talk11 练习与作业

目录

0.1 练习和作业说明

将相关代码填写入以 “‘{r} “‘ 标志的代码框中，运行并看到正确的结果；

完成后，用工具栏里的”Knit” 按键生成 PDF 文档；

将 PDF 文档改为：姓名**-**学号**-talk11** 作业**.pdf**，并提交到老师指定的平台/钉群。

0.2 talk11 内容回顾

待写..

0.3 练习与作业：用户验证

请运行以下命令，验证你的用户名。

如你当前用户名不能体现你的真实姓名，请改为拼音后再运行本作业！

```
Sys.info()[["user"]]
```

```
## [1] "lucas"
```

```
Sys.getenv("HOME")
```

```
## [1] "/Users/lucas"
```

0.4 练习与作业 1：linear regression

0.4.1 一元回归分析

用 `readr` 包的函数将 `Excercises and homework/data/talk11/` 目录下的 `income.data_.zip` 文件装入到 `income.dat` 变量中，进行以下分析：

1. 用线性回归分析 `income` 与 `happiness` 的关系；
2. 用点线图画出 `income` 与 `happiness` 的关系，将推导出来的公式写在图上；
3. 用得到的线性模型，以 `income` 为输入，预测 `happiness` 的值；
4. 用点线图画出预测值与真实 `happiness` 的关系，并在图上写出 R2 值。

```
## 代码写这里，并运行；
library(readr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
```

```
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

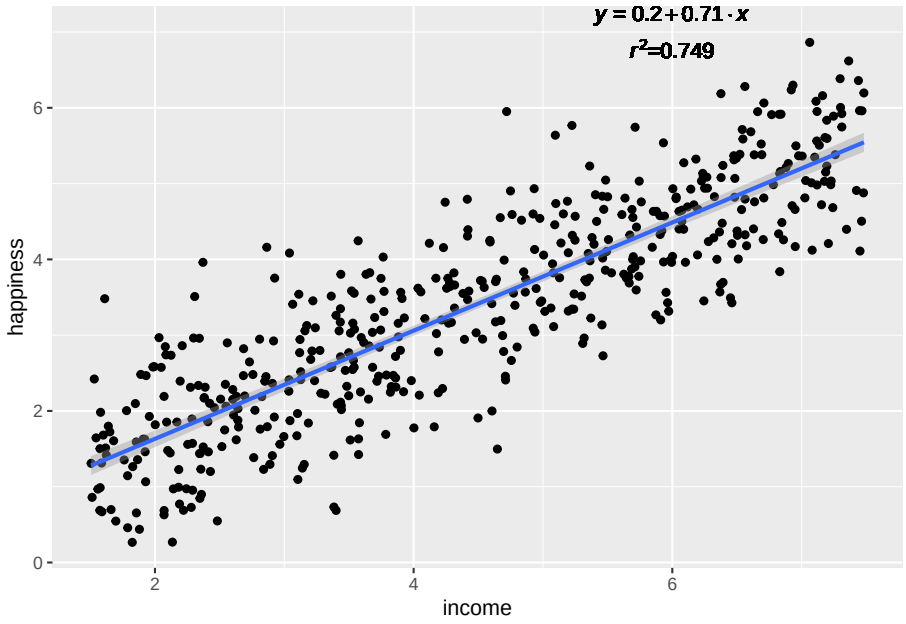
```
income.dat<-read_csv("data/talk11/income.data_.zip")
```

