UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
Campus CERRO LARGO

PROJETO DE EXTENSÃO
Software R:
Capacitação em análise estatística de dados utilizando um software livre.

Fonte: ⟨https://www.r-project.org/⟩

Módulo III
Regressão Logística

Ministrante: Felipe Smolski

Blog do projeto: ⟨https://softwarelivrer.wordpress.com/equipe/⟩

Equipe:
Coordenadora:
Profe. Iara Endruweit Battisti (iara.battisti@uffs.edu.br)
Colaboradores:
Profa. Denize Reis
Prof. Erikson Kaszubowski
Prof. Reneo Prediger
Profa. Tatiane Chassot
Mestrando Felipe Smolski (felipesmolski@hotmail.com)
Bolsista:
Djaina Rieger - aluna de Engenharia Ambiental (djaina.rieger@outlook.com)
Voluntárias:
Jaíne Frank
Jaqueline Caye

Sumário

1 Introdução

O modelo de regressão logística é utilizado quando a variável dependente é binária, categórica ordenada ou mesmo categórica desordenada (quando não há relação hierárquica entre elas). A Tabela 1 exemplifica as perguntas que levam a estes três tipos de variáveis:

Tabela 1: Tipos de variáveis dependentes

Variável dependente binária:
Você votou na última eleição? 0 - Não 1 - Sim
Você prefere transporte público ou dirigir um carro? 0 - Prefiro Carro 1 - Prefiro transporte público
Variável dependente categórica ordenada:
Você concorda ou desconcorda com o presidente? 1 - Discordo 2 - Neutro 3 - Concordo
Variável dependente categórica não ordenada:
Se as eleições fossem hoje, em que partido você votaria? 1 - Democratas 2 - Qualquer um 3 - Repulicanos

Fonte: Adaptado de (TORRES-REYNA, 2014).

A regressão logística, portanto, trata os grupos de interesse (variável dependente) com valores de 0 e 1, ao passo que sua funcionalidade se ocupa de prever a probabilidade de uma observação estar no grupo igual a 1 ("eventos"), em relação ao grupo igual a zero ("não eventos").

Para a estimação dos coeficientes das variáveis independentes, são utilizados o valor logit ou a razão de desigualdades (HAIR *et al.*, 2009):

$$Logit_i = ln\left(\frac{prob_{eventos}}{1 - prob_{eventos}}\right) = b_0 + b_1X_1 + \ldots + b_nX_n \quad (1)$$

ou

$$Logit_i = \left(\frac{prob_{eventos}}{1 - prob_{eventos}}\right) = e^{b_0+b_1X_1+\ldots+b_nX_n} \quad (2)$$

Mais detalhes sobre o modelo de regressão logística podem ser verificados na seção chamada **O Modelo de Regressão Logística**, bem como em Hosmer e Lemeschow (2000) e Gujarati2011.

2 Regressão Logística Simples - Exemplo 1

Este primeiro exemplo tratará da regressão logística simples, portanto, utilizando somente uma variável independente. Os dados são originados do livro de Hosmer e Lemeschow (2000), tratando-se de uma amostra com 100 pessoas. A variável dependente é a ocorrência ou não (1 ou 0) de doença coronária cardíaca (CHD), associando-se com a idade (AGE) dos indivíduos, criando assim um modelo de regressão logística.

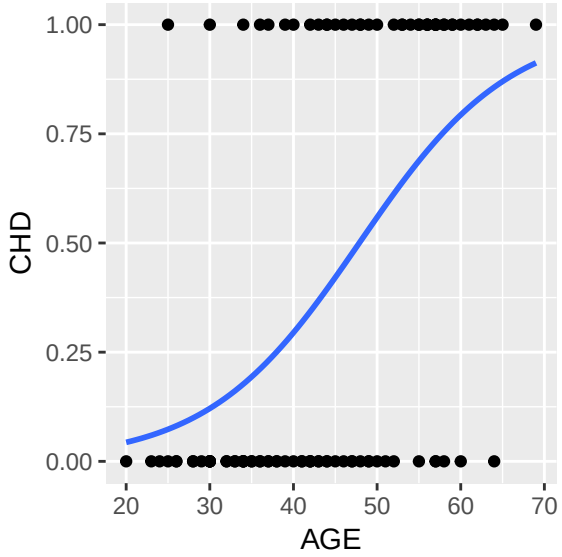
```
library(readr)
chd <- read_delim("https://raw.githubusercontent.com/Smolski/files/master/chd.csv",
                  ";", escape_double = FALSE, trim_ws = TRUE)
View(chd)
```

```
summary(chd)
```

```
##       AGE             AGRP            CHD
##  Min.   :20.00   Min.   :1.00   Min.   :0.00
##  1st Qu.:34.75   1st Qu.:2.75   1st Qu.:0.00
##  Median :44.00   Median :4.00   Median :0.00
##  Mean   :44.38   Mean   :4.48   Mean   :0.43
##  3rd Qu.:55.00   3rd Qu.:7.00   3rd Qu.:1.00
##  Max.   :69.00   Max.   :8.00   Max.   :1.00
```

