《基因组学数据分析》第一次作业

2023-9-26

考察内容:

1. 熟悉 R 语言的基础操作:读写数据，查看数据，应用基础函数等;
2. 简单的数据统计:求和，求均值，求相关系数等;
3. 简单线性回归分析;
4. 熟悉基本画图操作:散点图，箱线图等。

参考内容:

《统计学习导论 基于 R 应用》第二章和第三章。涉及的数据集可从教材官方网站下载: https://www.statlearning.com/resources-second-edition

作业内容:

A. 参考 38 页第 8 题，用教程推荐的 College 数据集，进行以下分析:

1. 从教材官方网站下载 College.csv 数据，用 read.csv() 函数将该数据读入 R。

提示:?read.csv 可帮助理解函数 read.csv() 的用法。

2. 观察数据，给出数据基本信息：数据类型，行列数。
3. 将数据的第一列设置为数据行名，之后删除第一列数据。

提示:rownames() 函数可以对行名进行操作。

4. 分别计算 Apps 和 Accept 两个变量的平均值以及两个变量的相关系数。
5. 使用 summary() 函数对前三列数据进行汇总，解释输出中 1st Qu. 和 3st Qu. 的含义。
6. 用 pairs() 函数对前五列数据产生一个散点图矩阵。
7. 分别用 plot() 和 boxplot() 函数产生 Outstate 对 Private 变量的箱线图。
8. 定义 Elite，对每一个大学，根据是否有超过 50%的学生来自排名在前 10%顶尖高中的情况将大学分为 Yes/No 两组;用 plot() 函数 产生 Outstate 对 Elite 的箱线图，并将 Yes 组 置于箱线图的左侧。

提示:factor() 函数中的 levels 变量可以设置 Yes/No 的顺序。

9. 用 hist() 函数对 Apps 和 Accept 两个变量制作直方图。
10. 将新的带有 Elite 信息的 College 数据输出到本地。

提示:可使用 write.table() 或 write.csv() 函数。

B. 参考 85 页第 8 题，用教程推荐的 Auto 数据集，完成以下分析:

11. 从教材官方网站下载 Auto.csv 数据，将该数据读入 R。

提示:该数据集有缺失值，读入时可设置 na.strings="?"，然后用 na.omit(Auto) 删除带有缺失值的行。

12. 当响应变量是 mpg，预测变量是 horsepower 时，使用 lm() 函数 完成一个简单的线性回归，并回答:预测变量和响应变量之间有关系吗?关系有多强?是正相关还是负相关?

当 horsepower=98 时，mpg 的预测值是多少?相应的 95%置信区和预测区间分别是多少?
13. 绘制响应变量和预测变量的关系图，用 abline() 函数显示最小二乘回归线。
14. 用 plot() 函数生成最小二乘回归拟合的诊断图，并分析拟合中的问题。

C. 解决赌徒谬误问题
一晚上手气不好的赌徒总认为再过几把之后就会风水轮流转，幸运降临。我们理智上知道这是错的，但在生活中 难以控制自己避免错误。假设一抛硬币问题，设定正面结果为+1，反面结果为-1,每一轮中抛出的硬币取值总和为 X。
15. 模拟 10 次抛硬币的取值总和 X。
提示:sample 函数和 sum 函数
16. 使用计算机模拟 200 轮(可设定变量名 Ntrials)，10 抛(可设定变量名 Nflips)序列，抽出连续抛出五次正面为开端的序列，求这些 序列的 X 的平均值。
提示:which 函数和 apply 函数(建议写成一个函数，变量分别为 Ntrials 和 Nflips)
17. 改参数 Ntrials=2000 和 8000 后,X 的均值收敛到 0 了吗?试从数学上解释该原因。
18. 固定 Ntrials=5000,当 Nflips=200，500，1000，5000，50000 时的结果如何?
使用分布直方图来展示 X 的分布，并计算 X 的均值和标准差。
19. 当序列越来越长时(即随 Nflips 数目增大)，与 X 的弥散度相比，X 的均值变得微不足道。体会“异常历史对后继行为的影响不会消失，而是随时间推移被稀释。”(无需作答)

作业要求:
1.请将 R 代码(hw1_name.R)和结果(hw1_name.pdf)打包成文件“作业 1-姓名-学号-学校”。教学网用户请在教学网通过教学网=>课程作业提交。非教学网用户发送到邮箱 genomics2023@163.com，邮件主题为《基因组学数据分析》第一次作业;
2.作业严禁照搬和抄袭;
3.本次作业占课程总成绩 20%;
4.请在 10 天内提交(2023 年 10 月 6 日 24:00 之前)