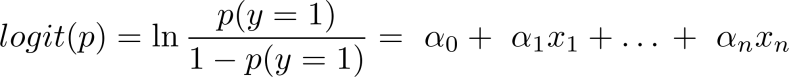
1. **Logistic回归模型**
2. **基本原理**

Logistic回归是研究因变量为二分类或多分类观察结果与影响因素（自变量）之间关系的一种多变量分析方法，属于概率型非线性回归。在医学领域中，经常遇到根据病人症状和体征进行能否治愈、能否生存、是否发病或者是否得某病的一个预测，这类问题的预测可以归纳为取值只有两个的二值问题，能（是）用1表示，否用0表示。但这类问题中存在对分类结果发生概率有影响的因素往往不止一个，对于这类问题可以使用Logistic回归建立预测模型，用于预测这类问题发生的概率。

Logistic回归其思想是基于线性回归模型，将线性回归模型函数映射到函数值介于0和1之间的sigmoid函数中，可以表示数据属于某一类型的概率。如用sigmoid函数值大于0.5将数据划分为A类，低于0.5则划分为B类。

Logistic回归模型作logit转换后的公式如下：



其中，表示因变量（y）发生的概率与不发生的概率之比的自然对数（也称为风险比率），，，…，为自变量，称为常数项或截距，，，…，为偏回归系数，在疾病相关二值预测问题中，就表示可能导致疾病的因素，通过logistic回归可以求出因变量y发生的概率。

1. **应用介绍**

采用logistic回归分析解决二值分类医学实际问题的时候，一般需要经历如下几个步骤：收集病历资料，清洗数据，建立logisitic回归模型，分析评价模型，应用模型。下面我们采用R语言以一个医学实例给大家展示Logisitic回归分析，通过计算，还可以通过建立的模型了解致病因素在疾病诊断中的贡献度，因此这种方法在某种程度上可以对疾病诊断进行定量分析，实现计量诊断。下面我们按从数据准备到分析的过程来说明Logistic回归模型的应用。

1. 数据准备和清洗

病历数据的收集可以来自于实际临床，对于中医临床数据的收集可能还存在症状、证型等术语的规范化问题，本文不做讨论。本节实验采用加州大学尔湾分校机器学习和智能系统中心的心脏病数据。

（1）下载数据：

url<-"http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data"

data<-read.csv(url,header = FALSE) #读取数据放入data中

head(data) #查看部分数据，此时看到的数据列名默认为V1至V14，如下表所示

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14  
## 1 63 1 1 145 233 1 2 150 0 2.3 3 0.0 6.0 0  
## 2 67 1 4 160 286 0 2 108 1 1.5 2 3.0 3.0 2  
## 3 67 1 4 120 229 0 2 129 1 2.6 2 2.0 7.0 1  
## 4 37 1 3 130 250 0 0 187 0 3.5 3 0.0 3.0 0  
## 5 41 0 2 130 204 0 2 172 0 1.4 1 0.0 3.0 0  
## 6 56 1 2 120 236 0 0 178 0 0.8 1 0.0 3.0 0

colnames(data) <- c("age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slope","ca","thal","hd") #将列名根据数据实际属性名进行命名

str(data) #列名更名后用str查看每个变量的类型如下所示，发现sex、cp、hd等等都是数值类型。

## 'data.frame': 303 obs. of 14 variables:  
## $ age : num 63 67 67 37 41 56 62 57 63 53 ...  
## $ sex : num 1 1 1 1 0 1 0 0 1 1 ...  
## $ cp : num 1 4 4 3 2 2 4 4 4 4 ...  
## $ trestbps: num 145 160 120 130 130 120 140 120 130 140 ...  
## $ chol : num 233 286 229 250 204 236 268 354 254 203 ...  
## $ fbs : num 1 0 0 0 0 0 0 0 0 1 ...  
## $ restecg : num 2 2 2 0 2 0 2 0 2 2 ...  
## $ thalach : num 150 108 129 187 172 178 160 163 147 155 ...  
## $ exang : num 0 1 1 0 0 0 0 1 0 1 ...  
## $ oldpeak : num 2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...  
## $ slope : num 3 2 2 3 1 1 3 1 2 3 ...  
## $ ca : Factor w/ 5 levels "?","0.0","1.0",..: 2 5 4 2 2 2 4 2 3 2 ...  
## $ thal : Factor w/ 4 levels "?","3.0","6.0",..: 3 2 4 2 2 2 2 2 4 4 ...  
## $ hd : int 0 2 1 0 0 0 3 0 2 1 ...

