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] "sicheng.wu"
```

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

```
## [1] "/home/vkorpela"
```

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(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
```

```
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

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

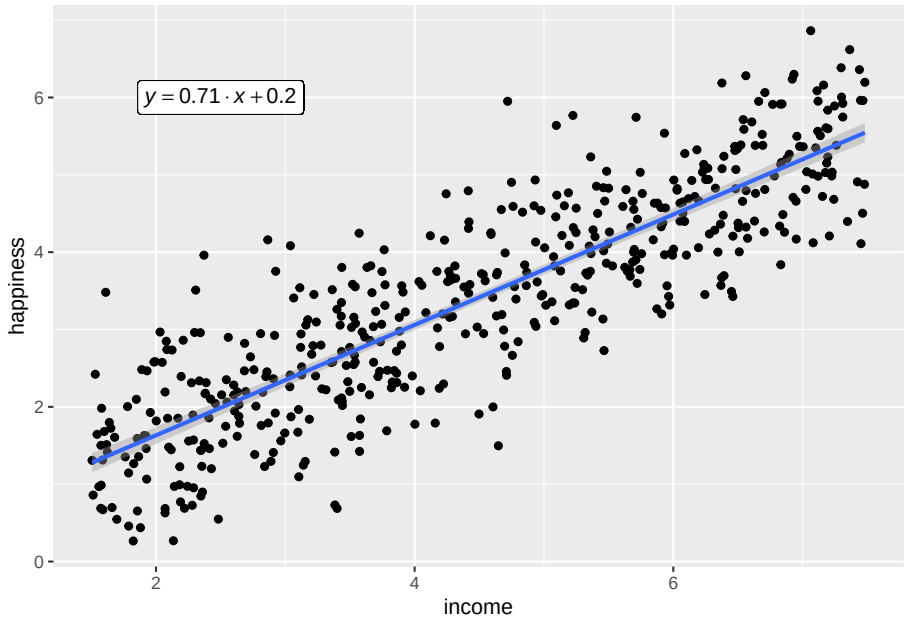
```
income.dat <- read_csv("./data/talk11/income.data_.zip")
```

```
## 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`
```

```
# 1. 用线性回归分析关系
income.lm <- lm(income.dat$happiness ~ income.dat$income)

# 2. 绘制 income 和 happiness 关系
eq_text <- substitute(
  italic(y) == a %.% italic(x) + b,
  list(
```

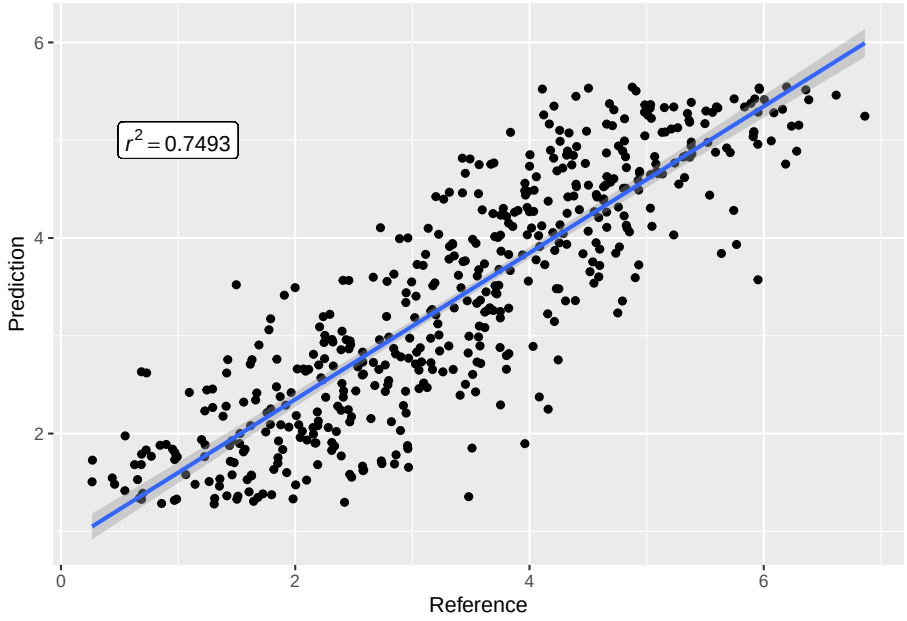
```
    a = format(coef(income.lm)[[2]], digits = 2),
    b = format(coef(income.lm)[[1]], digits = 2)
  )
) %>%
  as.expression() %>%
  as.character()

ggplot(income.dat, aes(x = income, y = happiness)) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_label(
    data = NULL,
    aes(x = 2.5, y = 6, label = eq_text),
    parse = TRUE,
    inherit.aes = FALSE
  )
```

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



