用 rmarkdown 写作实例

WISERCLUB 王泽贤

2016 年 *10* 月 *11* 日

简介

本文是用 *rmarkdown*书写的一个例子. 介绍的是如何用 R 进行简单的**线性回归**. 请将最终文档和 Rmd 文档配合学习.

本文仅仅介绍了用 R 进行数据分析的一个粗略流程, 目的是希望通过本文例子让初学者对 R 和 Rmarkdown 有初步的了解.

数据结构和函数仅简要介绍本文例子中用到的几个, 更多函数编程, 画图, 统计分析等内容请参考文中出现的**课后阅读**推荐的学习资料以及参加后续课程.

rmarkdown 的优势

1. 内嵌代码代码和写作协同完成, 无需在多个软件中复制粘贴. 代码灵活性强, 不止是 R.
2. 可重复性研究.
3. 生成图表格式统一由代码定义, 后续基本上无需调整. 可以和 R 强大的绘图功能结合.
4. 支持 latex 数学公式. 美观大方.
5. 沉浸式写作, 舍弃了繁多的排版设定.
6. 语法简单, 上手快速.
7. 输出格式多样, 一份文档, 多种呈现. 后续仍可用 html 语法或 word 进行二次细化美化编辑.

markdown 语法说明

1. “#“置于一段的开头用于定义标题, 有 n 个”#“表示第 n 级标题.
2. markdown 中回车不换行,2 个以上空格 + 回车, 或者 2 次回车才是换行.
3. 在一句话前面加上 “1. 空格” 则可以加上有序标记并自带缩进效果. 如本句.
4. 给一段字两端加上”*“或”_“则成为斜体字, 如简介中的”rmarkdown“.(导出中文 pdf 的时候可能失效, 问题来源于 tex 系统兼容性)
5. 两端加上”**“则为粗体字, 如正文的” 线性回归 “.
6. 两端加上”~~“则为删除线.
7. 需要分页时插入 \pagebreak

新建 Rmarkdown 的方法:File-New File-R Markdown

使用前需要先安装 knitr 包, 并把 Tools-Global Option-Sweave-Weave Rnw files using 选择 knitr, Typeset LaTex into PDF using 选择 XeLaTex.

数学公式与内嵌代码

数学公式

如

$$Y_i = e^{\beta_1+\beta_2X_i+\epsilon_i}$$

$$Y_i = \frac{1}{e^{\beta_1+\beta_2X_i+\epsilon_i}}$$

$$Y_i = \beta_1 + \beta_2(\frac{1}{X_i}) + \epsilon_i$$

markdown 语法说明

1. rmarkdown 支持 latex 的数学公式格式.
2. 用 $ 两端包围的公式可以显示在一个句子内部, 用 $$ 两端包围的公式会独占一行显示. 没用 $ 符号标记的代码不会被识别为公式.
3. 用 “-空格 “可以给一句话加上无序标记并自动缩进 (如下第 4 点）.
4. 上述方程的对应代码为

   $$Y_i=e^{\beta_1+\beta_2X_i+\epsilon_i}$$

   $$Y_i=\frac{1}{e^{\beta_1+\beta_2X_i+\epsilon_i}}$$

   $$Y_i=\beta_1+\beta_2(\frac{1}{X_i})+\epsilon_i$$

常用的公式对应代码

- 上标:a^x输出a^x, b^{xy}输出b^{xy}.
- 下标:a_x输出a_x, b_{xy}输出b_{xy}.
- 分数:\frac{a}{b}, 输出$\frac{a}{b}$.
- 积分号:\int_0^1, 输出$\int_0^1$.
- 求和号:\sum_{i=1}, 用单个 $输出$\sum_{i=1}$, 用 2 个 $输出

  $$\sum_{i=1}$$

  .
- 极限:\lim_{x \rightarrow 0}输出

  $$\lim_{x \rightarrow 0}$$

  .

课后阅读:

更多 *latex* 数学公式规则请见深入学习资料中的 *Latex-intro-ByTobias-cn.pdf(*第三章*)*

内嵌代码

在 rmarkdown 中可以轻松的内嵌代码, 输出最终文件的时候会自动将所有 Rmd 文件中的代码按顺序重新运行, 并将结果放在代码对应的位置.

