高级统计方法 第2次作业:

序号： 27 姓名： 王琪瑞 学号： 20202241014 班级：软2002

**概念**

1.问题

（a）问题（略）

光滑度更高的模型好，光滑度更高方法将使数据更接近并符合，在大样本量情况下，采用光滑度低的方法容易出现欠拟合。

样本量n非常大，预测变量数p很小时，光滑度更高的统计学习模型好。

（b）问题（略）

光滑度低的模型好，由于预测个数小，光滑度高的方法容易发生过拟合的情况。预测变量数p非常大，预测个数n很小时，光滑度更低的统计学习模型更合适。

（c）问题（略）

光滑度更高的好，光滑度更高方法将使数据更接近并符合，在非线性的情况下，高光滑度能让拟合效果更好。

（d）问题（略）

光滑度低的统计模型，光滑度低的统计模型方差小，光滑度高的统计模型方差更大，在方差很大的时候，采用光滑度低的统计学习模型。

2.问题（略）

（a）问题（略）

回归模型，最感兴趣的是推断。

n为500强公司。

p为利润，员工人数，产业类型。

推断利润，员工人数，产业类型与CEO的工资的关系。

（b）问题（略）

分类器，最感兴趣的是预测。

n为先前研发的20个相近产品。

p为价格成本，市场预算，竞争价格和其他的10个变量。

预测新的产品是成功还是失败。

（c）问题（略）

回归模型，最感兴趣的是预测。

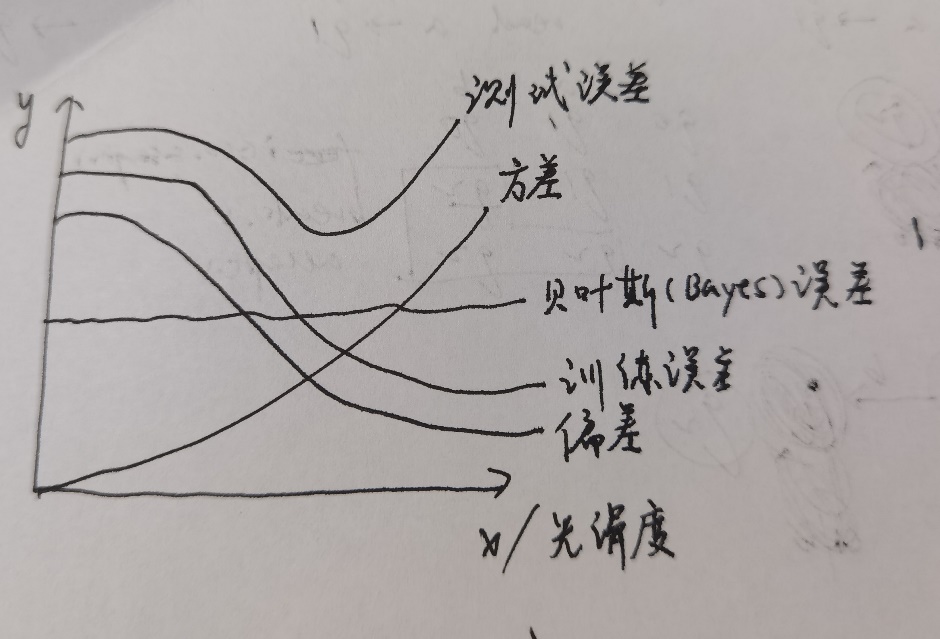
n为2021年的各周。

p为每周数据中的，美国，英国，德国市场百分比。

预测美元的百分比变化率。

3.问题（略）

（a）问题（略）



（b）问题（略）

1.偏差（平方）曲线是随着光滑度增加而单调减小的，拟合随着光滑度提升越来越接近。

2.方差随着光滑度增加单调增加，光滑度太大时发生过拟合现象。

3.训练误差曲线是随着光滑度增加而单调减小的，拟合随着光滑度提升越来越接近。

4.测试误差是上凹曲线，先随着光滑度增加，会有更好的拟合效果，光滑度不断增加会发生过拟合现象，拟合效果又会变差。

5.贝叶斯误差，定义下限，测试误差由于输出中的误差导致的不可约误差。当训练误差低于贝叶斯误差时就发生了过拟合。

4.问题（略）

（a）问题（略）

1.新冠核酸检测，目标是推断，响应变量是是否感染病毒，预测变量是核酸检测结果。

2.股票分析，目标是预测，响应变量是上涨，下跌，预测变量是近日走势。

