第一章 introduction

wxz

目录

1 introduction

我们相信上帝，其他的都得用数据证明。

量化社会科学研究中，社会学家用数据分析去理解人类和社会有关问题，他们的研究成果对社会的个体成员、政府政策和商业惯例有着重大影响。

近几十年由于数据革命和计算革命的发生，数据的数量和多样性激增，数据分析工具增加，各种数据集和数据来源增加，促进了量化社会科学的发展。

数据分析的三个要素：研究情境、编程技术、统计方法

1.1 本书概述

第一章：**介绍**。本章在于介绍如何最优地利用本书，介绍 **R 语言**（一种流行的开源统计编程环境）和 **Rstudio**（免费提供多种功能的软件包），两个

练习结尾。

第二章：**介绍因果关系**。我们必须推断观察不到的反事实的结果。

第三章：**度量的基本概念**。**精确的度量**对数据驱动的研究而言至关重要。度量的偏差有可能导致错误的结论和误导性的政策。**潜在的不可观测的度量**也是很重要。

第四章：**预测问题**。主要应用是利用选前民调预测大选结果。分析一个心理实验，即看候选人的面部照片并要求评估他的能力，引入线性回归模型，考察了”均值回归”的现象。讨论何时利用回归模型估计因果效应。介绍在观察研究中进行因果推断的断点回归设计。

第五章: **不同类型的数据中发现模式**。数据中**一致性**的模式，比如这里就可以用于《红楼梦》的前半部分作者和后半部分作者的创作风格一致性的研究上。介绍**地理空间数据**，讨论了斯诺的经典空间数据分析，研究 1854 年伦敦霍乱爆发的原因。使用美国选举数据为示例，演示如何通过创建地图进行可视化空间数据。

第六章：**概率**。如何进行估计参数和作出预测。介绍频数和贝叶斯法则、随机变量和概率分布、大数定律和极限中心定理。

第七章：**如何量化我们的估计和不确定性**。

1.2 如何使用本书

通过**实践**来学习数据分析。

http://press.princeton.edu/qss 去下载相应的代码和数据集

教学的时候应该采取”**特殊**——**一般**——**特殊**”的原则，指教师应该引入一个具体的例子来说明一个概念，然后提供一个对其一般的处理，最后将其运用于另一个具体的例子中。

在 R 中运行下列命令安装一个软件包：

```
install.packages("swirl")

library(swirl) # 加载软件包

install_course_github("kosukeimai","qss-swirl")
```

更多关于 swirl 的信息见 http://swirlstats.com/

1.3 R 的简介

R 语言是一个开源的统计编程环境（原生环境），可以用来处理各种数据集，绘制生动图形，广泛运用与学术界和工业界。

R 是用于统计分析、绘图的语言和操作环境。R 是属于 GNU 系统的一个自由、免费、开源的软件，它是一个用于统计计算和统计制图的优秀工具。R 语言是主要用于统计分析、绘图的语言和操作环境。

上官网下载适合自己电脑版本的 R 语言。

R 语言官方网站：https://cran.r-project.org/

官方镜像站列表：https://cran.r-project.org/mirrors.html

RStudio 是 R 的**集成开发环境 (IDE)**，它包括一个**控制台、支持直接执行代码的语法高亮编辑器**，以及用于绘图、历史、调试和工作空间管理的工具。

上官网下载适合自己电脑的 Rstudio 桌面版。

Rstudio 官网：https://www.rstudio.com/

1.3.1 算术运算

在 Rstudio 的控制台中我们可以直接输入算术的命令然后 enter 运行。

```
+ - * ^ sqrt()
```

1.3.2 对象

R 可以将信息存储为一个对象，这里我们要和参数赋值区分开来。利用 print() 函数打印。

(1) 对象名称

包含数字、字母、_ 等；

不能以数字开头；

不能含有特殊含义的字符；

不能有空格；

区分大小写；

避免使用 pi,if,for 等关键字；

(2) 对象的命名语句

对象名 <- 对象值

快捷键 alt + -

(3) 对象查询

```
ls()  # 返回所以内存中的对象名
ls(pat = 'b') # 返回所有对象名中包含 b 的对象名
ls.str() # 返回所有对象的具体信息
```

(4) 删除对象

```
rm(对象)/remove(对象名)
例如: rm(list=ls(pat='x',all.names = TRUE))
删除所有名字中含 x 的对象
all.names = TRUE 表示不显示以. 开头的这种特殊对象
```

(5) 对象基本类型

数值型 (numeric)

字符型 (character)

逻辑型 (logical)

因子型 (factor)

复数型 (complex)

1.3.3 向量

向量，是 R 中最重要的一个概念，它是构成其他数据结构的基础。R 中的向量概念与数学中向量是不同的，类似于数学上的集合的概念，由一个或多个元素所构成。

向量其实是用于**存储数值型、字符型或逻辑型数据**的**一维数组**。

用函数 c() 来创建向量。c 代表 concatenate 连接，也可以理解为收集 collect，或者合并 combine。

(1) 向量创建

第一种创建方法，应用 c() 函数

```
a <- c(1,2,3,4)  # 数值型向量

is.vector(a)  # 判断是否是向量

b <- c("one","two","three")  # 字符型向量

d <- c(TRUE,FALSE)  # 逻辑型向量
```

