近视预测模型开发与验证

张献伟

2022-07-25

献给……

呃，爱谁谁吧

目录

表格

插图

前言

你好，世界。我写了一本书。这本书是这样的，第 ?? 章介绍了见第??，第 ?? 章说见??，然后是请自己阅读吧!!!

我用了两个 R 包编译这本书，分别是 **knitr** (Xie, 2015) 和 **bookdown** (Xie, 2021)。以下是我的 R 进程信息：

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Chinese (Simplified)_People's Republic of China.936
## [2] LC_CTYPE=Chinese (Simplified)_People's Republic of China.936
## [3] LC_MONETARY=Chinese (Simplified)_People's Republic of China.936
## [4] LC_NUMERIC=C
## [5] LC_TIME=Chinese (Simplified)_People's Republic of China.936
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
## [6] methods   base
##
## loaded via a namespace (and not attached):
```

```
##  [1] compiler_4.1.1  magrittr_2.0.1  fastmap_1.1.0
##  [4] bookdown_0.24   cli_3.3.0       htmltools_0.5.2
##  [7] tools_4.1.1     rstudioapi_0.13 yaml_2.2.1
## [10] stringi_1.7.4   rmarkdown_2.11  knitr_1.34
## [13] stringr_1.4.0   digest_0.6.27   xfun_0.25
## [16] rlang_1.0.4     evaluate_0.14
```

致谢

感谢我自己的努力。

张献伟
于某角落

作者简介

尘世中一个迷途小书童

第一章　临床预测模型

常见的思路总结为三步：

1. 模型构建

2. 模型评价

3. 模型验证

1.1　模型构建

最熟悉的方法是：先单后多。大多数可行。但当变量数过多, 变量之间存在共线性或缺失值过多而又不愿舍弃掉缺失值的样本，有局限性。

如何解决?

1. 共线性问题——岭回归、*lasso*、弹性网络模型（正则技术）

2. 缺失值问题——随机森林模型

要解决的主要问题：变量筛选

方法 1：逐步回归（向后法、向前法、向前向后法）

方法 2：正则技术（岭回归、lasso、弹性网络模型)

方法 3：树模型

方法 4：随机森林模型（树模型的扩展）

方法 5：主成分分析

1.2 模型评价

为什么要评价模型?

- 欠拟合
- 过拟合

常见的评价指标主要有以下几种 1. 拟合优度检验（涉及卡方值及 P 值）

2.ROC（涉及 AUC,sen,spe,accuracy 等指标）

3.calibration（涉及 C-index 的计算）

4. 终极指标 MSE 的计算

5. 其他

通常来说，完成模型评价已经可以称之为" 完整" 的研究。

过拟合呢？或者是外推性如何？

1.3 模型验证

1.cross validation(简单交叉;K-fold corss validation;N-fold cross validation) 2.bootstrap 3.crossvalidation+bootstrap（最常用）

ps: 三个过程可能需要多次操作，才可以得到最终的结果。

第二章　数据

2.1　数据集简介

该数据集是来自 Orinda 近视纵向研究 (OLSM) 的数据子集，这是一项眼科队列研究儿童近视发病的成分发育和危险因素。数据收集始于 1989-1990 年学年，并每年持续到 2000-2001 学年。有关构成眼睛的部分的所有数据（眼部成分）是在上学期间的一次考试中收集的。家族史数据和在家长或监护人完成的一项调查中，每年都会收集视觉活动。

本文中使用的数据集来自 618 名至少接受五年随访且非近视的受试者当他们进入队列时，所有数据都来自他们的初始检查，数据集包括 17 个变量。此外眼睛数据有关于进入年龄，进入年份，近视家族史和各种视觉小时数的信息活动。眼部数据来自受试者的右眼。

```
data <- read.csv('myopia.csv')
library(knitr)
knitr::kable(head(data))
```

ID	STUDYYEAR	MYOPIC	AGE	GENDER	SPHEQ	AL	ACD	LT
1	1992	1	6	1	-0.052	21.89	3.690	3.498
2	1995	0	6	1	0.608	22.38	3.702	3.392
3	1991	0	6	1	1.179	22.49	3.462	3.514
4	1990	1	6	1	0.525	22.20	3.862	3.612
5	1995	0	5	0	0.697	23.29	3.676	3.454
6	1995	0	6	0	1.744	22.14	3.224	3.556

2.2 用 SPHEQ 预测近视

```
library(tidyverse)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'hms'
```

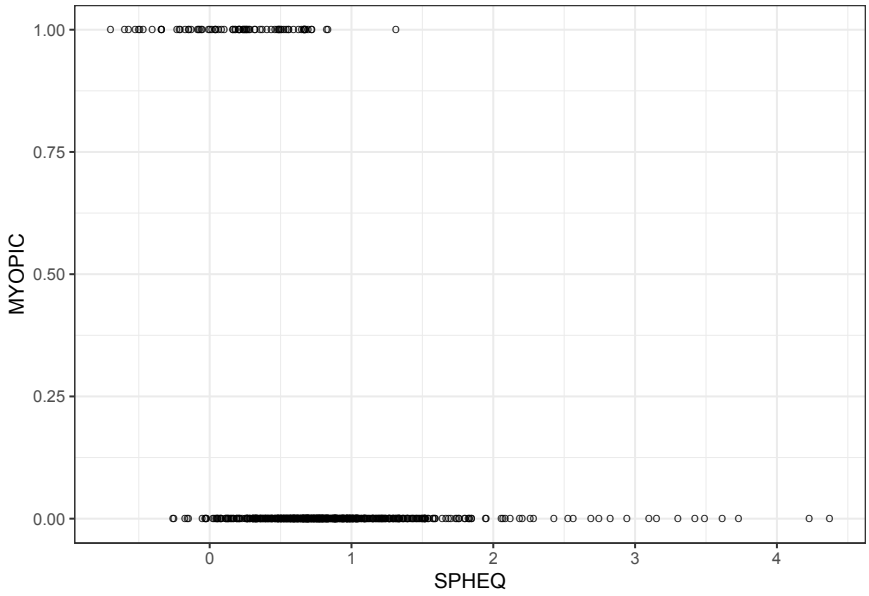
```
## -- Attaching packages --------------------------------------- tidyverse 1.3
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: 程辑包'tibble'是用R版本4.1.3 来建造的
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data %>%
  ggplot(.,aes(x=SPHEQ,y=MYOPIC))+
  geom_jitter(shape="O",position = position_jitter(height = 0))+
  theme_bw()
```



在这种情况下，“SPHEQ” 显然会影响近视的存在，但不足以准确预测。需要向模型添加更多属性以改进预测。为此，需要检查每个属性与近视存在之间的相关性。

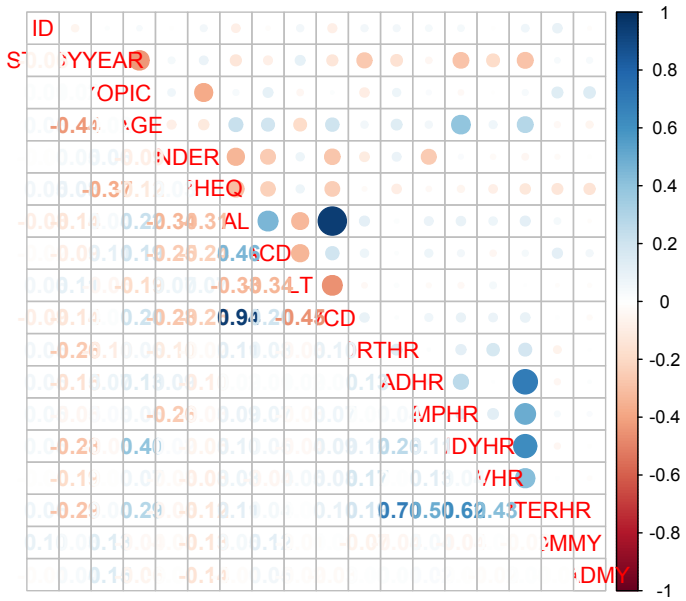
2.3 绘制所有变量之间的相关性

