Regressão logística

Baseado em: “*An introduction to statistical learning with applications in R*” & “*Practical data science with R*”

Por: Pedro de Araújo Ribeiro

Considerações:

- Dataset usado: Natal Risk
- Bibliotecas do R Studio usadas:

R studio

Motivação:

A regressão logística existe para contemplar casos onde um modelo de regressão é desejado porém as variáveis resposta são qualitativas e não quantitativas.

Dentre os motivos para usar um modelo diferente da regressão normal é o comportamento das variáveis comparadas com uma dada reta de regressão.

Especificamente, há consequências indesejadas tanto a atribuir uma resposta qualitativa a uma variável numérica quanto a calcular uma regressão para a mesma.

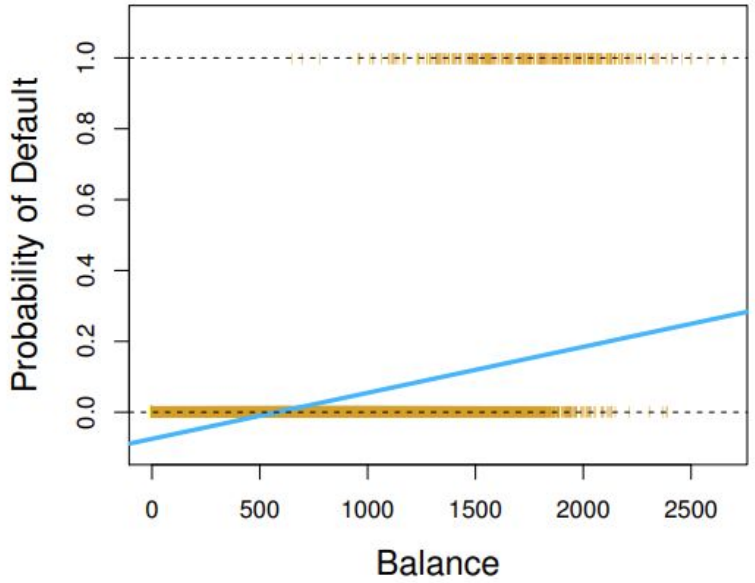
Motivação:

Problema com a quantificação de variáveis qualitativas:

$$Y = \begin{cases} 1 & \text{if } \texttt{epileptic seizure}; \\ 2 & \text{if } \texttt{stroke}; \\ 3 & \text{if } \texttt{drug overdose}. \end{cases}$$

ISRLv2 - sessão 4.2

Problema com o uso de regressão linear:

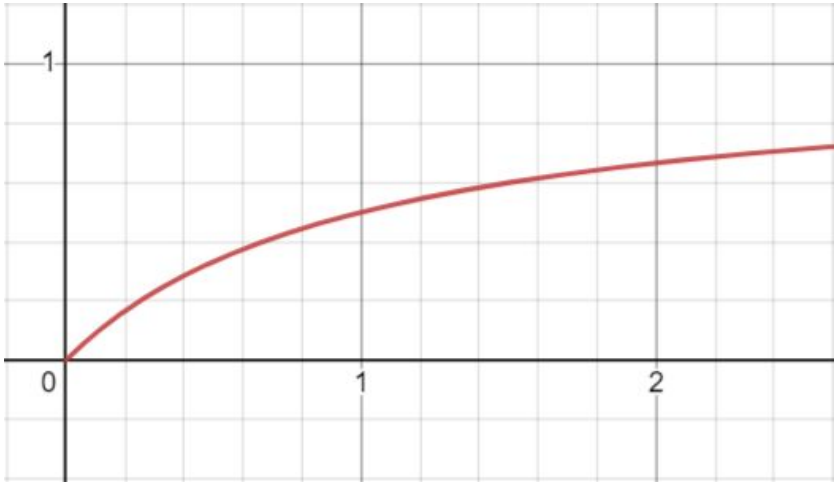


ISRLv2 - sessão 4.2

Solução:

Calcularemos então a probabilidade de uma observação, dadas suas variáveis de estudo, se encontrar em uma das classificações disponíveis.

Inicialmente tomamos a função ao lado, onde p representa a probabilidade da variável resposta e está contido no intervalo 0 < p < 1.

