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！

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

```
## [1] "mingyuwang"
```

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

```
## [1] "C:/Users/rhong/Documents"
```

```r
library("tidyverse")
library("readr")
library("relaimpo")
library("interactions")
library("caret")
library("vip")
library("gridExtra")
library("earth")
```

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。

```r
## 代码写这里，并运行；
income_dat <- read_csv("data/talk11/income.data_.zip", show_col_types = FALSE)
```

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

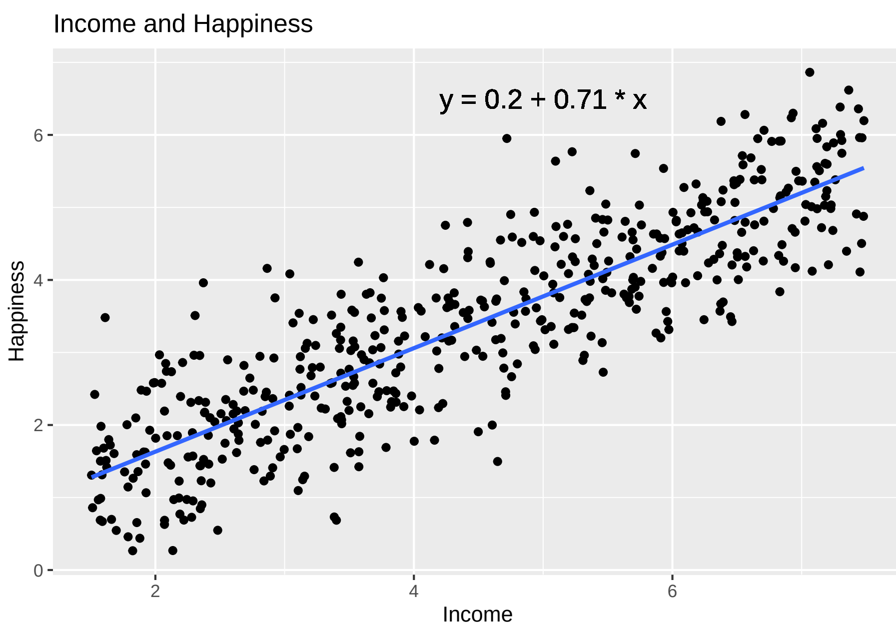
```r
# 线性回归分析，检验 income 与 happiness 的相关性
income_lm <- lm(happiness ~ income, data = income_dat)
summary(income_lm)
```

```
##
## Call:
## lm(formula = happiness ~ income, data = income_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02479 -0.48526  0.04078  0.45898  2.37805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20427    0.08884   2.299   0.0219 *
## income       0.71383    0.01854  38.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7181 on 496 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
## F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

```r
# 画出 income 与 happiness 的关系
(income_plot <- ggplot(income_dat, aes(x = income, y = happiness)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
```

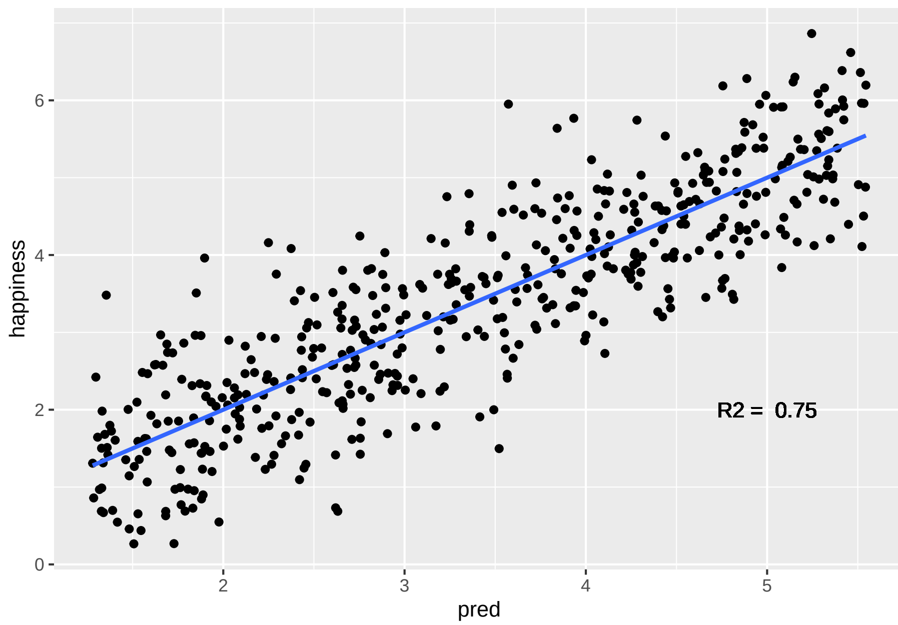
```
  geom_text(aes(label = paste("y = ", round(income_lm$coefficients[1], 2),
    " + ", round(income_lm$coefficients[2], 2), " * x", sep = "")),
    x = 5, y = 6.5, size = 5) +
  labs(title = "Income and Happiness", x = "Income", y = "Happiness"))
```

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