```
## New names:
## Rows: 498 Columns: 3
## -- Column specification
```

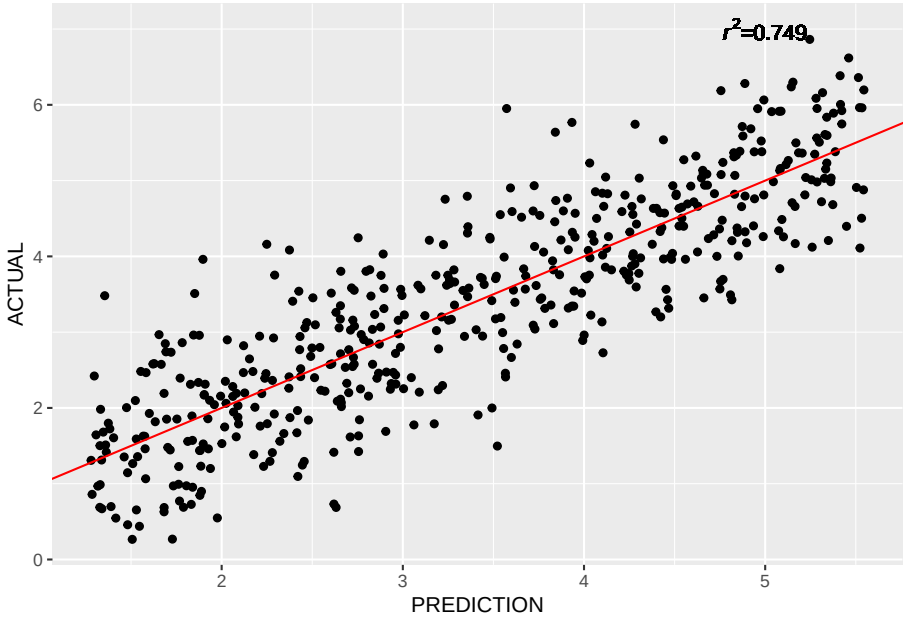
```
## ------------------------------------------------------ Delimiter: "," dbl
## (3): ...1, income, happiness
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
resM=lm(happiness~income,income.dat)



eq<-substitute(atop(paste( italic(y), " = ", a + b %.% italic(x), sep = ""),
paste(italic(r)^2,"=",r2)),
list(a = as.vector( format(coef(resM)[1], digits = 2) ),
b = as.vector( format(coef(resM)[2], digits = 2) ),
r2 = as.vector( format(summary(resM)$r.squared, digits = 3) ))
)

eq<-as.character(as.expression(eq))

ggplot(income.dat,aes(x=income,y=happiness))+geom_point()+geom_smooth(method = "lm")+ge
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
predictions<-resM%>%predict(income.dat)
pre_act<-data.frame(PREDICTION=predictions,ACTUAL=income.dat$happiness)
resC=lm(ACTUAL~PREDICTION,pre_act)

eq2<-substitute(paste(italic(r)^2,"=",R2),list(R2=as.vector( format(summary(resC)$r.squ
eq2<-as.character(as.expression(eq2))
pre_act%>%
  ggplot(aes(PREDICTION,ACTUAL))+geom_point()+geom_abline(intercept = 0,slope = 1,color
```



0.4.2 多元回归分析

用 `readr` 包的函数将 `Excercises and homework/data/talk11/` 目录下的 `heart.data_.zip` 文件装入到 `heart.dat` 变量中，进行以下分析：

1. 用线性回归分析 `heart.disease` 与 `biking` 和 `smoking` 的关系；
2. 写出三者间关系的线性公式；
3. 解释 `biking` 和 `smoking` 的影响（方向和程度）；
4. `biking` 和 `smoking` 能解释多少 `heart.disease` 的 variance？这个值从哪里获得？
5. 用 `relaimpo` 包的函数计算 `biking` 和 `smoking` 对 `heart.disease` 的重要性。哪个更重要？
6. 用得到的线性模型预测 `heart.disease`，用点线图画出预测值与真实值的关系，并在图上写出 R2 值。
7. 在建模时考虑 `biking` 和 `smoking` 的互作关系，会提高模型的 R2 值吗？如果是，意味着什么？如果不是，又意味着什么？

```
## 代码写这里，并运行;
library(relaimpo)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
##
##     melanoma