```
library(corrplot)
```

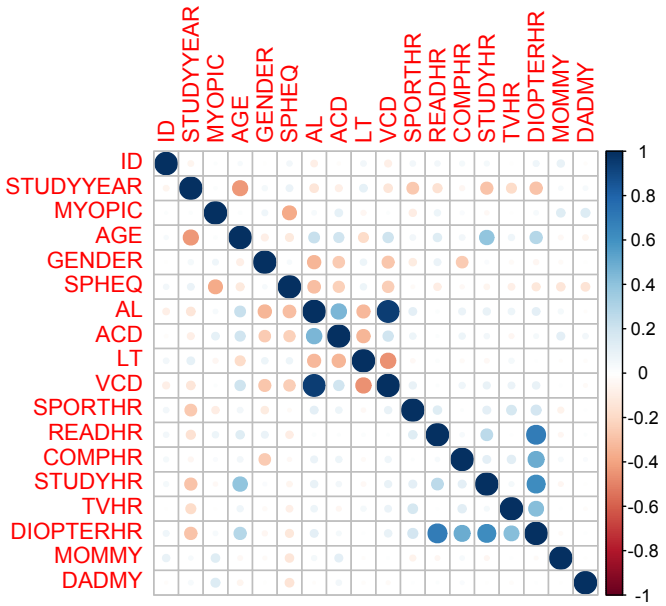
```
## Warning: 程辑包'corrplot'是用R版本4.1.2 来建造的
```

```
## corrplot 0.92 loaded
```

```
corrplot.mixed(cor(data))
```




```
corrplot(cor(data))
```



很明显,例如“DIOPTERHR”与“SPORTHR”、“TVHR”、“STUDYHR”、

“COMPHR” 和 “READHR” 高度相关。因此，“DIOPTERHR” 变量不会包含在预测模型中为了防止共线性问题

2.4 变量与近视之间的相关性

再来看每个属性与近视的关联性如何

```
library(corrplot)
correlations <- cor(data)

pcor <- correlations[,3] %>%
  print()
```

```
##           ID     STUDYYEAR        MYOPIC           AGE        GENDER
##  0.012242256  0.016330987  1.000000000  0.018525875  0.061556801 -0.37
##           AL          ACD           LT          VCD       SPORTHR
##  0.037752311  0.107952757 -0.045704451  0.011854862 -0.098282028  0.07
##       COMPHR      STUDYHR         TVHR    DIOPTERHR        MOMMY
##  0.025874323 -0.031858867 -0.004032443  0.036983991  0.134032827  0.14
```

```
 # corrplot(correlations)
```

根据图中与近视高度相关的属性是 “SPHEQ”、“ACD”、“MOMMY”、“DADMY”、“SPORTHR”、“READHR”、“GENDER”。

第三章　变量筛选

3.1　先单后多

3.1.1　分类变量处理

```
data <- read.csv('myopia.csv')
data$MYOPIC <- factor(data$MYOPIC,levels = c(0,1),labels = c("非近视","近
data$GENDER <- factor(data$GENDER,levels = c(0,1),labels = c("女性","男性
data$MOMMY <- factor(data$MOMMY,levels = c(0,1),labels = c("母亲不近视","
data$DADMY <- factor(data$DADMY,levels = c(0,1),labels = c("父亲不近视","
#年龄不设标签
# str(data)
# summary(data)
```

3.1.2　制作 table1

```
#连续性自变量
x1 <- c("SPHEQ","AL","ACD","LT","VCD","SPORTHR","READHR","COMPHR","STUDYH
#分类变量
x2 <- c("MYOPIC","AGE","GENDER","MOMMY","DADMY")
```

```
library(tableone)
```

```
## Warning: 程辑包'tableone'是用R版本4.1.2 来建造的
```

```
table1 <- CreateTableOne(vars = c(x1,x2),
                         data = data,
                         factorVars = x2,
                         strata=c("MYOPIC"),addOverall = F)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'hms'
```

```
results1 <- print(table1,shouwAllLevels=FALSE)
```

```
##                          Stratified by MYOPIC
##                           非近视          近视            p      test
##   n                         537             81
##   SPHEQ (mean (SD))        0.89 (0.60)    0.20 (0.40)   <0.001
##   AL (mean (SD))          22.49 (0.69)   22.56 (0.61)    0.349
##   ACD (mean (SD))          3.57 (0.23)    3.64 (0.20)    0.007
##   LT (mean (SD))           3.54 (0.16)    3.52 (0.14)    0.257
##   VCD (mean (SD))         15.37 (0.67)   15.40 (0.61)    0.769
##   SPORTHR (mean (SD))     12.26 (7.93)    9.94 (8.00)    0.015
##   READHR (mean (SD))       2.71 (2.97)    3.37 (3.62)    0.071
##   COMPHR (mean (SD))       2.07 (2.99)    2.31 (3.50)    0.521
##   STUDYHR (mean (SD))      1.52 (2.29)    1.31 (1.68)    0.429
##   TVHR (mean (SD))         8.96 (5.79)    8.89 (5.26)    0.920
##   DIOPTERHR (mean (SD)) 25.79 (15.92) 27.54 (16.75)    0.359
##   MYOPIC = 近视 (%)          0 ( 0.0)      81 (100.0)   <0.001
##   AGE (%)                                               0.497
##      5                     17 ( 3.2)       4 (  4.9)
```