Observa-se na figura abaixo a dispersão dos "eventos" da CHD com a idade (AGE).

```
library(ggplot2)
ggplot(chd, aes(x=AGE, y=CHD)) + geom_point() +
  stat_smooth(method="glm", method.args=list(family="binomial"), se=FALSE)
```



Monta-se então o modelo de regressão logística com a variável dependente CHD e a variável independente AGE. Abaixo demonstra-se a descrição da equação utilizando o comando *summary()* para o modelo m1 com a sintaze básica: *glm(Y modelo, family=binomial(link="logit"))*. Assim é obtida a função de ligação estimada do modelo:

$$\hat{g}(CHD) = -5,309 + 0,1109AGE \tag{3}$$

```
m1=glm(CHD~AGE, family = binomial(link="logit"), data = chd)
summary(m1)
```

```
##
## Call:
## glm(formula = CHD ~ AGE, family = binomial(link = "logit"), data = chd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9718  -0.8456  -0.4576   0.8253   2.2859
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
## AGE          0.11092    0.02406   4.610 4.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 136.66  on 99  degrees of freedom
## Residual deviance: 107.35  on 98  degrees of freedom
## AIC: 111.35
##
## Number of Fisher Scoring iterations: 4
```

Se observa o intercepto com o valor de -5,309, sendo que para a análise aqui proposta da relação entre CHD e AGE não obtém-se um significado prático para este resultado. No entanto, a variável de interesse é idade, que no modelo de regressão obteve o coeficiente de 0,1109 e pelo fato de ser positivo informa que quando a idade (AGE) se eleva, elevam-se as chances de ocorrência de CHD. De igual forma, nota-se que há significância estatística a $p = 0,001$ na utilização da variável AGE para o modelo, mostrando que possui importância ao modelo de regressão proposto.

2.1 Estimando a Razão de Chances

O modelo de regressão logística, porém, traz os resultados dos estimadores na forma logarítma, ou seja, o log das chances da variável idade no modelo é 0,1109. No entanto, para uma interpretação mais enriquecida da relação da idade com o CHD é necessária a transformação deste coeficiente, ou seja, que seja efetuada a exponenciação da(s) variavel(eis) da regressão. Assim, obtém-se a razão das chances (OR - Odds Ratio em inglês) para as variáveis independentes.

Uma maneira prática de se obter a razão de chances no RStudio é utilizando o pacote mfx. Novamente o intercepto não nos interessa nesta análise mas sim a variável AGE. Como demonstrado abaixo, o resultado da razão de chances da variável AGE foi de 1,1173, o que pode assim ser interpretado: para cada variação unitária na idade (AGE), as chances de ocorrência de CHD aumentam 1,1173 vezes. Dito de outra forma, para cada variação unitária em AGE, aumentam-se 11,73% ((1,1173-1)*100) as chances da ocorrência de CHD.

```
require(mfx)
logitor(CHD~AGE,data = chd)

## Call:
## logitor(formula = CHD ~ AGE, data = chd)
##
## Odds Ratio:
##     OddsRatio Std. Err.      z     P>|z|
## AGE  1.117307  0.026882 4.6102 4.022e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.2 Determinando o Intervalo de Confiança

A determinação do intervalo de confiança do modelo proposto é relevante para que seja analizada a estimativa do intervalo de predição do coeficiente da variável independente, a um nível de confiança de 95%. Desta forma, em 95% dos casos, o parâmetro dos coeficientes estará dentro deste intervalo.

De forma prática é possível determinar os intervalos de confiança com o comando $confint()$ commo observado abaixo, sendo que o coeficiente AGE toma o valor de 1,1173, podendo variar de 1,0692 a 1,1758.