3.产品检测，目标是预测，响应变量是成功，失败，预测变量是先前研发产品数据。

（b）问题（略）

1.CEO工资分析，目标是推断，响应变量是CEO工资，预测变量是行业，员工等。

2.疾病分析，目标是预测，响应变量是死亡年龄，预测变量是年龄，性别，身体各指标水平。

3.某货币市场变化率分析，目标是预测，响应变量是货币变化率，预测变量是各市场的变化。

（c）问题（略）

1.文物分类。根据化学成分进行分类。

2.物品销售。根据消费者的消费偏爱来针对进行产品销售。

3.视频推荐。根据用户的兴趣特征来推荐短视频。

5.问题（略）

1.光滑度高的回归模型或者分类模型的优点是，会获得更好的拟合效果，来减小偏差；缺点则是，需要估计更多的参数，还会容易发生过拟合现象和增大方差。

2.当模型的预测结果更需要准确性而不是方法的可解释性时，选择一个光滑度更高的模型会比光滑度低的模型更好。

3.相反，当模型的预测结果更加注重方法的可解释性而不是结果的准确性时，光滑度低的模型会更适合。

6.问题（略）

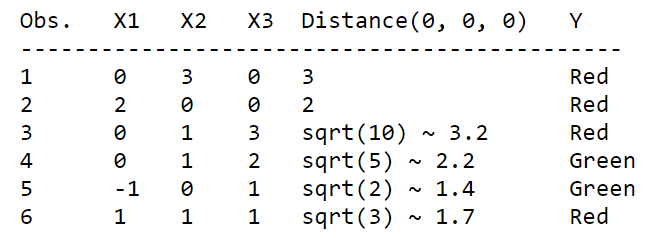
1.不同：参数模型是假设函数f具有一定的形式或形状，再通过估计一组参数的方式来估计f，观察的数据量较小；非参数方法不需要对函数f的形式事先做明确的假设，但为了获得更为精确的估计，往往需要大量的观测数据。

2.优点：将函数f简化为少量参数，而不需要那么多的观测结果，也更容易得出结果。

3.缺点：如果假设函数f的形式或形状有错误或者不够准确就会让估计结果不准确，并在假设模型的时候可能会出现欠拟合或者过拟合的问题。

7.问题（略）

（a）问题（略）



（b）问题（略）

绿色。5号观测点数据是K=1的最近邻，为绿色。

（c）问题（略）

红色。2号，5号，6号观测点数据是K=3的最近邻，分别是红色，绿色，红色，所以结果是红色。

（d）问题（略）

期望最优的K值应该是小。越大的K值会越接近线性，而越小的K值会越光滑。

**应用**

8.问题（略）

（a）问题（略）

脚本：

install.packages("ISLR2")

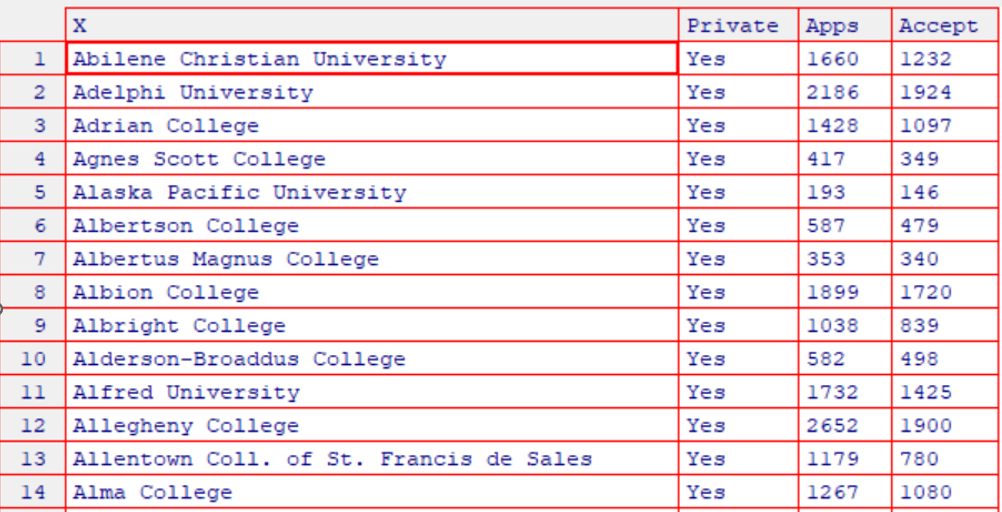
library(ISLR2)

write.csv(College,"College.csv")

