

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 $Y_i \in \{0, 1\}$.

Em busca da pamonha perfeita

Palmirinha¹ 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 `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 visualização de dados e de avaliação de modelos que façam sentido para o tipo de dado disponível.

¹Palmira 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!

Exercícios de fixação

1. Suponha que $Z_i \sim \text{Poisson}(\mu_i)$ para $i = 1, 2, \dots, n$ são amostras independentes. Suponha ainda que $E[Z_i] = \mu_i = \exp(\mathbf{X}\boldsymbol{\beta})$. 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. $\log(-\log(1 - \theta_i)) = \mathbf{X}\boldsymbol{\beta}$. 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.