```
exp(cbind(OR=coef(m1), confint(m1)))

##                        OR        2.5 %    97.5 %
## (Intercept) 0.004944629 0.0004412621 0.0389236
## AGE         1.117306795 1.0692223156 1.1758681
```

2.3 Predição de Probabilidades

A partir dos coneficientes do modelo de regressão logística é possível, portanto, efetuar a predição da variável categórica CHD, ou seja, saber a chance de ocorrer CHD com relação à uma determinada idade (AGE). No exemplo abaixo, primeiramente utilizamos a idade média das observações (44,38 anos), criando assim um novo data.frame chamadio media. Para utilizar o valor da idade média na função de regressão obtida ($m1$), utiliza-se a função $predict()$, de acordo com valor da média encontrada (data.frame media). O resultado mostra que para a idade média da amostra, 44,38 anos, há uma probabilidade de 40,44% na ocorrência da doença CHD. Esta ferramenta permite também a comparação pelo pesquisador das diferentes probabilidades entre as diversas idades (variável AGE).

```
media = data.frame(AGE=mean(chd$AGE))
media

##     AGE
## 1 44.38
```

```
media$pred.prob = predict(m1, newdata=media, type="response")
media

##     AGE pred.prob
## 1 44.38 0.4044944
```

2.4 Matriz de Confusão

Uma maneira prática de qualificar o ajuste do modelo de regressão logística é pela projeção do modelo na tabela de classificação (ou Matriz de Confusão). Para isto, precisa-se criar uma tabela com o resultado da classificação cruzada da variável resposta, de acordo com uma variável dicotômica em que os valores se derivam das probabilidades logísticas estimadas na regressão (HOSMER; LEMESCHOW, 2000). No entanto, é preciso definir uma regra de predição, que dirá se houve acerto ou não da probabilidade estimada com os valores reais, pois as probabilidades variam de 0 a 1 enquanto os valores reais binários possuem valores fixos de 0 “ou” 1.

É intuitivo supor que se as probabilidades aproximam-se de 1 o índivíduo estimado pode ser classificado como $\hat{Y}_i = 1$, bem como de forma contrária, se o modelo estimar probabilidades perto de 0, classificá-la como $\hat{Y}_i = 0$. Mas qual nível utilizar? Para resolver este problema, é preciso em primeiro lugar determinar um ponto de corte para classificar a estimação como 0 ou 1. Usualmente na literatura se utiliza o valor de 0,5 mas dependendo do estudo proposto pode não ser limitado a este nível (HOSMER; LEMESCHOW, 2000).

Após determinado o ponto de corte, é importante avaliar o poder de discriminação do modelo, pelo seu desempenho portanto em classificar os “eventos” dos “não eventos”. Cria-se a Matriz de Confusão (vide Tabela 2) com as observações de Verdadeiro Positivo (VP), Falso

Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN) e em seguida determinam-se alguns parâmetros numéricos, a serem descritos abaixo:

Precisão: representa a proporção das predições corretas do modelo sobre o total:

$$ACC = \frac{VP + VN}{P + N} \tag{4}$$

onde P representa o total de "eventos" positivos (Y=1) e N é o total de "não eventos" (Y=0, ou negativo).

Sensibilidade: representa a proportação de verdadeiros positivos, ou seja, a capacidade do modelo em avaliar o evento como $\hat{Y} = 1$ (estimado) dado que ele é evento real $Y = 1$:

$$SENS = \frac{VP}{FN} \tag{5}$$

Especificidade: a proporção apresentada dos verdadeiros negativos, ou seja, o poder de predição do modelo em avaliar como "não evento" $\hat{Y} = 0$ sendo que ele não é evento $Y = 0$:

$$SENS = \frac{VN}{VN + FP} \tag{6}$$

Verdadeiro Preditivo Positivo: se caracteriza como proporção de verdadeiros positivos com relação ao total de predições positivas, ou seja, se o evento é real $Y = 1$ dada a classificação do modelo $\hat{Y} = 1$:

$$VPP = \frac{VPP}{VN + FP} \tag{7}$$

Verdadeiro Preditivo Negativo: se caracteriza pela proporção de verdadeiros negativos comparando-se com o total de predições negativas, ou seja, o indivíduo não ser evento $Y = 0$ dada classificação do modelo como "não evento" $\hat{Y} = 0$:

$$VPN = \frac{VN}{VN + FN} \tag{8}$$

Tabela 2: Matriz de Confusão

		Valor Observado	
		Y=1	**Y=0**
Valor Estimado	$\hat{Y}_i = 1$	VP	FP
	$\hat{Y}_i = 0$	FN	VN

Fonte: Adaptado de Fawcett (2006).