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

fix(college)

截图：



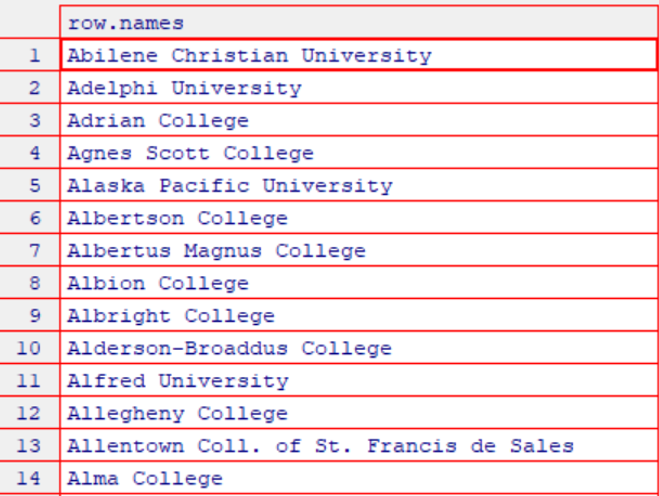
（b）问题（略）

脚本1：

rownames(college)=college[,1]

fix(college)

截图：



脚本2：

college=college[,-1]

fix(college)

截图：



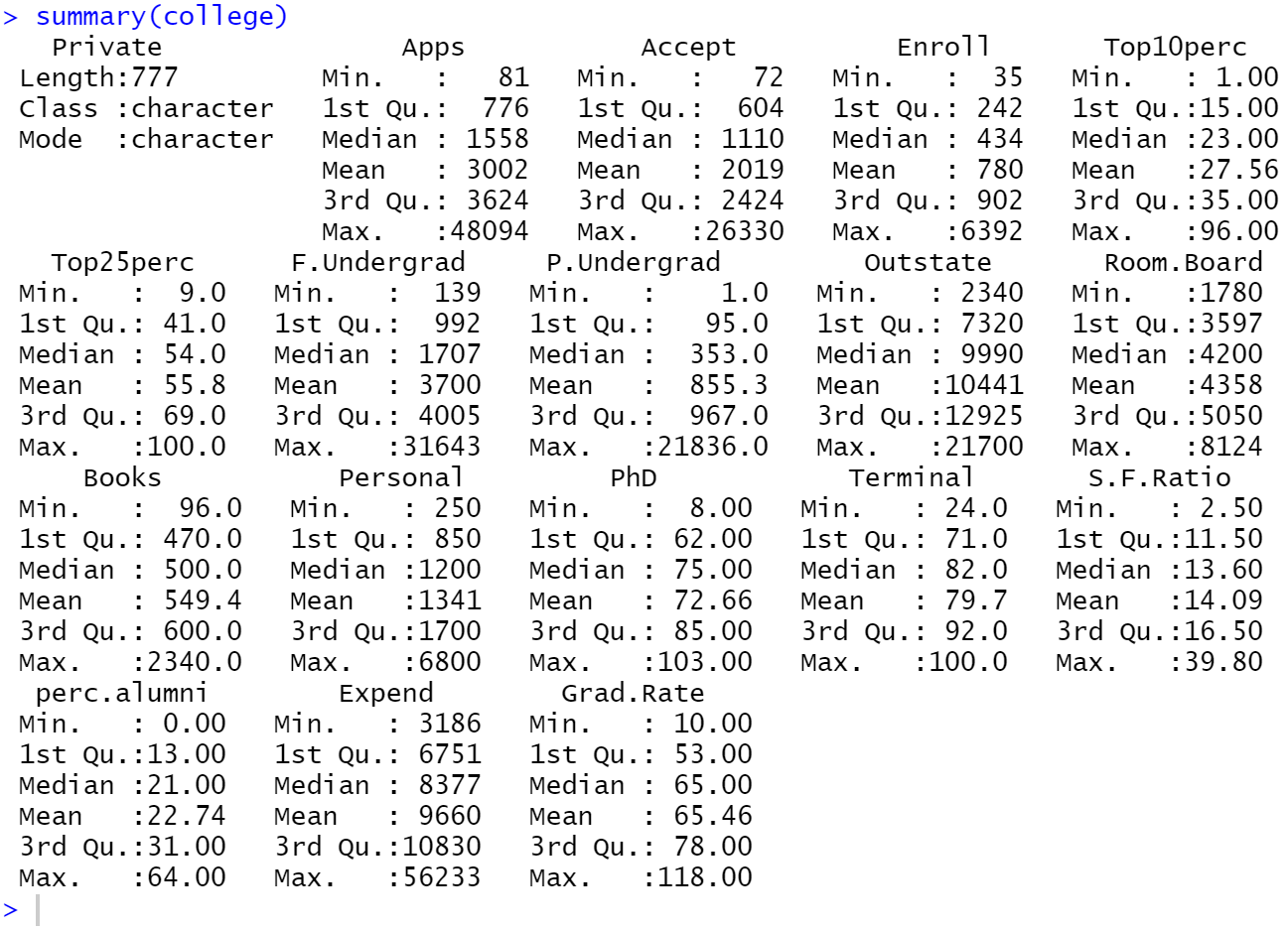
（c）问题（略）

i．

脚本：

summary(college)

截图：

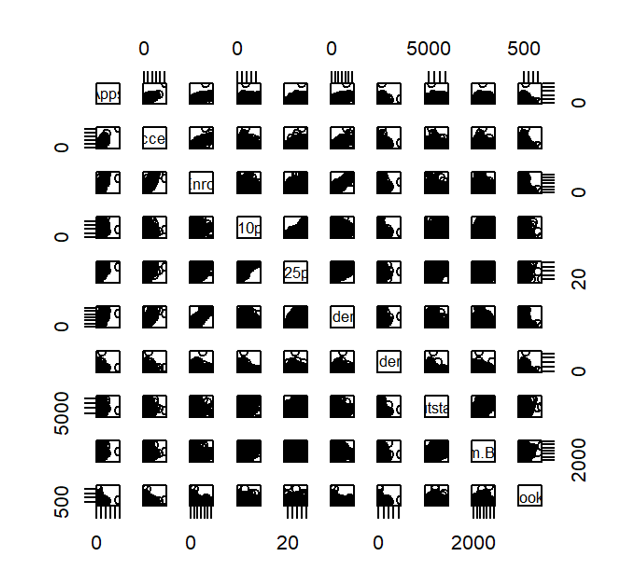


ii．

脚本：

pairs(college[,2:11])

截图：