## Loading required package: survey

## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:boot':
##
##     aml

## The following object is masked from 'package:caret':
##
##     cluster

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

## Loading required package: mitools

## This is the global version of package relaimpo.

## If you are a non-US user, a version with the interesting additional metric pmvd is a

## from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.
```

```
heart.dat<-read_csv("data/talk11/heart.data_.zip")
```

```
## New names:
## * `` -> `...1`
```

```
## Rows: 498 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): ...1, biking, smoking, heart.disease
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
model<-lm(heart.disease~biking+smoking,data = heart.dat)

# 2. 写出三者间关系的线性公式
coef(model)
```

```
## (Intercept)      biking     smoking
##  14.9846580  -0.2001331   0.1783339
```

```
# heart.disease = 14.9846580 - 0.2001331 * biking + 0.1783339 * smoking

# 3. 解释 biking 和 smoking 的影响（方向和程度）
# biking: 每增加 1 个单位的 biking, heart.disease 平均减少 0.2001331 个单位
# smoking: 每增加 1 个单位的 smoking, heart.disease 平均增加 0.1783339 个单位

# 4. biking 和 smoking 能解释多少 heart.disease 的 variance? 这个值从哪里获得?
res<-summary(model)
R2<-res$r.squared
R2
```

```
## [1] 0.9796175
```

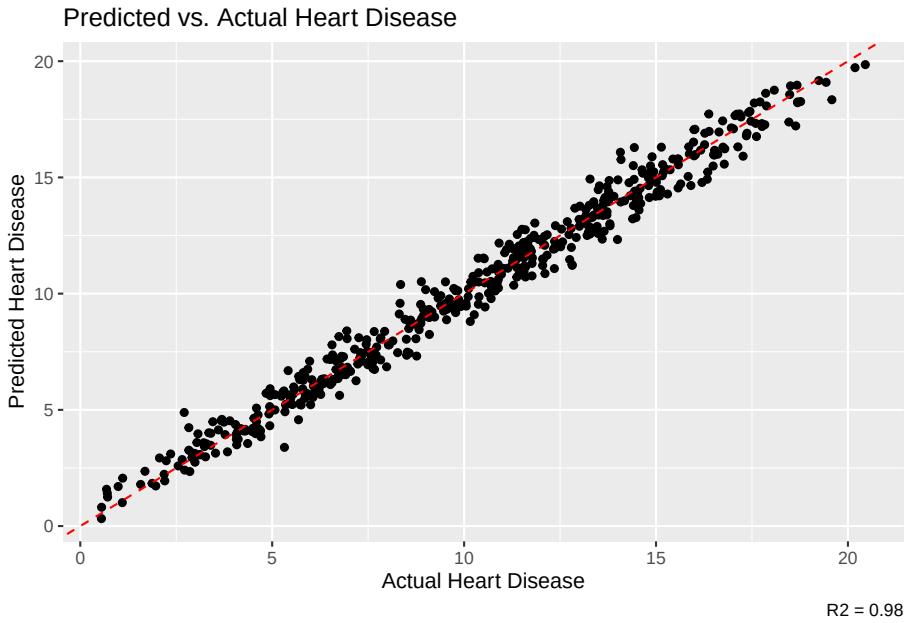
```
# 可以通过 summary(model) 函数获得，R2 为 0.9796175，即 biking 和 smoking 能解释 97.9617

# 5. 用 relaimpo 包的函数计算 biking 和 smoking 对 heart.disease 的重要性。哪个更重要?
# 可以使用 calc.relimp() 函数计算
library(relaimpo)
```

```r
relimp <- calc.relimp(model)
relimp
```

```
## Response variable: heart.disease
## Total response variance: 20.90203
## Analysis based on 498 observations
##
## 2 Regressors:
## biking smoking
## Proportion of variance explained by model: 97.96%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##               lmg
## biking  0.8795662
## smoking 0.1000512
##
## Average coefficients for different model sizes:
##
##                  1X        2Xs
## biking  -0.1990914 -0.2001331
## smoking  0.1704843  0.1783339
```