在数据处理过程中，我们发现数据大多是数值型数据，但对于性别、胸痛级别等变量的数值只是表示分类，而没有数值实际意义，因此这类变量就转换为factor类型。

（2）根据数据样本查看数据，并对数据进行清洗。

1.资料审核（1）资料的性质；（2）样本数量特征

2.离散型变量的数量化及其连续型变量的规格化

（二）logistic回归方程的建立

（三）logistic回归模型的假设检验

（四）根据样本资料估计参数

（五）变量的筛选，并对各自变量的作用大小作出评价

url<-"http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data"

data<-read.csv(url,header = FALSE)

str(data)

#cp类型应该用factor类型，因为cp是指chest pain，分为三级1，2，3；4表示没有chest pain

#ca and thal are correctly called factors, but one of the levels is "?" when we need it to be NA

data[data=="?"]<- NA

str(data)

data[data$sex==0,]$sex <- "F"

data[data$sex==1,]$sex <- "M"

data$sex <- as.factor(data$sex)

data$cp <- as.factor(data$cp)

data$fbs <- as.factor(data$fbs)

data$restecg <- as.factor(data$restecg)

data$exang <- as.factor(data$exang)

data$slope <- as.factor(data$slope)

data$ca <- as.integer(data$ca)

data$thal <- as.integer(data$thal)

data$hd <- ifelse(test = data$hd == 0, yes = "Healthy", no = "unhealthy")

data$hd <- as.factor(data$hd)

str(data)

nrow(data[is.na(data$ca) | is.na(data$thal),])

data[is.na(data$ca) | is.na(data$thal),]

nrow(data)

data <- data[!(is.na(data$ca) | is.na(data$thal)),] #删除Na值的行

nrow(data)

xtabs(~hd + age, data = data)

xtabs(~hd + sex, data = data)

xtabs(~hd + cp, data = data)

xtabs(~hd + trestbps, data = data)

xtabs(~hd + chol, data = data)

xtabs(~hd + fbs, data = data)

xtabs(~hd + restecg, data = data)

xtabs(~hd + age, data = data)

logistic <- glm(hd~., data = data, family = binomial(link = "logit"))