```
require(caret)
pdata <- predict(m1, newdata = chd, type = "response")
confusionMatrix(data = as.numeric(pdata>0.5), reference = chd$CHD)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 45 14
##          1 12 29
##
##                Accuracy : 0.74
##                  95% CI : (0.6427, 0.8226)
##     No Information Rate : 0.57
```

```
##     P-Value [Acc > NIR] : 0.0003187
##
##                   Kappa : 0.4666
##  Mcnemar's Test P-Value : 0.8445193
##
##             Sensitivity : 0.7895
##             Specificity : 0.6744
##          Pos Pred Value : 0.7627
##          Neg Pred Value : 0.7073
##              Prevalence : 0.5700
##          Detection Rate : 0.4500
##    Detection Prevalence : 0.5900
##       Balanced Accuracy : 0.7319
##
##        'Positive' Class : 0
##
```

2.5 Curva ROC

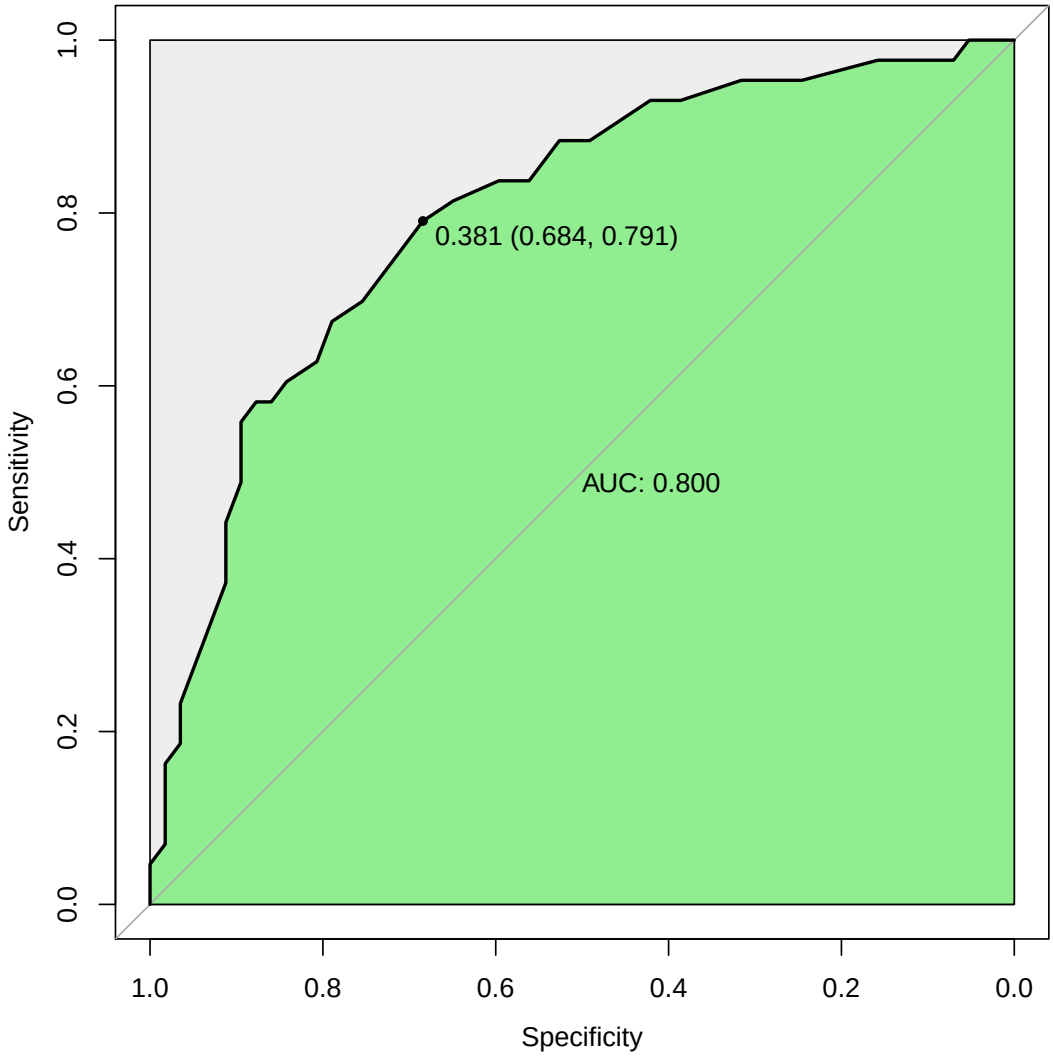
A Curva ROC (Receiver Operating Characteristic Curve) associada ao modelo logístico mensura a capacidade de predição do modelo proposto, através das predições da sensibilidade e da especificidade.

1º Passo:

library(pROC) roc1=plot.roc(chdCHD, $fitted(m1)$)

2º Passo:

```
plot(roc1,print.auc=TRUE, auc.polygon=TRUE,grud=c(0.1,0.2),
grid.col=c("green","red"),max.auc.polygon=TRUE,auc.polygon.col="lightgreen",print.thres=TRUE)
```



2.6 O teste Hosmer e Lemeshow

