高级统计方法 第二次作业:

序号：1 姓名：杨程瑜 学号：20202241107 班级：网2001

**概念**

1.问题

（a）问题（略）

光滑度高的模型适合。预测变量数p很小时，这样容易欠拟合，样本量大时光滑度高更接近真实的函数f，同时变量数小不容易过拟合。

（b）问题（略）

光滑度低的模型适合。样本量小预测变量数p很大时光滑度高容易过拟合。

（c）问题（略）

光滑度高的模型适合。光滑度低的模型不好拟合非线性关系，光滑度高的模型更容易拟合非线性关系。

（d）问题（略）

光滑度低的模型适合。光滑度高的模型会拟合很多误差项，导致方差变大。

2.问题（略）

（a）问题（略）

需要一个回归模型，因为因变量是公司CEO的工资。感兴趣的是关于美国500强公司的推断。

n：美国500强公司

p：美国500强公司的利润、员工人数、产业类型

（b）问题（略）

需要一个分类器，原因是其为关于是否成功的二分类问题。感兴趣的是产品是否成功的预测。

n：先前研发的20个相近产品

p：先前研发的20个相近产品的数据：失败或成功的状态、产品的价格成本、市场预算、竞争价格等其他10个变量

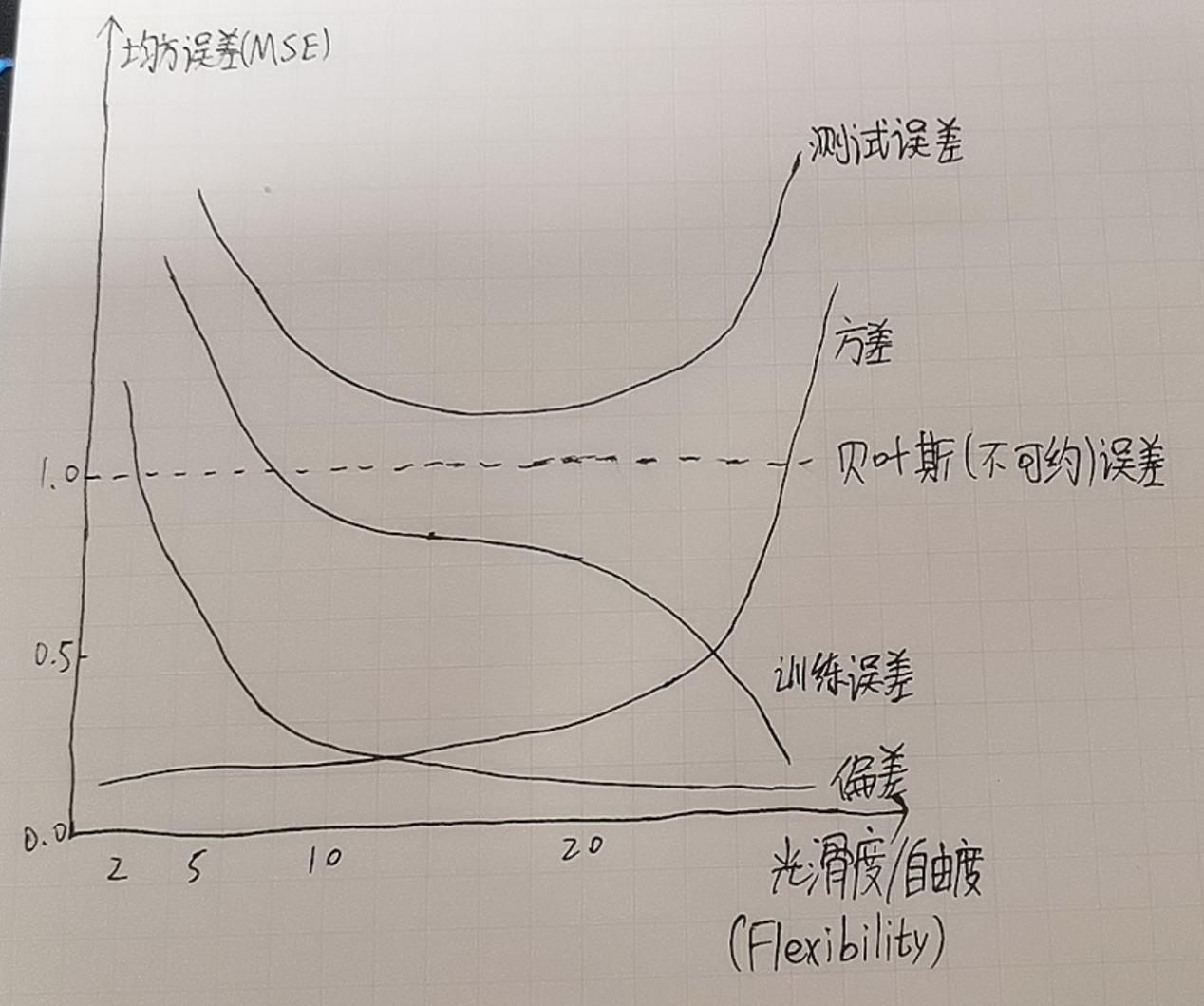
（c）问题（略）

需要一个回归模型，因为研究的是美元百分比变化的规律。感兴趣的是预测。

n：2012年的52个星期

p：每周的美元百分比变化：美国市场百分比变化、英国市场百分比变化、德国市场百分比变化

1. 问题（略）
2. 问题（略）



1. 问题（略）

偏差：光滑度/自由度增加而单调减少，现象是会产生更紧密更精确的拟合

方差：光滑度/自由度增加而单调增加，现象是会产生过拟合

训练误差：由于光滑度/自由度增加二单调减少，现象是会产生更紧密更精确的拟合

测试误差：光滑度/自由度增加而先单调减少后单调增加（上凹型曲线），现象是先产生更紧密更精确的拟合，然后产生过拟合

贝叶斯（不可约）误差：不随光滑度/自由度变化，是均方误差的下限

1. 问题（略）
2. 问题（略）

识别手写数字：响应变量是数字类别（0-9），预测变量为扁平化的图片像素值特征。应用的目的是进行预测，原因是我们只对预测结果感兴趣。

泰坦尼克号乘客生存预测：响应变量为是否生还，预测变量为乘客的身高体重年龄等信息。应用的目的是进行推断，原因是我们对哪些预测变量使得乘客生还率高感兴趣。