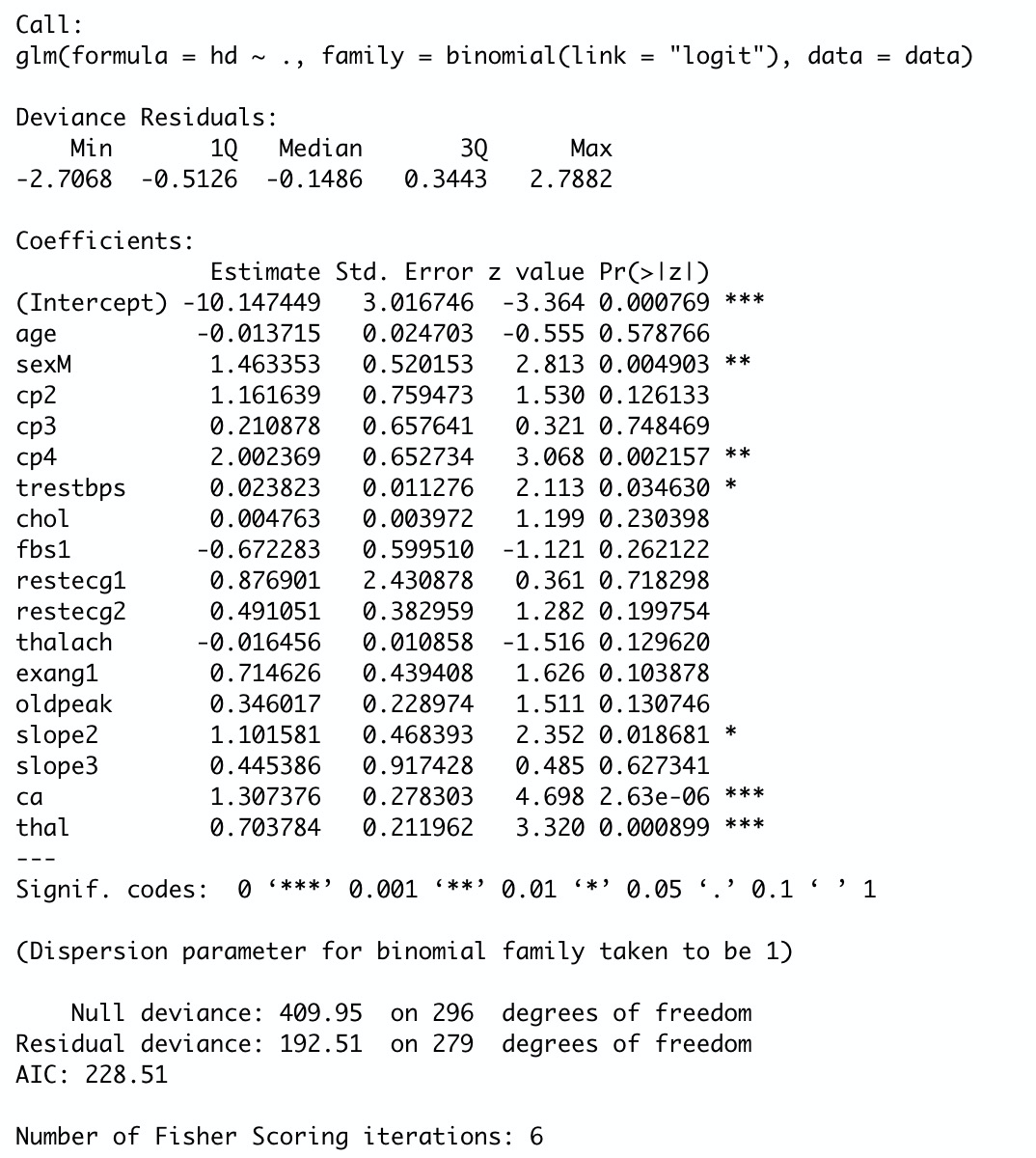
summary(logistic)

predicted.data <- data.frame(probability.of.hd= logistic$fitted.values, hd = data$hd)

predicted.data <- predicted.data[order(predicted.data$probability.of.hd,decreasing = FALSE),]

predicted.data

predicted.data$rank <- 1:nrow(predicted.data)



#绘制概率图

library(ggplot2)

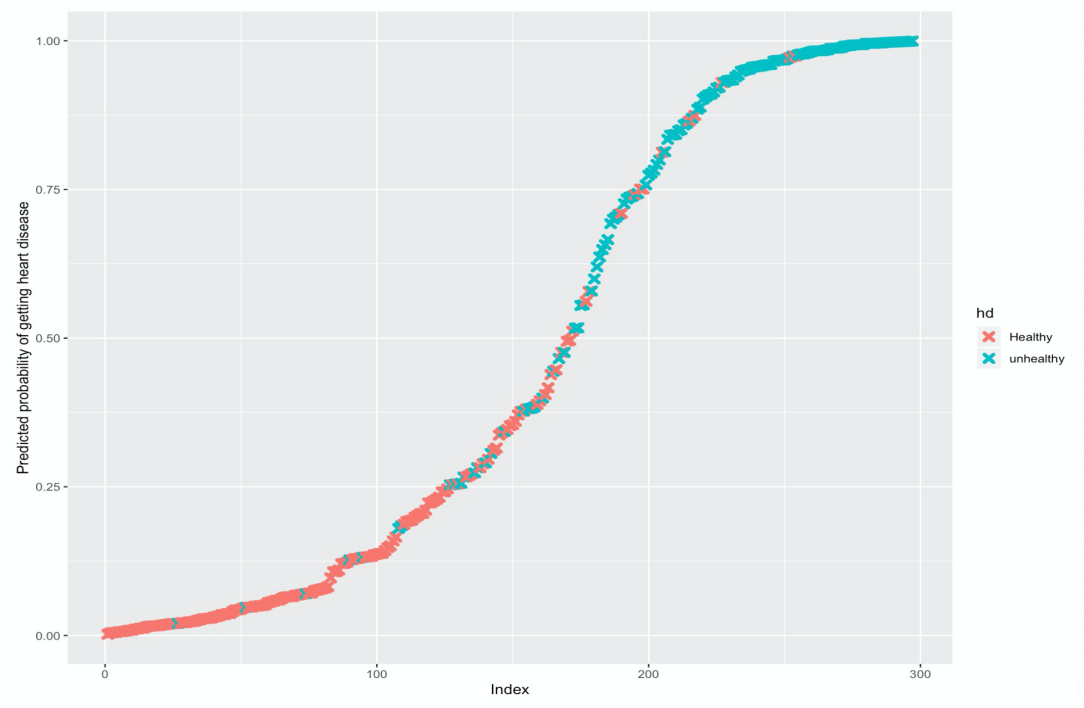
library(cowplot)

