丁香园

丁香公开课
CLASS.DXY.CN

临床预测模型构建&机器学习(R语言进阶)

第2章 无序多分类与有序分类Logistic回归

周支瑞

CONTENT

01 | 无序多分类Logistic回归

02 | 有序分类Logistic回归

无序多分类 logistic 回归

- 多元Logistic回归模型被用来建立有多个输出变量的模型，且这些预测变量通过一个线性组合变成为一个最终的预测变量。Multinomial Logistic 回归模型中因变量可以取多个水平值.

https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

Analysis methods you might consider

- Multinomial probit regression, similar to multinomial logistic regression with independent normal error terms.
- Multiple-group discriminant function analysis. A multivariate method for multinomial outcome variables
- Multiple logistic regression analyses, one for each pair of outcomes: One problem with this approach is that each analysis is potentially run on a different sample. The other problem is that without constraining the logistic models, we can end up with the probability of choosing all possible outcome categories greater than 1.

https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

无序多分类 logistic 回归 案例

- [案例] 进入高中的学生在做计划选择时，面临着**一般计划、职业计划及学术计划**(general, vocational, academic)三种选择。他们的选择往往是通过对他们的写作成绩及社会经济地位进行模型化决定。hsbdemo.dta这个数据集包含了200个学生的记录，结果变量为prog(program type)三种项目类型，预测变量为：ses(social economic status)社会经济地位，3个分类变量（id, female, schtyp)及其他连续变量。

https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

无序多分类 logistic 回归模型

$$ln\left(\frac{P(prog=general)}{P(prog=academic)}\right)=b_{10}+b_{11}(ses=2)+b_{12}(ses=3)+b_{13}write$$

$$ln\left(\frac{P(prog=vocation)}{P(prog=academic)}\right)=b_{20}+b_{21}(ses=2)+b_{22}(ses=3)+b_{23}write$$

https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

无序多分类 logistic 回归 案例代码

```
## 利用 nnet 包中的函数multinom， 建立多元logistic回归模型：
library(foreign)
library(nnet)
library(ggplot2)
library(reshape2)
ml <- read.dta("hsbdemo.dta")
with(ml, table(ses, prog))
with(ml, do.call(rbind, tapply(write, prog, function(x) c(M = mean(x), SD = sd(x)))))
ml$prog2 <- relevel(ml$prog, ref = "academic")
test <- multinom(prog2 ~ ses + write, data = ml)
summary(test)
# 2-tailed z test
z <- summary(test)$coefficients/summary(test)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

无序多分类 logistic 回归 案例代码

```
# extract the coefficients from the model and exponentiate
exp(coef(test))
head(pp <- fitted(test))
dses <- data.frame(ses = c("low", "middle", "high"), write = mean(ml$write))
predict(test, newdata = dses, "probs")
dwrite <- data.frame(ses = rep(c("low", "middle", "high"), each = 41), write = rep(c(30:70),3))
# store the predicted probabilities for each value of ses and write
pp.write <- cbind(dwrite, predict(test, newdata = dwrite, type = "probs", se = TRUE))
# calculate the mean probabilities within each level of ses
by(pp.write[, 3:5], pp.write$ses, colMeans)
#melt data set to long for ggplot2
lpp <- melt(pp.write, id.vars = c("ses", "write"), value.name = "probability")
head(lpp)  # view first few rows
# plot predicted probabilities across write values for each level of ses
# facetted by program type
ggplot(lpp, aes(x = write, y = probability, colour = ses)) + geom_line() + facet_grid(variable ~., scales =
"free")
```

无序多分类 logistic 回归 案例代码

```
## 程序包mlogit提供了多项logit的模型拟合函数
install.packages("mlogit")
library(Formula)
library(maxLik)
library(miscTools)
library(mlogit)
data("Fishing", package = "mlogit")
Fish <- mlogit.data(Fishing,shape = "wide",choice = "mode")
summary(mlogit(mode ~ 0|income, data = Fish))
```

程序包mlogit多项logit的模型拟合函数

