Modelos de Misturas de Regressão:
Conceitos e Aplicação

Juan Sotomayor, Larissa Castilho e Sofia Marinho

Introdução

O objetivo deste projeto é oferecer uma introdução aos modelos de mistura de regressão, explicando seus conceitos principais e ilustrando algumas de suas aplicações mais relevantes. Em seguida, apresentamos a aplicação desses modelos ao banco de dados selecionado, analisando os resultados obtidos e discutindo suas implicações.

Metodologia

DEFINIÇÃO

Mistura de regressão é uma abordagem flexível à modelagem estatística, e caracteriza a heterogeneidade observada em dados ao assumir que a população é composta por várias subpopulações (ou componentes) que se relacionam de maneira diferente com cada variável. Nesses modelos, a variável resposta é modelada como uma mistura de diferentes distribuições, cuja densidade é dada por:

$$f(y) = \sum_{i=1}^{g} \pi_i f_i(y_j), \qquad \begin{cases} \sum_{i=1}^{g} \pi_i = 1 \\ i = (1, 2, \ldots, g) \end{cases}, \quad 0 < \pi_i < 1$$

Assim, seu modelo de regressão é dado por:

$$y_i = X^T \beta_j + \varepsilon_{ij}\,, \qquad \begin{cases} \varepsilon_{ij} \sim N\left(0, \sigma_j^2\right) \\ \beta_j = \left(\beta_{1j,} \beta_{2j}, \ldots, \beta_{pj}\right)^T \end{cases}$$

Isso permite que o modelo identifique diferentes padrões de comportamento entre subgrupos, mesmo quando esses subgrupos são desconhecidos. O modelo é comumente ajustado usando o algoritmo de Expectation-Maximization (EM) para estimar os parâmetros das distribuições componentes e suas proporções.

Além disso, os modelos de mistura de regressão são amplamente aplicados em diversas áreas, como biologia, marketing e análise de imagem, onde a flexibilidade dos modelos é essencial para capturar nuances nos dados complexos.

A observação de dados assimétricos durante a análise de um gráfico de dispersão é um forte indicativo de que a melhor abordagem para modelá-los envolve uma combinação de várias distribuições.

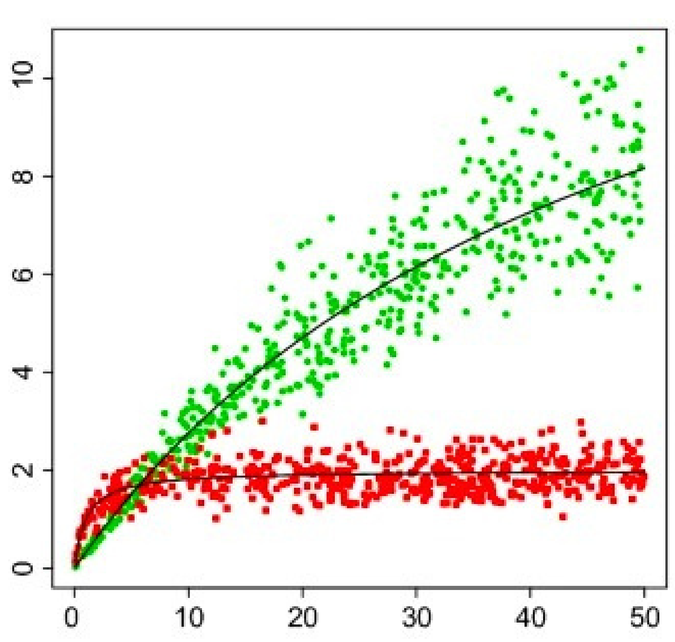

Gráfico 1

O Gráfico 1 apresenta um exemplo de modelo de mistura de regressão, onde os subgrupos são tais que:

$$Y_k \sim Gama\left(\mu_k, \theta_k\right), k = [1, 2]$$

$$\begin{cases} \mu_1 = \dfrac{x}{3 + \frac{x}{16}} \\ \mu_2 = \dfrac{x}{0,5 + \frac{x}{2}} \\ x \sim U\left(0, 50\right) \end{cases} \qquad \begin{cases} \theta_1 = \dfrac{1}{50} \\ \theta_2 = \dfrac{1}{30} \end{cases}$$

Para aplicar estes conceitos às nossas análises, utilizamos o pacote flexmix, do software R.