```
# 3. 预测 happiness 值
happiness.pred <- predict(income.lm, income.dat['income'])

# 4. 绘制预测值与真实值的关系
happiness.cmp <- data.frame(
  ref = income.dat$happiness,
  pred = happiness.pred
)

happiness.r2 <- with(happiness.cmp, R2(pred, ref))
r2_text <- substitute(
  italic(r)^2 == r2,
  list(r2 = format(happiness.r2, digits = 4))
) %>%
  as.expression() %>%
  as.character()

ggplot(happiness.cmp, aes(x = ref, y = pred)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("Reference") +
  ylab("Prediction") +
  geom_label(
    data = NULL,
    aes(x = 1, y = 5, label = r2_text),
    parse = TRUE,
    inherit.aes = FALSE
  )
```

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



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.
```

```r
# 1. 线性回归分析
heart.lm <- lm(
  heart.disease ~ biking + smoking,
  data = heart.dat
)

# 2. 线性回归公式
heart.coef <- coef(heart.lm)

paste0(
  "heart.disease = ",
  format(heart.coef["biking"], digits = 3),
  " * biking + ",
  format(heart.coef["smoking"], digits = 3),
  " * smoking + ",
  format(heart.coef[1], digits = 3)
)
```

```
## [1] "heart.disease = -0.2 * biking + 0.178 * smoking + 15"
```

```r
# 3. biking 和 smoking 的影响
# 由系数可见，biking 与 heart.disease 负相关
# 而 smoking 与 heart.disease 正相关
# 二者的影响程度接近，相对而言 biking 更大
heart.coef
```

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

```
# 4. 两个系数对变化的解释
# 根据 R^2 的值，可以得到结论：
# biking 能解释 87.5% 的 variance
# smoking 能解释 9.6% 的 variance
with(
  heart.dat,
  data.frame(
    biking = R2(biking, heart.disease),
    smoking = R2(smoking, heart.disease)
  )
)
```

```
##      biking    smoking
## 1 0.8750769 0.09556196
```

```
# 5. 计算重要性
# 结论：biking 比 smoking 更重要
calc.relimp(heart.disease ~ biking + smoking, data = heart.dat)
```

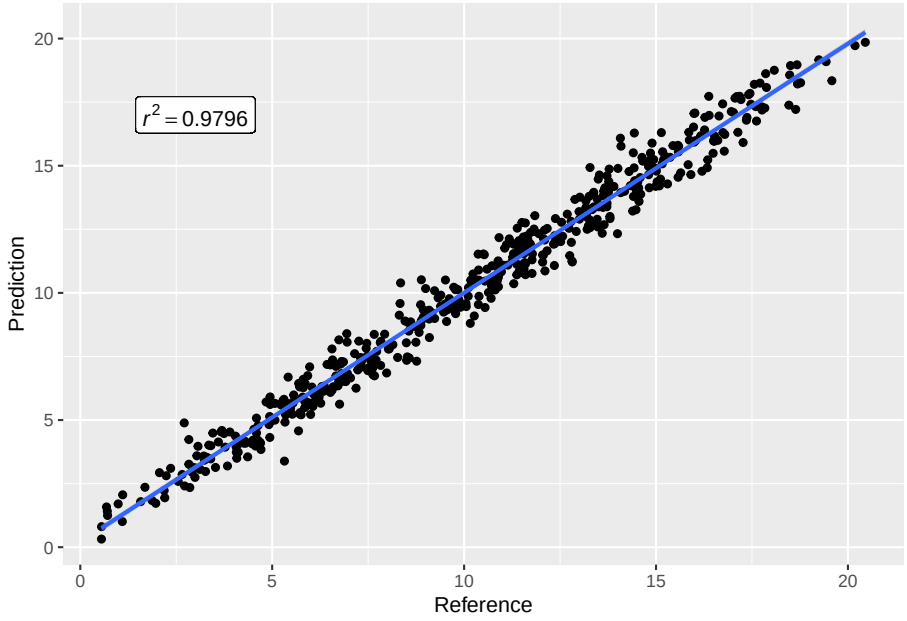
```
## Response variable: heart.disease
## Total response variance: 20.90203
## Analysis based on 498 observations
##
## 2 Regressors:
## biking smoking
## Proportion of variance explained by model: 97.96%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##                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
```

```
# 6. 预测
heart.pred <- predict(heart.lm, heart.dat)

heart.comp <- data.frame(
  ref = heart.dat$heart.disease,
  pred = heart.pred
)

heart.pred.r2 <- with(heart.comp, R2(ref, pred))
r2_text <- substitute(
  italic(r)^2 == r2,
  list(r2 = format(heart.pred.r2, digits = 4))
) %>%
  as.expression() %>%
  as.character()

ggplot(heart.comp, aes(x = ref, y = pred)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("Reference") +
  ylab("Prediction") +
  geom_label(
    data = NULL,
    aes(x = 3, y = 17, label = r2_text),
    parse = TRUE,
```

```
    inherit.aes = FALSE
  )
```