内嵌代码块的方式是用一个特殊标记将代码段包括在内, 具体标记方式请看 Rmd 文件中的写法.

比如前文的公式, 放入标记符号中的范围中就可以显示成代码:

$$Y_i=e^{\beta_1+\beta_2X_i+\epsilon_i}$$

$$Y_i=\frac{1}{e^{\beta_1+\beta_2X_i+\epsilon_i}}$$

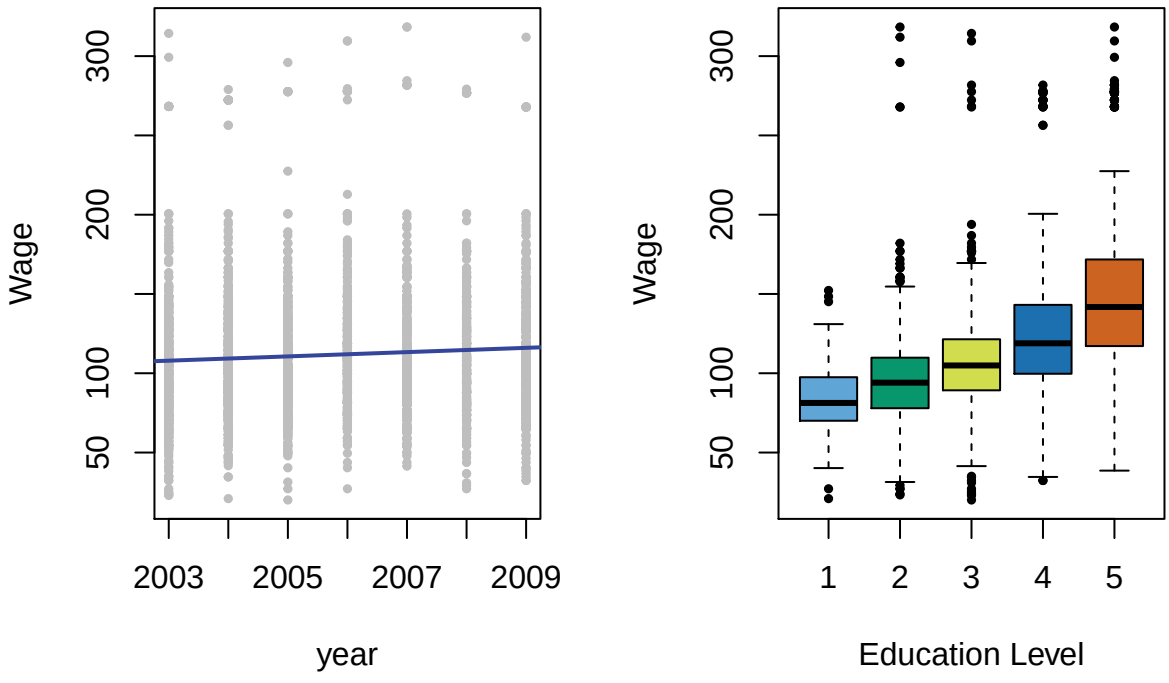
$$Y_i=\beta_1+\beta_2(\frac{1}{X_i})+\epsilon_i$$

如果需要内嵌 R 代码则需要在开头的标记符后面接上 {r}, 则成为以下结果 (具体标记方法请见 Rmd 文件中对应位置)

```
1+2
```

```
## [1] 3
```

默认是显示代码加上输出结果. 可以通过点击代码块右上角的齿轮按钮对代码块命名, 选择 output 设置可以设定更多细节, 如



这是选择了 show output only 的结果, 绘图代码可以查看 Rmd 文件. 具体用法请参加我们后续课程的绘图部分.