```
require(ResourceSelection)
hl=hoslem.test(chd$CHD,fitted(m1),g=10)
hl

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  chd$CHD, fitted(m1)
## X-squared = 2.2243, df = 8, p-value = 0.9734
```

3 Pseudo R^2

```
library(modEvA)
RsqGLM(m1)

## $CoxSnell
## [1] 0.2540516
##
## $Nagelkerke
## [1] 0.3409928
##
## $McFadden
## [1] 0.2144684
##
## $Tjur
## [1] 0.2705749
##
## $sqPearson
## [1] 0.2725518
```

4 Regressão Logística Múltipla - Exemplo 2

O exemplo abaixo abordado foi extraído de Torres-Reyna (2014), onde observa-se o banco de dados criado chamado *mydata*, possuindo as variáveis *country*, *year*, *y*, *y_bin*, *x1*, *x2*, *x3* e *opinion*. A variável dependente é *y_bin*, da qual foi categorizada entre 0 e 1 conforme a ocorrência de valores negativos em *y*. As variáveis independentes do modelo serão *x1*, *x2*e *x3*.

```
library(foreign)
mydata <- read.dta("http://dss.princeton.edu/training/Panel101.dta")
summary(mydata)

##  country      year            y                  y_bin
##  A:10    Min.   :1990   Min.   :-7.863e+09   Min.   :0.0
##  B:10    1st Qu.:1992   1st Qu.: 2.466e+08   1st Qu.:1.0
##  C:10    Median :1994   Median : 1.898e+09   Median :1.0
##  D:10    Mean   :1994   Mean   : 1.845e+09   Mean   :0.8
##  E:10    3rd Qu.:1997   3rd Qu.: 3.372e+09   3rd Qu.:1.0
##  F:10    Max.   :1999   Max.   : 8.941e+09   Max.   :1.0
##  G:10
##        x1                x2                x3                opinion
##  Min.   :-0.5676   Min.   :-1.6218   Min.   :-1.16539   Str agree:20
##  1st Qu.: 0.3290   1st Qu.:-1.2156   1st Qu.:-0.07931   Agree    :15
##  Median : 0.6413   Median :-0.4621   Median : 0.51419   Disag    :19
##  Mean   : 0.6480   Mean   : 0.1339   Mean   : 0.76185   Str disag:16
##  3rd Qu.: 1.0958   3rd Qu.: 1.6078   3rd Qu.: 1.15486
##  Max.   : 1.4464   Max.   : 2.5303   Max.   : 7.16892
##
```

Utiliza-se uma função para Modelos Lineares Generalizados (glm - em inglês Generalized Linear Models), determinando a variável dependente (y_bin), as variáveis independentes (x1+x2+x3), a base de dados a ser utilizada (data=mydata) e a família dos modelos (family = binomial(link="logit")).

Abaixo os resultados da estimação do modelo utilizando o comando *summary*. Observa-se que os valores **estimados** mostram os coeficientes em formato logarítmo de chances. Assim, quando x3 eleva-se em 1 (uma) unidade, o log das chances esperado para x3 altera-se em 0,7512. Neste

ponto, observa-se que as três variáveis independentes possuem efeitos positivos para determinação das chances do preditor ser igual a 1, caso contrário constariam com sinal negativo. A coluna $Pr(>|z|)$ traz os p-valores das variáveis indicando o teste da hipótese nula. Como resultado a variável x3 revelou significância estatística a 10% (0,10), no entanto o valor usual para considerá-la estatísticamente significante é 5% (0,05). Para fins de explanação do modelo, neste trabalho, serão efetuadas as demais análises do modelo de forma explicativa.

```
logit=glm(y_bin~x1+x2+x3, data=mydata, family = binomial(link="logit"))
summary(logit)

##
## Call:
## glm(formula = y_bin ~ x1 + x2 + x3, family = binomial(link = "logit"),
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0277   0.2347   0.5542   0.7016   1.0839
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4262     0.6390   0.667   0.5048
## x1            0.8618     0.7840   1.099   0.2717
## x2            0.3665     0.3082   1.189   0.2343
## x3            0.7512     0.4548   1.652   0.0986 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.056  on 69  degrees of freedom
## Residual deviance: 65.512  on 66  degrees of freedom
## AIC: 73.512
##
## Number of Fisher Scoring iterations: 5
```

A razão de chances (OR - odds ratio em inglês) estimada no modelo terá de ser transformada por estar apresentada na forma logarítma conforme o modelo de regressão logística o estima. Assim, utiliza-se o pacote *mfx* para efetuar esta transformação para todo o modelo de forma automátizada(logitor($y_bin\ x1 + x2 + x3, data = mydata$)) :

