Lista 5 - MAE0217

Exercício 18 - Capítulo 6

```
par <- c(0.69,0.33,-0.03)
od <-exp(par)
se <- c(0.12,0.10,0.005)
```

Considerando o seguinte modelo:

$$log\left\{\frac{\pi_i(x_i,w_i)}{1-\pi_i(x_i,w_i)}\right\} = \alpha + \beta x_i + \gamma(w_i - 5)$$

Onde sabemos que x_i e w_i representam, respectivamente, o gênero e idade da i-éisma criança.

a)

Podemos interpretar os parâmetros α, β e γ da seguinte forma:

1. α corresponde ao logaritmo da chance de preferência das crianças do gênero feminino com 5 anos de idade.
2. β corresponde ao lagaritmo da razão entre a chance de preferência das crianças do gênero masculino com e a chance corresponde para crianças do gênero feminino com a mesma idade.
3. γ corresponde ao logaritmo da razão entre a chance de preferência para crianças com diferença de 1 ano e do mesmo gênero.

b)

Denominaremos a razão de chances por OD por convenção (Odds Ratio).

Sabemos que a a variação em uma unidade em x_i atera γ no log da chance de preferência, logo se queremos saber a razão de chances com a variação de 5 unidades em x_i basta fazer o seguinte cálculo:

$$OD = exp\{5\hat{\gamma}\} = 0.861$$

c)

Para a construção de intervalos de confiança usaremos os erros padrões dados no enunciado de exercício. Inicialemnte encontraremos um intervalo de confiança para

$$IC(\beta, 95\%) = [0.134, 0.526]$$

$$IC(\gamma, 95\%) = [-0.0398, -0.0202]$$

Agora para a construção dos intervalos de confiança pedidos iremos exponenciar os limites dos respectivos intervalos de confiança(com 3 casas decimais),

$$IC(\exp\{\beta\}, 95\%) = [1.143, 1.692]$$

Esse intervalo pode ser interpretado como um intervalo de confiança para a razão entre chance da preferência por Kcola por crianças do gênero masculino e a chance corresponde para crianças do gênero feminino, ou seja, espera-se que em 95 amostras de 100 de mesmo tamanho, um intervalo construído da mesma maneira, contenha a razão de chances descrita.

Tomando este intervalo como base, podemos dizer que espera-se a preferência por Kcola por crianças do gênero masculino é maior do que para crianças do gênero feminino, para crianças de mesmas idades.

$$IC(\exp\{\gamma\}, 95\%) = [0.961, 0.98]$$

Esse intervalo pode ser interpretado como um intervalo de confiança para a razão entre chance da preferência por Kcola por crianças que diferem em um ano, ou seja, espera-se que em 95 amostras de 100 de mesmo tamanho, um intervalo construído da mesma maneira, contenha a razão de chances descrita.

Tomando este intervalo como base, podemos dizer que espera-se a preferência por Kcola por crianças mais velhas, seja menor do que para crianças mais novas, para crianças do mesmo sexo.

d)

$$\hat{\pi}(x = 1, w = 15) = \frac{\exp\{0.72\}}{1 + \exp\{0.72\}} = 0.673$$

Exercício 19 - Capítulo 6

Considere as seguintes expressões:

$$\log\left\{\frac{P(Y_i = 1|X = x_i)}{P(Y_i = 0|X = x_i)}\right\} = \alpha + \beta x_i \qquad e \qquad P(Y_i = 1|X = x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

Considere a equação do lado esquerdo, então se as duas expressões são equivalentes eu devo ser capaz de chegar na expressão da equação do lado direito,

$$\log\left\{\frac{P(Y_i = 1|X = x_i)}{P(Y_i = 0|X = x_i)}\right\} = \alpha + \beta x_i$$