```
##      6                    398 (74.1)     58 ( 71.6)
##      7                     73 (13.6)      9 ( 11.1)
##      8                     45 ( 8.4)      8 (  9.9)
##      9                      4 ( 0.7)      2 (  2.5)
##   GENDER = 男性 (%)        256 (47.7)     46 ( 56.8)   0.158
##   MOMMY = 母亲近视 (%)     258 (48.0)     55 ( 67.9)   0.001
##   DADMY = 父亲近视 (%)     252 (46.9)     56 ( 69.1)  <0.001
```

```
write.csv(results1,"table1.csv")
```

```
#如果不服从正态分布秩和检验、卡方检验
library(tableone)
table1 <- CreateTableOne(vars = c(x1,x2),
                         data = data,
                         factorVars = x2,
                         strata = "MYOPIC",addOverall = F)
results2 <- print(table1,shouwAllLevels=FALSE,
                  nonnormal=x1)#指定非阐述检验的变量
```

```
##                            Stratified by MYOPIC
##                             非近视                近视                 p
##   n                           537                  81
##   SPHEQ (median [IQR])       0.79 [0.55, 1.10]    0.23 [-0.07, 0.50]  <
##   AL (median [IQR])         22.45 [22.02, 22.97] 22.56 [22.07, 22.94]
##   ACD (median [IQR])         3.57 [3.41, 3.72]    3.68 [3.50, 3.74]
##   LT (median [IQR])          3.54 [3.44, 3.65]    3.51 [3.42, 3.63]
##   VCD (median [IQR])        15.36 [14.92, 15.83] 15.33 [14.96, 15.89]
##   SPORTHR (median [IQR])    10.00 [6.00, 16.00]   8.00 [3.00, 15.00]
##   READHR (median [IQR])      2.00 [0.00, 4.00]    3.00 [1.00, 5.00]
##   COMPHR (median [IQR])      1.00 [0.00, 3.00]    1.00 [0.00, 3.00]
##   STUDYHR (median [IQR])     1.00 [0.00, 2.00]    1.00 [0.00, 2.00]
##   TVHR (median [IQR])        8.00 [4.00, 12.00]   8.00 [5.00, 12.00]
##   DIOPTERHR (median [IQR]) 22.00 [14.00, 34.00] 24.00 [16.00, 36.00]
##   MYOPIC = 近视 (%)             0 ( 0.0)             81 (100.0)       <
```

```
##   AGE (%)                                                                0.497
##      5                    17 ( 3.2)            4 (  4.9)
##      6                   398 (74.1)           58 ( 71.6)
##      7                    73 (13.6)            9 ( 11.1)
##      8                    45 ( 8.4)            8 (  9.9)
##      9                     4 ( 0.7)            2 (  2.5)
##   GENDER = 男性 (%)      256 (47.7)           46 ( 56.8)         0.158
##   MOMMY = 母亲近视 (%)   258 (48.0)           55 ( 67.9)         0.001
##   DADMY = 父亲近视 (%)   252 (46.9)           56 ( 69.1)        <0.001
##                          Stratified by MYOPIC
##                           test
##   n
##   SPHEQ (median [IQR])     nonnorm
##   AL (median [IQR])        nonnorm
##   ACD (median [IQR])       nonnorm
##   LT (median [IQR])        nonnorm
##   VCD (median [IQR])       nonnorm
##   SPORTHR (median [IQR])   nonnorm
##   READHR (median [IQR])    nonnorm
##   COMPHR (median [IQR])    nonnorm
##   STUDYHR (median [IQR])   nonnorm
##   TVHR (median [IQR])      nonnorm
##   DIOPTERHR (median [IQR]) nonnorm
##   MYOPIC = 近视 (%)
##   AGE (%)
##      5
##      6
##      7
##      8
##      9
##   GENDER = 男性 (%)
##   MOMMY = 母亲近视 (%)
##   DADMY = 父亲近视 (%)
```

```
#exact 可以指定确切概论检验的变量，这里忽略（数据大，不需要）
write.csv(results2,"results2.csv")
```

3.1.3 单因素 logistic

```
# 自变量
model <- glm(MYOPIC ~ SPHEQ,data = data,family = binomial())
# 查看模型结果
summary(model)$coefficients
```

```
##                Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)  0.05397315  0.2067483  0.2610573 7.940483e-01
## SPHEQ       -3.83309762  0.4183696 -9.1619878 5.094998e-20
```

```
# 计算OR及其可信区间
exp(cbind("OR"=coef(model),confint(model)))
```

```
## Waiting for profiling to be done...
```

```
##                      OR      2.5 %     97.5 %
## (Intercept) 1.05545626 0.70776986 1.59472261
## SPHEQ       0.02164247 0.00910867 0.04716271
```

取单因素 P<0.1

3.1.4 多因素模型

```
# 多因素模型
model_1<- glm(factor(MYOPIC)~MOMMY+DADMY+SPHEQ+ACD+SPORTHR+READHR,family
# 查看结果
summary(model_1)$coefficients
```

```
##                   Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)   -3.53288934 2.42831871 -1.454871 1.457051e-01
## MOMMY母亲近视  0.74791347 0.31126257  2.402838 1.626840e-02
## DADMY父亲近视  0.86552437 0.30735580  2.816034 4.862053e-03
## SPHEQ         -3.79733834 0.43556193 -8.718251 2.825317e-18
## ACD            0.82711271 0.66485466  1.244050 2.134810e-01
## SPORTHR       -0.05461693 0.02041790 -2.674953 7.473971e-03
## READHR         0.06843900 0.04577874  1.494995 1.349157e-01
```

```
# 计算OR及其可信区间
exp(cbind("OR"=coef(model),confint(model_1)))
```

```
## Waiting for profiling to be done...
```

```
## Warning in cbind(OR = coef(model), confint(model_1)): number of rows of res
## is not a multiple of vector length (arg 1)
```

```
##                       OR        2.5 %     97.5 %
## (Intercept)   1.05545626 0.0002380046 3.33425391
## MOMMY母亲近视 0.02164247 1.1604118328 3.95048116
## DADMY父亲近视 1.05545626 1.3157000566 4.41021361
## SPHEQ         0.02164247 0.0090945183 0.05041805
## ACD           1.05545626 0.6237053229 8.51268747
## SPORTHR       0.02164247 0.9077904406 0.98385439
## READHR        1.05545626 0.9784535086 1.17140573
```

3.1.4.1 PS: 引入变量

如果我们已经确定是研究自变量 X 与因变量 Y 之间的关系，这时候协变量应该如何筛选进入多因素模型呢? 如果某个协变量 Z 与因变量 Y 之间单因素分析的 P 值小于 0.1，并且协变量 Z 与自变量 X 同时分析与因变量 Y 的关系，由于 Z 的存在，使 X 的系数较单因素分析时变化超过 10%，这时候协变量 Z 应该纳入多因素分析中。

我们这里主要（关心）的 X 变量时 SPHEQ 情况，其他自变量都是协变量 Z