```
require(mfx)
logitor(y_bin~x1+x2+x3,data=mydata)

## Call:
## logitor(formula = y_bin ~ x1 + x2 + x3, data = mydata)
##
## Odds Ratio:
##    OddsRatio Std. Err.      z   P>|z|
## x1   2.36735   1.85600 1.0992 0.27168
## x2   1.44273   0.44459 1.1894 0.23427
## x3   2.11957   0.96405 1.6516 0.09861 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O resultado acima evidencia que para uma alteração em 1 (uma) unidade em x3, a chance de que y seja igual a 1 aumenta em 112% ((2,12-1)*100). Dito de outra forma, a chance de y=1 é 2,12 vezes maior quando x3 aumenta em uma unidade (sendo que aqui mantêm-se as demais variáveis independentes constantes).

Como visto, para cada variação unitária em x3 o log das chances varia 0,7512. É possível estimar, portanto, a alteração das chances em função das médias dos valores de cada variável x1 e x2, e utilizar como exemplo os valores de 1, 2 e 3 para x3, para assim alcançar os preditores do log das chances nesta simulação, como segue abaixo:

Para facilitar a interpretação do modelo, se torna mais fácil depois de transformado a sua exponenciação dos coeficientes logísticos utilizando o comando exp(coef(logit)). Desta forma, para cada incremento unitário em x2 e mantendo as demais variáveis constantes, conclui-se que é 1,443 vezes provável que y seja igual a 1 em oposição a não ser (igual a zero), ou seja, as chances aumentam em 44,30%.

```
exp(coef(logit))
```

```
## (Intercept)          x1          x2          x3
##    1.531417    2.367352    1.442727    2.119566
```

O **intervalo de confiança** do modelo pode ser exposto utilizando o comando *confint* para os coeficientes estimados, como segue abaixo:

```
exp(cbind(OR=coef(logit), confint(logit)))
```

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

```
##                    OR     2.5 %    97.5 %
## (Intercept) 1.531417 0.4387468  5.625299
## x1          2.367352 0.5129380 11.674641
## x2          1.442727 0.8041221  2.737965
## x3          2.119566 1.0038973  5.718637
```

A partir do modelo logístico, podemos realizar **predições das probabilidades** de se encontrar o resultado y=1 conforme visto acima. Para isto, como exercício utilizaremos as médias das observações de cada variável independente do modelo. Em primeiro lugar deve ser criado um data.frame com os valores médios, como segue:

```
allmean = data.frame(x1=mean(mydata$x1),
                     x2=mean(mydata$x2),
                     x3=mean(mydata$x3))
allmean
```

```
##          x1        x2       x3
## 1 0.6480006 0.1338694 0.761851
```

Utiliza-se o comando *predict()* para predição do modelo, como segue abaixo, informando o objeto criado com a equação do modelo (logit), a base de dados com as condições dos valores médios (allmean) e o tipo de teste requerido (”response”) para predizer as probabilidades. Como resultado, o modelo informa que constando os valores médios das variáveis independentes, obtêm-se a probabilidade de 83% em y se constituir igual a 1.

```
allmean$pred.prob = predict(logit, newdata=allmean, type="response")
allmean
```

```
##          x1        x2       x3 pred.prob
## 1 0.6480006 0.1338694 0.761851 0.8328555
```

4.1 Método Stepwise

O método Stepwise auxilia o pesquisador em selecionar as variáveis importantes ao modelo:

```
step(logit, direction = 'both')

## Start:  AIC=73.51
## y_bin ~ x1 + x2 + x3
##
##        Df Deviance    AIC
## - x1    1   66.736 72.736
## - x2    1   66.996 72.996
## <none>      65.512 73.512
## - x3    1   69.402 75.402
##
## Step:  AIC=72.74
## y_bin ~ x2 + x3
##
##        Df Deviance    AIC
## - x2    1   67.330 71.330
## <none>      66.736 72.736
## + x1    1   65.512 73.512
## - x3    1   70.032 74.032
##
## Step:  AIC=71.33
## y_bin ~ x3
##
##        Df Deviance    AIC
## <none>      67.330 71.330
## - x3    1   70.056 72.056
## + x2    1   66.736 72.736
## + x1    1   66.996 72.996
##
## Call:  glm(formula = y_bin ~ x3, family = binomial(link = "logit"),
##     data = mydata)
##
## Coefficients:
## (Intercept)           x3
##      1.1339       0.4866
##
## Degrees of Freedom: 69 Total (i.e. Null);  68 Residual
## Null Deviance:     70.06
## Residual Deviance: 67.33  AIC: 71.33
```

5 Regressão Logística Múltipla - exemplo 3

Abaixo segue um exemplo com uma variável dependente categórica:

admin: Variável dependente = 0 (não admitido) e 1 (admitido)

Rank: Variável independente = ranking da escola de proveniência do candidato

Gre: Variável independente = exames prévios do candidato.