```
# 用得到的线性模型，以 income 为输入，预测 happiness 的值
income_pred <- predict(income_lm, income_dat)
# 画出预测值与真实 happiness 的关系，并在图上写出 R2 值
income_comp <- income_dat %>% mutate(pred = income_pred)
ggplot(income_comp, aes(x = pred, y = happiness)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_text(aes(label = paste("R2 = ", round(summary(income_lm)$r.squared, 2))),
    x = 5, y = 2, size = 4)
```

```
## `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？

```r
## 代码写这里，并运行；
heart_dat <- read_csv("data/talk11/heart.data_.zip", show_col_types = FALSE)
```

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

```r
# 1. 线性回归分析，检验 heart.disease 与 biking 和 smoking 的相关性
heart_lm1 <- lm(heart.disease ~ biking + smoking, data = heart_dat)
summary(heart_lm1)
```

```
##
## Call:
## lm(formula = heart.disease ~ biking + smoking, data = heart_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1789 -0.4463  0.0362  0.4422  1.9331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.984658   0.080137  186.99   <2e-16 ***
## biking      -0.200133   0.001366 -146.53   <2e-16 ***
## smoking      0.178334   0.003539   50.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.654 on 495 degrees of freedom
## Multiple R-squared:  0.9796, Adjusted R-squared:  0.9795
## F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

```r
coef(heart_lm1)
```

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

```
heart_lm2 <- lm(heart.disease ~ biking, data = heart_dat)
heart_lm3 <- lm(heart.disease ~ smoking, data = heart_dat)
data.frame(Biking = summary(heart_lm2)$r.squared,
    Smoking = summary(heart_lm3)$r.squared)
```

```
##      Biking    Smoking
## 1 0.8750769 0.09556196
```

```
# 2. 写出三者间关系的线性公式
paste0("HeartDisease = ", round(heart_lm1$coefficients[1], 2),
    " + ", round(heart_lm1$coefficients[2], 2), " * Biking",
    " + ", round(heart_lm1$coefficients[3], 2), " * Smoking")
```

```
## [1] "HeartDisease = 14.98 + -0.2 * Biking + 0.18 * Smoking"
```

```
# 3. 解释 biking 和 smoking 的影响（方向和程度）
coef(heart_lm1)
```

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

```
# biking 值的升高对应 heart.disease 值的下降，系数为 0.2,
# smoking 值的升高对应 heart.disease 值的升高，系数 0.17
```