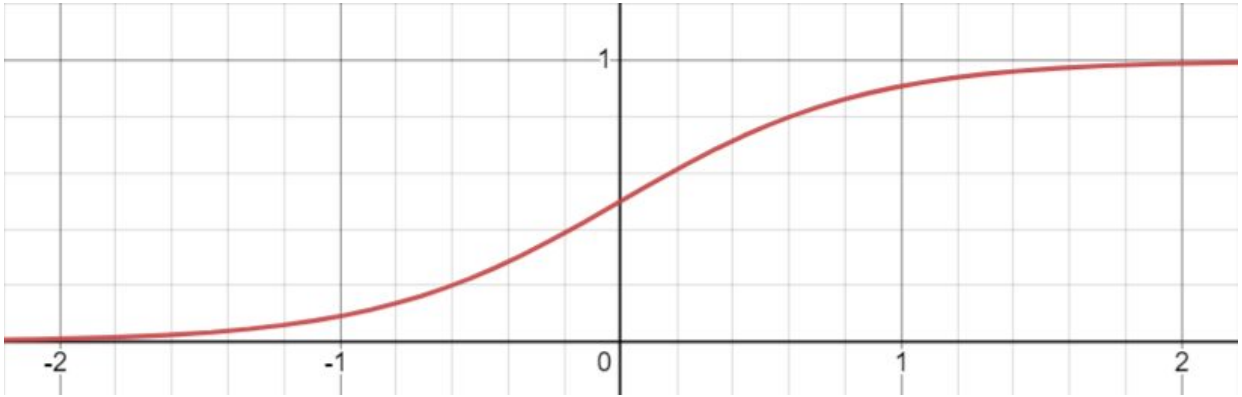


$$x = \frac{p}{1-p} \{0 < p < 1\}$$

Desmos

Solução:

Em seguida aplicamos o logaritmo da função anterior para chegar a uma função que define melhor o comportamento de problemas de classificação qualitativa.

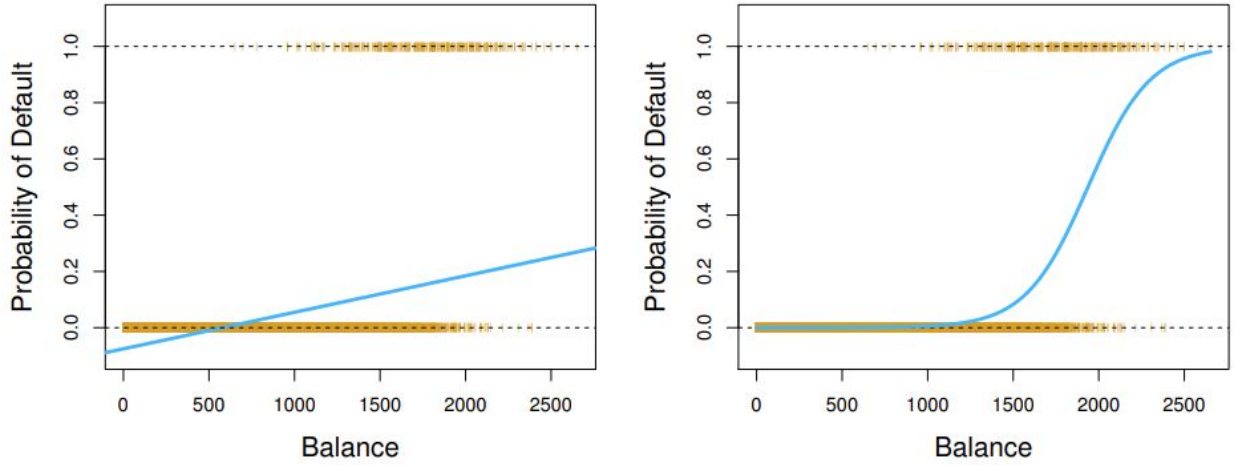


$$x = \log\left(\frac{p}{1-p} \{0 < p < 1\} \right)$$

Desmos

Solução:

Podemos ver que a nova função se adequa muito melhor ao comportamento das variáveis qualitativas.



ISRLv2 - sessão 4.2

Solução:

Porém neste estágio ainda não é possível expressar essa função como um modelo de regressão, então a igualamos com a função de regressão linear básica:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

ISRLv2 - sessão 4.3.1

E em seguida aplicamos a função exponencial dos dois lados:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

ISRLv2 - sessão 4.3.1

Solução:

Agora aplicamos transformações aritméticas simples e obtemos uma função de regressão capaz de descrever uma probabilidade:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

ISRLv2 - sessão 4.3.2

Em seguida nossa próxima prioridade será estimar os coeficientes da função, para isso buscaremos parâmetros que, dado uma observação, resultarão em um número mais próxima de 0 se negativo e 1 se positivo.

Solução:

Em termos simples, os coeficientes são derivados da função de máxima verossimilhança, onde aplicamos o multiplicando da probabilidade de uma observação ser positiva dado que ela é positiva e da probabilidade de ser negativa dado que é negativa.

