Universidade Cruzeiro do Sul
Pós Graduação Estatística Aplicada
Leandro Sampaio Silva
RGM: 19225818

Atividade Final – Multivariada II

install.packages('xlsx') library(xlsx) dados <- read.xlsx('Dados_ativ_final201802.xlsx', 1) credito <- read.xlsx('credito_ativ_final201802.xlsx', 1)

1) Com base na tabela enviada (Dados_ativ_final201802.xlsx), elabore um modelo de regressão logística para avaliar se as variáveis estado civil, idade e sexo podem determinar a probabilidade de uma pessoa pagar ou não um empréstimo, sendo:

- Estado Civil = 0 (solteiro)
- Estado Civil = 1 (casado)
- Sexo = 0 (masculino)
- Sexo = 1 (feminino)

Resultados a serem avaliados:

- código usado para a construção do modelo de regressão logística.
- Quais variáveis são significativas no modelo?
- interpretação dos resultados (não esquecer de calcular o ODDS RATIO).

In [105]:

```
str(dados)
summary(dados[,2:5])
```

```
'data.frame':   180 obs. of  5 variables:
 $ id         : num  85 86 87 88 89 91 92 93 94 95 ...
 $ pagamento  : num  1 1 1 1 1 1 1 1 1 1 ...
 $ estadocivil: num  0 0 0 0 0 0 0 0 0 0 ...
 $ idade      : num  20 34 21 22 22 22 23 30 30 27 ...
 $ sexo       : num  0 1 1 1 1 1 0 1 1 1 ...
```

```
   pagamento       estadocivil         idade            sexo       
 Min.   :0.0000   Min.   :0.0000   Min.   :16.00   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:22.00   1st Qu.:0.0000  
 Median :1.0000   Median :0.0000   Median :25.00   Median :1.0000  
 Mean   :0.7222   Mean   :0.1611   Mean   :26.21   Mean   :0.5222  
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:30.00   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :55.00   Max.   :1.0000  
```

In [45]:

```
mod_pagamento <- glm(pagamento ~ estadocivil
                                + idade
                                + sexo, data = dados,
                                 family =binomial(link='logit'))
```

In [46]:

```
summary(mod_pagamento)
```

```
Call:
glm(formula = pagamento ~ estadocivil + idade + sexo, family = binom
ial(link = "logit"),
    data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4892  -0.4015   0.4166   0.5905   2.1662

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.96591    1.12267  -1.751  0.07993 .
estadocivil -2.95095    0.58293  -5.062 4.14e-07 ***
idade        0.11614    0.04432   2.621  0.00877 **
sexo         1.30123    0.43861   2.967  0.00301 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212.70  on 179  degrees of freedom
Residual deviance: 146.65  on 176  degrees of freedom
AIC: 154.65

Number of Fisher Scoring iterations: 5
```

In [39]:

```
##Odds Ratio
OR1 <- exp(mod_pagamento$coefficients);

#Intervalo de confiança para odds ratio
IC1 <- exp(confint(mod_pagamento))
#IC1;

round(cbind(OR1, IC1),3);
```

```
Waiting for profiling to be done...
```

	OR1	2.5 %	97.5 %
(Intercept)	0.140	0.014	1.154
estadocivil	0.052	0.015	0.152
idade	1.123	1.036	1.233
sexo	3.674	1.598	9.072

R: As três variáveis (idade, sexo e estado civil) são significativas para o modelo.

Contudo, o estado civil, apresenta maior significancia, onde as pessoas casadas tem uma chance de ~95% menor de fazer o pagamento, considerando constantes as variáveis idade e sexo.
A chance de realizar o pagamento aumenta em 12% a cada ano que o pagador fica mais velho, mantendo constante as outras variáveis preditoras. A chance da pessoa ser boa pagadora sendo mulher é 267% maior que sendo homem.

2) A planilha “credito_ativ_final21802” apresenta os dados de default (falta de pagamento) de crédito a partir de várias variáveis distintas. Os dados referem-se a 500 pessoas que são clientes de uma financeira. Por meio de uma regressão logística pede-se:

a) Quais variáveis são significativas para se elaborar uma boa previsão de risco de default?
b) Calcule a probabilidade de default de um indivíduo com as seguintes características:

- Idade = 40 anos
- Nível de educação = 3
- Emprego atual = 3 anos
- Endereço atual = 5 anos
- Outras dívidas (em milhares) $30,00

c) Interprete os resultados e anexe/envie o código utilizado para a construção do modelo.