```
# 4. biking 和 smoking 能解释多少 heart.disease 的 variance? 这个值从哪里获得?
# 97.96% 的 variance 能被解释, 可以从 summary(heart_lm1)$r.squared 获得
summary(heart_lm1)$r.squared
```

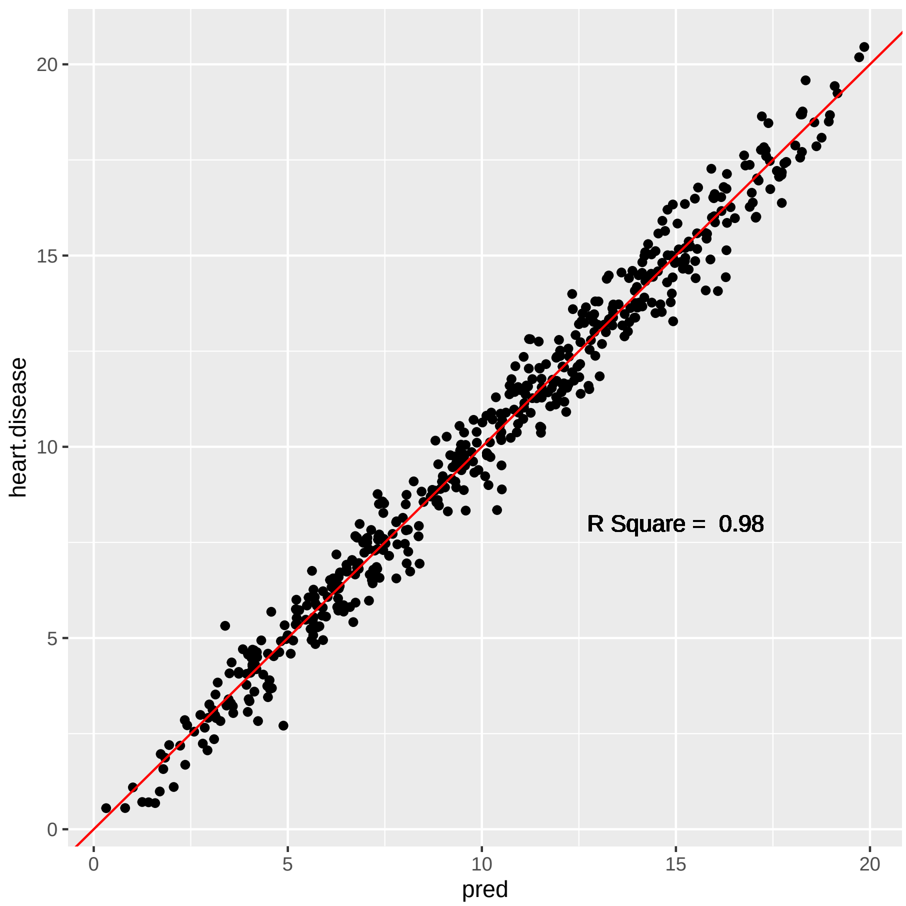
```
## [1] 0.9796175
```

```
# 5. 用 relaimpo 包的函数计算 biking 和 smoking 对 heart.disease 的重要性。哪个更重要?
calc.relimp(heart_lm1)
```

```
## 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
```

```
# 从 relative importance matrix 来看， biking 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
heart_pred <- predict(heart_lm1, heart_dat)
heart_comp <- heart_dat["heart.disease"] %>%
  mutate(pred = heart_pred)
ggplot(heart_comp, aes(x = pred, y = heart.disease)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  geom_text(aes(label = paste("R Square = ",
    round(summary(heart_lm1)$r.squared, 2))),
    x = 15, y = 8, size = 4)
```




```r
# 7. 在建模时考虑 biking 和 smoking 的互作关系
heart_lm4 <- lm(heart.disease ~ biking * smoking, data = heart_dat)
summary(heart_lm4)
```

```
##
## 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
```

```
anova(heart_lm1, heart_lm4)
```

```
## Analysis of Variance Table
##
## Model 1: heart.disease ~ biking + smoking
## Model 2: heart.disease ~ biking * smoking
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    495 211.74
## 2    494 211.52  1   0.21692 0.5066  0.477
```

```
data.frame(heart_lm1 = summary(heart_lm1)$r.squared,
    heart_lm4 = summary(heart_lm4)$r.squared)
```

```
##   heart_lm1 heart_lm4
## 1 0.9796175 0.9796383
```

没有提高模型的 R2 值。另外，从 anova 方差分析表中可以看出，P 值为 0.477，说明 biking 和 smoking 的互作关系对模型的解释力没有提高。

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"))
m_gml <- glm(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,
  data = iris_dat, family = binomial)
set.seed(1231)
data.frame(predicted = m_gml %>%
    predict(iris_dat, type = "link"), original = iris_dat$Species) %>%
    arrange(original) %>%
    sample_n(6)
```

```
##     predicted  original
## 74  24.81853 virginica
## 47 -30.11375    setosa
## 12 -27.25675    setosa
## 60  38.68465 virginica
## 36 -32.00469    setosa
## 53  35.29448 virginica
```

```
data.frame(predicted = m_gml %>%
    predict(iris_dat, type = "response"), original = iris_dat$Species) %>%
    arrange(original) %>%
    sample_n(6)
```

```
##        predicted  original
## 60 1.000000e+00 virginica
## 40 1.337875e-13    setosa
## 62 1.000000e+00 virginica
## 46 1.861773e-12    setosa
## 61 1.000000e+00 virginica
```

