**Regressão Logística aplicada ao dataset Titanic**

**Flávia Guimarães Gaia Paula**

Centro Universitário Instituto de Educação Superior de Brasília (IESB)

Brasília – DF – Brazil

flavia.paula@iesb.edu.br

1. **Introdução**

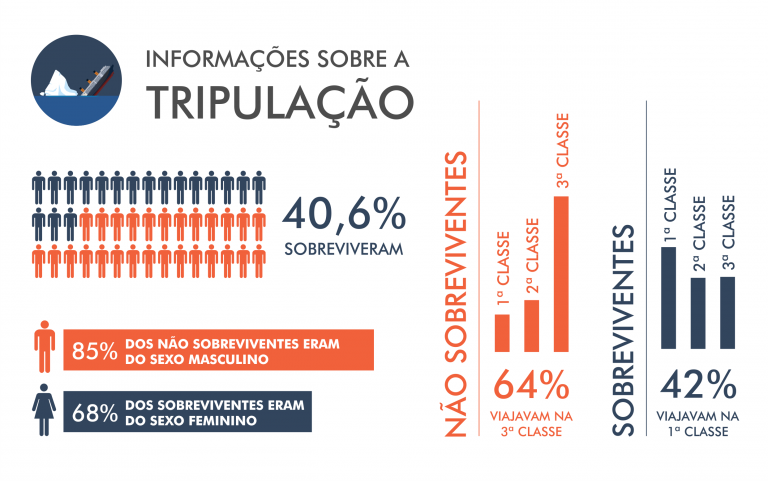
O naufrágio do Titanic, aconteceu em abril de 1912 e é considerado um dos acidentes marítimos mais famosos da história. O navio tinha sido um símbolo das conquistas industriais de Belfast e quando construiu prometia ser o mais luxuoso e seguro da época. O tema fez tanto sucesso que foi para o cinema em 1997 como uma história de ficção do naufrágio real do RMS Titanic, estrelando Leonardo DiCaprio como Jack Dawson, e Kate Winslet como Rose DeWitt Bukater.

Neste trabalho será utilizado o dataset do software R do caso Titanic considerando como variável resposta a morte ou a sobrevivência do passageiro no acidente com o navio RMS Titanic. Para isso será aplicada a técnica de regressão logística múltipla, que faz parte da família de Modelos Lineares Generalizados (MLG) e é adequada quando a variável de interesse (resposta) é binária, isto é, “sim” ou “não”. Através da Regressão Logística é possível avaliar os fatores que influenciam a ocorrência de determinado evento e a probabilidade do mesmo acontecer ou não.

1. **Desenvolvimento**

A Regressão Logística utiliza a função de ligação “logit”, que conecta os fatores de influência à variável resposta, esta função limita os valores previstos pelo modelo ao intervalo 0 e 1, o que possibilita a interpretação dos resultados em função da Razão de Chances (Odds Ratio). A Razão de Chances pode ser definida como a razão de um evento ocorrer em um grupo A em função de um grupo B. Como mulheres e crianças eram prioridade na evacuação do navio é esperado que a chance de sobrevivência dos indivíduos do sexo Feminino (grupo A) seja superior a chance de sobrevivência do sexo Masculino (grupo B).

As estatísticas aplicadas no dataset mostram que 424 indivíduos não sobreviveram, sendo 85% do sexo masculino e 290 sobreviveram, destes cerca de 68% eram do sexo feminino. Em relação a classe, 42% dos sobreviventes eram da 1ª Classe e 64% dos não sobreviventes eram da 3ª Classe. A idade média dos indivíduos que não sobreviveram foi de 30,62 anos, sendo a idade mínima de 1 ano e a máxima de 74 anos e daqueles que sobreviveram foi de 28,24 anos, sendo a idade mínima menor que 1 ano e a máxima de 80 anos, como mostra a figura abaixo:



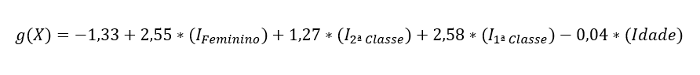
***Figura 1: Características da tripulação***

Quanto ao ajuste do modelo é possível observar que todas as variáveis foram significativas ao nível de 5% de significância (valor-p < 0,050). Logo, a chance de sobrevivência para indivíduos do sexo Feminino foi 12,46 [8,3; 18,71] vezes maior que a chance de sobrevivência para os indivíduos do sexo Masculino. Em relação a chance de sobrevivência dos indivíduos da 3ª a Classe, a Chance de sobrevivência para os indivíduos da 2ª Classe foi 3,36 [2,21; 5,75] vezes maior e para os indivíduos da 1ª Classe a chance de sobrevivência foi 13,21 [7,61; 22,93] vezes maior. Sobre a idade, a cada ano acrescido na idade do indivíduo a chance de sobrevivência foi 0,96 [0,95; 0,98] vezes menor.

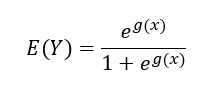


***Figura 2: Ajuste do Modelo de Regressão Logística***

A equação do modelo de Regressão Logística é dada por:



E a probabilidade de sobrevivência de um indivíduo, E(Y) é representada por:



1. **Desenvolvimento**

Dessa forma, conclui-se que uma mulher de 42 anos que estava na 1ª classe tem probabilidade de sobrevivência de 0,90, enquanto que uma mulher de mesma idade na 3ª classe tem probabilidade de sobrevivência de 0,41. No caso dos homens de 30 anos de idade da 1ª classe tem probabilidade de sobrevivência de 0,42, enquanto que um homem de mesma idade na 2ª classe tem probabilidade de sobrevivência de 0,16.

No filme Rose DeWitt Bukater, era uma garota de 17 anos que estava na primeira classe e Jack Dawson, interpretado por Leonardo DiCaprio, tinha 20 anos e estava na terceira classe. Então, segundo o modelo logístico a probabilidade de sobrevivência de Rose é de aproximadamente 0,96, enquanto que a probabilidade de sobrevivência de Jack é de aproximadamente 0,11.

**Referências**

SMOLSKI, Felipe Micail da Silva. Capítulo 7 Regressão Logística. **Smolski**, 2017. Disponível em: <https://smolski.github.io/livroavancado/reglog.html>. Acesso em: 10 de junho de 2022.

ALICE, Michy. How to Perform a Logistic Regression in R. **Datascienceplus**, 2015. Disponível em: <https://datascienceplus.com/perform-logistic-regression-in-r/>. Acesso em: 10 de junho de 2022.

RICKERT, Joseph. Some R Packages for ROC Curves. **R Views**, 2019. Disponível em: <https://rviews.rstudio.com/2019/03/01/some-r-packages-for-roc-curves/>. Acesso em: 10 de junho de 2022.

MARIA, Júlia. Análise descritiva - Banco titanic. **R Pubs**, 2019. Disponível em: <https://rpubs.com/JulhinhaM/ProjTitanic>. Acesso em: 13 de junho de 2022.