Técnicas Estatísticas de Predição

Exercício 4

João Pedro Gentil da Silveira

Questões

A) Retire uma amostra de 300 observações utilizando a sintaxe com seed.

R:

```
# Leitura da base de dados
base = read.csv2("titanic_data.csv", dec=".")

# Tira a amostra de tamanho 300 da base de dados
set.seed(29072003)
amostra = base[sample(nrow(base), 300),]
```

B) Classifique as variáveis independentes que forem qualitativas como fatores.

R:

```
# Classifica as variáveis qualitativas independentes como fatores
amostra$Pclass = as.factor(amostra$Pclass)
amostra$Sex = as.factor(amostra$Sex)
amostra$Embarked = as.factor(amostra$Embarked)
```

C) Utilize o método de seleção de variáveis Forward para determinar o modelo final.

R:

```
# Cria um modelo zero e constrói um modelo adequado
modelo_zero = glm(Survived ~ 1, family = binomial(), data=amostra)
modelo_adequado = step(modelo_zero,list(lower = ~ 1,
                                        upper = ~
Pclass+Sex+Age+SibSp+Parch+Fare+Embarked),
                       direction="forward")

## Start:  AIC=390.47
## Survived ~ 1
##
##             Df Deviance    AIC
## + Parch      1   374.28 378.28
## + Sex        1   378.40 382.40
## + SibSp      1   379.12 383.12
## + Pclass     2   383.08 389.08
## <none>           388.47 390.47
## + Embarked   2   385.95 391.95
## + Age        1   388.34 392.34
## + Fare     135   189.18 461.18
```

```
##
## Step:  AIC=378.28
## Survived ~ Parch
##
##             Df Deviance    AIC
## + Sex        1   367.03 373.03
## + SibSp      1   370.72 376.72
## + Pclass     2   368.85 376.85
## + Embarked   2   369.68 377.68
## <none>           374.28 378.28
## + Age        1   374.28 380.28
## + Fare     135   188.85 462.85
##
## Step:  AIC=373.03
## Survived ~ Parch + Sex
##
##             Df Deviance    AIC
## + SibSp      1   363.90 371.90
## + Embarked   2   362.05 372.05
## + Pclass     2   362.34 372.34
## <none>           367.03 373.03
## + Age        1   367.02 375.02
## + Fare     135   185.27 461.27
##
## Step:  AIC=371.9
## Survived ~ Parch + Sex + SibSp
##
##             Df Deviance    AIC
## + Embarked   2   358.70 370.70
## + Pclass     2   359.47 371.47
## <none>           363.90 371.90
## + Age        1   363.90 373.90
## + Fare     135   184.99 462.99
##
## Step:  AIC=370.7
## Survived ~ Parch + Sex + SibSp + Embarked

## Warning: glm.fit: probabilidades ajustadas numericamente 0 ou 1
ocorreu

##           Df Deviance    AIC
## + Pclass   2   354.07 370.07
## <none>         358.70 370.70
## + Age      1   358.40 372.40
## + Fare   135   183.64 465.64
##
## Step:  AIC=370.07
## Survived ~ Parch + Sex + SibSp + Embarked + Pclass

## Warning: glm.fit: algoritmo não convergiu
```

```
## Warning: glm.fit: probabilidades ajustadas numericamente 0 ou 1
ocorreu

##         Df Deviance    AIC
## + Age    1    351.9  369.9
## <none>        354.1  370.1
## + Fare 134   5478.6 5762.6
##
## Step:  AIC=369.92
## Survived ~ Parch + Sex + SibSp + Embarked + Pclass + Age

## Warning: glm.fit: algoritmo não convergiu

## Warning: glm.fit: probabilidades ajustadas numericamente 0 ou 1
ocorreu

##         Df Deviance    AIC
## <none>        351.9  369.9
## + Fare 134   4613.6 4899.6

summary(modelo_adequado)

##
## Call:
## glm(formula = Survived ~ Parch + Sex + SibSp + Embarked + Pclass +
##     Age, family = binomial(), data = amostra)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.472768   0.523784   0.903   0.3667
## Parch        0.486972   0.193592   2.515   0.0119 *
## Sexmale     -0.659797   0.260600  -2.532   0.0113 *
## SibSp        0.302150   0.173844   1.738   0.0822 .
## EmbarkedQ    0.940960   0.480427   1.959   0.0502 .
## EmbarkedS   -0.146084   0.328502  -0.445   0.6565
## Pclass2     -0.894933   0.421853  -2.121   0.0339 *
## Pclass3     -0.856243   0.357492  -2.395   0.0166 *
## Age         -0.013292   0.009156  -1.452   0.1466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 388.47  on 299  degrees of freedom
## Residual deviance: 351.92  on 291  degrees of freedom
## AIC: 369.92
##
## Number of Fisher Scoring iterations: 4
```

D) Determine as razões de chances com os intervalos de confiança para as variáveis selecionadas no modelo e faça a interpretação das razões de chances.