```r
# 结果显示，biking 对 heart.disease 的重要性为 0.8795662，smoking 对 heart.disease 的重要

# 6. 用得到的线性模型预测 heart.disease，用点线图画出预测值与真实值的关系，并在图上写出 R2
# 预测
heart.dat$pred <- predict(model)

# 绘图
library(ggplot2)
ggplot(heart.dat, aes(x = heart.disease, y = pred)) +
```

```
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed",color="red") +
  labs(title = "Predicted vs. Actual Heart Disease",
       x = "Actual Heart Disease",
       y = "Predicted Heart Disease",
       caption = paste0("R2 = ", round(summary(model)$r.squared, 3)))
```




```
# 7. 在建模时考虑 biking 和 smoking 的互作关系，会提高模型的 R2 值吗？如果是，意味着什么？
# 可以通过添加交互项来考虑 biking 和 smoking 的互作关系
model2 <- lm(heart.disease ~ biking * smoking, data = heart.dat)
summary(model2)
```

```
##
## Call:
## lm(formula = heart.disease ~ biking * smoking, data = heart.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.20619 -0.44862  0.02892  0.44099  1.94142
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.0527397  0.1248112 120.604   <2e-16 ***
## biking         -0.2019916  0.0029472 -68.536   <2e-16 ***
## smoking         0.1740065  0.0070359  24.731   <2e-16 ***
## biking:smoking  0.0001177  0.0001653   0.712    0.477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6544 on 494 degrees of freedom
## Multiple R-squared:  0.9796, Adjusted R-squared:  0.9795
## F-statistic:  7922 on 3 and 494 DF,  p-value: < 2.2e-16
```

```
# 结果显示，模型的 R2 值从 0.9796 变为到 0.9795 ，说明考虑了 biking 和 smoking 的互作关系
```

0.4.3 glm 相关问题

用 glm 建模时使用 family=binomial；在预测时，type= 参数可取值 link（默认）和 response。请问，两者的区别是什么？请**写代码**举例说明。

```
## 代码写这里，并运行；

#`type="link"`返回的是预测值的对数几率，`type="response"`返回的是预测值的概率。

# 加载数据集
data(iris)

# 将鸢尾花数据集转换为二元分类问题
iris$Species <- ifelse(iris$Species == "setosa", "setosa", "non-setosa")
```

```r
#dat <- iris %>% filter( Species %in% c("setosa", "virginica") )

iris$Species <- as.factor(iris$Species)

# 划分训练集和测试集
trainIndex <- sample(1:nrow(iris), 0.7*nrow(iris))
trainData <- iris[trainIndex,]
testData <- iris[-trainIndex,]

# 使用 glm 建立二元分类模型
model2 <- glm(Species ~., data = iris, family = binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# 预测测试集
linkPred <- predict(model2, iris, type = "link")
responsePred <- predict(model2, iris, type = "response")

# 输出前 10 个测试集样本的预测结果
head(data.frame(linkPred, responsePred), 10)
```

```
##    linkPred responsePred
## 1  38.02709            1
## 2  31.75445            1
## 3  32.97973            1
## 4  27.00205            1
## 5  37.63532            1
## 6  34.34355            1
## 7  29.20256            1
## 8  34.05821            1
## 9  25.09069            1
```

```
## 10 32.69058            1
```

练习与作业2: non-linear regression

分析 `swiss` ，用其它列的数据预测`Fertility`

1. 使用`earth`包建模，并做 10 times 10-fold cross validation;
2. 使用`lm`方法建模，同样做 10 times 10-fold cross validation;
3. 用 `RMSE` 和 `R2` 两个指标比较两种方法，挑选出较好一个;
4. 用 `vip` 包的函数查看两种方法中 feature 的重要性，并画图（如下图所示）：

```
![](images/talk11-feature-importance.png)
```

```r
## 代码写这里，并运行;
library(earth)
```

```
## Loading required package: Formula
```

```
## Loading required package: plotmo
```

```
## Loading required package: plotrix
```

```
## Loading required package: TeachingDemos
```