Trabalho de dados Binários

Acidentes de carro

Laís Hoffmam, Simone Matsubara, Yasmin Fernandes, Willian Meira

2020-09-14

´´´{r, include=FALSE} library(lattice) library(readxl) library(readr) library(tidyverse) library(MASS) library(DAAG) library(gridExtra) library(corrplot) library(carData) library(car) library(statmod) library(effects) library(ROCR) library(hnp) library(faraway) library(ggplot2) ´´´

1. Base de Dados

1.1 Descrição dos dados

Os dados foram retirados do pacote "DAAG", sendo dados dos EUA, entre 1997-2002, de acidentes de carro relatados pela polícia nos quais há um evento prejudicial (pessoas ou propriedade) e do qual pelo menos um veículo foi rebocado. Os dados são restritos aos ocupantes do banco da frente, incluem apenas um subconjunto das variáveis registradas e são restritos de outras maneiras também.

A base original possui uma base de dados com 26.217 observações nas 15 variáveis a seguir.

1 - *veloc*: velocidades estimadas do impacto do acidente: 1-9km/h, 10-24, 25-39, 40-54, 55+
2 - *pesos*: Pesos de observação
3 - *sobrev*: Classificação se sobreviveu ao acidente: 1 = sobreviveu ou 0 = morreu
4 - *airbag*: Se o carro possui airbag: com ou sem airbag
5 - *cinto*: uso do cinto de segurança: com ou sem cinto
6 - *frontal*: impacto do acidente: 0 = não frontal, 1 = impacto frontal
7 - *sexo*: Sexo: 0 = Feminino ou 1 = Masculino
8 - *idade*: Idade dos ocupantes do veículo
9 - *anoaci*: Ano do acidente (1997-2002)
10 - *anovei*: Ano do veículo (1953-2003)
11 - *airbagcat*: Se Airbags foram acionados: deploy, nodeploy, unavail
12 - *ocupantes*: Posição do airbag acionado: driver, pass
13 - *abfunc*: Airbag acionados: 0: Se não possuia airbag ou não foi acionado, 1: Um ou mais airbags foram acionados
14 - *grav*: Gravidade do acidente: 0:none, 1 = Possível Lesão, 2:no incapacity, 3:incapacity, 4:killed; 5:unknown, 6:prior death
15 - *numcaso*: Número do caso.

O escopo da análise tem como variável respota a sobrevivencia após o acidente e ás demais variáveis serão as covariáveis explicativas.

Antes da analise começar foi verificado que algumas das variáveis presentes na base são irrelevantes para o modelo.

As variáveis são: anoaci: ano do acidente anovei: ano do veiculo peso das observações: sem descrição airbagcat e abfunc: pois ja temos na base numcaso: id grav

{r, include=FALSE} ## Carregando e ajustando a base de dados #********** dados <- read.csv("C:\Users\Ketlin\Desktop\basevivomorto.csv", header = T, sep = ',') dados<- dados[,c(-1,-10)] dados$dvcat < -ifelse(dados$dvcat == '1-9km/h',1, ifelse(dados$dvcat ==' 10 − 24', 2, ifelse(dados$dvcat == '25-39',3, ifelse(dados$dvcat ==' 40−54', 4, 5))))dados$dvcat <- as.factor(dados$dvcat)

2 Análise Descritiva

2.1 Medidas de Resumo

{r} names(dados)<-c('veloc','sobrev','cinto','frontal','sexo','idade','ocupantes','airbag') summary(dados)

Nota-se na varável velocidade uma frequência maior de acidentes na faixa de 25-39 milhas. A maioria estava com cinto de segurança e os acidentes foram a maioria frontais.