R: Obtidas as razões de chances das variáveis selecionadas no modelo, é possível depreender o seguinte:

Parch: Uma vez que seu intervalo de confiança vai de 1.14 até 2.44, a variável é dita estatisticamente significativa - pois o seu IC não inclui 1. Ademais, com OR de 1.62, é viável dizer que, para cada unidade adicional de Parch - isto é, para cada filho a bordo do Titanic -, as chances de sobrevivência aumentam em cerca de 62%.

Sex: Tomando o sexo feminino como referência, uma vez que o IC de “Sexmale” vai de 0.30 até 0.85, a variável é considerada estatisticamente significativa - pois o IC de “Sexmale” não inclui 1. Além disso, com “Sexmale” possuindo um OR de 0.51, é possível dizer que ser do sexo masculino diminui as chances de sobrevivência em cerca de 49% (1 - 0.51).

SibSp: Uma vez que seu intervalo de confiança vai de 0.98 até 1.94, a variável não é dita estatisticamente significativa - pois o seu IC inclui 1. Dessa forma, é possível afirmar que o número de irmãos ou cônjuges a bordo provavelmente não influencia no aumento ou na diminuição das chances de sobrevivência.

Embarked: Tomando o porto de embarque Cherbourg (C) como referência, uma vez que o IC de “EmbarkedS” - porto de embarque Southampton (S) - vai de 0.45 até 1.65, a variável não é dita estatisticamente significativa - pois o IC de “EmbarkedS” inclui 1. Dessa forma, é possível dizer que o porto de embarque provavelmente não influencia no aumento ou na diminuição das chances de sobrevivência.

Pclass: Tomando a primeira classe como referência, uma vez que o IC de “Pclass2” vai de 0.17 até 0.92 e o IC de “Pclass3” vai de 0.20 até 0.85, a variável é rotulada estatisticamente significativa - pois os ICs avaliados não incluem 1. Ademais, com “Pclass2” possuindo um OR de 0.40, é possível dizer que ser da segunda classe diminui as chances de sobrevivência em cerca de 60% (1 - 0.40), e, com “Pclass3” possuindo um OR de 0.42, é possível dizer que ser da terceira classe minimiza as chances de sobrevivência em cerca de 58% (1 - 0.42).

Age: Uma vez que seu intervalo de confiança vai de 0.96 até 1.00, a variável não é dita estatisticamente significativa - pois seu IC inclui 1. Dessa forma, é possível dizer que a idade provavelmente não impacta no aumento ou na diminuição das chances de sobrevivência.

```
# Razão de Chances
OR = data.frame(exp(modelo_adequado$coefficients))
IC = data.frame(exp(confint(modelo_adequado)))
IC_OR = cbind(OR[-1,],IC[-1,])
colnames(IC_OR) = c("OR","2.5%","97.5%")
IC_OR
```

```
##                   OR      2.5%     97.5%
## Parch     1.6273814 1.1466327 2.4446142
```

```
## Sexmale   0.5169564 0.3089424 0.8597841
## SibSp     1.3527647 0.9802784 1.9458240
## EmbarkedQ 2.5624403 1.0031377 6.6462443
## EmbarkedS 0.8640854 0.4559149 1.6596022
## Pclass2   0.4086349 0.1762110 0.9261363
## Pclass3   0.4247547 0.2088928 0.8521268
## Age       0.9867958 0.9689210 1.0044340
```