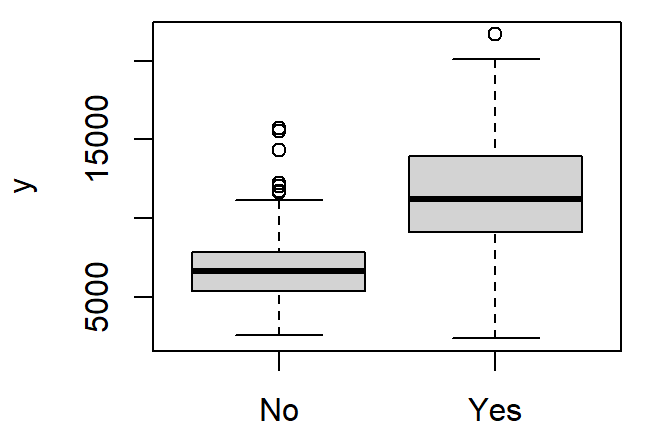


iii．

脚本

plot(factor(college$Private), college$Outstate)

截图：



iv：

脚本：

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

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

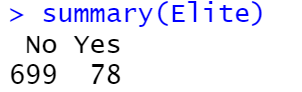
Elite = as.factor(Elite)

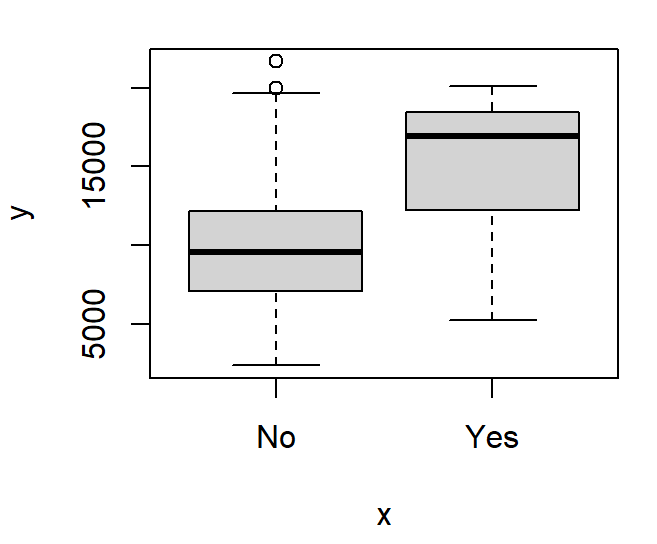
college = data.frame(college, Elite)

summary(college$Elite)

plot(college$Elite, college$Outstate)