鸢尾花种类预测：响应变量为花的种类，预测变量为花的四种特征。应用的目的是进行预测，原因是我们只对预测结果感兴趣。

1. 问题（略）

房价预测：响应变量为房价，预测变量为目标房屋的地理位置、房屋结构等多个特征。应用的目的是进行预测，原因是我们只对预测结果感兴趣。

用户活跃度预测：响应变量为用户活跃度，预测变量为用户的上线时间、浏览时长等多个特征。应用的目的是进行推断，原因是我们对那些使得活跃度高的预测变量感兴趣。

季度销售量预测：响应变量为季度销售量，预测变量为前若干年各个季度的销售量。应用的目的是进行预测，原因是我们只对预测结果感兴趣。

1. 问题（略）

新闻类别聚类、用户类别聚类、城市类型聚类

1. 问题（略）

光滑度高的回归/分类模型：

优点是能够更紧凑更精确的拟合模型，偏差更小

缺点是需要计算的参数更多，并且容易产生过拟合，方差偏大

光滑度低的回归/分类模型：

优点是需要计算的参数比较少，不容易产生过拟合，方差偏小

缺点是拟合的模型不够精确，偏差较大

当自变量和因变量的关系大概率可能为非线性，以及数据样本点很多、参数数量少时，光滑度较高的模型更好

当自变量和因变量的关系大概率可能为非线性，以及数据样本点很多、参数数量少时，光滑度较高的模型更好

6.问题（略）

参数模型的优点：把将要估计的f限定为一组参数中，简化了具体函数的形式，不需要拟合任意一个函数f。

参数模型的缺点：选定的模型不一定和实际的函数f形式上一致，从而模型不准确，同时也有过拟合的可能。

非参数模型的优点：不对模型做出明确的假设，不限定函数f的形式，可以从更大范围中选择模型，同时模型也更接近真实的函数情况。

非参数模型的缺点：无法将函数限定为一定数量的参数中，一般需要大量的观测点才能进行拟合。

7.问题（略）

（a）问题（略）

点1：

点2：

点3：

点4：

点5：

点6：

1. 问题（略）

当K=1时，最近的点为点5，预测结果为Green

1. 问题（略）

当K=3时，最近的3个点为点5、点6、点2，占多数的为Red，所以预测结果为Red

（d）问题（略）

所期望的最优K值应该比较小。对于非线性决策边界，较小的K将是更灵活的，

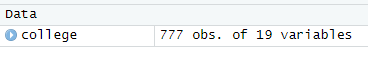
而较大的K会尝试拟合更线性的边界。

**应用**

8.问题（略）

（a）问题（略）

college = read.csv("College.csv")