Sintaxe Completa

```r
# Leitura da base de dados
base = read.csv2("titanic_data.csv", dec=".")

# Tira a amostra de tamanho 300 da base de dados
set.seed(29072003)
amostra = base[sample(nrow(base), 300),]

# Classifica as variáveis qualitativas independentes como fatores
amostra$Pclass = as.factor(amostra$Pclass)
amostra$Sex = as.factor(amostra$Sex)
amostra$Embarked = as.factor(amostra$Embarked)

# Cria um modelo zero e constrói um modelo adequado
modelo_zero = glm(Survived ~ 1, family = binomial(), data=amostra)
modelo_adequado = step(modelo_zero,list(lower = ~ 1,
                                        upper = ~
Pclass+Sex+Age+SibSp+Parch+Fare+Embarked),
                       direction="forward")
```

```
## Start:  AIC=390.47
## Survived ~ 1
##
##             Df Deviance    AIC
## + Parch      1   374.28 378.28
## + Sex        1   378.40 382.40
## + SibSp      1   379.12 383.12
## + Pclass     2   383.08 389.08
## <none>           388.47 390.47
## + Embarked   2   385.95 391.95
## + Age        1   388.34 392.34
## + Fare     135   189.18 461.18
##
## Step:  AIC=378.28
## Survived ~ Parch
##
##             Df Deviance    AIC
## + Sex        1   367.03 373.03
## + SibSp      1   370.72 376.72
## + Pclass     2   368.85 376.85
## + Embarked   2   369.68 377.68
## <none>           374.28 378.28
## + Age        1   374.28 380.28
## + Fare     135   188.85 462.85
##
## Step:  AIC=373.03
## Survived ~ Parch + Sex
##
##             Df Deviance    AIC
## + SibSp      1   363.90 371.90
## + Embarked   2   362.05 372.05
## + Pclass     2   362.34 372.34
```

```
## <none>           367.03 373.03
## + Age        1   367.02 375.02
## + Fare     135   185.27 461.27
##
## Step:  AIC=371.9
## Survived ~ Parch + Sex + SibSp
##
##             Df Deviance    AIC
## + Embarked   2   358.70 370.70
## + Pclass     2   359.47 371.47
## <none>           363.90 371.90
## + Age        1   363.90 373.90
## + Fare     135   184.99 462.99
##
## Step:  AIC=370.7
## Survived ~ Parch + Sex + SibSp + Embarked

## Warning: glm.fit: probabilidades ajustadas numericamente 0 ou 1
ocorreu

##           Df Deviance    AIC
## + Pclass   2   354.07 370.07
## <none>         358.70 370.70
## + Age      1   358.40 372.40
## + Fare   135   183.64 465.64
##
## Step:  AIC=370.07
## Survived ~ Parch + Sex + SibSp + Embarked + Pclass

## Warning: glm.fit: algoritmo não convergiu

## Warning: glm.fit: probabilidades ajustadas numericamente 0 ou 1
ocorreu

##         Df Deviance    AIC
## + Age    1    351.9  369.9
## <none>        354.1  370.1
## + Fare 134   5478.6 5762.6
##
## Step:  AIC=369.92
## Survived ~ Parch + Sex + SibSp + Embarked + Pclass + Age

## Warning: glm.fit: algoritmo não convergiu

## Warning: glm.fit: probabilidades ajustadas numericamente 0 ou 1
ocorreu

##         Df Deviance    AIC
## <none>        351.9  369.9
## + Fare 134   4613.6 4899.6

summary(modelo_adequado)
```

```
##
## Call:
## glm(formula = Survived ~ Parch + Sex + SibSp + Embarked + Pclass +
##     Age, family = binomial(), data = amostra)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.472768   0.523784   0.903   0.3667
## Parch        0.486972   0.193592   2.515   0.0119 *
## Sexmale     -0.659797   0.260600  -2.532   0.0113 *
## SibSp        0.302150   0.173844   1.738   0.0822 .
## EmbarkedQ    0.940960   0.480427   1.959   0.0502 .
## EmbarkedS   -0.146084   0.328502  -0.445   0.6565
## Pclass2     -0.894933   0.421853  -2.121   0.0339 *
## Pclass3     -0.856243   0.357492  -2.395   0.0166 *
## Age         -0.013292   0.009156  -1.452   0.1466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 388.47  on 299  degrees of freedom
## Residual deviance: 351.92  on 291  degrees of freedom
## AIC: 369.92
##
## Number of Fisher Scoring iterations: 4
```

```
# Razão de Chances
OR = data.frame(exp(modelo_adequado$coefficients))
IC = data.frame(exp(confint(modelo_adequado)))
IC_OR = cbind(OR[-1,],IC[-1,])
colnames(IC_OR) = c("OR","2.5%","97.5%")
IC_OR
```

```
##                  OR      2.5%     97.5%
## Parch     1.6273814 1.1466327 2.4446142
## Sexmale   0.5169564 0.3089424 0.8597841
## SibSp     1.3527647 0.9802784 1.9458240
## EmbarkedQ 2.5624403 1.0031377 6.6462443
## EmbarkedS 0.8640854 0.4559149 1.6596022
## Pclass2   0.4086349 0.1762110 0.9261363
## Pclass3   0.4247547 0.2088928 0.8521268
## Age       0.9867958 0.9689210 1.0044340
```