Exponenciando ambos os membros obtemos,

$$\frac{P(Y_i = 1|X = x)}{P(Y_i = 0|X = x)} = \exp\{\alpha + \beta x_i\}$$

Como sabemos que a variavel Y_i só assume os valores 0 e 1, pode-se afirmar que:

$$P(Y_i = 0|X = x_i) = P(Y_i = 1|X = x_i)^C = 1 - P(Y_i = 1|X = x_i)$$

Então,

$$\frac{P(Y_i = 1|X = x_i)}{1 - P(Y_i = 1|X = x_i)} = \exp\{\alpha + \beta x_i\}$$

$$P(Y_i = 1|X = x_i) = \exp\{\alpha + \beta x_i\} - P(Y_i = 1|X = x_i)\exp\{\alpha + \beta x_i\}$$

$$P(Y_i = 1|X = x) + P(Y_i = 1|X = x_i)\exp\{\alpha + \beta x_i\} = \exp\{\alpha + \beta x_i\}$$

$$P(Y_i = 1|X = x_i)(1 + \exp\{\alpha + \beta x_i\}) = \exp\{\alpha + \beta x_i\}$$

$$P(Y_i = 1|X = x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

Sabemos ainda que

$$0 < \exp(a) < +\infty \qquad \forall a \in \mathbb{R}$$

logo,

$$\lim_{exp(\alpha+\beta x_i)\to+\infty} \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} = 1 \qquad e \qquad \lim_{exp(\alpha+\beta x_i)\to 0} \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} = 0$$

Então, independente dos valoresα,β e x_i,

$$0 < P(Y_i = 1|X = x_i) < 1$$

Exercício 20 - Capítulo 6

Considere novamente o modelo

$$\log\left\{\frac{P(Y_i = 1|X = x_i)}{P(Y_i = 0|X = x_i)}\right\} = \alpha + \beta x_i$$

É fácil ver que

$$\beta = \alpha - \alpha + \beta x_i - \beta x_i + \beta = [\alpha + \beta(x_i + 1)] - (\alpha + \beta x_i)$$

$$\beta = \log\left\{\frac{P(Y_i = 1|X = x_i + 1)}{P(Y_i = 0|X = x_i + 1)}\right\} - \log\left\{\frac{P(Y_i = 1|X = x_i)}{P(Y_i = 0|X = x_i)}\right\}$$

$$\beta = \log\left\{\frac{\frac{P(Y_i=1|X=x_i+1)}{P(Y_i=0|X=x_i+1)}}{\frac{P(Y_i=1|X=x_i)}{P(Y_i=0|X=x_i)}}\right\}$$

Seja C_{x_i} a chance de resposta positiva para um paciente com x_i unidades na variável explicativa e definida por

$$C_{x_i} = \frac{P(Y_i = 1|X = x_i)}{P(Y_i = 0|X = x_i)}$$

Então é fácil ver que,

$$\beta = \log\left\{\frac{C_{x_i+1}}{C_{x_i}}\right\}$$

Dessa maneira fica evidente que β representa o logaritmo da razão entre as chances de resposta positiva para pacientes com diferença de uma unidade na variável explicativa.

Exercício 21 - Capítulo 6

```
diab <- c(1,4,7,10,13,16,19,22)
sim <- c(17,26,39,27,35,37,26,23)
nao <- c(215,218,137,62,36,16,13,15)
df <- data.frame(diab,sim,nao)
m <- matrix(data=c(sim,nao),ncol=2)
mod <- glm(formula = m~diab,family = binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = m ~ diab, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6128  -0.5363  -0.1574   0.7274   2.0178
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.62936    0.15780  -16.66   <2e-16 ***
## diab         0.18021    0.01456   12.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.66  on 7  degrees of freedom
## Residual deviance:  13.83  on 6  degrees of freedom
## AIC: 54.514
##
## Number of Fisher Scoring iterations: 4
```

Desta forma podemos obter

$$\exp(\hat{\alpha}) = 0.072$$

$$\exp(\hat{\beta}) = 1.197$$

Onde

1. $\exp(\hat{\alpha})$ representa a chance de ocorrência de retinopatia para uma duração de 0 anos de diabetes.
2. $\exp(\hat{\beta})$ represeta a razão entre a chance de ocorrência de retinopatia para pacientes com um ano de diferença na duração do diabetes. Podemos concluir que para a variação de um ano na duração da diabete a chance de ocorrência de retinopatia aumenta em 19,7%.