```
## 程序包mlogit提供了多项logit的模型拟合函数
install.packages("mlogit")
library(Formula)
library(maxLik)
library(miscTools)
library(mlogit)
data("Fishing", package = "mlogit")
Fish <- mlogit.data(Fishing, shape = "wide", choice = "mode")
summary(mlogit(mode ~ 0|income, data = Fish))
```

mlogit参数说明

- formula：mlogit提供了条件logit，多项logit，混合logit多种模型，对于多项logit的估计模型应写为：因变量~0|自变量,如：mode ~ 0|income
- data：使用mlogit.data函数使得数据结构符合mlogit函数要求。
- choice：确定分类变量是什么
- shape：如果每一行是一个观测，我们选择wide，如果每一行是表示一个选择，那么就应该选择long。
- alt.var：对于shape为long的数据，需要标明所有的选择名称

CONTENT

01 | 无序多分类Logistic回归

02 | 有序分类Logistic回归

等级logistic回归案例

Example 1: A study looks at factors that **influence the decision of whether to apply to graduate school**. College juniors are asked if they are **unlikely, somewhat likely, or very likely** to apply to graduate school. Hence, our outcome variable has three categories. Data on **parental educational status**, whether the undergraduate institution is **public or private**, and **current GPA** is also collected. The researchers have reason to believe that the “distances” between these three points are not equal. For example, the “distance” between “unlikely” and “somewhat likely” may be shorter than the distance between “somewhat likely” and “very likely”.

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

Analysis methods you might consider

- Ordered logistic regression: the focus of this page.
- OLS regression: This analysis is problematic because the assumptions of OLS are violated when it is used with a non-interval outcome variable.
- ANOVA: If you use only one continuous predictor, you could “flip” the model around so that, say, gpa was the outcome variable and apply was the predictor variable. Then you could run a one-way ANOVA. This isn’t a bad thing to do if you only have one predictor variable (from the logistic model), and it is continuous.
- Multinomial logistic regression: This is similar to doing ordered logistic regression, except that it is assumed that there is no order to the categories of the outcome variable (i.e., the categories are nominal). The downside of this approach is that the information contained in the ordering is lost.
- Ordered probit regression: This is very, very similar to running an ordered logistic regression. The main difference is in the interpretation of the coefficients.

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

程序包MASS提供polr()函数

```
polr(formula, data, weights, start, ..., subset, na.action,
     contrasts = NULL, Hess = FALSE, model = TRUE,
     method = c("logistic", "probit", "cloglog", "cauchit"))
```

参数说明：

formula：回归形式，与lm()函数的formula参数用法一致

data：数据集

method：默认为order logit，选择probit时变为order probit模型。

等级logistic回归案例 代码

```
## 程序包MASS提供polr()函数可以进行ordered logit或probit回归
require(foreign)
require(ggplot2)
require(MASS)
require(Hmisc)
require(reshape2)
```

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

等级logistic回归案例 代码续

```
dat <- read.dta("ologit.dta")
head(dat)
## one at a time, table apply, pared, and public
lapply(dat[, c("apply", "pared", "public")], table)
## three way cross tabs (xtabs) and flatten the table
ftable(xtabs(~ public + apply + pared, data = dat))
summary(dat$gpa)
sd(dat$gpa)
```

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

等级logistic回归案例 代码续

```
ggplot(dat, aes(x = apply, y = gpa)) +
        geom_boxplot(size = .75) +
        geom_jitter(alpha = .5) +
        facet_grid(pared ~ public, margins = TRUE) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

等级logistic回归案例 代码续

```
## fit ordered logit model and store results 'm'
m <- polr(apply ~ pared + public + gpa, data = dat, Hess=TRUE)
## view a summary of the model
summary(m)
```

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

等级logistic回归案例 代码续

```
## store table
(ctable <- coef(summary(m)))
## calculate and store p values
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
## combined table
(ctable <- cbind(ctable, "p value" = p))
```

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

等级logistic回归案例 代码续

```
(ci <- confint(m)) # default method gives profiled CIs
confint.default(m) # CIs assuming normality
## odds ratios
exp(coef(m))
## OR and CI
exp(cbind(OR = coef(m), ci))
```

https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

需要考虑的问题

- Perfect prediction: Perfect prediction means that one value of a predictor variable is associated with only one value of the response variable. If this happens, Stata will usually issue a note at the top of the output and will drop the cases so that the model can run.
- Sample size: Both ordered logistic and ordered probit, using maximum likelihood estimates, require sufficient sample size. How big is big is a topic of some debate, but they almost always require more cases than OLS regression.
- Empty cells or small cells: You should check for empty or small cells by doing a crosstab between categorical predictors and the outcome variable. If a cell has very few cases, the model may become unstable or it might not run at all.
- Pseudo-R-squared: There is no exact analog of the R-squared found in OLS. There are many versions of pseudo-R-squares. Please see Long and Freese 2005 for more details and explanations of various pseudo-R-squares.
- Diagnostics: Doing diagnostics for non-linear models is difficult, and ordered logit/probit models are even more difficult than binary models. For a discussion of model diagnostics for logistic regression, see Hosmer and Lemeshow (2000, Chapter 5). Note that diagnostics done for logistic regression are similar to those done for probit regression.

小结：各种logistic回归方法选择

- 1. 普通二分类 logistic 回归用广义线性模型 glm()
- 2. 因变量多分类 logistic 回归
  - a. 有序分类因变量：用 MASS 包里的 polr
  - b. 无序分类因变量：用 nnet 包里的 multinom
- 3. 条件logistic回归 用 survival 包里的 clogit

请在此处输入小标题

丁香公开课
CLASS.DXY.CN

感谢观看

THANKS

丁香园特邀讲师 周支瑞

丁香园