# Regressão binária: classificação à moda estatística

## Motivação

Como vimos, os modelos lineares generalizados (GLMs) são ferramentas que permitem estender os modelos de regressão para dados em vários domínios, como dados estritamente positivos, proporções e dados composicionais e, como vamos ver aqui, dados binários, onde $Y_i \in \{0, 1\}$.

Vimos na lição anterior como ajustar esses modelos e agora vamos aprender a interpretar e avaliar os resultados obtidos para um problema de classificação.

## Métricas de avaliação

Para avaliar a qualidade de um modelo de classificação, é comum utilizar métricas como a matriz de confusão, acurácia, a sensibilidade, a especificidade, a precisão, o recall e a F1-score. Vamos definir cada uma delas:

- **Matriz de confusão**: é uma tabela que mostra a frequência de classificações corretas e incorretas feitas pelo modelo. A matriz de confusão para um problema de classificação binária é dada por:

| | **Predito 0** | **Predito 1** |
|---|---|---|
| **Real 0** | Verdadeiro Negativo (VN) | Falso Positivo (FP) |
| **Real 1** | Falso Negativo (FN) | Verdadeiro Positivo (VP) |

- **Acurácia**: é a proporção de classificações corretas feitas pelo modelo. É dada por $\frac{VP+VN}{VP+VN+FP+FN}$.
- **Sensibilidade (Recall)**: é a proporção de verdadeiros positivos em relação ao total de positivos reais. É dada por $\frac{VP}{VP+FN}$.
- **Especificidade**: é a proporção de verdadeiros negativos em relação ao total de negativos reais. É dada por $\frac{VN}{VN+FP}$.

- **Precisão**: é a proporção de verdadeiros positivos em relação ao total de positivos preditos. É dada por $\frac{VP}{VP+FP}$.
- **Recall**: é a proporção de verdadeiros positivos em relação ao total de positivos reais. É dada por $\frac{VP}{VP+FN}$.
- **F1-score**: é a média harmônica entre precisão e recall. É dada por $2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$.

Outra métrica comum é a curva ROC (Receiver Operating Characteristic), que é um gráfico da sensibilidade em função da especificidade para diferentes pontos de corte. A área sob a curva ROC (AUC) é uma medida da qualidade do modelo, onde um valor de 1 indica um modelo perfeito e um valor de 0.5 indica um modelo aleatório.

## Interpretação dos coeficientes

### Regra da divisão por 4

A curva logística é mais íngrime no ponto médio entre 0 e 1, ou seja, quando $\boldsymbol{x}_i^T\boldsymbol{\beta} = 0$. Assim, a inclinação é maximizada neste ponto e atinge o valor de $\beta_j/4$. Logo, a esse valor corresponde a mudança máxima em $\Pr(Y_i = 1)$ para uma mudança unitária em $x_{ij}$.

### Odds ratio

Outra maneira de interpretar os coeficientes de um modelo logístico é por meio de *odds*. Se $\Pr(Y_i = 1 \mid \boldsymbol{X}) = p_i$, então a *odds* é dada por $\frac{p_i}{1-p_i}$. É possível mostrar que esse valor é dado por $\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Além disso, o *odds ratio* é a razão entre duas *odds*. Logo, se quisermos saber o efeito de mudar em uma unidade a variável $x_{ij}$ na *odds* de $Y_i = 1$, basta calcular

$$\Pr(Y_i = 1 \mid X_{ij} = x_{ij} + 1) / \Pr(Y_i = 1 \mid X_{ij} = x_{ij}) = \exp(\beta_j).$$

## Em busca da pamonha perfeita

Palmirinha[1] quer estudar os fatores que fazem uma batelada de pamonha ser classificada como boa ou ruim.

Ao longo de sua longa carreira Palmirinha – sendo extremamente meticulosa – anotou os resultados de milhares de experimentos de degustação, disponíveis em `qualidade_pamonha.csv`. Nestes experimentos, temos informações sobre o `ano` em que o experimento foi feito, a

[1] Palmira Nery da Silva Onofre (Bauru, 29 de junho de 1931 – São Paulo, 7 de maio de 2023) foi uma grande apresentadora de programas culinários. No *StatVerso* da EMAp, é também uma grande estatística *old school*, com treinamento clássico **e** bayesiano!

`temperatura` (em graus Celsius) em que a pamonha foi servida, o potencial de hidrogênio (`pH`) da pamonha, o tipo de `milho` que foi utilizado para a pamonha, o teor de `sacarose` na pamonha (em `%`) e, finalmente, se foi considerada boa (`1`) ou ruim (`0`).

Com a ajuda de seu fiel escudeiro, Guinho, ela pretende utilizar esses dados para entender quais os fatores fazem com que a pamonha seja classificada como boa, de modo a criar a pamonha perfeita.

Ajude Palmirinha e Guinho nesta tarefa. Lembre-se de empregar ferramentas de vizualização de dados e de avaliação de modelos que façam sentido para o tipo de dado disponível.

## Análise dirigida

1. Vizualize a relação entre cada covariável e a variável-resposta. Se preciso, discretize as covariáveis contínuas para obter sumários mais suaves.
2. Vizualize a relação entre as covariáveis.
3. Ajuste um modelo de regressão logística (com intercepto) para cada covariável e compare os resultados. Se estiver usando o R, explore a função `confint()`.
4. Agora desenvolva modelos mais complexos: que covariáveis você incluiria em um modelo conjunto? Porquê?
5. Considere a necessidade de interações.
6. Analise o poder preditivo de cada modelo proposto – considere validação cruzada e indicadores como AIC.

## Exercícios de fixação

1. Suponha que $Z_i \sim \text{Poisson}(\mu_i)$ para $i = 1, 2, \ldots, n$ são amostras independentes. Suponha ainda que $E[Z_i] = \mu_i = \exp(\boldsymbol{X}\boldsymbol{\beta})$. Defina $Y_i = \mathbb{I}(Z_i > 0)$ e defina $\theta_i = \Pr(Y_i = 1)$. Mostre que:

    a. $Y_i \sim \text{Binomial}(\theta_i)$;
    b. $\log\left(-\log(1-\theta_i)\right) = \boldsymbol{X}\boldsymbol{\beta}$. Esta função de ligação chama-se *complementary log-log (cloglog)*, *Gompertz* ou ainda valor extremo (*extreme value*).

2. Ajuste seu modelo preferido aos dados acima usando a função de ligação `cloglog`. Discuta os resultados.

### Referências


- Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and other stories. Cambridge University Press.