第二种创建方法，应用冒号生成等差数列

```
e <- c(1:6)

f <- 1:6   # 同 c(1:6)
```

第三种方法，rep() 函数生成重复向量

```
g <- rep(1,4)

rep(1:4,each = 2) # 对每个向量个体重复

rep(1:4,times = 2) # 对向量整体进行重复
```

第四种方法，seq() 函数生成等差向量

```
h <- seq(1,5)  # 默认是等差数列

h <- seq(1,6,by = 2)  # 向量中元素间距为 2

h <- seq(1,6,length.out = 3)   # 向量长度为 3

h <- seq(from = 1950,to = 2010,by = 10)
```

(2) 定义向量值中的名称

```
names(向量) <- 另一个向量值
```

为向量中每一个值赋予名称，就像是字典中的 keys

(3) 向量获取 (索引）

```
world.pop[1]  # 获取向量中第一个值
```

(4) 向量删除

```
world.pop[-2]  # 减去第二个值
```

(5) 向量值替换

```
world.pop[2] <- 5  # 替换第二个值为 5
world.pop[c(1,2)] <- c(5,6)   # 替换第一和二个值
```

1.3.4 函数

(1) 定义函数名

```
函数名 <- function(参数){具体内容}
```

(2) 常见函数

```
length() # 获取长度
mean() # 获取数据的平均值
sum() # 获取数据值的总和
max() # 获取数据的最大值
min() # 获取数据的最小值
range() # 获取数据的范围
```

1.3.5 数据文件

(1) 数据文件类型

主要用两种数据文件类型。

CSV: 表格数据

RData：包含数据集的 R 对象的集合

(2) 工作路径

下拉菜单 file->open file

或者在 file 窗口中直接选择

`setwd(" 工作文件夹路径") #` 打开文件夹作为工作路径

`getwd() #` 获取当前工作路径

(3) 表格数据和 RData 的相关操作

`read.csv() #` 获取表格数据

`load() #` 获取数据

`names() #` 获取变量名的向量

`nrows() #` 获取行数

`ncol(UNpop) #` 获取列数

`dim(UNpop) #` 返回行数和列数

`summary() #` 获取数据集的统计值

获取具体行和列的内容

`UNpop$world.pop #` 获取 world.pop 的列数据

`UNpop[, "world.pop"] #` 获取 world.pop 列的数据

`UNpop[c(1,2,3)] #` 获取第 1,2,3 列数据

`UNpop[1:3, "world.pop"] #` 获取 world.pop 列的前三个数据

1.3.6 保存对象

方法一：

在 environment 窗口中的 save 图标保存工作区，或者单击对话-> 将工作区另存为，然后选择位置保存，注意要修改扩展名为".RData"。

再次编写的时候要打开文件，单击对话-> 加载工作区，或者使用 `load()`。

例如：`load("Chapter.RData")` # 加载对象

方法二：

`save.image("xxx.RData")`

方法三：保存特定对象到 RData 文件中

`save(对象名称 1, 对象名称 2,file = " 文件名")`

方法四：保存特定对象到表格中

`write.csv(对象名称,file = " 文件名")`

spss 和 dta 文件相关获取和保存方法同 CSV

1.3.7 软件包

1.3.7.1 安装 一般有线上安装和本地安装两种方法

1. 可以从 R 的原生环境中下载，不常用

2. 可以从 Rstudio 中的 packages 窗口的 install 按钮中选择 CRAN（线上）或者 Packages Achive file(本地)

3.`install.packages(" 软件包")`

1.3.7.2 载入 `library(软件包)` # 注意不要加引号

1.3.7.2.1 单独加载包内某个函数 `car::vif()`

1.3.7.3 更新包 `update.package()` 更新所有包，逐个提示

更新指定包就以包名称作为参数即可

1.3.7.4 移除包 `remove.package()`

1.3.7.5 获取帮助

1.3.7.5.1 获取某个函数的帮助 `help("library")`

1.3.7.5.2 获取某个关键词的帮助 `??help`

`help.search("library")`

1.3.7.5.3 获取某个 package 的帮助 例如:`help(package = "gglot2")`

1.3.8 编程及学习技巧

(1) 运行

运行：选择点击 run 按钮或者 `ctrl+enter`(windows 系统中)

后台运行：`source("UNpop.R")`

(2) 注释

双注释字符 ## 是注释整行

单注释字符 # 是注释该行后面的内容

注意提示文件和作者以及功能

```
##
## File: UNpop.R
## Author: Wxz
## the code loads the UN population data and
## saves it as a Stata file
##
```

(3) 检查

`lint(".R 文件")` # 检查文件错误

(4) 代码说明

注意要学会使用 RMarkdown

1.4 总结

没啥，都在上面。

1.5 练习

自我汇报是否参加投票的偏差

了解世界动态