ggplot(data=predicted.data,aes(x=rank, y=probability.of.hd))+

geom\_point(aes(color=hd),alpha=1, shape=4, stroke=2)+

xlab("Index")+

ylab("Predicted probability of getting heart disease")



ggsave("heart\_disease\_probability.pdf")

# testing overall contribution of chol fbs1....

#install.packages("aod")

library(aod)

wald.test(b=coef(logistic), Sigma = vcov(logistic), Terms = 2) #Chi-squared test

wald.test(b=coef(logistic), Sigma = vcov(logistic), Terms = 4:5)

wald.test(b=coef(logistic), Sigma = vcov(logistic), Terms = 8:14)

#asses model using cross validation...

install.packages("caret")

library(lattice)

library(caret)

crossValSettings <- trainControl(method = "repeatedcv",number = 10,

savePredictions = TRUE)

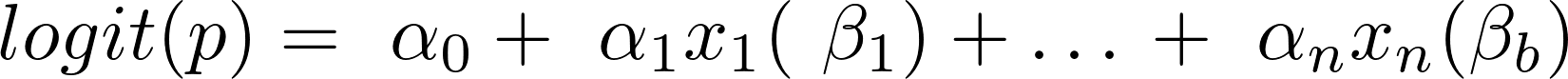
crossVal <- train(hd ~ sex + cp + trestbps + slope + ca + thal, data=data,

family = "binomial", method = "glm", trControl = crossValSettings)

pred <- predict(crossVal, newdata = data)

confusionMatrix(data = pred, data$hd)

针对数据库都是二值变量的情形，采用国际通用的sas6.12统计软件中的非条件Logistic多元逐步回归方法筛选变量，最终获得回归方程数学模型，表达为：



这里的“非条件”是针对成组设计的资料进行的分析，Logistic回归是处理二值型变量的回归方法，逐步回归包括两个相反的过程：（1）不断从模型以为的变量中挑选“重要”的引入模型。（2）对引入模型的变量进行综合分析，蒋哪些“不重要”的变量从模型中剔除。在回归方程中，logit(p)表示因变量发生的概率与不发生的概率之比的自然对数，，，…，为自变量，，，…，为偏回归系数，，，…，为标准化系数。

在变量的筛选中，（1）以某一症状、证型、病位、病因为被研究对象（因变量），从198个常见证型中筛选临床证型（或伴随证型），从80个常见病位中筛选发病部位，从88个常见病因中筛选病因，从817个常见症状中筛选症状（或伴随症状），从781种常用药物中筛选用药，被选中的对象称为自变量。（2）坚持自变量较少而你和优度较高的原则，即简约性。随着被纳入的自变量的数量增加，AIC（Akaike信息标准）减少，表示拟合优度增加。容易发现AIC先快后慢地减少，那么AIC刚进入缓慢减小阶段的自变量组成的模型正是我们寻找的简约模型。见图1.（3）进入模型的自变量在进行Wald 检验时，应至少在水平上统计显著。（4）对于与因变量在数据结构上产生空单元（Zero Cell Count）或准完全分离（quasi-complete separation）的自变量，则根据计算机提示，将其从数据库中临时删除并留待进一步研究。总之，整个变量筛选过程不做任何中医专业性分析和取舍。