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




```
# 7. 考虑互作关系
# 结论：在建模时考虑互作关系，不会明显提高模型的 R2，
# 这说明收集到的数据中 biking 与 smoking 并不存在强关联。
heart.lm2 <- lm(
  heart.disease ~ biking * smoking,
  data = heart.dat
)

data.frame(
  no_interaction = summary(heart.lm)$r.squared,
  with_interaction = summary(heart.lm2)$r.squared
)
```

```
##   no_interaction with_interaction
## 1      0.9796175        0.9796383
```

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；
iris.dat <- iris %>% filter(Species %in% c("setosa", "virginica"))
iris.binom <- glm(Species ~ Sepal.Length + Sepal.Width +
                  Petal.Length + Petal.Width,
                  data = iris.dat, family = binomial)
```

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

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

```
# 当 type 取值为 link 时，给出的分类是 log-odds
# （即通过 logit 函数表示的属于某一类的概率）
# 当 type 取值为 response 时，给出的则是直接的概率值
data.frame(
  ref = iris.dat$Species,
  pred.link = predict(iris.binom, iris.dat, type = "link"),
  pred.response = predict(iris.binom, iris.dat, type = "response")
) %>%
  sample_n(6) %>%
  arrange(ref)
```

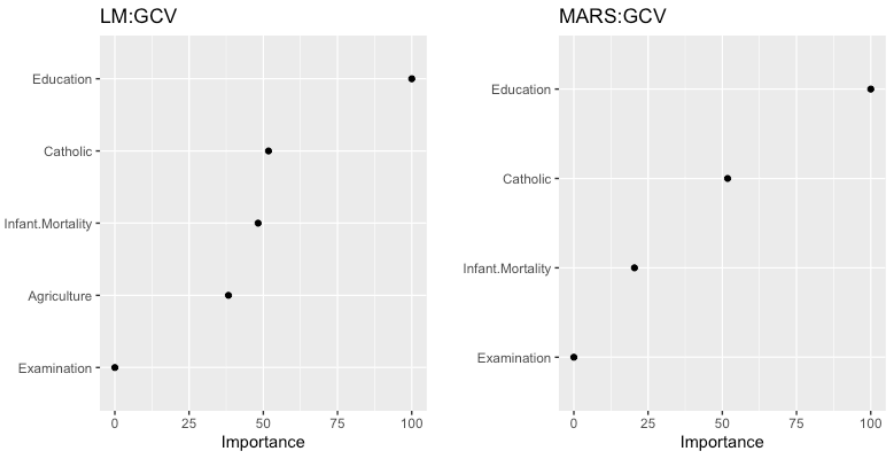
```
##          ref pred.link pred.response
## 16    setosa -34.24374  2.220446e-16
## 29    setosa -31.23614  2.220446e-16
```

```
## 24    setosa -23.90952  4.132652e-11
## 91 virginica  35.78971  1.000000e+00
## 55 virginica  37.44698  1.000000e+00
## 92 virginica  28.03628  1.000000e+00
```

0.5 练习与作业 2：non-linear regression

0.5.1 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）:




```
## 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
```

```
library(vip)
```

```
##
## Attaching package: 'vip'

