Data Analysis Using R: Chapter02

罗智超 (ROKIA.ORG)　　1814347@qq.com

1　通过本章你将学会

- 通过案例了解 R 语言的基本特征
- 了解数据集的结构
- 简单线性回归
- 掌握 tidy data 的基本形
- 一些小程序练习

2　R 语言自带数据

- R 里面自带了很多数据集，这样方便研究人员验证算
- 通过 data() 可以查阅所有数据集名称
- 通过 data(package="packagename") 来查阅 R 包里面自带的数据集名

3　Anscombe 数据

- 1973 年，统计学家 F.J. Anscombe 在Graphs in Statistical Analysis构造出了四组奇特的数据。它告诉人们，在分析数据之前，描绘数据所对应的图像有多么的重要
- 本章通过对 R 自带的的 Anscombe 数据进行简单的处理，让同学了解 R 语言的基本特

4 查看数据

- 看看数据集的结构有什么特征

```
attach(anscombe)
str(anscombe)
detach(anscombe)
head(anscombe)
tail()
```

- 计算下各个变量的统计

```
summary(anscombe)
```

- 基本绘图

```
boxplot(anscombe)
hist(anscombe$x4)

plot(y1~x1,data=anscombe)
lm.a<-lm(y1~x1,data=anscombe)

abline(lm.a)
plot(lm.a$residuals~x1,data=anscombe)
```

5 如何更改 Anscombe 数据的格

- 方法一（R BASE 函数）
- 方法二（library(dplyr)）
- 方法（循环）
- 方法四（library(reshape2)

6 如何绘图

- 方法一（基于原始数据格式）
- 方法二（基于 tidy 数据格式

7 代码分析

- anscombe.r

8 扩展案例

- 简单线性回

```
# 模拟成绩数据
str(cars)
cars.lm<-lm(dist~speed,data=cars)

cars.lm
summary(cars.lm)
coefficients(cars.lm)
confint(cars.lm)
fitted(cars.lm)
residuals(cars.lm)
anova(cars.lm)
vcov(cars.lm)
AIC(cars.lm)
plot(cars.lm)
predict(cars.lm)
```

- 收入和声望的关系

```
#bc: blue collar,wc:white collar, prof:professional

library(car)
str(Prestige)
head(Prestige)
scatterplot(prestige~income|type,data=Prestige,span=1,legend.coords="bottomright")
```

9 数据集的基本元素

- 变量（variable name）
- 记录（column name）
- 变量类型（数值、字符、因子、时间）

10 什么是 tidy 数据格式

- Hadley 发表在 Journal of statistical software 上的文章Tidy Data
- 该文章的源码地址
- 宽数据 VS 窄数

11 tidy data Rule

- Data that is easy to model, visualise and aggregate
- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

12 Tidy data Case

- Values in column names

  – Income distribution within U.S. religious groups

```
#Collected by Pew Research Center
#Examines the relationship between income and religion in the US
#i.e, which religions have the wealthiest adherents?
# 计算不同宗教哪个水平工资的人数最多
library(reshape2)
raw <- read.csv("data/pew.csv", check.names = F)
tidy <- melt(raw, id = "religion")
head(tidy)
# We can now fix the column names
names(tidy) <- c("religion", "income", "n")

# Alternatively
tidy <- melt(raw, id = "religion",variable.name = "income", value.name = "n")
head(tidy)
```

- Variable names in cells

  – Weather data

```
#Daily temperatures in Cuernavaca, Mexico for 2010
#1 - 31, days of month
#tmax, tmin, maximum and minimum temperatures
# 计算每天最低气温和最高气温的差值
raw <- read.delim("data/weather.txt",check.names = F, na.strings = ".")
# na.rm = TRUE is useful if the missing values don't have any meaning
raw.tidy <- melt(raw,id = c("year", "month", "element"),variable.name = "day", na.rm =

# reordering columns
```

```
raw <- raw[, c("year", "month", "day","element", "value")]
head(raw)

tidy <- dcast(raw, year + month + day ~ element,value.var = "value")
head(tidy)
```

– titanic2

```
# 思考下如何通过原始表，计算存活率
titanic2 <-read.csv("data/titanic2.csv",stringsAsFactors = FALSE)
head(titanic2)
# 计算不同类别的存活率 = 存活人数/（存活人数 + 死亡人数）

#Step 1
tidy <- melt(titanic2, id = c("class", "age", "fate"),variable.name = "gender")
head(tidy)
#Step 2
tidy <- dcast(tidy, class + age + gender ~ fate,value.var = "value")
head(tidy)
#Step 3
tidy$rate <- round(tidy$survived /(tidy$survived + tidy$perished), 2)
head(tidy)
```

13 小练习

```
#Forbes2013 top 2000 company
#Statistics by Wuxizhi P13
w<-read.csv("data/Forbes2000.csv")
names(w)
summary(w)
View(w)
```

```
w[w[,3]=="China",2]


par(mfrow=c(2,2))

for(i in 4:7)
{
  hist(log(w[,i]),main=paste("Log",names(w)[i]),xlab="")
  rug(log(w[,i]))

}

C<-w[w[,3]=="China",]
G<-w[w[,3]=="Germany",]
par(mfrow=c(1,2))
hist(C$Market.Value,20,main="Histogram of Market Value (China)",col=3,prob=T,ylim=c(0,0
lines(density(C$Market.Value),lwd=2)
hist(G$Market.Value,20,,main="Histogram of Market Value (German)",col=2,prob=T,ylim=c(0
lines(density(G$Market.Value),lwd=2)
```

14 本周 “大牛

- 弗朗西斯·高尔顿（Francis Galton1822216 日－ 1911117 日），英国科学家和探险家。他曾到西南非洲探险，因树立功绩而知名并被选为英国皇家地理学会会员，三年后又入选英国皇家学会，晚年受封为爵士。他的学术研究兴趣广泛，包括人类学、地理、数学、力学、气象学、心理学、统计学等方面。他是查尔斯·达尔文的表弟，深受其进化论思想的影响，把该思想引入到人类研究。他着重研究个别差异，从遗传的角度研究个别差异形成的原因，开创了优生学。他关于人类官能的研究开辟了个体心理和心理测验研究的新途径