Aplicação ao Banco de Dados

Descrição

O banco de dados foi desenvolvido a partir do conceito de “Programas de Estudo” e desempenho de provas. É um banco artificial de autoria do grupo, de 100 observações e 2 variáveis (Horas de Estudo e Nota de Prova) em sua versão final.

Motivação

O banco foi feito para poder apresentar a performance da Mistura de Regressão em um dos principais problemas ao lidar com regressão em um conjunto de dados com subgrupos latente (não observável), que é o Paradoxo de Simpson.

Metodologia

Para, de fato, ajustar os dados a partir de um modelo de de mistura de regressão, foram realizados dois passos: o primeiro, diz respeito à *clusterização*, a seleção do número de componentes aos quais pertence a população analisada; e o segundo, analisa os resultados obtidos no passo anterior, e utiliza este para obter os coeficientes de cada componente, para que assim seja possível realizar a modelagem.

A seleção de componentes foi realizada utilizando o método stepFlexmix, que ajusta modelos de mistura para diferentes números de componentes (neste caso, foi testado de 1 a 5 componentes) e seleciona o melhor com base em critérios como BIC e ICL, garantindo robustez por meio de múltiplas inicializações (10 repetições).

A partir deste passo, uma vez selecionado o melhor modelo (com 2 componentes), foi utilizado os coeficientes presentes na Figura 1 para realizar a representação gráfica.

```
> summary(refit(best_model))
$Comp.1
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  50.7315     1.9057 26.6217 < 2.2e-16 ***
HorasEstudo   1.8855     0.3054  6.1738 6.665e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$Comp.2
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 89.58503    1.91882 46.6877 < 2.2e-16 ***
HorasEstudo -2.94476    0.34838 -8.4527 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 1

Resultados

Os resultados do modelo de mistura de regressão linear revelaram dois grupos distintos com comportamentos diferentes em relação à relação entre a nota dos alunos e as horas de estudo, como é possível perceber no Gráfico 2.

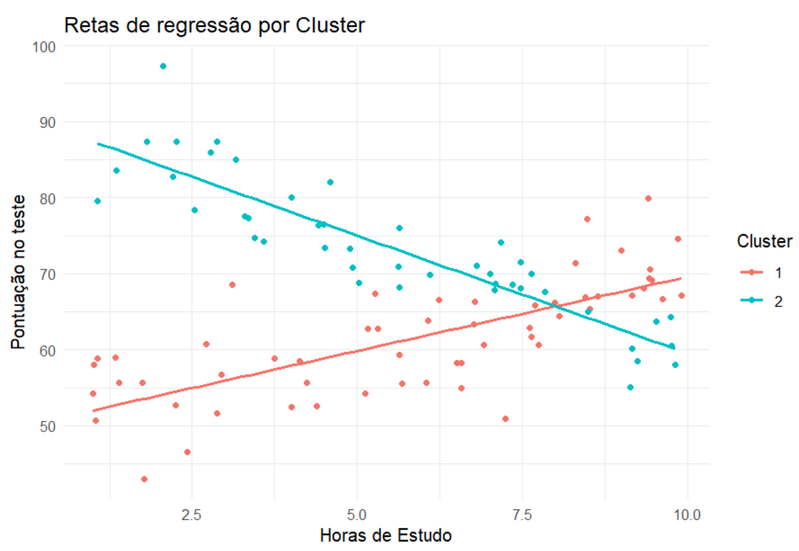

Gráfico 2

Grupo 1

A estimativa do coeficiente de StudyHours foi de 1,8856, o que indica que, para cada hora adicional de estudo, a nota no teste aumenta, em média, em 1,89 pontos. Essa inclinação positiva sugere que, para este grupo, um maior número de horas dedicadas ao estudo está associado a um melhor desempenho na prova.