```
uni_methods<-function(xvar){
  model<-glm(MYOPIC~SPHEQ,data = data,family = binomial())
  coef <- coef(model)[2]
  form <- as.formula(paste0("MYOPIC~SPHEQ+",xvar))
  model2 <- glm(form,data = data,family = binomial())
  coef2 <- coef(model2)[2]
  ratio <- abs(coef2-coef)/coef>0.1
  if(ratio){
    return(xvar)
  }
}
```

```
xvar <- c(x1,x2)
xvar <- xvar[-which(xvar=="SPHEQ")]
xvar
```

```
##  [1] "AL"        "ACD"       "LT"        "VCD"       "SPORTHR"   "READ
##  [7] "COMPHR"    "STUDYHR"   "TVHR"      "DIOPTERHR" "MYOPIC"    "AGE"
## [13] "GENDER"    "MOMMY"     "DADMY"
```

```
lapply(xvar,uni_methods)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): 在公式右手的反应略

## Warning in model.matrix.default(mt, mf, contrasts): 模型矩阵的2项有问
## 定的列

## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
```

```
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
##
## [[6]]
## NULL
##
## [[7]]
## NULL
##
## [[8]]
## NULL
##
## [[9]]
## NULL
##
## [[10]]
## NULL
##
## [[11]]
## NULL
##
## [[12]]
## NULL
##
## [[13]]
## NULL
##
## [[14]]
## NULL
##
```

```
## [[15]]
## NULL
```

第四章　LASSO-Logistic

lasso 回归不能把数据处理成 *factor*，如果涉及多分类，手动设置哑变量

4.1　读入数据

```
library(tidyverse)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'hms'
```

```
## -- Attaching packages --------------------------------------- tidyvers
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: 程辑包'tibble'是用R版本4.1.3 来建造的
```

```
## -- Conflicts ------------------------------------------ tidyverse_conf
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
myopia <- read.csv("myopia.csv") %>%
  mutate(PARENTS=MOMMY+DADMY) %>%
  select(ID:DIOPTERHR,PARENTS)
```

4.2 将数据集分成训练集和测试集

```
library(caret)
```

```
## Warning: 程辑包'caret'是用R版本4.1.3 来建造的
```

```
## 载入需要的程辑包：lattice
```

```
##
## 载入程辑包：'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(1000)
index <- createDataPartition(
  myopia$MYOPIC,
  p = 0.7,
  list = FALSE
)
train <- myopia[index, ]
test <- myopia[-index, ]
```

4.3 加载包

```
library(corrplot)
```

```
## Warning: 程辑包'corrplot'是用R版本4.1.2 来建造的
```

```
## corrplot 0.92 loaded
```

```
library(car)
```

```
## 载入需要的程辑包：carData
```

```
##
## 载入程辑包：'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(leaps)
```

```
## Warning: 程辑包'leaps'是用R版本4.1.3 来建造的
```

```
library(glmnet)#岭回归、lasso、弹性网络模型
```

```
## Warning: 程辑包'glmnet'是用R版本4.1.2 来建造的
```

```
## 载入需要的程辑包：Matrix
```

```
##
## 载入程辑包：'Matrix'
```

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

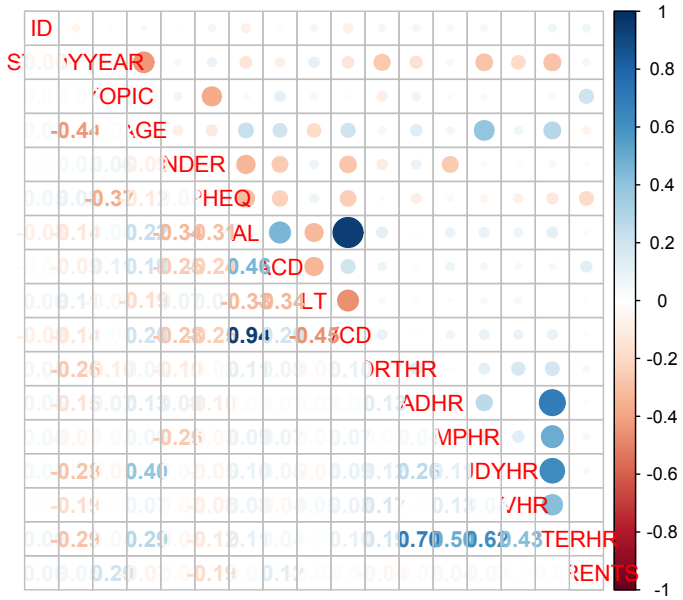
```
## Loaded glmnet 4.1-3
```

```
library(caret)
#### 在glmnet()语法中alpha=0为岭回归，alpha=1表示lasso回归
```

```
correlations <- cor(myopia)
pcor <- correlations[,3] %>%
  print()
```

```
##           ID    STUDYYEAR       MYOPIC          AGE       GENDER         SPH
##  0.012242256  0.016330987  1.000000000  0.018525875  0.061556801 -0.373639(
##           AL          ACD           LT          VCD      SPORTHR       READ
##  0.037752311  0.107952757 -0.045704451  0.011854862 -0.098282028  0.072749(
##       COMPHR      STUDYHR         TVHR    DIOPTERHR      PARENTS
##  0.025874323 -0.031858867 -0.004032443  0.036983991  0.201417458
```

```
corrplot.mixed(cor(myopia))
```



“SPHEQ”, “ACD”, ’’MOMMY”, “DADMY”, ”SPORTHR” , ”READHR”,”GENDER”

4.4 定义自变量，因变量