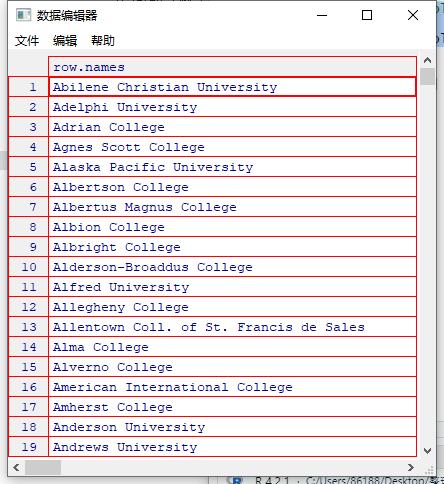


（b）问题（略）

college = read.csv("College.csv")

rownames(college) = college[,1]

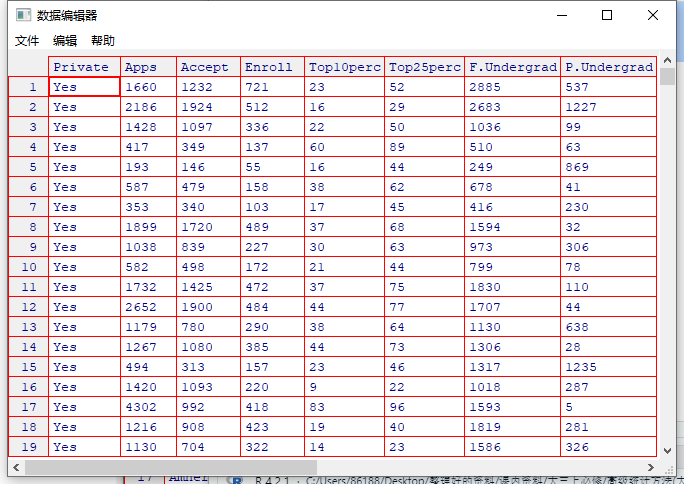
fix(college)



college = read.csv("College.csv")

college = college[,-1]

fix(college)

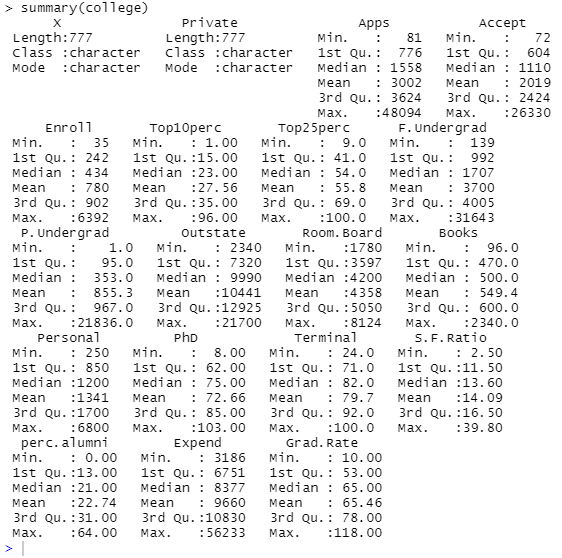


（c）问题（略）

i.问题（略）

college = read.csv("College.csv")

summary(college)