Exercício 22 - Capítulo 6

Considere o seguinte modelo

$$log\left\{\frac{\pi_i(\mathbf{x})}{1-\pi_i(\mathbf{x})}\right\} = \beta_0 + < \mathbf{x}, \beta' >$$

$$\mathbf{x} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8]$$

$$\beta' = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8]$$

Onde x_1 representa a idade (anos),x_2 representa a presença ou não de dormenstrual, valendo 1 em caso positivo e 0 caso contrário, as variáveis x_3, x_4, x_5, x_6 idicam na ordem, a intensidade da dismenorreia, leve, moderada, intensa e incapacitante, valendo 1 caso a intensidade seja a correspondente e 0 caso contrário e por último as variáveis x_7, x_8 indicam o tipo de esterilidade, primária ou secundária, valendo 1 caso a esterilidade seja a correspondente e 0 caso contrário.

A tabela abaixo expressa um resumo dos dados considerados para o ajuste do modelo,

```r
endo <- na.omit(readxl::read_excel(path="/Users/kevin/Downloads/endometriose2.xls"))
endo2 <- endo %>%
  filter(tipoesteril!="sim") %>%
  select(endometriose,idade,dormenstrual,dismenorreia,tipoesteril) %>%
  mutate(endometriose=as.factor(as.numeric(endometriose=="sim"))) %>%
  mutate(dormenstrual=as.factor(as.numeric(dormenstrual=="sim")))
endo2$dismenorreia<- as.factor(endo2$dismenorreia)
endo2$tipoesteril<- as.factor(endo2$tipoesteril)
endo2$endometriose<- as.factor(endo2$endometriose)
endo2$dismenorreia <- factor(as.character(endo2$dismenorreia),levels=c("nao","leve","moderada","intensa
endo2$tipoesteril <- factor(endo2$tipoesteril,levels=c("nao","primaria","secundaria"))
summary(endo2)
```

```
##  endometriose     idade       dormenstrual        dismenorreia
##  0:1374       Min.   :14.00   0: 645       nao          :644
##  1: 489       1st Qu.:29.00   1:1218       leve         :379
##               Median :35.00                moderada     :228
##               Mean   :34.98                intensa      :413
##               3rd Qu.:42.00                incapacitante:199
##               Max.   :50.00
##      tipoesteril
##  nao       :1401
##  primaria  : 221
##  secundaria: 241
##
##
##
```

```r
endo2$idade <- endo2$idade-14
```

Observe que a idade minima observada é de 14 anos, então para maior compatibilidade com o modelo, o intercepto do modelo β_0 será interpretado como o logaritmo da chance de ocorrência de endometriose para indivíduos de 14 anos sem dormenstrual, que não apresentam dismenorreia e não estéreis. Os outros coeficiente serão explicados com o auxílio do seguinte código **R**:

```r
logistic_model <- glm(formula = endometriose~.,data=endo2,family = binomial(link = "logit"))
summary(logistic_model)
```

```
##
## Call:
## glm(formula = endometriose ~ ., family = binomial(link = "logit"),
##     data = endo2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9072  -0.7906  -0.4206   0.7201   2.3318
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.860297   0.218733  -8.505  < 2e-16 ***
## idade                     -0.024688   0.007607  -3.246  0.00117 **
## dormenstrual1              1.899458   1.432983   1.326  0.18500
```

```
## dismenorreialeve          -1.773009   1.442755  -1.229  0.21911
## dismenorreiamoderada       0.254864   1.434259   0.178  0.85896
## dismenorreiaintensa        0.200996   1.436595   0.140  0.88873
## dismenorreiaincapacitante -0.424796   1.439380  -0.295  0.76790
## tipoesterilprimaria        1.347624   0.168948   7.977  1.5e-15 ***
## tipoesterilsecundaria     -0.051478   0.180756  -0.285  0.77580
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2144.8  on 1862  degrees of freedom
## Residual deviance: 1742.8  on 1854  degrees of freedom
## AIC: 1760.8
##
## Number of Fisher Scoring iterations: 5
```

```
coef <- exp(logistic_model$coefficients)
ep <- summary(logistic_model)$coefficients[,2]
```

