生成表格：

library(dplyr)

表格

描述已自动生成marks <- c(80,60,34,56,70,56,65,95)

gender <- c("M","F","M","M","F","M","M","F")

name <- c("John", "Emma", "Peter", "Dave", "Jane", "Rob", "Chris", "Emily")

unit <- c("CITS4009","CITS5503","CITS1401","CITS4009","CITS5503","CITS4009","CITS1401","CITS1401")

df <- data.frame(name=name, gender=gender, unit=unit, marks=marks)

插入新列：

df[,"isPass"] <- ifelse(df[,"marks"] >= 50, "Pass","Fail")

df[,"isHigh"] <- ifelse(df[,"unit"]=="CITS4009" | df[,"unit"]=="CITS5503", 1, 0)

表格

描述已自动生成

切分数据集:

set.seed(729375)

df$rgroup <- runif(dim(df)[1])

dTrain <- subset(df, rgroup<=0.8)

dCalibration <- subset(df, rgroup>0.8)

dTrain

dCalibration

计算AIC的函数：

calAIC <- function(ytrue, ypred){

#log\_likelihood = sum(ytrue\*log(ypred) + (1-ytrue)\*log(1-ypred))

log\_likelihood = sum(ifelse(ytrue == ypred, log(ypred), log(1-ypred)))

deviance = (-2)\*(log\_likelihood - 0)

AIC = deviance + 2\*4

AIC

} 【注意ytrue为0或1；ypred为模型预测后的预测概率，在0-1之间】

用逻辑回归模型预测isHigh的值：

y <- "isHigh"

x <- c("name", "gender", "unit", "marks")

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

model <- glm(fmla, data=dTrain, family=binomial(link="logit"))

dCalibration$pred <- predict(model, newdata= dCalibration, type="response")

calAIC(dCalibration[,isHigh], dCalibration[,pred])