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

ISRLv2 - sessão 4.3.1

Os valores de B0 e B1 serão aqueles que maximizam a função.

Solução:

Podemos extender a lógica de coeficientes para englobar observações com mais de uma variável de informação, resultando na seguinte fórmula:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

ISRLv2 - sessão 4.3.3

Realizando previsões:

Tendo calculado os coeficientes e os substituído na fórmula, para prever uma dada observação inserimos os valores de suas variáveis também na fórmula e o resultado será a probabilidade da observação ser positiva.

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576,$$

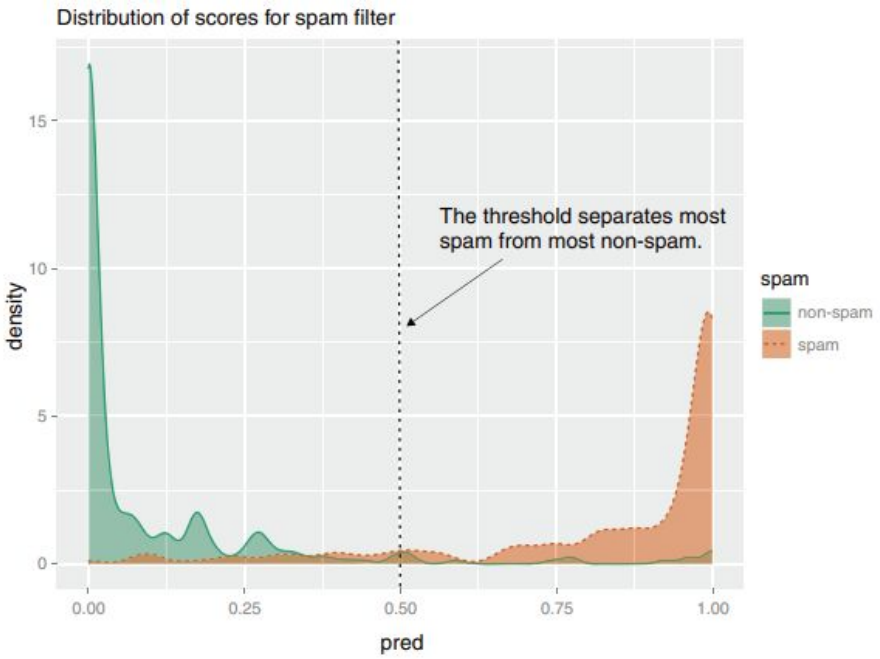
ISRLv2 - sessão 4.3.3

Contudo, em um cenário de problemas de classificação não basta ter apenas a probabilidade de algo ser positivo ou negativo, devemos realizar a asserção de que ela é de fato.

Realizando previsões:

Portanto iremos calcular a probabilidade para o conjunto de treino e comparar seus resultados entre si, tentando diferenciar as observações baseado na variável resposta.

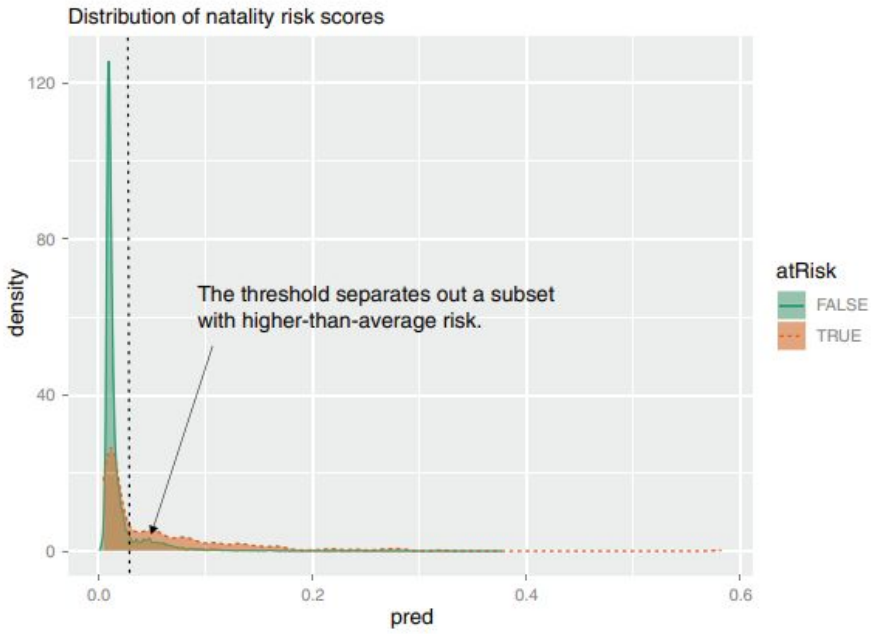
Idealmente encontraremos resultados como o da figura ao lado, porém normalmente não é o caso.