```
x <- as.matrix(train[,4:17])
y <- train[,3]
lambdas <- 10 ^ seq(8,-4,length=250)
```

第五章

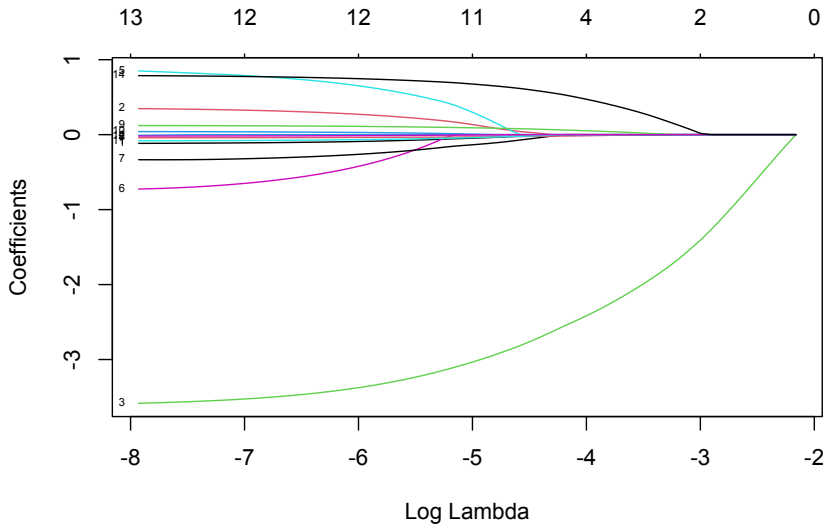
```
lasso <- glmnet(x,y,family = "binomial",alpha = 1)
print(lasso)
```

```
##
## Call:  glmnet(x = x, y = y, family = "binomial", alpha = 1)
##
##    Df  %Dev   Lambda
## 1   0  0.00 0.115200
## 2   1  2.76 0.105000
## 3   1  5.36 0.095630
## 4   1  7.78 0.087140
## 5   1  9.99 0.079400
## 6   1 11.99 0.072340
## 7   1 13.78 0.065920
## 8   1 15.35 0.060060
## 9   1 16.73 0.054720
## 10  2 18.06 0.049860
## 11  2 19.49 0.045430
## 12  2 20.74 0.041400
## 13  3 21.91 0.037720
## 14  3 23.02 0.034370
## 15  3 23.98 0.031320
## 16  4 24.92 0.028530
## 17  4 25.79 0.026000
```

```
## 18  4 26.54 0.023690
## 19  4 27.19 0.021580
## 20  4 27.75 0.019670
## 21  4 28.23 0.017920
## 22  4 28.64 0.016330
## 23  5 29.01 0.014880
## 24  8 29.43 0.013560
## 25  8 29.82 0.012350
## 26  8 30.15 0.011250
## 27  8 30.44 0.010250
## 28  9 30.73 0.009343
## 29 10 31.00 0.008513
## 30 11 31.26 0.007757
## 31 11 31.48 0.007068
## 32 11 31.68 0.006440
## 33 11 31.84 0.005868
## 34 12 31.99 0.005347
## 35 12 32.12 0.004872
## 36 12 32.23 0.004439
## 37 12 32.32 0.004045
## 38 12 32.40 0.003685
## 39 12 32.47 0.003358
## 40 12 32.53 0.003060
## 41 12 32.57 0.002788
## 42 12 32.61 0.002540
## 43 12 32.65 0.002314
## 44 12 32.68 0.002109
## 45 12 32.70 0.001921
## 46 12 32.72 0.001751
## 47 12 32.74 0.001595
## 48 12 32.75 0.001454
## 49 12 32.76 0.001324
## 50 12 32.77 0.001207
## 51 12 32.78 0.001100
## 52 12 32.79 0.001002
```

```
## 53 12 32.79 0.000913
## 54 12 32.80 0.000832
## 55 12 32.80 0.000758
## 56 12 32.81 0.000691
## 57 13 32.81 0.000629
## 58 13 32.81 0.000573
## 59 13 32.81 0.000522
## 60 13 32.81 0.000476
## 61 13 32.82 0.000434
## 62 13 32.82 0.000395
## 63 13 32.82 0.000360
```

```
plot(lasso,xvar="lambda",label = TRUE)
```



传入一个 lambda 值看看

```
loss.coef <- predict(lasso,s=0.05,type = 'coefficients')
loss.coef
```

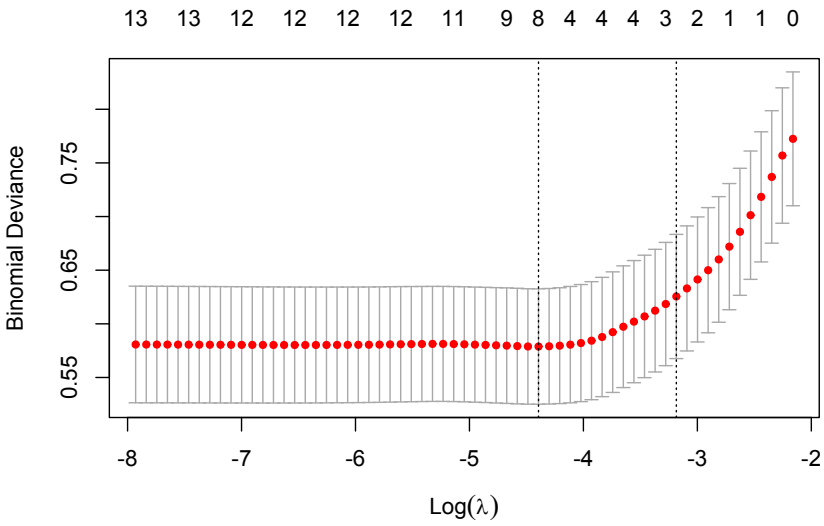
```
## 15 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                       s1
## (Intercept) -1.04549610
## AGE           .
## GENDER        .
## SPHEQ        -1.40167370
## AL            .
## ACD           .
## LT            .
## VCD           .
## SPORTHR       .
## READHR        .
## COMPHR        .
## STUDYHR       .
## TVHR          .
## DIOPTERHR     .
## PARENTS       0.01645111
```

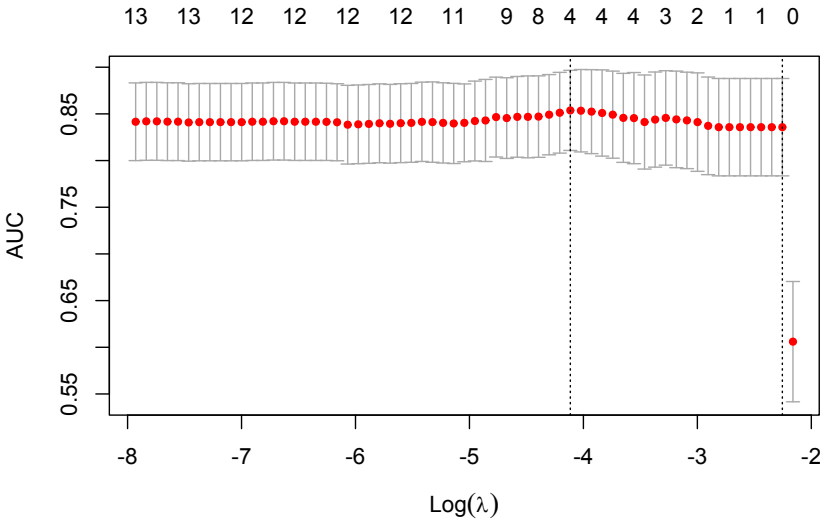
type=c("link", "response", "class", "coefficients", "nonzero")。link 给出的是线性预测值，即进行 logit 变化前的值，函数默认值；response 给出的是概率预测值，即进行 logit 变换之后的值；clase 给出 0/1 预测值；coefficients 给出的是指定 值的模型系数；nonzero 给出指定的定 值时系数不为 0 的模型变量。

5.1 cv 交叉验证