Grupo 2

A estimativa do coeficiente de StudyHours foi de -2,94476, indicando que, para cada hora adicional de estudo, a nota no teste diminui, em média, 2,94 pontos. Em contrapartida ao grupo 1, onde a relação é positiva, no grupo 2 essa relação é negativa, sugerindo que, nesse caso, mais horas de estudo estão associadas a um desempenho inferior nas provas.

Além disso, a proporção dos grupos foi de 55,6% para o Grupo 1 e 44,4% para o Grupo 2, evidenciando uma leve predominância do Grupo 1.

Essas descobertas destacam a heterogeneidade do conjunto de dados, confirmando que o modelo de mistura é mais adequado do que a regressão linear simples para capturar esses padrões complexos.

Ao ajustar o mesmo conjunto de dados com uma regressão linear simples, apresentada no Gráfico 3, observa-se que o modelo produz um único coeficiente médio para toda a amostra, desconsiderando as diferenças entre os grupos. Isso pode levar a uma interpretação equivocada, já que o comportamento geral do conjunto de dados seria uma média dos padrões distintos. Como resultado, variações importantes nos subgrupos são mascaradas, comprometendo a precisão e a utilidade das inferências realizadas.

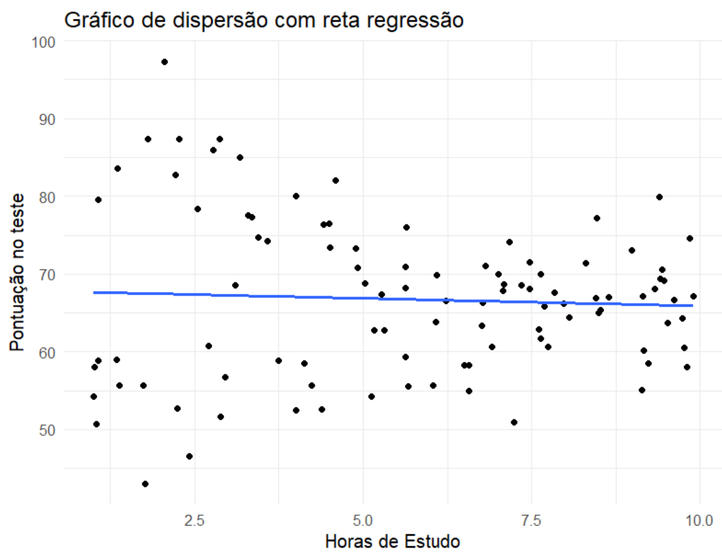

Gráfico 3

Conclusão

Este estudo explorou os Modelos de Mistura de Regressão, destacando sua capacidade de identificar subgrupos em populações heterogêneas e modelar as relações entre variáveis em cada segmento. A aplicação prática demonstrou que essa abordagem, composta por dois subgrupos, é superior à regressão linear tradicional para os dados analisados, especialmente em cenários que exigem segmentação para compreender padrões complexos. Assim, os Modelos de Mistura de Regressão se mostram uma ferramenta poderosa para análises modernas, permitindo interpretar e aproveitar melhor a diversidade de informações em populações heterogêneas.

Bibliografia

Leisch F (2004b). “FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R.”

Benaglia, T et al: "Mixtools: Tools for Analyzing Finite Mixture Models

McLachlan G e Peel D: “Mixtures: Estimation ans Applications”

Greco, L. Robust fitting of mixtures of GLMs by weighted likelihood. AStA Adv Stat Anal 106, 25–48 (2022).