PDSWR - sessão 7.2.3

Realizando previsões:

Em casos como o da figura, deveremos fazer uma análise do objetivo da previsão e qual taxa de erro/acerto seria ideal maximizar e a partir daí testar valores de corte que atende essa necessidade.

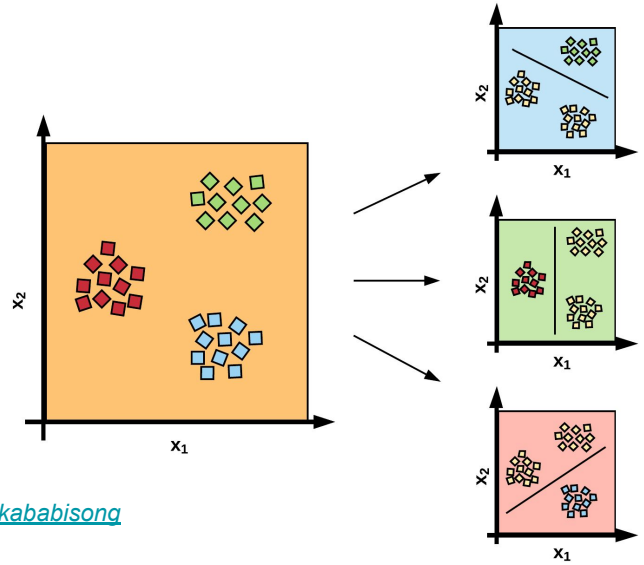


PDSWR - sessão 7.2.3

Método para respostas não binárias:

Uma forma de realizar previsões para mais que duas classificações é montar diferentes modelos para cada classificação. Especificamente, realizamos a chance de ser ou não do tipo 1, ser ou não do tipo 2, ser ou não do tipo 3 e etc.

Feito isso, a maior probabilidade dentre elas será a classificação escolhida



ekababisong

Análise de coeficientes:

A fórmula da regressão logística associa cada coeficiente a uma variável do objeto de estudo, porém diferente da regressão linear um aumento de X unidades na variável não significa um aumento de beta no resultado, e sim uma multiplicação por e^beta para cada X unidades.

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Além disso, observa-se que coeficientes negativos direcionam o resultado para classificar/prever a observação como falsa (0), enquanto coeficientes positivos levam ao verdadeiro (1)

Exemplo:

```
> summary(sdata)
      PWGT          UPREVIS         CIG_REC               GESTREC3                  DPLURAL
 Min.   : 74.0   Min.   : 0.00   Mode :logical   >= 37 weeks:23308   single            :25440
 1st Qu.:126.0   1st Qu.: 9.00   FALSE:23928     < 37 weeks : 3005   triplet or higher:   44
 Median :145.0   Median :11.00   TRUE :2385                          twin              :  829
 Mean   :153.7   Mean   :11.17
 3rd Qu.:172.0   3rd Qu.:13.00
 Max.   :375.0   Max.   :49.00
  ULD_MECO        ULD_PRECIP      ULD_BREECH       URF_DIAB        URF_CHYPER      URF_PHYPER
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:25084     FALSE:25642     FALSE:24662     FALSE:24900     FALSE:25991     FALSE:25171
 TRUE :1229      TRUE :671       TRUE :1651      TRUE :1413      TRUE :322       TRUE :1142


 URF_ECLAM         atRisk            DBWT      ORIGRANDGROUP
 Mode :logical   Mode :logical   Min.   : 227   Min.   : 0.000
 FALSE:26256     FALSE:25831     1st Qu.:2977   1st Qu.: 2.000
 TRUE :57        TRUE :482       Median :3316   Median : 5.000
                                 Mean   :3273   Mean   : 5.059
                                 3rd Qu.:3629   3rd Qu.: 8.000
                                 Max.   :6165   Max.   :10.000
```

R studio

Exemplo:

Agora faremos uma análise da significância dos coeficientes e do modelo como um todo.

```
> model <- glm(fmla, data = train, family = binomial(link = "logit"))
> summary(model)

Call:
glm(formula = fmla, family = binomial(link = "logit"), data = train)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -4.412189   0.289352 -15.249  < 2e-16 ***
PWGT                     0.003762   0.001487   2.530 0.011417 *
UPREVIS                 -0.063289   0.015252  -4.150 3.33e-05 ***
CIG_RECTRUE              0.313169   0.187230   1.673 0.094398 .
GESTREC3< 37 weeks       1.545183   0.140795  10.975  < 2e-16 ***
DPLURALtriplet or higher 1.394193   0.498866   2.795 0.005194 **
DPLURALtwin              0.312319   0.241088   1.295 0.195163
ULD_MECOTRUE             0.818426   0.235798   3.471 0.000519 ***
ULD_PRECIPTRUE           0.191720   0.357680   0.536 0.591951
ULD_BREECHTRUE           0.749237   0.178129   4.206 2.60e-05 ***
URF_DIABTRUE            -0.346467   0.287514  -1.205 0.228187
URF_CHYPERTRUE           0.560025   0.389678   1.437 0.150676
URF_PHYPERTRUE           0.161599   0.250003   0.646 0.518029
URF_ECLAMTRUE            0.498064   0.776948   0.641 0.521489
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2698.7  on 14211  degrees of freedom
Residual deviance: 2463.0  on 14198  degrees of freedom
AIC: 2491

Number of Fisher Scoring iterations: 7
```

R studio

Análise de coeficientes: Significância

A análise se torna muito mais conveniente graças ao R, que já verifica o p-valor de cada coeficiente para diferentes níveis de significância. A existência de pelo menos um coeficiente significativo já nos permite concluir que o modelo em si é estatisticamente significativo.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -4.412189   0.289352 -15.249  < 2e-16 ***
PWGT                     0.003762   0.001487   2.530 0.011417 *
UPREVIS                 -0.063289   0.015252  -4.150 3.33e-05 ***
CIG_RECTRUE              0.313169   0.187230   1.673 0.094398 .
GESTREC3< 37 weeks       1.545183   0.140795  10.975  < 2e-16 ***
DPLURALtriplet or higher 1.394193   0.498866   2.795 0.005194 **
DPLURALtwin              0.312319   0.241088   1.295 0.195163
ULD_MECOTRUE             0.818426   0.235798   3.471 0.000519 ***
ULD_PRECIPTRUE           0.191720   0.357680   0.536 0.591951
ULD_BREECHTRUE           0.749237   0.178129   4.206 2.60e-05 ***
URF_DIABTRUE            -0.346467   0.287514  -1.205 0.228187
URF_CHYPERTRUE           0.560025   0.389678   1.437 0.150676
URF_PHYPERTRUE           0.161599   0.250003   0.646 0.518029
URF_ECLAMTRUE            0.498064   0.776948   0.641 0.521489
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R studio

Análise de coeficientes: Melhora no modelo

Para o modelo anterior podemos realizar a previsão, obter a taxa de acerto e depois remover os coeficientes não significativos e recriar o modelo para verificar se houve uma melhora na previsão.

```
> ( ctab.test <- table(pred = test$pred > 0.02, atRisk = test$atRisk) )
       atRisk
pred    FALSE TRUE
  FALSE  9487   93
  TRUE   2405  116
> (ctab.test[1,1] + ctab.test[2,2])/sum(ctab.test)
[1] 0.7935708
> ctab.test[2,2]/(ctab.test[2,2]+ctab.test[1,2])
[1] 0.5550239
```

R studio

Análise de coeficientes: Melhora no modelo

```
> ctab.test
       atRisk
pred     FALSE TRUE
  FALSE   9523   97
  TRUE    2369  112
> (ctab.test[1,1] +
[1] 0.7962152
> ctab.test[2,2]/(ct
[1] 0.5358852
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.438467   0.280966 -15.797  < 2e-16 ***
PWGT                0.004110   0.001428   2.878 0.003998 **
UPREVIS            -0.062373   0.015177  -4.110 3.96e-05 ***
GESTREC3< 37 weeks  1.637163   0.131869  12.415  < 2e-16 ***
ULD_MECOTRUE        0.791332   0.235797   3.356 0.000791 ***
ULD_BREECHTRUE      0.847204   0.170086   4.981 6.32e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R studio

Conclusão:

- Regressão logística oferece uma nova alternativa para problemas de classificação, baseada no cálculo da probabilidade de uma observação pertencer a uma categoria.
- É aplicável a modelos com várias variáveis de estudo e também com mais que duas classificações, porém quanto menos melhor.
- Podemos analisar a significância estatística dos coeficientes da regressão e por consequência do modelo inteiro, e feito isso podemos remover os não significativos e reconstruir o modelo na esperança de obter melhores resultados.

FIM.