## The following object is masked from 'package:utils':
##
##     vi
```

```
set.seed(1129)

# 1. 使用 earth 包建模
cv.mars <- train(
  Fertility ~ .,
  data = swiss,
  method = "earth",
  trControl = trainControl(method = "cv", number = 10)
)
cv.mars
```

```
## Multivariate Adaptive Regression Spline
##
## 47 samples
##  5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 41, 41, 42, 42, 43, 42, ...
```

```
## Resampling results across tuning parameters:
##
##   nprune  RMSE       Rsquared   MAE
##    2      10.552262  0.3608626  8.963709
##    8       8.963881  0.5406864  8.022220
##   14       8.963881  0.5406864  8.022220
##
## Tuning parameter 'degree' was held constant at a value of 1
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nprune = 8 and degree = 1.
```

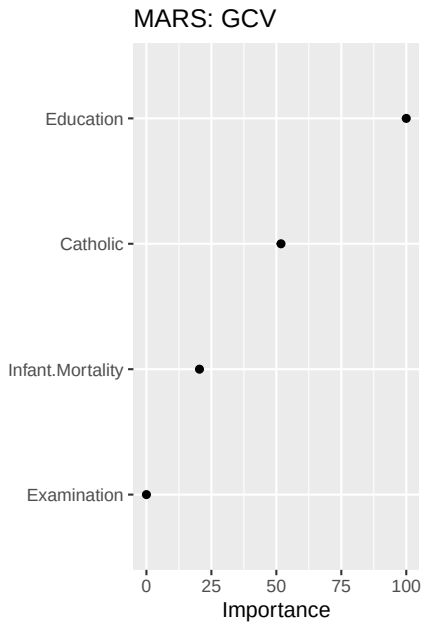
```
# 2. 使用 lm 建模
cv.lm <- train(
  Fertility ~ .,
  data = swiss,
  method = "lm",
  trControl = trainControl(method = "cv", number = 10)
)
cv.lm
```

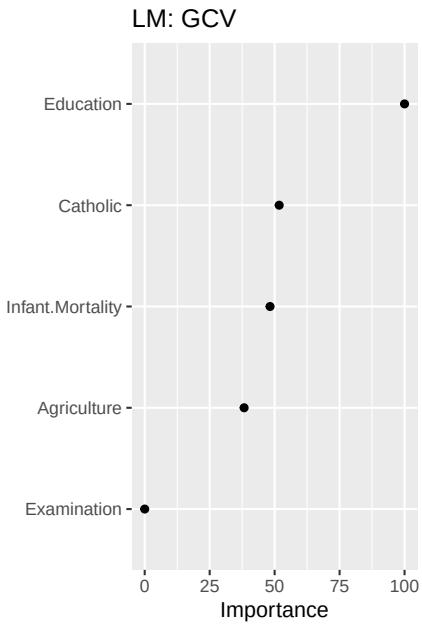
```
## Linear Regression
##
## 47 samples
##  5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 41, 41, 43, 43, 42, 43, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   7.713672  0.7228716  6.355422
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
# 3. 比较 RMSE 和 R^2 的指标
# 可以得到线性模型（lm）更优的结论
data.frame(
  name = c("MARS", "LM"),
  r.squared = c(
    cv.mars[["results"]][["Rsquared"]][2],
    cv.lm[["results"]][["Rsquared"]]
  ),
  RMSE = c(
    cv.mars[["results"]][["RMSE"]][2],
    cv.lm[["results"]][["RMSE"]]
  )
)
```

```
##   name r.squared     RMSE
## 1 MARS 0.5406864 8.963881
## 2   LM 0.7228716 7.713672
```

```r
# 4. 查看 feature 重要性
plot.mars <- vip(cv.mars, num_features = 5, geom = "point", value = "gcv") +
  ggtitle("MARS: GCV")
plot.lm <- vip(cv.lm, num_features = 5, geom = "point", value = "gcv") +
  ggtitle("LM: GCV")

gridExtra::grid.arrange(plot.mars, plot.lm, ncol = 2)
```


MARS: GCV
Education
Catholic
Infant.Mortality
Examination
0
25
50
75
100
Importance


LM: GCV
Education
Catholic
Infant.Mortality
Agriculture
Examination
0
25
50
75
100
Importance