1. β_1 representa o logaritmo da razão entre a chance de ocorrência de endometriose em indivíduos com diferença de 1 ano de idade com as mesmas características(dor menstrual,tipo de esterilidade e dismenorréia).
2. β_2 representa o logaritmo da razão entre a chance de ocorrência de endometriose em indivíduos que apresentam dor menstrual e indivíduos que não e com as mesmas características(idade,tipo de esterilidade e dismenorréia).
3. β_3 representa o logaritmo da razão entre a chance de ocorrência de endometriose em indivíduos com dismenorréia leve e indivíduos sem dismenorréia com as mesmas características(idade, dor menstrual,tipo de esterilidade).
4. β_4 representa o logaritmo da razão entre a chance de ocorrência de endometriose em indivíduos com dismenorréia moderada e indivíduos sem dismenorréia com as mesmas características(idade, dor menstrual,tipo de esterilidade).
5. β_5 representa o logaritmo da razão entre a chance de ocorrência de endometriose em indivíduos com dismenorréia intensa e indivíduos sem dismenorréia com as mesmas características(idade, dor menstrual,tipo de esterilidade).
6. β_6 representa o logaritmo da razão entre a chance de ocorrência de endometriose em indivíduos com dismenorréia incapacitante e indivíduos sem dismenorréia com as mesmas características(idade, dor menstrual,tipo de esterilidade).
7. β_7 representa o logaritmo da razão entre a chance de ocorrência de endometriose em indivíduos com esterilidade primária e indivíduos não estéreis com as mesmas características(idade, dor menstrual,dismenorréia).
8. β_8 representa o logaritmo da razão entre a chance de ocorrência de endometriose em indivíduos com esterilidade secundária e indivíduos não estéreis com as mesmas características(idade, dor menstrual,dismenorréia).

Exercício 5 - Capítulo 8

O objetivo é escolher o melhor modelo com base na acurácia, desta forma, iremos ajustar os modelos pedidos em cada item, e criar uma tabela onde cada linha representa um modelo, além disso essa tabela terá três colunas, a coluna da esquerda especifica as variáveis explicativas utilizadas no modelo, a coluna do meio e a coluna da direita indicam respectivamente, a acurácia e o coeficiente Kappa de Cohen(κ) obtidos com uma validação cruzada de ordem 5.

```r
# Já foi criado um data frame com as variaveis que serão utilizadas. No exercício anterior.
#Vamos usar os pacotes caret e pROC
set.seed(1234)
var <- c("idade","dismenorreia","dormenstrual","tipoesteril")
kappa <- vector(mode = "numeric",4)
```

i)

Para modelos com apenas uma variável explicativa:

```r
acurr <- vector(mode = "numeric",4)
kappa <- vector(mode = "numeric",4)
cont <- 1
for(v in var){
  train_control = trainControl(method="repeatedcv", number=5,repeats=5)
  frmla <- as.formula(str_c("endometriose~",v))
  model1 = train(frmla, data=endo2,method="glm", family=binomial, trControl=train_control)
  acurr[cont] <- round(model1$results$Accuracy,3)
  kappa[cont] <- round(model1$results$Kappa,3)
  prev <- predict(model1,newdata = endo2,type = "raw")
  cont <- cont+1
}

knitr::kable(data.frame("Variavel explicativa do modelo"=str_to_title(var),"Acurácia"=acurr,"Kappa"=kap
```

Variavel.explicativa.do.modelo	Acurácia	Kappa
Idade	0.738	0.000
Dismenorreia	0.729	0.084
Dormenstrual	0.738	0.000
Tipoesteril	0.751	0.219

ii)