In [66]:

```
str(credito)
```

```
'data.frame':   500 obs. of  5 variables:
 $ idade     : num  41 27 40 41 24 41 39 43 24 36 ...
 $ educação  : num  3 1 1 1 2 2 1 1 1 1 ...
 $ t_emprego : num  18 10 15 15 2 5 20 12 3 0 ...
 $ outras_dív: num  5.01 4 2.17 0.82 3.06 ...
 $ default   : num  1 0 0 0 1 0 0 0 1 0 ...
```

In [44]:

```
summary(credito)
```

```
     idade          educação       t_emprego       outras_dív    
 Min.   :20.00   Min.   :1.000   Min.   : 0.00   Min.   : 0.050  
 1st Qu.:29.00   1st Qu.:1.000   1st Qu.: 3.00   1st Qu.: 1.018  
 Median :34.00   Median :1.000   Median : 7.00   Median : 1.945  
 Mean   :34.71   Mean   :1.716   Mean   : 8.29   Mean   : 3.024  
 3rd Qu.:40.00   3rd Qu.:2.000   3rd Qu.:13.00   3rd Qu.: 3.665  
 Max.   :56.00   Max.   :5.000   Max.   :29.00   Max.   :27.030  
    default      
 Min.   :0.000  
 1st Qu.:0.000  
 Median :0.000  
 Mean   :0.258  
 3rd Qu.:1.000  
 Max.   :1.000  
```

In [57]:

```
mod_credito_full <- glm(default ~ ., data =credito, family = "binomial")
```

In [46]:

```
summary(mod_credito_full)
```

```
Call:
glm(formula = default ~ ., family = "binomial", data = credito)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7534  -0.7694  -0.4719   0.3561   2.3829

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.463970   0.540957  -0.858    0.391
idade       -0.006992   0.016299  -0.429    0.668
educação     0.011350   0.120760   0.094    0.925
t_emprego   -0.193733   0.028580  -6.779 1.21e-11 ***
outras_dív   0.311104   0.048140   6.462 1.03e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.95  on 499  degrees of freedom
Residual deviance: 475.57  on 495  degrees of freedom
AIC: 485.57

Number of Fisher Scoring iterations: 5
```

In [64]:

```
mod_credito_menor <- step(mod_credito_full, direction='both', trace = 0);mod_cre
dito_menor
```

```
Call:  glm(formula = default ~ t_emprego + outras_dív, family = "bin
omial",
    data = credito)

Coefficients:
(Intercept)    t_emprego   outras_dív
    -0.6464      -0.1977       0.3086

Degrees of Freedom: 499 Total (i.e. Null);  497 Residual
Null Deviance:      571
Residual Deviance: 475.8        AIC: 481.8
```

In [65]:

```
summary(mod_credito_menor)
```

```
Call:
glm(formula = default ~ t_emprego + outras_dív, family = "binomial",

    data = credito)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7303  -0.7782  -0.4672   0.3637   2.3881

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.64638    0.17309  -3.734 0.000188 ***
t_emprego   -0.19771    0.02667  -7.414 1.22e-13 ***
outras_dív   0.30855    0.04612   6.691 2.22e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 570.95  on 499  degrees of freedom
Residual deviance: 475.76  on 497  degrees of freedom
AIC: 481.76

Number of Fisher Scoring iterations: 5
```

In [61]:

```
anova(mod_credito_full, mod_credito_menor, test="Chisq")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
495	475.5662	NA	NA	NA
497	475.7567	-2	-0.1905375	0.9091286

In [106]:

```
dados_previsao <- data.frame(idade=40, educação = 3, t_emprego=3, outras_dív= 30
);dados_previsao
```

idade	educação	t_emprego	outras_dív
40	3	3	30

In [107]:

```
dados_previsao$Prob <- predict(mod_credito_full, dados_previsao, type='response'
);dados_previsao
```

idade	educação	t_emprego	outras_dív	Prob
40	3	3	30	0.9996785

R:
Observando, percebemos que somente as variáveis t_emprego e outras_dív são significativas, o stepwise mostra que o melhor modelo sugerido também utiliza somente as variáveis preditoras t_emprego e outras_dív. O teste anova confirma que o modelo completo e o resumido não são significativamente diferentes. Usando o modelo completo, a previsão da probabilidade de default (falta de pagamento) com os dados informados é de 99,9%, ou seja, é muito provável que uma pessoa com esse perfil não realize o pagamento.