Gpa: Variável independente = exames prévios do candidato.

```
library(readr)
binary <- read_csv("http://www.karlin.mff.cuni.cz/~pesta/prednasky/NMFM404/Data/binary.csv")

binary$rank <- factor(binary$rank)
mylogit <- glm(admit ~ gre + gpa + rank, data = binary, family = binomial(link="logit"))
```

```
exp(cbind(OR = coef(mylogit), confint(mylogit)))

##                      OR       2.5 %    97.5 %
## (Intercept) 0.0185001 0.001889165 0.1665354
## gre         1.0022670 1.000137602 1.0044457
## gpa         2.2345448 1.173858216 4.3238349
## rank2       0.5089310 0.272289674 0.9448343
## rank3       0.2617923 0.131641717 0.5115181
## rank4       0.2119375 0.090715546 0.4706961
```

6 O Modelo de Regressão Logística

A regressão *logit* é, dentre os métodos de regressão existentes, aquela em que a variável resposta é binária/dicotômica. O modelo paramétrico e as premissas básicas refletem a diferença existente do modelo de regressão logística para os modelos lineares segundo Hosmer e Lemeschow (2000). "*Once this diference is accounted for, the method employed in an analysis using logistic regression follow the same general principles used in liner regression. Thus, the teqniques used in linear regression analysis will motivate our approach to logistic regression*" (HOSMER; LEMESCHOW, 2000, p.1). Três questões sãoimportantes para a utilização da regressão logística: a variável dependente deve ser zero ou um; é utilizada a distribuição binomial para descrever a distribuição dos erros em vez da distribuição normal e; utiliza dos princípios de análises da regressão linear.

A probabilidade de que um "evento" ocorra (1) em comparação a que não ocorra ("não evento" - 0) é dada por:

$$P_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X_i)}} \tag{9}$$

Sendo melhor representado por:

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e^Z}{1 + e^Z} \tag{10}$$

onde $Z_i = \beta_1 + \beta_2 X_i$.

Assim, a Equação 10 mostra o que se chama de função de **distribuição logística acumulada** (GUJARATI; PORTER, 2011). Se P_i é a probabilidade de um evento ocorrer, então a probabilidade de um evento não ocorrer é dado por $(1 - P_i)$:

$$1 - P_i = \frac{1}{1 + e^{Z_i}} \tag{11}$$

sendo reescrita:

$$\frac{P_i}{1 - P_i} = \frac{1 + e^{Z_i}}{1 + e^{-Z_i}} = e^{Z_i} \tag{12}$$

Desta forma, $P_i/(1-P_i)$ se torna a razão de chances em favor de um evento ocorrer menos a razão de probabilidade de que um evento ocorra, contra a probabilidade de um evento não ocorrer.

Obtém-se em seguida o logaritmo natural desta equação, sendo o L o logarítmo da razão de chances (**logit**), e para fins de estimação, reescreve-se a equação:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i + \mu_i \tag{13}$$

Algumas características importantes do modelo de regressão logística são tratadas por Gujarati e Porter (2011):

- (a) P varia entre 0 e 1, logo o logit L vai de $-\infty$ a $+\infty$ e então os logits não são limitados;
- (b) L é linear em X, no entanto as probabilidades em si não são, o que contrasta com o Modelo de Probabilidade Linear (MPL);
- c) na regressão logística podem ser acrescentados vários regressores, tanto quanto a teoria permita;
- (d) caso o logit L for positivo, significa que em caso de aumento no valor do regressor, as chances de que o regressando seja igual a 1 aumentam. Caso contrário (regressor L negativo), a medida que o regressor aumenta diminuem as chances de que o regressando seja igual a 1;
- (e) a interpretação do modelo da Equação 13: o coeficiente angular β_2 mede a variação em L para cada unidade de variação em X, ou seja, as chances faroráveis em encontrar-se o regressando 1 em comparação a zero. o intercepto β_1 representa o valor do logarítmo das chances favoráveis ao regressando quando o regressor for igual a zero (quando se interpreta o intercepto na maioria dos casos não possui sentido físico);
- (f) é possível calcular não somente as chances favoráveis à encontrar o regressando igual a 1 mas a própria probabilidade de que o regressando seja 1, conforme dado pela Equação 13, visto que β_1 e β_2 estão dados;
- (g) supõe-se que o log da razão das chances está linearmente relacionadao a X_i. Para estimar-se a equação, em estando os dados em nível individual ou micro, não se pode utilizar do método dos Mínimos Quadrados Ordinários (MQO). Recorre-se então ao método da máxima verossimilhança (MV), que está detalhada a seguir.