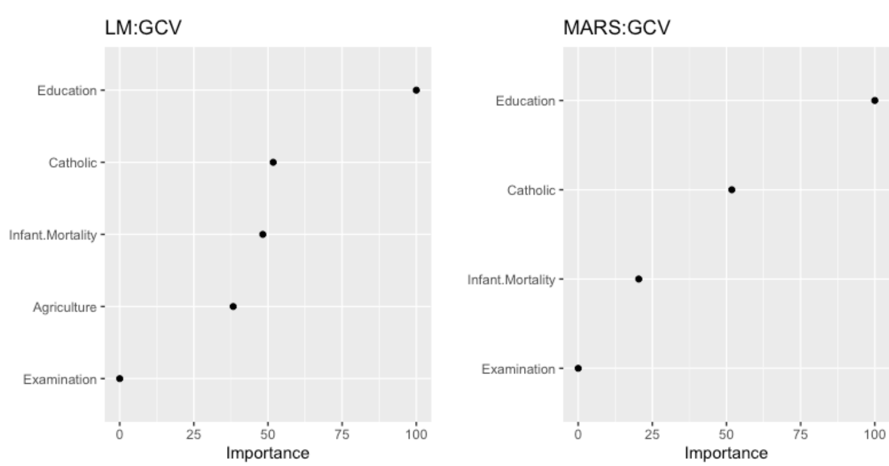
```
## 10 3.900993e-13     setosa
```

`type = "link"` 返回的是 logit 函数的值，也就是 log(odd)；`type = "response"` 返回的是概率值。

0.5 练习与作业 2：non-linear regression

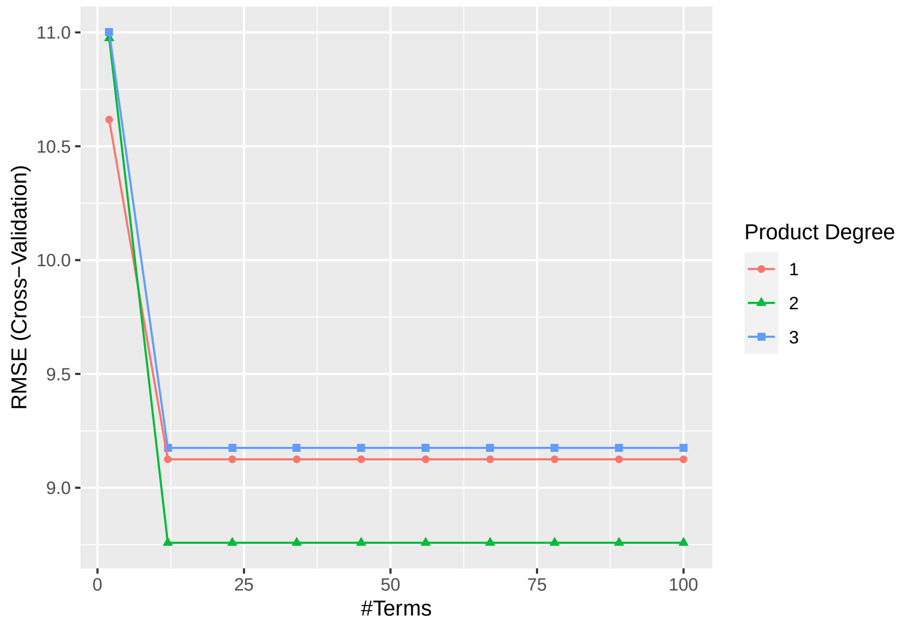
0.5.1 分析 `swiss`，用其它列的数据预测 `Fertility`

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；
hyper_grid <- expand.grid(
  degree = 1:3, ## number of interaction degrees
  nprune = seq(2, 100, length.out = 10) %>% floor() ## number of features to select
)
```