除了代码块以外, 也可以嵌入行内代码比如以下内容:

1+2 的值为 3

实际上是通过:

```
# 1+2 的值为 `r 1+2`
```

(由于转码显示问题, 上式以 Rmd 文件中形式为准)

再比如:

```
x = 5
```

然后直接在 markdown 中书写就有:

5 的平方是 25

实际上, 上面这句话是通过以下代码实现:

```
# `r x`的平方是 `r x^2`
```

这一点在论文模型解释的时候相当方便. 如果你所有模型解释中涉及到的系数都是通过行内代码引用得到的, 那么当你需要更改数据, 或者改变估计方法, 只要参数个数和位置没变, 那么只需要更换一下数据源, 再点一下上面的 knit, 所有的参数结果就都自动更正过来了.

对于 r 语言代码块, 右上角的下箭头可以运行前面所有的 r 语言代码块, 右箭头则是运行当前代码块. 或者可以像在 R script 中一样选中后运行

课后阅读:

除了 *R* 也可以运行 *python,js* 等, 有兴趣的同学可以参考

knitr Language Engines

http://rmarkdown.rstudio.com/authoring_knitr_engines.html

数据读取

除了在 rstudio 基本操作中介绍的函数可以读取用户数据外,R 和 R 的 packages 中也内置了一些数据可供使用, 如

```
dataclass1 <- women
```

则将 women 数据载入到 dataclass1.

课后阅读:

更多关于数据读取的函数请阅读并尝试课后推送的深度阅读材料中*:*

Basic_Data_Structures 文件夹内的 *"Import_data.pdf"* 或 *"Import_data.html"* 的具体函数代码. 作者为 *2015* 年 *WIS-ERCLUB* 负责人丑高武

除了直接在 environment 窗口中点击 dataclass1 来预览数据, 还可以

```
head(dataclass1)
```

```
##   height weight
## 1     58    115
## 2     59    117
## 3     60    120
## 4     61    123
## 5     62    126
## 6     63    129
```

head 函数的作用是查看目标的前几行 (默认 6 行).

如果想要取得具体某一个变量

```
h <- dataclass1$height
head(h)
```

```
## [1] 58 59 60 61 62 63
```

$ 可以作为引用符, 这样 height 变量中的数据就被赋值到 h 变量中了

数据结构

上述的 dataclass1 是 R 的基本数据结构中的一种:data.frame. 其中的数据以 R 的数据类型之一 numeric 储存.

查看数据结构可以通过:

```
str(dataclass1)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ height: num  58 59 60 61 62 63 64 65 66 67 ...
##  $ weight: num  115 117 120 123 126 129 132 135 139 142 ...
```

R 中包含了以下几种基本数据结构, 按照纬度和内含元素是否同质可以分为

纬度	Homogeneous	Heterogeneous
1d	Atomic vector	List
2d	Matrix	Data frame
3d	Array	

数据结构中可以容纳的数据类型包括 logical(逻辑), integer(整数), double(双精度小数, 一般也叫 numeric), character(字符), 还有 complex(复数) 和 raw.

课后阅读:

具体数据结构细节请阅读并尝试课后推送的深度阅读材料中*:*

Basic_Data_Structures 文件夹内的 *"Basic_Data_Structures.html"* 和 *"Basic_Data_Structure_2.html"* 的具体函数代码. 作者为 *2015* 年 *WISERCLUB* 主负责人邓光宏.

也可以直接阅读由 *R* 语言大师 *Hadley Wickham* 所著《*Advanced R*》的 *Data structures* 章节 *(*英文*)*

Advanced R-Data structures

"http://adv-r.had.co.nz/Data-structures.html"

rmarkdown 语法说明

对于简单的表格可以通过用 | 和 -分割的方法进行制表. 具体请看 Rmd 文件.

描述性统计与探索性分析

拿到数据以后在建模之前我们需要对数据形态进行观察, 进而选择适合的模型. 最简单的操作包括计算描述性统计量, 画图等.

简单的一些统计量

```
summary(dataclass1)
```

```
##      height         weight
##  Min.   :58.0   Min.   :115.0
##  1st Qu.:61.5   1st Qu.:124.5
##  Median :65.0   Median :135.0
##  Mean   :65.0   Mean   :136.7
##  3rd Qu.:68.5   3rd Qu.:148.0
##  Max.   :72.0   Max.   :164.0
```

summary 函数用在一个数据集上可以获得各个变量的一些描述性统计量, 如均值, 极值等.