```
lasso.cv <- cv.glmnet(x,y, alpha=1,  family="binomial")
plot(lasso.cv)
```




```
lasso.cv_auc <- cv.glmnet(x,y,alpha=1,family="binomial",type.measure = "a
plot(lasso.cv_auc)
```



横坐标是 lambda 的对数值，也就是惩罚力度，值越大，惩罚力度越大。纵坐标是模型的 MSE(均方误差)。图形上方横坐标是自变量数量。随着

lambda 的增加，MSE 不断变化。第一条虚线表示 MSE 最小值对应的 lambda 值，第二条虚线表示距离均方误差一个标准误时的 lambda 值（最优解）

```
coef(lasso.cv , s = c(1,0.1,0.01,0.001))
```

```
## 15 x 4 sparse Matrix of class "dgCMatrix"
##                      s1         s2          s3          s4
## (Intercept) -1.906893 -1.7282362  0.16212541  4.02046678
## AGE          .          .         -0.02171082 -0.10849682
## GENDER       .          .          0.04626298  0.32335354
## SPHEQ        .         -0.2400235 -2.82801562 -3.51734389
## AL           .          .          .           .
## ACD          .          .          0.01674432  0.78087139
## LT           .          .          .          -0.63712500
## VCD          .          .         -0.07605748 -0.31816768
## SPORTHR      .          .         -0.02514996 -0.03991801
## READHR       .          .          0.07966571  0.11800176
## COMPHR       .          .          .           0.03702366
## STUDYHR      .          .         -0.02353405 -0.07749223
## TVHR         .          .          .          -0.01674881
## DIOPTERHR    .          .          .           .
## PARENTS      .          .          0.61989773  0.77437955
```

```
lasso.cv_min <- lasso.cv$lambda.min %>%
  print()
```

```
## [1] 0.01235143
```

```
lasso.coef <- coef(lasso.cv$glmnet.fit,s=lasso.cv_min,exact = F)
lasso.coef
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) -0.470162310
```

```
## AGE          -0.009058156
## GENDER        0.015668468
## SPHEQ        -2.694167376
## AL            .
## ACD           .
## LT            .
## VCD          -0.036936193
## SPORTHR      -0.022001607
## READHR        0.071207248
## COMPHR        .
## STUDYHR      -0.012574109
## TVHR          .
## DIOPTERHR     .
## PARENTS       0.577869537
```

我们可以试一下如果选择 1s 是什么情况

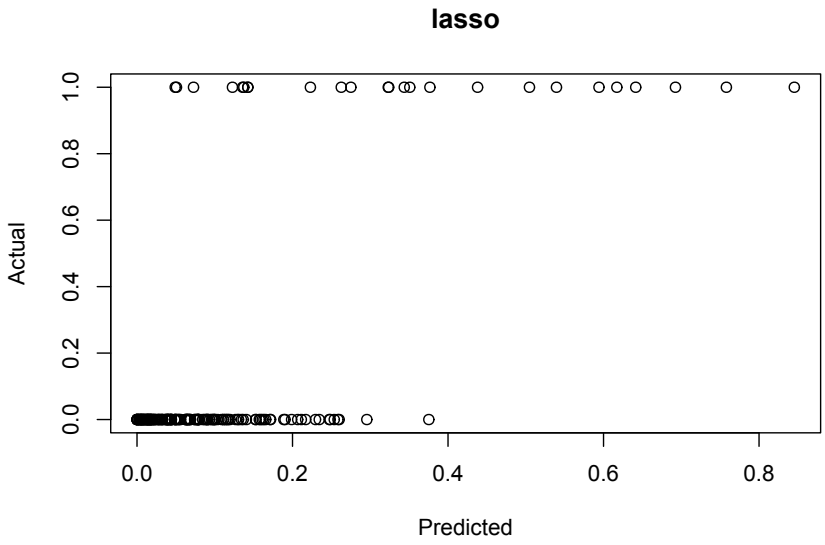
```
lasso.cv_1se <- lasso.cv$lambda.1se#通常使用距离MSE最小一个标准差lambda作

lasso.coef_1se<- coef(lasso.cv$glmnet.fit,s=lasso.cv_1se,exact = F)
lasso.coef_1se
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) -1.0378542
## AGE           .
## GENDER        .
## SPHEQ        -1.6514186
## AL            .
## ACD           .
## LT            .
## VCD           .
## SPORTHR       .
## READHR        .
## COMPHR        .
```

```
## STUDYHR       .
## TVHR          .
## DIOPTERHR     .
## PARENTS       0.1220882
```

5.2 lasso 在测试集上的表现

```
newx=as.matrix(test[4:17])
lasso.y <- predict(lasso,newx = newx,type = "response",s=0.01235)
plot(lasso.y,test$MYOPIC,xlab="Predicted",ylab="Actual",main="lasso")
```



5.3 建立模型并绘制列线图

5.3.1 建立一个模型吧

```
library(rms)
```

```
## 载入需要的程辑包: Hmisc
```

```
## 载入需要的程辑包: survival
```

```
##
## 载入程辑包: 'survival'
```

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

```
## 载入需要的程辑包: Formula
```

```
##
## 载入程辑包: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## 载入需要的程辑包: SparseM
```

```
##
## 载入程辑包: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

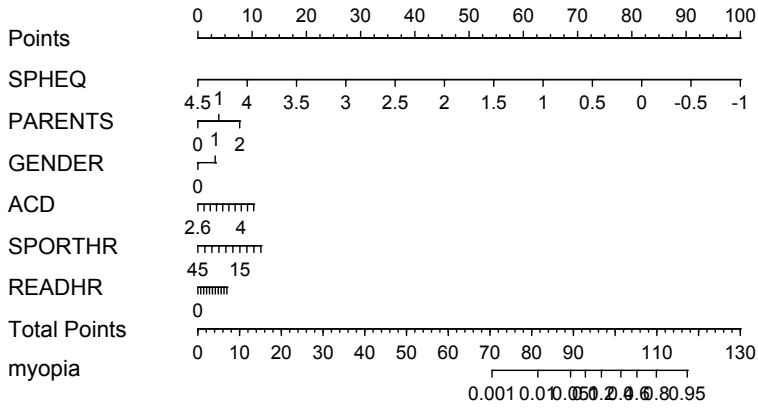
```
##
## 载入程辑包: 'rms'
```

```
## The following objects are masked from 'package:car':
##
##     Predict, vif
```

```
dd <- datadist(myopia)
options(datadist=dd)
model<- lrm(MYOPIC~SPHEQ+PARENTS+GENDER+ACD+SPORTHR+READHR,data=myopia,x=TRUE,
```

