第六章作业

计算机2001 袁子宸 20205962

题目与要求：

在新生儿出生率数据集上使用Logistic回归模型对新生儿是否需要急救进行预测（程序代码已下发）。回答以下问题：  
（1）通过调用系数函数和概要函数，尝试对自变量系数进行解释，并通过残差概要、伪R-平方、AIC准则对模型质量进行评价；  
（2）通过准确率和召回率、输出概率的双密度图对分类器性能进行评价。  
要求通过附件提交作业，作业中需要有程序执行结果，除了以上评价指标，还可以选择其他指标（第四章中分类模型和打分模型的度量标准）对模型性能进行评价（这部分酌情加分）。

程序执行：

加载数据：

load("D://user//yzc//Desktop//NatalRiskData.rData")

train - sdata[sdataSoRIGRANDGRoUP<-5,]

test <- sdata[sdataSORIGRANDGROUP>5,]

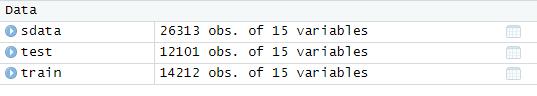


图1

Complications <- c("ULD MECO","ULD\_PRECIP","ULD\_BREECH")  
riskfactors <- C("URF\_DIAB", "URF\_CHYPER", "URF\_PHYPER",

y <- "atRisk"

X<-c("PWGT",

"UPREVIS",

"CIG\_REC",

"GESTREC3",

"DPLURAL",

complications,

riskfactors)

fmla <- paste(y, paste(x,collapse="+"), sep = "～")

print(fmla)

使用 summary 命令，输出模型概要：

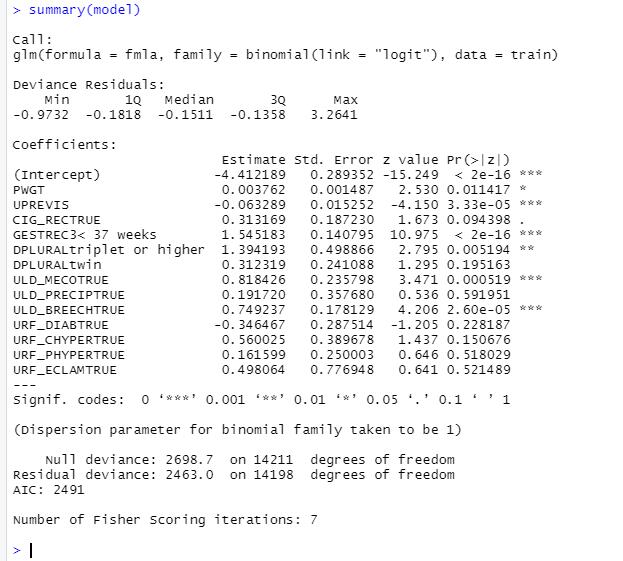


图 2

使用coefficients 函数，输出自变量的系数：

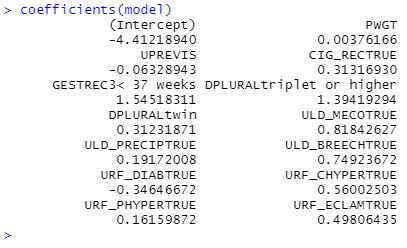


图 3

使用glm ()命令构建Logistic回归模型：

logit <- function(p) {log(p/(1-p)) }

s <- function(x) {1/(1 + exp(-x))}

s(logit(0.6))

logit(s(-4))

如果在变量pred中存储了预测值，并且没有指定type="response"，那么默认的predict()函数会返回link()函数的输出，即logit(y)。为了确保predict()函数返回被预测的y的概率，你必须指定type="response"。

library(ggplot2)

ggplot(train, aes(x=pred, color=atRisk, linetype=atRisk)) + geom density()