2.3 Histogramas {r} x11() par(mfrow = c(1,3)) barplot(table(dados$sobrev, dados$cinto), beside=T, las = 1, xlab = 'Cinto', ylab = 'Frequência', legend = c('Não','Sim')) barplot(table(dados$sobrev, dados$frontal), beside=T, las = 1, xlab = 'Frontal', ylab = 'Frequência', legend = c('Não','Sim')) barplot(table(dados$sobrev, dados$sexo), beside=T, las = 1, xlab = 'Sexo', ylab = 'Frequência', legend = c('Não','Sim'))

3. Ajuste do Modelo de Regressão

##3.1 Ligação Logito

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Logito. A expressão do modelo é dada por:

$ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 airbag_i$

No R, o modelo é declarado da seguinte forma:

{r} ajuste1 <- glm(sobrev ~ .,family=binomial(link='logit'),data = dados)

##3.2 Ligação Probito

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Probito. A expressão do modelo é dada por:

$\phi^{-1}(\pi_i) = \beta_0 + \beta_1 Veloc_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 airbag_i$

No R, o modelo é declarado da seguinte forma:

{r} ajuste2 <- glm(sobrev ~ .,family=binomial(link = 'probit'),data = dados)

##3.3 Ligação Complemento log-log

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Complemento Log Log. A expressão do modelo é dada por:

$ln[-ln(1-\pi_i)] = \beta_0 + \beta_1 Veloc_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 airbag_i$

No R, o modelo é declarado da seguinte forma:

{r} ajuste3 <- glm(sobrev ~ .,family=binomial(link='cloglog'),data = dados)

##3.4 Ligação Cauchy

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Cauchy. A expressão do modelo é dada por:

$tan[\pi_i(\mu_i - 0,5)] = \beta_0 + \beta_1 Veloc_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 airbag_i$

No R, o modelo é declarado da seguinte forma:

{r} ajuste4 <- glm(sobrev ~ .,family=binomial(link='cauchit'),data = dados)

4. Escolha do Modelo

O critério de informação AIC pode também ser utilizado, porém o AIC penaliza o número de parâmetros do modelo. Como os modelos tem o mesmo número de parâmetros, o critério aponta para a mesma direção da verossimilhança pois todos são penalizados da mesma forma.

{r, echo=FALSE, eval=TRUE, results="hide"} selec <- data.frame(ajuste=c('logito', 'probito', 'cloglog', 'cauchy'), aic=c(AIC(ajuste1), AIC(ajuste2), AIC(ajuste3), AIC(ajuste4)), logLik=c(logLik(ajuste1),logLik(ajuste2),logLik(a selec

O modelo que apresentou menor AIC e maior verossimilhança foi o modelo Binomial com função de ligação Cauchy.

5. Análise do Modelo Ajustado Selecionado

##5.1 Resumo do Modelo

{r, echo=FALSE, eval=TRUE, results="hide"} summary(ajuste4)

##5.2 Reajuste do Modelo

Como próximo passo será usado o algoritmo stepwise para seleção de variáveis do modelo.

O novo modelo fica da seguinte forma:

{r, echo=TRUE, eval=TRUE, results="hide"} ajuste4.1 <- step(ajuste4, direction = "both")

{r, echo=FALSE, eval=TRUE, results="hide"}

summary(ajuste4)

{r, echo=FALSE, eval=TRUE, results="hide"}

summary(ajuste4.1)

Vamos testar a seguinte hipotese: H_0: modelos não diferem H_1: modelos diferem

{r} anova(ajuste4, ajuste4.1, test = 'Chisq') summary(ajuste4.1)

P-valor é de 0.26 os modelos não diferem ou seja airbag pode ser retirado do modelo

O modelo final ficou da seguinte forma :

$tan[\pi\_i(\mu\_i - 0,5)] = \beta\_0 + \beta\_1 Veloc\_i + \beta\_4 Cinto\_i + \beta\_5 Frontal\_i + \beta\_6 Sexo\_i + \beta\_7 Idade\_i + \beta\_8 Ocupantes\_i$*##5.3 Análise de Resíduos*

{r, echo=FALSE, eval=TRUE, results="hide"}

par(mfrow=c(2,2)) plot(ajuste4.1, 1:4) sq par(mfrow=c(2,2)) plot(ajuste4, 1:4)