5.3.2 列线图 1

```
nom1 <- nomogram(model,fun = plogis,fun.at=c(0.001,0.01,0.05,0.1,seq(0.2,0.8,b
plot(nom1)
```



列线图 2

```
# # install.packages("DynNom")
# library(DynNom)
# model_dynnom <- glm(MYOPIC~SPHEQ+PARENTS+GENDER+ACD+SPORTHR+READHR,data=myop
```

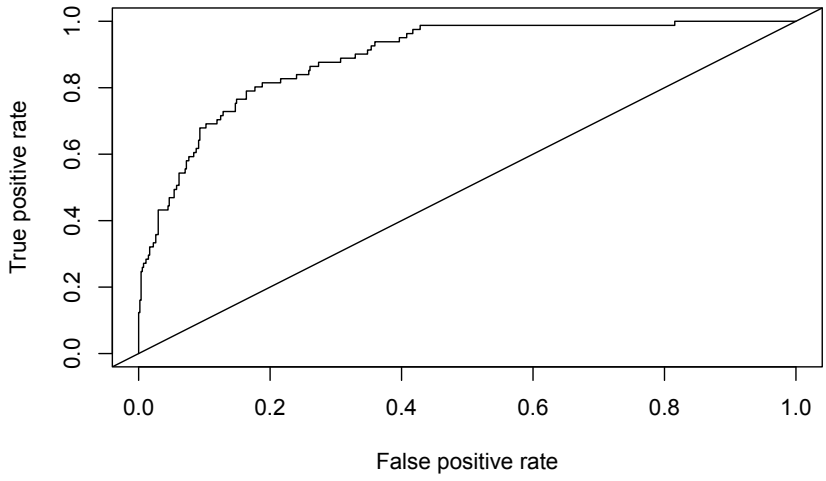
```
# DynNom(model_dynnom,DNtitle = "Nomogram",DNxlab = "Probability")
```

5.3.3 C 统计量

```
myopia$predvalue <- predict(model)
# install.package('ROCR')
library(ROCR)
```

```
## Warning: 程辑包'ROCR'是用R版本4.1.3 来建造的
```

```
pred <- prediction(myopia$predvalue,myopia$MYOPIC)
perf <- performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)
```



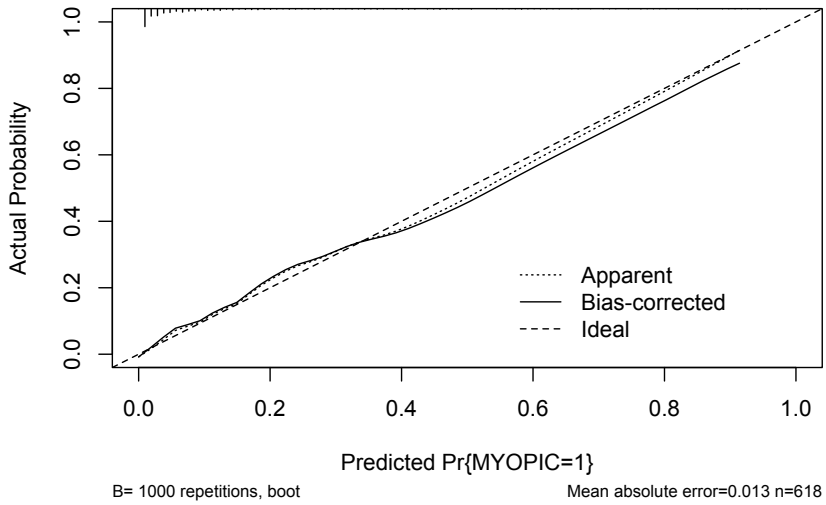

```
auc <- performance(pred,"auc")
auc@y.values
```

```
## [[1]]
## [1] 0.8914178
```

```
# model_glm <- glm(MYOPIC~SPHEQ+PARENTS+GENDER+ACD+SPORTHR+READHR,data = myopi
# myopia$predvalue <- predict(model_glm)
# # install.package('ROCR')
# library(ROCR)
# pred <- prediction(myopia$predvalue,myopia$MYOPIC)
# perf <- performance(pred,"tpr","fpr")
# plot(perf)
# abline(0,1)
# auc <- performance(pred,"auc")
# auc@y.values
```

5.3.4 校正曲线

```
cal1 <- calibrate(model,method = "boot",B=1000)
plot(cal1,xlim=c(0,1.0),ylim=c(0,1.0))
```




```
##
## n=618   Mean absolute error=0.013   Mean squared error=3e-04
## 0.9 Quantile of absolute error=0.032
```

5.3.5 ps：同时绘制多条

```
formula1 <- as.formula(MYOPIC~SPHEQ)

formula2 <- as.formula(MYOPIC~SPHEQ+PARENTS+GENDER+ACD+SPORTHR+READHR)

formula3 <- as.formula(MYOPIC~PARENTS+GENDER+ACD+SPORTHR+READHR)

DD=datadist(myopia)
options(datadist='DD')
```

```
fit1 = glm(formula1, data=myopia,family = binomial())
fit2 = glm(formula2, data=myopia,family = binomial())
```

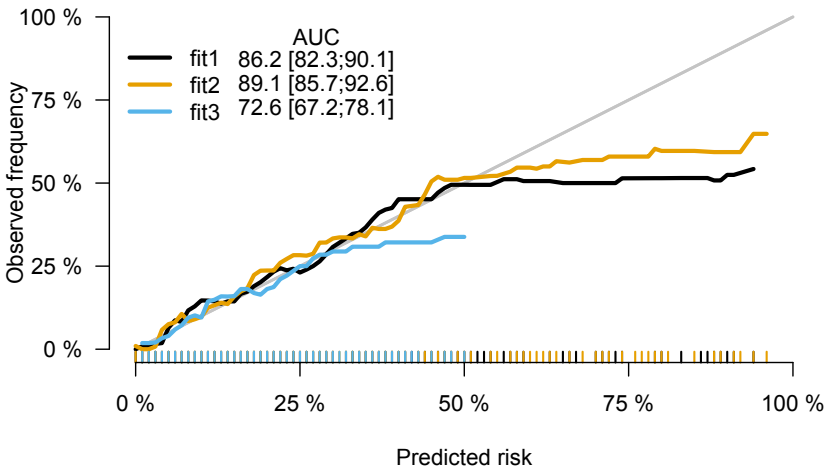
```
fit3 = glm(formula3, data=myopia,family = binomial())


library(riskRegression)
```

```
## Warning: 程辑包'riskRegression'是用R版本4.1.3 来建造的
```

```
## riskRegression version 2022.03.22
```