1. **应用要求**
2. **资料的性质[１]**

因变量P必须是限于0～1之间的数据或二分类数据（0,1）

自变量的要求则随所用的Logistic回归分析方法不同而有所区别。

若采用Logistic判别法，则要求自变量X1,X2,…,XP服从下列①或②或③或④。

若采用Logistic最大似然法，则要求自变量X1,X2,…,XP服从下列②或③或④或⑤。

①要服从多元正态分布且协方差距阵相同

②自变量X1,X2,…,XP为多元独立的二分类变量

③自变量X1,X2,…,XP都服从对数正态的离散变量

④为①和③两项同时存在

⑤自变量X1,X2,…,XP不一定符合正态分布

1. **样本数量特征**

①样本含量主要取决于进入模型待分析变量的多少、每个变量 所需分的层数、研究所需的显著性水平三个要素。

②应用非条件Logistic回归分析，要求样本含量大。

③应用条件Logistic回归分析，所需样本含量一般少于非条件 Logistic回归分析。原则上应多于50对。

④原则上要保证每个待分析的变量在分层后，每层内的频数数 量以不少于5例为宜。

**Logistic回归模型的适用条件：**

1 因变量为二分类的分类变量或某事件的发生率，并且是数值型变量。但是需要注意，重复计数现象指标不适用于Logistic回归。

2 残差和因变量都要服从二项分布。二项分布对应的是分类变量，所以不是正态分布，进而不是用最ƒ小二乘法，而是最大似然法来解决方程估计和检验问题。

3 自变量和Logistic概率是线性关系

4 各观测对象间相互独立。

参考文献：

1. <http://cc.shutcm.edu.cn/G2S/eWebEditor/uploadfile/20150301181753287.pdf>

Poisson regression

1. **其他方法（综述形式）**

近年来人工智能发展迅速，概率统计方法在机器学习中的运用越来越广泛，并且不再是单纯的某种方法的应用，往往会结合多种算法提高分类和聚类的效果。除了本章前11节详细介绍的常用的几种可用于中医计量诊断的方法外，还有一些方法也可用于症候分类和症候选择，这些方法都可以在定性的基础上进行一定程度的定量分析。BPML,MLKNN,BSVM,ECC,LIFT等多标记算法，CART和C4.5算法

1. 随机森林

随机森林是一个包含多个决策树的分类器，属于集成学习算法中的Bagging类型，通过组合多个若分类器，通过投票或取均值获得最终结果，具有较高的精度和泛化能力。该算法名称由贝尔实验室Tin Kam Ho提出的随机决策森林而来，由Leo Breinman于2001提出。。由于中医临床证型多样且多兼证，采用随机森林可以得到较好的分类效果，并可获得每个症状/体征在模型中的分值，同时还具有较好的解释性。

洪燕珠等[1]引入随机森林方法研究慢性疲劳症候要素分类，通过对样本中的95个症状编码，并选取慢性疲劳常见症候要素的中医症状，通过随机森林算法进行计算，选取5次实验中的症状交集作为最终结果，所得分类模型预测准确率超过94%，具有良好的分类效果。王华珍等[2]采用随机森林构建“内生五邪”病症智能诊断模型，并通过对比实验证明该模型在稳健性和准确性方面都优于支持向量机和贝叶斯神经网络模型的效果。徐玮斐等[3]运用随机森林结合多标记学习算法对症候建模，挑选出15个特征症状，相比与采用信息增益算法选出的特征症状更加精简，提高了诊断模型的识别率。颜建军等[4]采用机制随机森林法对919例慢性胃炎数据样本的胃湿热、湿浊中阻等6个证型进行分析，发现该算法在6个多标记评价中均优于MLKNN、BPMLL等算法，具有较好的分类效果，同时还具有良好的可解释性。

1. Bridge回归
2. Lasso回归

参考文献：

1. 洪燕珠, 周昌乐, 张志枫, 等. 基于随机森林法的慢性疲劳证候要素特征症状的选择[J]. 中医杂志, 2010, 51(7): 634-638.
2. 王华珍, 胡雪琴. 中医 “内生五邪” 的智能证型分类[D]. , 2011.
3. 徐玮斐, 顾巍杰, 刘国萍, 等. 基于随机森林和多标记学习算法的慢性胃炎实证特征选择和证候分类识别研究[J]. 中国中医药信息杂志, 2016 (2016 年 08): 18-23.
4. 颜建军, 胡宗杰, 刘国萍, 等. 基于极值随机森林的慢性胃炎中医证候分类[J]. 华东理工大学学报 (自然科学版), 2017 (5): 698-703.
5. 胡宗杰. 基于极值随机森林的慢性胃炎中医问诊证候分类研究[D]. 华东理工大学, 2017.