5.4 Medidas de Influencia

{r} influenceIndexPlot(ajuste4.1, vars=c("Cook"), main="Distância de Cook")

{r} influenceIndexPlot(ajuste4.1, vars=c("Studentized"), main="Resíduos Padronizados")

5.5 Resíduos Quantílicos Aleatoriazados

##5.6 Gráfico Normal de Probabilidades com Envelope Simulado

Lineu O gráfico de resíduos simulados permite verificar a adequação do modelo ajustado mesmo que os resíduos não tenham uma aproximação adequada com a distribuição Normal. Neste tipo de gráfico espera-se, para um modelo bem ajustado, os pontos (resíduos) dispersos aleatoriamente entre os limites do envelope.

Deve-se ficar atento à presença de pontos fora dos limites do envelope ou ainda a pontos dentro dos limites porém apresentando padrões sistemáticos.

Vamos utilizar a função envelope implementada pelo professor Cesar Augusto Taconeli :

{r, echo=FALSE, eval=TRUE, results="hide"} envSim <- function(model, data, nsim = 100){ dados <- na.omit(data) n <- .subset2(model, "df.null") + 1 resM <- matrix(0, nrow = n, ncol = nsim) sim <- simulate(model, nsim) for (i in 1:nsim){ dados$y <- .subset2(sim, i) mSim <- update(model, y ~ ., data = dados) resM[,i] <- sort.default(rstandard(mSim, type = 'deviance'), na.last = TRUE) } qS <- apply(resM, 1 , quantile, c(0.025, 0.5, 0.975), na.rm = TRUE) qN <- qnorm((1:n-0.5)/n) plot(rep(qN, 2), c(qS[1,], qS[3,]), type = 'n', xlab = 'Percentil da N(0,1)', ylab = 'Resíduos Padronizados', main = 'Gráfico Normal de Probabilidades') lines(qN, qS[1,], type = 'l') lines(qN, qS[2,], type = 'l', lty = 2, col = 4) lines(qN, qS[3,], type = 'l') points(qN, sort.default(rstandard(model, type = 'deviance'), na.last = TRUE), pch = 16, cex = 0.75) }

envSim(model = ajuste4.1, data = ajuste4.1$data)

5.7 Gráficos de Efeitos

{r, echo=FALSE, eval=TRUE, results="hide"} plot(allEffects(ajuste4.1), type = 'response', main = '')

6. PREDIÇÃO

7. AVALIAÇÃO DO PODER PREDITIVO DO MODELO

Como temos uma base de tamanho razoável para fins preditivos, uma alternativa é separar a base em duas: uma para o ajuste do modelo, com 70% dos dados (com 477 observações) e outra para validação, com 30% (com 203 observações).

7.1 Divisão da Base de dados

{r} set.seed(1909) indices <- sample(1:680, size = 477) dadosajuste <- dados[indices,] dadosvalid <- dados[-indices,]

7.2 Ponto de Corte

Como estamos modelando a probabilidade de tumor maligno, vamos estabelecer o ponto de corte 0.5, isso é, se a probabilidade estimada for maior que este valor o tumor será classificado como maligno. Vamos armazenar os valores preditos do modelo para os dados de validação:

{r} pred <- predict(ajuste4.1, newdata = dadosvalid, type = 'response') corte <- ifelse(pred > 0.5, 'maligno', 'benigno')

7.3 Sensibilidade e Especificidade

Para fazer uso dos dados de validação, dois conceitos são necessários: sensibilidade e especificidade.

Define-se por sensibilidade a capacidade do modelo de detectar tumores malignos, ou seja, de classificar como malignos os tumores que de fato o são .

Já a especificidade é a capacidade do modelo de detectar classificar como benignos tumores verdadeiramente benignos.

A sensibilidade é dada por

{r} sens <- dados[2,2]/sum(dados[,2]) sens

7.4 Curva ROC

7.5 Outra Alternativa de validação

8. REFERÊNCIAS