```r
em_swiss <- earth(swiss$Fertility ~ ., data = swiss, degree = 2)
set.seed(1231)
cv_em_swiss <- train(
  x = subset(swiss, select = -Fertility),
  y = swiss$Fertility,
  method = "earth",
  metric = "RMSE",
  trControl = trainControl(method = "cv", number = 10),
  tuneGrid = hyper_grid
)

lm_swiss <- lm(swiss$Fertility ~ ., data = swiss)
set.seed(1231)
cv_lm_swiss <- train(
  x = swiss[,-1],
  y = swiss$Fertility,
  method = "lm",
  trControl = trainControl(method = "cv", number = 10),
  tuneLength = 10
)
ggplot(cv_em_swiss)
```

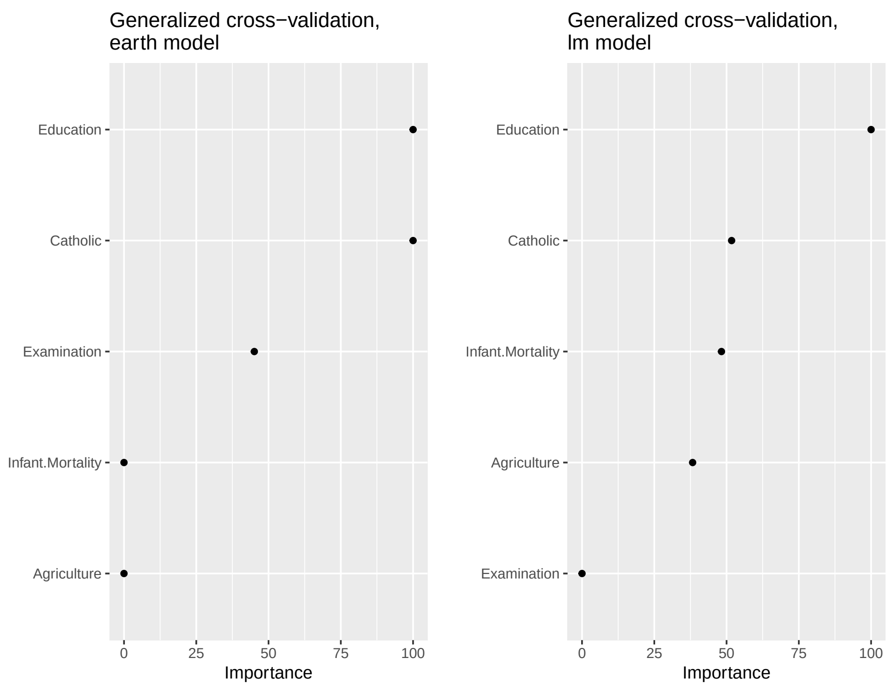



```
bt <- cv_em_swiss$bestTune

res1 <- cv_em_swiss$results %>%
  filter(degree == bt$degree & nprune == bt$nprune) %>%
  subset(select = c(RMSE, Rsquared)) %>%
  mutate(method = "earth")
res2 <- cv_lm_swiss$results %>%
  subset(select = c(RMSE, Rsquared)) %>%
  mutate(method = "lm")
bind_rows(res1, res2)
```

```
##       RMSE  Rsquared method
## 1 8.758456 0.5866177  earth
## 2 7.541866 0.7198014     lm
```

相比于 earth 方法，lm 方法的 RMSE 值更低，R2 值更高，可以说 lm 方法更好。

```r
p1 <- vip(cv_em_swiss, geom = "point", value = "gcv") +
  ggtitle("Generalized cross-validation,\nearth model")
p2 <- vip(cv_lm_swiss, geom = "point", value = "gcv") +
  ggtitle("Generalized cross-validation,\nlm model")

grid.arrange(p1, p2, ncol = 2)
```



从上图中可以看出，`earth` 模型中 Education 和 Catholic 的重要性达到 100%，Examination 的重要性为 40% 左右，Infant.Mortality 和 Agriculture 的重要性为 0%；而 `lm` 模型中 Education 的重要性为 100%，Catholic Infant.Mortality 和 Agriculture 的重要性为 40% 左右，Examination 的重要性为 0%。只有 Education 在两个模型中的重要性一致，另外四个因素的重要性在两个模型中不一致。