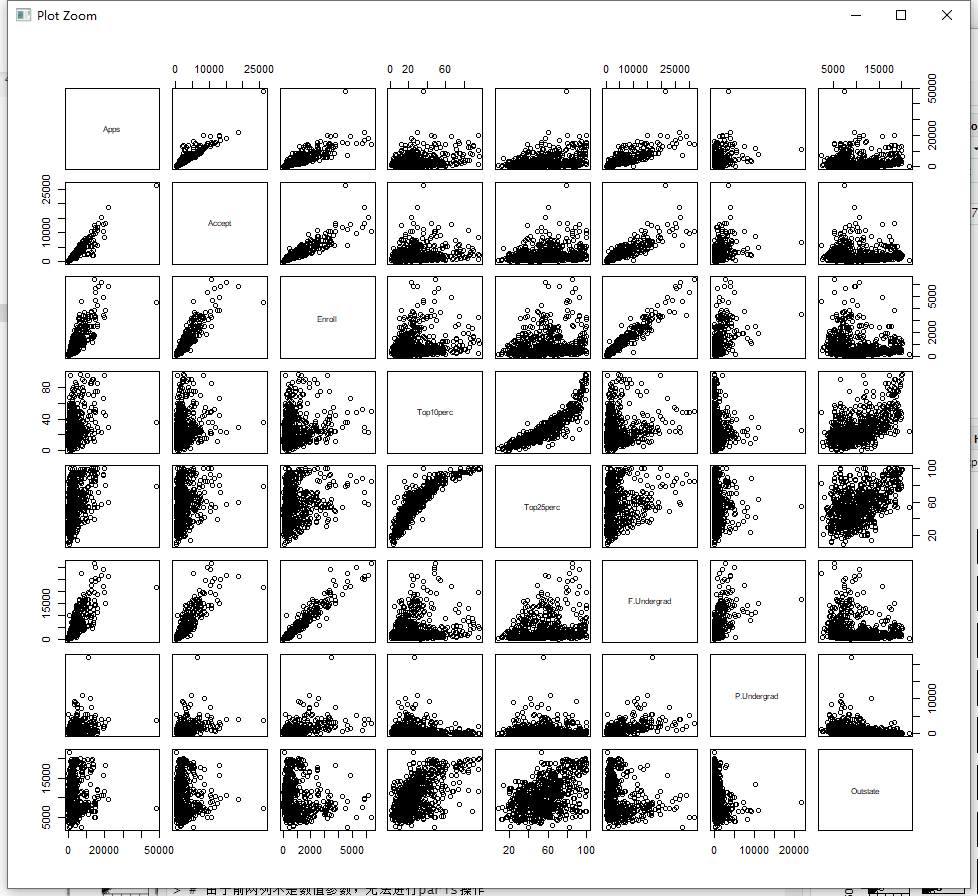
ii.问题（略）

college = read.csv("College.csv")

# 由于前两列不是数值参数，无法进行paris操作

# 故paris从第三列开始

pairs(college[,3:10])



iii.问题（略）

画箱线图需要使用到ggplot2库，首先需要下载ggplot2库

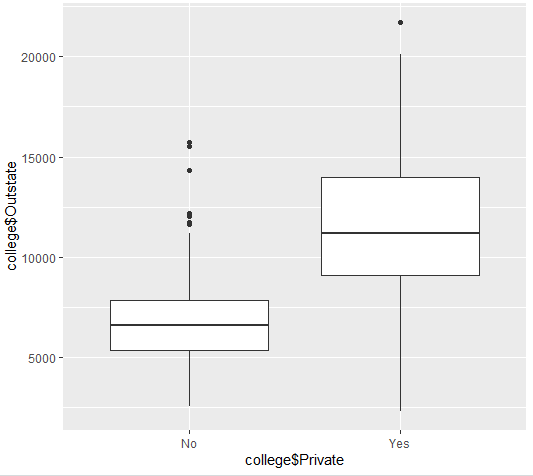
install.packages("ggplot2")

下载完成后使用library(ggplot2)加载ggplot2库

library(ggplot2)

college = read.csv('College.csv')

ggplot(college, aes(x=college$Private, y=college$Outstate)) + geom\_boxplot()



iv.问题（略）

college = read.csv('College.csv')

Elite = rep("No", nrow(college))

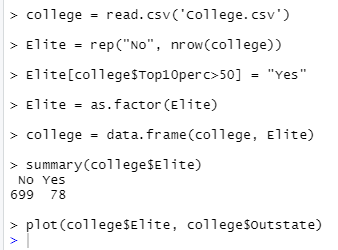
Elite[college$Top10perc>50] = "Yes"

Elite = as.factor(Elite)

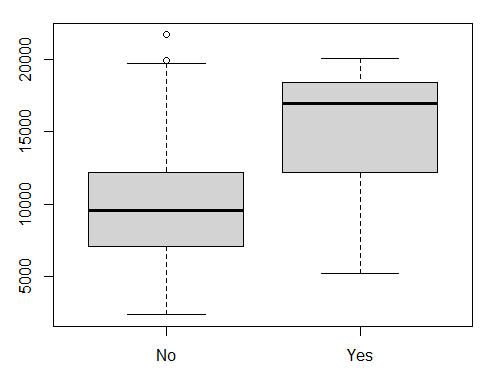
college = data.frame(college, Elite)

summary(college$Elite)

plot(college$Elite, college$Outstate)