6.1 Máxima Verossimilhança e sua Estimativa para o Modelo Logit

Gujarati e Porter (2011) contribuem para a explicação da estimativa da máxima verossimilhança para os modelos de regressão logística. O interesse do modelo de regressão logística é determinar o cálculo da probabilidade de que um evento ocorra dada outra variável X, que pode ser expressa pela função logística:

$$P_i = \frac{1}{1 - e^{\beta_1 + \beta_2 X_i}} \tag{14}$$

Não se observa diretamente P_i mas o resutlado de $Y = 1$ caso exista a ocorrência do indicador e zero se não existir. Pelo fato de que cada Y_i é uma variável aleatória de Bernoulli, se escreve:

$$Pr(Y_i = 1) = P_i \tag{15}$$

$$Pr(Y_i = 0) = (1 - P_i) \tag{16}$$

Caso seja relacionada uma amostra aleatória de n observações, e considera-se que $f_i(Y_i)$ indica uma probabilidade de que $Y_i = 1$ ou 0, então a probabilidade conjunta de observação dos n valores Y, ou seja, $f(Y_1, Y_2, \ldots, Y_n)$ é dada por:

$$f(Y_1, Y_2, \ldots, Y_n) = \prod_1^n f_i(Y_i) = \prod_1^n P_i^{Y_i}(1 - P_i)^{1-Y_i} \tag{17}$$

onde $\prod$ se configura no operador de produtório. Pode ser escrita a função de densidade da probabilidade conjutna como produto das funções de densidade individuais, pois cada Y_i possui a mesma função densidade (logística). Portanto, se conhece a probabiliadade conjunta da Equação (17) como **função de verossimilhança (FV)**, podendo tomar seu logarítmo natural e obtendo a **função de verossimilhança logística (FVL)**:

$$\begin{aligned} f(Y_1, Y_2, \ldots, Y_n) &= \sum_1^n [Y_i \ln P_i + (1 - Y_i) \ln (1 - P_i)] \\ &= \sum_1^n [Y_i \ln P_i - Y_i \ln (1 - P_i) + \ln (1 - P_i)] \\ &= \sum_1^n \left[Y_i \ln \left(\frac{P_i}{1 - P_i} \right) \right] + \sum_1^n \ln(1 - P_i) \end{aligned} \tag{18}$$

Da Equação (1) verifica-se que

$$(1 - P_i) = \frac{1}{1 + e^{\beta_1 + \beta_2 X_i}} \tag{19}$$

bem como

$$\ln \left(\frac{P_i}{1 - P_i} \right) = 1 + \beta_1 + \beta_2 X_i \tag{20}$$

Ao usar as equações (6) e (7), pode-se reescrever a FLV (5):

$$f(Y_1, Y_2, \ldots, Y_n) = \sum_1^n Y_i(\beta_1 + \beta_2 X_i) - \sum_1^n \ln \left[1 + e^{(\beta_1 + \beta_2 X_i)} \right] \tag{21}$$

A equação demonstra que a função de verossimilhança logarítma é uma função dos parâmetros β_1 e β_2, ao passo que os X_i são conehecidos. O objetivo é maximizar a FV (ou FVL), obtendo os valores dos parâmetros desconhecidos de maneira que a probabilidade de observar Y nos dados seja a mais alta possível. Diferencia-se a equação (8) parcialmente com relação a cada incógnita e iguala-se as expressões resultantes a zero para resolver. Aplica-se, em seguida, a condição de maximização de segunda ordem para verificar se os valores destes parâmetros obtidos maximizam a FV.

Referências

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, p. 861–874, 2006.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica**. 5a. ed. New York: Mc Graw Hill, 2011.

HAIR, J. F. *et al.* **Análise Multivariada de Dados**. 6a. ed. São Paulo: Bookman, 2009.

HOSMER, D. W.; LEMESCHOW, S. **Applied Logistic Regression**. 2a. ed. New York: Wiley, 2000. 397 p.

TORRES-REYNA, O. **Logit, Probit and Multinomial Logit models in R**. 2014. Disponível em: ⟨http://dss.princeton.edu/training/LogitR101.pdf⟩.