输出概率的双密度图：

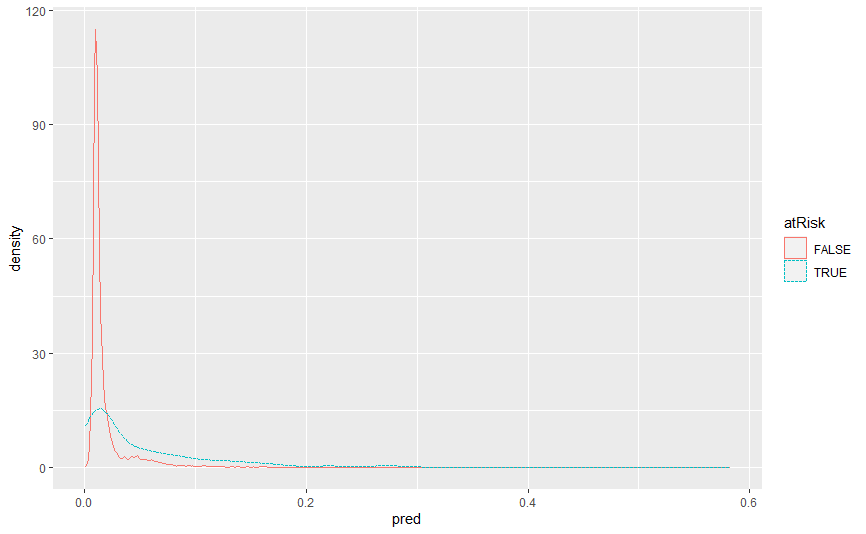


图 4

图4显示了正例（风险新生儿）和负例（非风险新生儿）的分数分布。正例（风险新生儿）在数据集中很少，只占总体的1.8%。两种分数分布都比较集中在左侧，表明正负例的分数都较低。但负例（非风险新生儿）的分布比正例（风险新生儿）的分布降得更快。这意味着这个模型能够在数据中发现一些子集，在这些子集中，风险新生儿的比例高于总体平均值。如果要将这个模型用作分类器，就需要选择一个阈值。分数高于这个阈值的就是正例，低于这个阈值的就是负例。在选择阈值时，需要在准确率（分类器预测为正例的实际上是多少正例）和召回率（分类器发现了多少正例）之间取得平衡。分类器的准确率与正例的比率之比称为浓缩率。分类器越准确，就需要选择越高的阈值。

准确率和召回率：

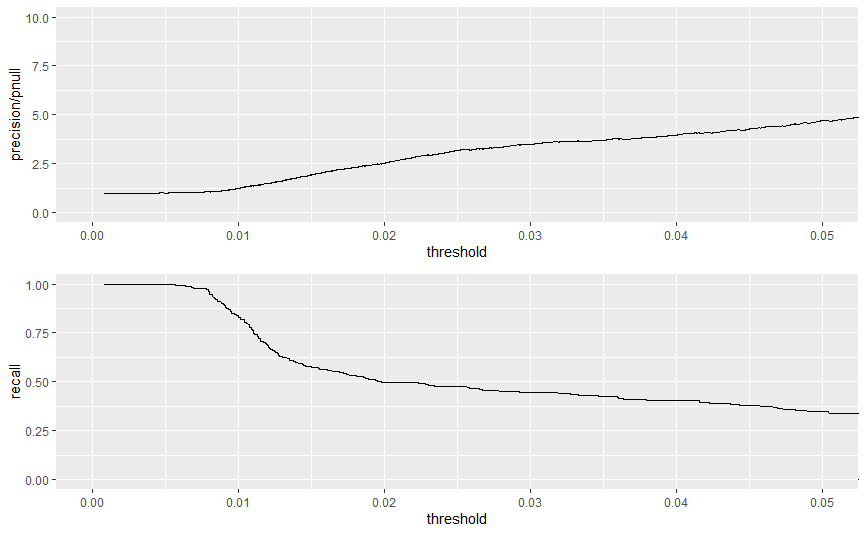


图 5

评价：

对自变量系数进行解释：例如，变量DPLURAL有3个类别值，分别为单胞胎、双胞胎和三胞胎及以上，logistic回归模型有两个对应的系数：DPLURALtwin 和DPLURALtriple or higher。参考类别值是单胞胎，DPLURAL的两个系数均为正，表明在所有其他变量都相等的情况下，多胞胎比单胞胎具有更高的风险概率。

残差概要 ：在Logistic回归模型中，偏差残差与观察到的真实结果的对数似然值是相关的。对数似然值体现的思想是：对于正例 y，在模型中发生的概率 py 应该较高，而负例的发生概率应该较低。对数似然函数会奖励预测概率 py 与真实结果 y 之间的匹配，而惩罚不匹配（如负例的高 py 或正例的低 py）。与线性回归模型通过最小化残差平方和来求解不同，逻辑斯谛回归模型是通过最小化偏差残差平方和来求解的，这相当于在给定模型的条件下最大化模型数据的对数似然值。

伪R-平方 ：该模型仅仅解释了偏差的7.7%~8.7%左右，它不是一个高预测模型，这表明我们仍然还未识别出实际上用于预测风险新生儿的所有因素。

对分类器性能进行评价:

输出概率的双密度图并不符合我们的期望。我们希望的是负例的打分分布集中在左侧，而正例的打分分布集中在右侧，这样两种分布就分离了。然而，在这个例子中，两种分布都集中在左侧，表明正负例的打分都较低。这表明该模型能够在数据中找到一些子集，在这些子集中，风险新生儿的比率高于总体平均值。