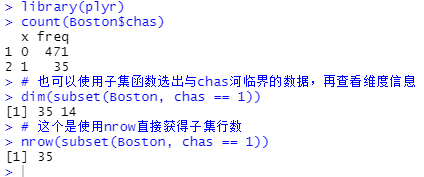
count(Boston$chas)

# 也可以使用子集函数选出与chas河临界的数据，再查看维度信息

dim(subset(Boston, chas == 1))

# 这个是使用nrow直接获得子集行数

nrow(subset(Boston, chas == 1))



(f)

median(Boston$ptratio)



(g)

# 通过使用子集函数选出中值最小的那一条/几条数据

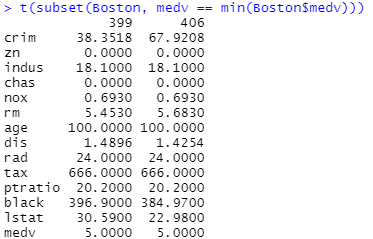
t(subset(Boston, medv == min(Boston$medv)))

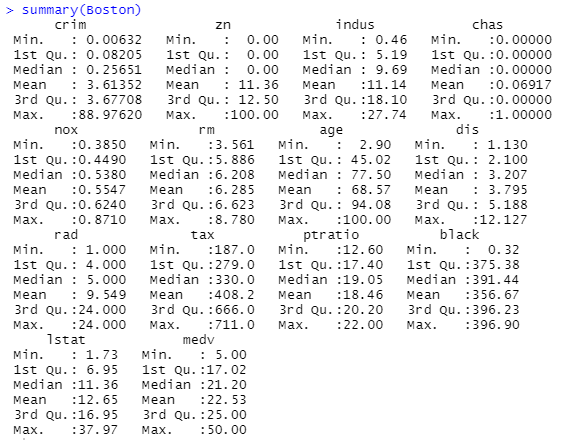
# 最小的两条数据是399河406

summary(Boston)

# 对比其他的总体数据的分布来说，399的黑人比例是全局最高

# 而其他的数据对比来说不是最好也不是最差





(h)

# 平均居住房间数超过7个

dim(subset(Boston, rm > 7))

# 平均居住房间数超过8个

dim(subset(Boston, rm > 8))

summary(Boston)

summary(subset(Boston, rm > 8))

# 比较发现犯罪率和低地位人口占比都比较低

# 某种程度上说明住宅房间数越多，地区居民越富足优渥