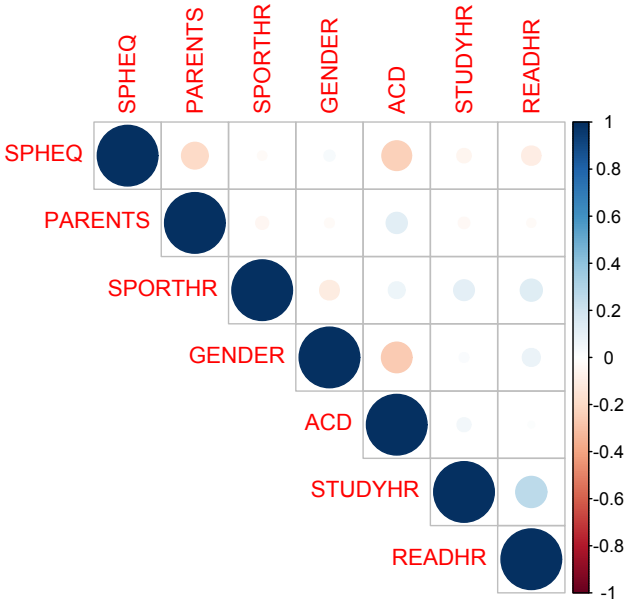
```
xb <- Score(list("fit1"=fit1,
                 "fit2"=fit2,
                 "fit3"=fit3),
            formula=MYOPIC~1,
            null.model = FALSE,
            conf.int =TRUE,
            plots =c("calibration","ROC"),
            metrics = c("auc"),
            B=1000,M=50,
            data=myopia)
plotCalibration(xb)
```

```
## Warning in getLegendData(object = x, models = models, times = tp, auc.in.le
## = auc.in.legend, : Cannot show Brier score as it is not stored in object. S
## metrics='brier' in the call of Score.
```



5.4 共线性讨论

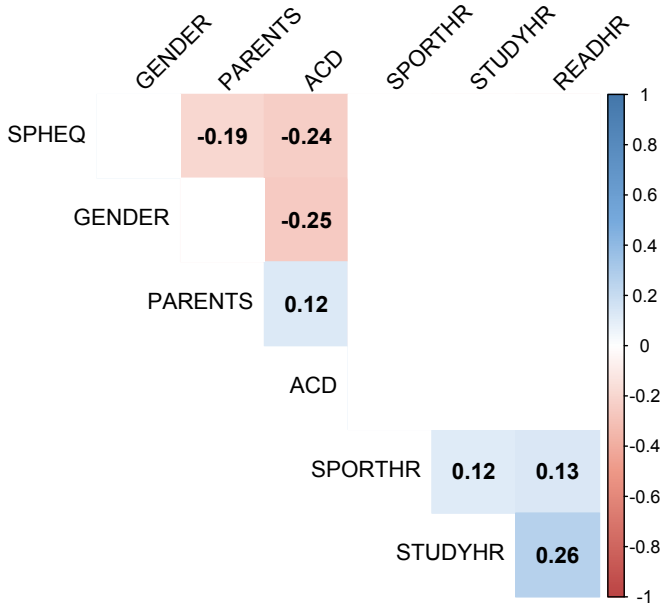
```
#collinearity
collin <- cor(subset(myopia, select=c(SPHEQ, PARENTS, SPORTHR, GENDER ,AC
# dev.off()
library(corrplot)
corrplot(collin , type="upper")
```




```
M <- collin
cor.mtest <- function(mat, ...) {
  mat <- as.matrix(mat)
  n <- ncol(mat)
  p.mat<- matrix(NA, n, n)
  diag(p.mat) <- 0
  for (i in 1:(n - 1)) {
    for (j in (i + 1):n) {
      tmp <- cor.test(mat[, i], mat[, j], ...)
      p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
    }
  }
  colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
  p.mat
}
# matrix of the p-value of the correlation
p.mat <- cor.mtest(subset(myopia, select=c(SPHEQ, PARENTS,
                                           SPORTHR, GENDER ,ACD , STUDYHR,READ
head(p.mat[, 1:6])
```

```
##                SPHEQ      PARENTS     SPORTHR       GENDER           AC
## SPHEQ   0.000000e+00 1.351303e-06 0.577189469 4.206886e-01 1.841368e-0
## PARENTS 1.351303e-06 0.000000e+00 0.270789297 5.333279e-01 2.718984e-0
## SPORTHR 5.771895e-01 2.707893e-01 0.000000000 1.025312e-02 6.185484e-0
## GENDER  4.206886e-01 5.333279e-01 0.010253119 0.000000e+00 1.757790e-1
## ACD     1.841368e-09 2.718984e-03 0.061854843 1.757790e-10 0.000000e+0
## STUDYHR 1.730743e-01 3.891322e-01 0.003930083 5.391071e-01 1.982234e-0
##             STUDYHR
## SPHEQ   0.173074294
## PARENTS 0.389132155
## SPORTHR 0.003930083
## GENDER  0.539107149
## ACD     0.198223365
## STUDYHR 0.000000000
```

```
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF",
                          "#77AADD", "#4477AA"))

#create correlation plot
corrplot(M, method="color", col=col(200),
         type="upper", order="hclust",
         addCoef.col = "black", # Add coefficient of correlation
         tl.col="black", tl.srt=45, #Text label color and rotation
         # Combine with significance
         p.mat = p.mat, sig.level = 0.01, insig = "blank",
         # hide correlation coefficient on the principal diagonal
         diag=FALSE )
```




```
vif(model)
```

```
##     SPHEQ   PARENTS    GENDER       ACD   SPORTHR    READHR
## 1.034080 1.032454 1.101447 1.092771 1.054899 1.060157
```

VIF 值 >= 10 表示高共线性。在这种情况下，所有 vif 值都接近小于 10

附录 A　余音绕梁

呐，到这里朕的书差不多写完了，但还有几句话要交待，所以开个附录，再🄴嗦几句，各位客官稍安勿躁、扶稳坐好。

参考文献

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2021). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.24.

索引