截图：





v．

脚本：

par(mfrow=c(2,2))

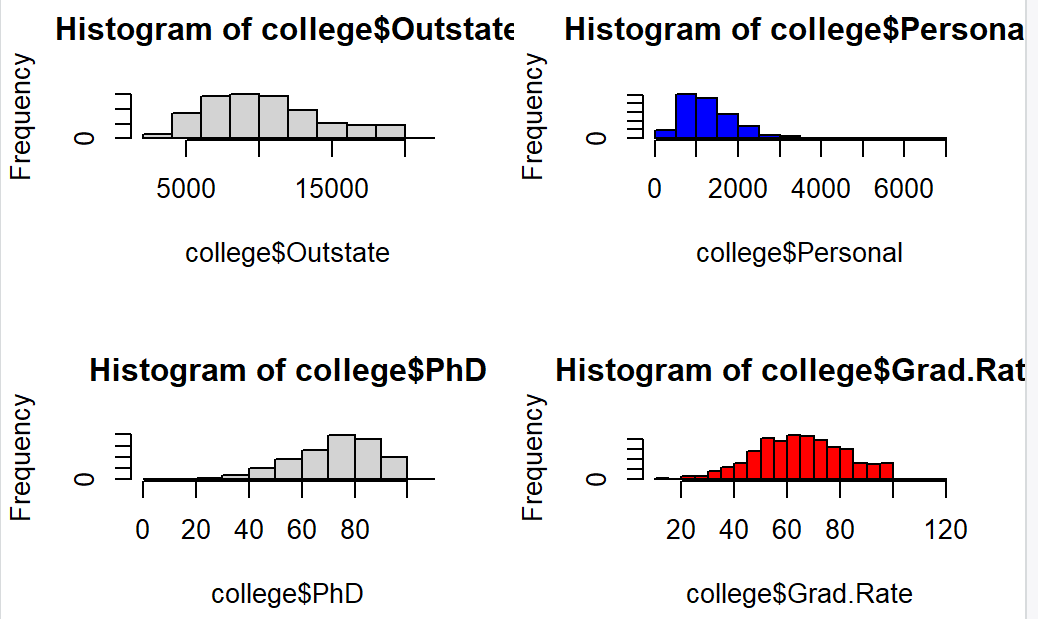
hist(college$Enroll)

hist(college$P.Undergrad,col=3)

hist(college$PhD,breaks=13)

hist(college$Grad.Rate,breaks=20,col=4)

截图：

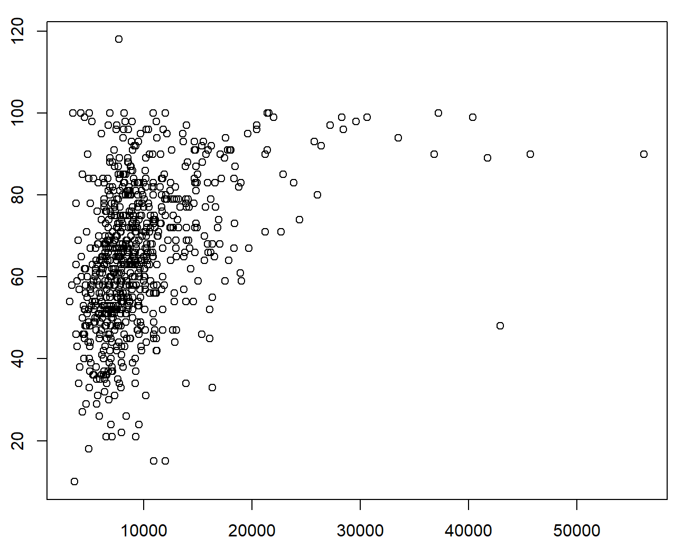


vi：

脚本：

plot(college$Expend,college$Grad.Rate)

截图：



当学校的人均支出越多，高校学生的毕业率也越高。