Para modelos com duas variáveis explicativas:

```r
control <- vector("character",4)
acurr <- vector(mode = "numeric",6)
kappa <- vector(mode = "numeric",6)
cont <- 1
cont2 <- 1
variaveis <- vector("character",6)
for(v in var){
  control[cont2] <- v
  for(v2 in var){
    if(!(v2 %in% control)){
      train_control = trainControl(method="repeatedcv", number=5,repeats=5)
      frmla <- as.formula(str_c("endometriose~",v,"+",v2))
      model1 = train(frmla, data=endo2,method="glm", family=binomial,trControl=train_control)
      acurr[cont] <- round(model1$results$Accuracy,3)
      kappa[cont] <- round(model1$results$Kappa,3)
```

```
      prev <- predict(model1,newdata = endo2,type = "raw")
      variaveis[cont] <- str_c(v," + ",v2)
      cont <- cont+1
    }
  }
  cont2 <- cont2+1
}

knitr::kable(data.frame("Variáveis explicativas do modelo"=str_to_title(variaveis),"Acurácia"=acurr,"Ka
```

Variáveis.explicativas.do.modelo	Acurácia	Kappa
Idade + Dismenorreia	0.743	0.234
Idade + Dormenstrual	0.738	0.000
Idade + Tipoesteril	0.749	0.194
Dismenorreia + Dormenstrual	0.727	0.073
Dismenorreia + Tipoesteril	0.770	0.222
Dormenstrual + Tipoesteril	0.760	0.211

iii)

Para modelos com três variáveis explicativas:

```
control1 <- vector("character",4)
control2 <- vector("character",4)
acurr <- vector(mode = "numeric",4)
kappa <- vector(mode = "numeric",4)
cont <- 1
cont2 <- 1
cont3 <- 1
variaveis <- vector("character",4)
for(v in var){
  control1[cont3] <- v
  control2 <- vector("character",4)
  for(v2 in var){
    control2[cont2] <- v2
    if(!(v2 %in% control1)){

      for(v3 in var){
        if(!(v3 %in% control2)&!(v3 %in% control1)){
          train_control = trainControl(method="repeatedcv", number=5,repeats=5)
          frmla <- as.formula(str_c("endometriose~",v,"+",v2,"+",v3))
          model1 = train(frmla,data=endo2,method="glm",family=binomial,trControl=train_control)
          acurr[cont] <- round(model1$results$Accuracy,3)
          kappa[cont] <- round(model1$results$Kappa,3)
          prev <- predict(model1,newdata = endo2,type = "raw")
          variaveis[cont] <- str_c(v," + ",v2," + ",v3)
          cont <- cont+1
        }
      }
    }
    cont2 <- cont2+1
```

```
  }
  cont3 <- cont3+1
}

knitr::kable(data.frame("Variáveis explicativas do modelo"=str_to_title(variaveis),"Acurácia"=acurr,"Ka
```

Variáveis.explicativas.do.modelo	Acurácia	Kappa
Idade + Dismenorreia + Dormenstrual	0.744	0.238
Idade + Dismenorreia + Tipoesteril	0.757	0.220
Idade + Dormenstrual + Tipoesteril	0.760	0.211
Dismenorreia + Dormenstrual + Tipoesteril	0.769	0.220

iv)

Para o modelo com todas as quatro variáveis explicativas:

```
train_control = trainControl(method="repeatedcv", number=5,repeats=5)
model1 = train(endometriose~., data=endo2,method="glm", family=binomial, trControl=train_control)
acurr <- round(model1$results$Accuracy,3)
kappa <- round(model1$results$Kappa,3)
prev <- predict(model1,newdata = endo2,type = "raw")
variaveis <- str_c(var,collapse = " + ")

knitr::kable(data.frame("Variáveis explicativas do modelo"=str_to_title(variaveis),"Acurácia"=acurr,"Ka
```

Variáveis.explicativas.do.modelo	Acurácia	Kappa
Idade + Dismenorreia + Dormenstrual + Tipoesteril	0.757	0.215

Logo podemos concluir que o modelo com maior acurácia é aquele que possui a Dismenorréia e o tipo de esterilidade como variáveis explicativas. $\mathbb{R}^8$