可以看出共有78所精英大学



v.问题（略）

college = read.csv('College.csv')

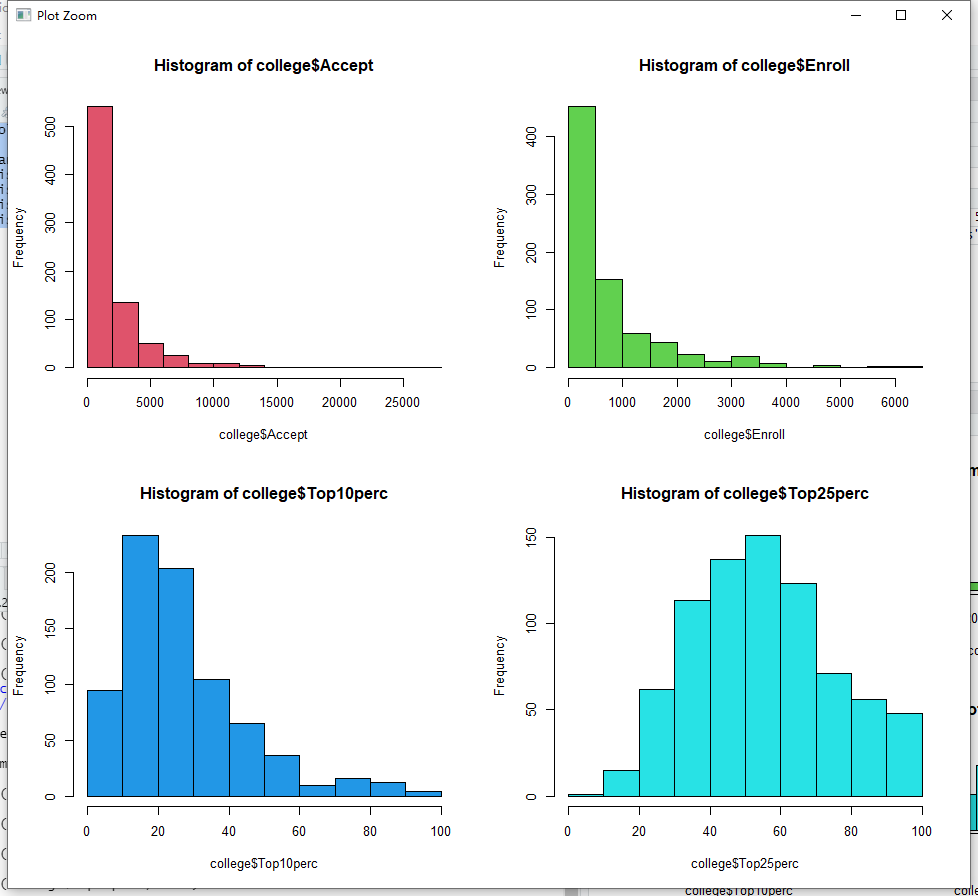
par(mfrow=c(2,2))

hist(college$Accept,col=2)

hist(college$Enroll,col=3)

hist(college$Top10perc,col=4)

hist(college$Top25perc,col=5)



vi.问题（略）

下面对大学的接受率与注册率，以及前25%和前10%的新生之间的关系画散点图进行研究

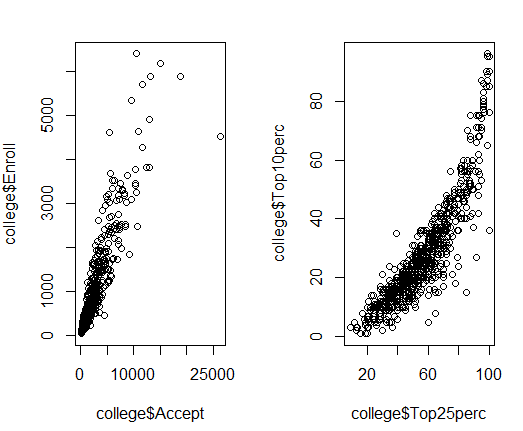
college = read.csv('College.csv')

fix(college)

par(mfrow=c(1,2))

plot(college$Accept, college$Enroll)

plot(college$Top25perc, college$Top10perc)



从图中能够看到随着大学的接收率增长，大学的注册率基本上单调增加。

同时大学中随着是高中班前25%的新生的增长，其是高中班前10%的新生的概率也是单调增加，并且增加的幅度越来越大。