其他函数还有

```
mean(h) # 求 H 变量的均值
```

```
## [1] 65
```

```
cor(dataclass1$height,dataclass1$weight) # 求 height 和 weight 变量的相关系数
```

```
## [1] 0.9954948
```

```
cov(dataclass1$height,dataclass1$weight) # 求 height 和 weight 变量的协方差
```

```
## [1] 69
```

```
sd(h) # 求 h 变量的标准差
```

```
## [1] 4.472136
```

可以看到 weight 和 height 变量相关性很高.

表格生成

我们可以把描述性统计量整合成一个表格

```
stattable <- rbind(c("weight",mean(dataclass1$weight),sd(dataclass1$weight)),
                   c("height",mean(dataclass1$height),sd(dataclass1$height)))
library(knitr)
kable(stattable, col.names = c(" 变量"," 均值"," 标准差"),caption = "Table 1", align = "c")
```

Table 2: Table 1

变量	均值	标准差
weight	136.733333333333	15.4986942614378
height	65	4.47213595499958

kable 函数中 caption 参数定义的是表的名字. 该函数会自动统计 pdf 文件中有多少个表格, 并默认在表名前面加上 Table1,Table2.

前面的表格因为不是用 kable 函数生成的, 所以没有自动加 Table1.

kable 只是一个简单的出表函数, 此外还有 xtable, tables, ascii, pander 等函数可以用于出表和对表进行修改. 具体可在安装对应包后在 R 控制台中 help 查看.

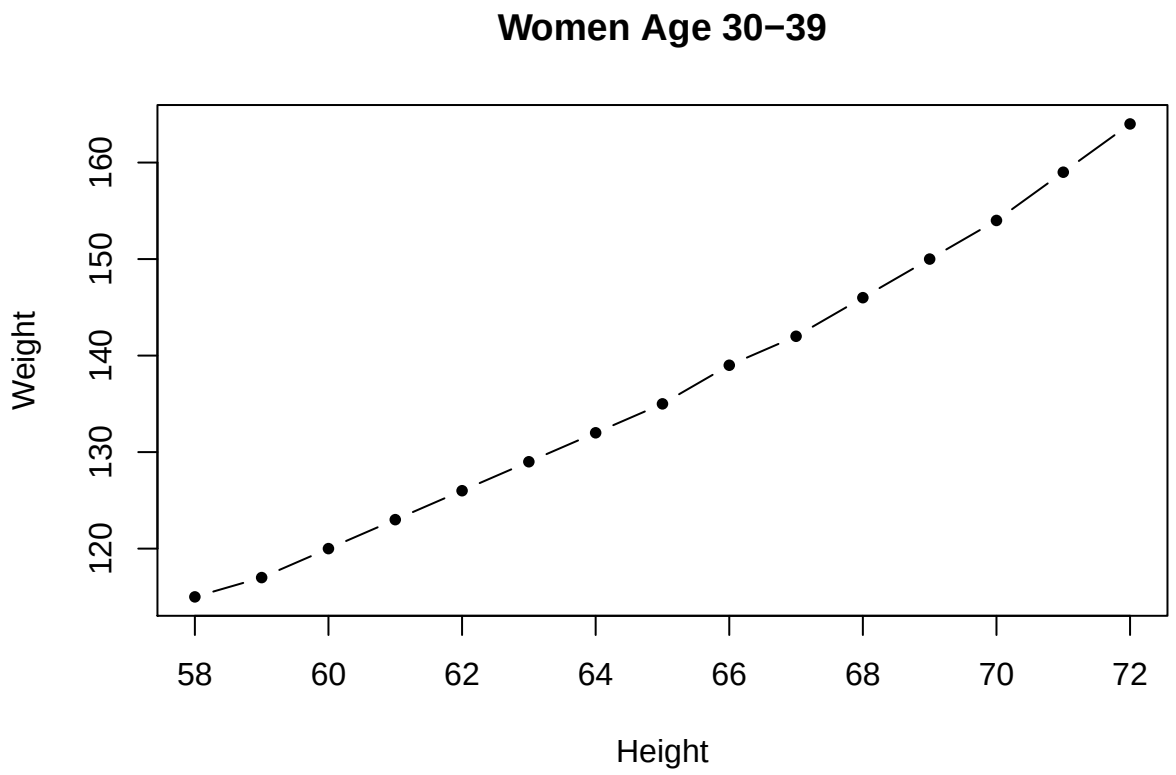
更多数据处理相关内容请参加我们后续的数据处理模块课程学习

画图

plot 函数为最基本的绘图函数, 也提供了最基本的绘图功能. 在下面的例子中, 使用格式为:

plot(x 轴变量,y 轴变量,main="图名",xlab="横轴名称",ylab="纵轴名称",pch = 点形状, cex = 缩放倍数, type = 线点类型)

```
plot(women$height,women$weight,main="Women Age 30-39",xlab="Height",ylab="Weight", pch = 16,cex =0.8, type = "b")
```



更多绘图技巧请参加我们后续的绘图模块课程学习

线性回归

lm 函数

R 中自带了 lm 函数可以用于线性回归. 用法:lm(y 变量 ~x1 变量 +x2 变量 +…. , data = 变量所在的数据集), 对 lm 回归后的对象 (fit) 使用 summary 可以得到线性回归的各个统计量

```
fit <- lm(weight ~ height, data = women)
summary(fit)
```

```
##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7333 -1.1333 -0.3833  0.7417  3.1167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
## height        3.45000    0.09114   37.85 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 on 13 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9903
## F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

我们可以尝试加入 height 变量的平方项:

```
fit2 <- lm(weight ~ height + I(height^2), data=women)
summary(fit2)
```

```
##
## Call:
## lm(formula = weight ~ height + I(height^2), data = women)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50941 -0.29611 -0.00941  0.28615  0.59706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 261.87818   25.19677  10.393 2.36e-07 ***
## height       -7.34832    0.77769  -9.449 6.58e-07 ***
## I(height^2)   0.08306    0.00598  13.891 9.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3841 on 12 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic: 1.139e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

获取结果

我们可以用 $ 引用符直接引用其中的参数如:

```
fit$coefficients[2]
```

```
## height
##   3.45
```

写出模型

比如我们要写出 fit 对应的模型:

代码：

```
# $$weight = `r fit$coefficients[1]` + `r fit$coefficients[2]` \times height$$
```

效果：

$$weight = -87.5166667 + 3.45 \times height$$

之后可以进行模型评价与解释等

更多统计建模相关内容请参加我们后续的统计分析模块课程学习

其他

如何输出最终文件

如果需要输出新建 rmarkdown 文件时指定格式意外的格式, 点击上面 knit 按钮右边的小三角形即可选择.

对于 html 和 word 格式无需额外配置, 中英文均可直接输出.

对于英文 pdf 则需要安装 tex 系统.ctex, texlive 等均可. 建议安装 ctex

ctex 下载地址, 建议下载稳定完整版 (1.34G)

http://www.ctex.org/CTeXDownload

对于中文 pdf, 除了安装好 ctex 等包含中文组建的 tex 系统外, 还需要进行以下配置:

1. 把课程资料中的 Chinese template 文件夹内的 “header.tex” 文件和需要输出中文的 Rmd 文件放在同一个文件夹下.
2. 把 Rmd 开头的 output 下属选项改成以下内容再输出 pdf：

```
output:
  pdf_document:
    latex_engine: xelatex
    includes:
        in_header: header.tex
```

3. 或者安装完 ctex 后直接用本节课资料中 Chinese template 文件夹中附带的模版 Rmd 文件 “Chinese template.Rmd” 进行修改 (需确保 “header.tex” 文件和 Rmd 文件在同一个文件夹内).

在 rmarkdown 中插入网址:

代码:

```
[厦门大学经济学院](http://economic.xmu.edu.cn/)
```

中括号 [] 内的是网址超链接要显示的文字.

效果：

厦门大学经济学院

在 rmarkdown 中插入图片

```
![wiserclub](http://i.imgbox.com/qNJQcj7P.jpg)
```

效果：

Figure 1: wiserclub