22 de julho de 2025

TÉCNICA DE REGRESSÃO DE MODELO MISTO

ME613 - ANÁLISE DE REGRESSÃO

Juan Sotomayor
Larissa Castilho
Sofia Marinho

Tabela de Conteúdos

1 Introdução

A personalização das plataformas de redes sociais, a oferta de crédito bancário e as propagandas online compartilham um elemento essencial: todas essas aplicações utilizam a construção de perfis de usuários para oferecer serviços de forma direcionada e eficiente. Esse processo baseia-se na identificação de padrões nos comportamentos, preferências ou características dos indivíduos, agrupando-os em categorias ou segmentos homogêneos. Esse agrupamento, denominado *clustering*, é uma etapa fundamental em diversas análises de ciência de dados e estatística, permitindo segmentações que servem de base para técnicas mais avançadas, como os modelos de mistura de regressão.

Os modelos de mistura de regressão representam uma ferramenta analítica poderosa, especialmente em contextos onde há heterogeneidade nas populações analisadas. Esses modelos têm aplicações em áreas tão diversas quanto marketing, saúde pública e planejamento urbano, permitindo não apenas a identificação de subgrupos com comportamentos distintos, mas também a análise detalhada das relações entre variáveis dentro de cada segmento.

Este relatório tem como objetivo apresentar uma análise aprofundada desses modelos, abordando seus conceitos fundamentais, ilustrando exemplos práticos de aplicação e explorando uma implementação detalhada em um conjunto de dados específico. Por meio dessa abordagem, espera-se fornecer ao leitor uma compreensão abrangente dessa técnica estatística e de sua relevância em cenários reais de análise e tomada de decisão, destacando sua importância em um mundo onde a personalização e a análise de dados são centrais para o progresso tecnológico.

2 Metodologia

Nesta seção, serão abordados os principais elementos que definem, de fato, modelos de mistura de regressão. Além disso, exemplos de tal modelo serão apresentados, juntamente com sua aplicação no software R.

2.1 Definição

Compreender os conceitos fundamentais que embasam os modelos de mistura de regressão é essencial. Nesta seção, serão apresentadas suas definições e princípios, fornecendo o suporte teórico necessário para sua aplicação.

Os modelos de mistura de regressão constituem uma abordagem altamente flexível para a modelagem de diferentes tipos de dados, especialmente em situações em que a amostra apresenta heterogeneidade não observada ou problemas relacionados à superdispersão. Tais modelos são definidos como aqueles que assumem a forma:

$$q(y; x, \tau) = \sum_{k=1}^{K} \pi_k m(y; \mu_k, \theta_k), \tag{1}$$

onde:

- K denota o número de componentes,
- $m(y; \mu_k, \theta_k)$ é uma função densidade de probabilidade membro de uma família de distribuições exponenciais,

- μ_k são esperanças de componentes, tais que
  - $g(\mu_k) = \eta_k = x\beta_k$,
  - para alguma função $g(.)$, $\beta_k \in R^p$,
  - p>1 são coeficientes de regressão,
  - x é um vetor p-dimensional de covariâncias
- π_k são probabilidades de pertencimento ao subconjunto k, tal que $\sum_{k=1}^{K} \pi_k = 1$,
- θ_k são parâmetros de dispersão, e
- $\tau = (\pi_1, ...\pi_K, \beta_1, ..., \beta_K, \theta_1, ..., \theta_K)^T$ denota o vetor de todos os parâmetros.

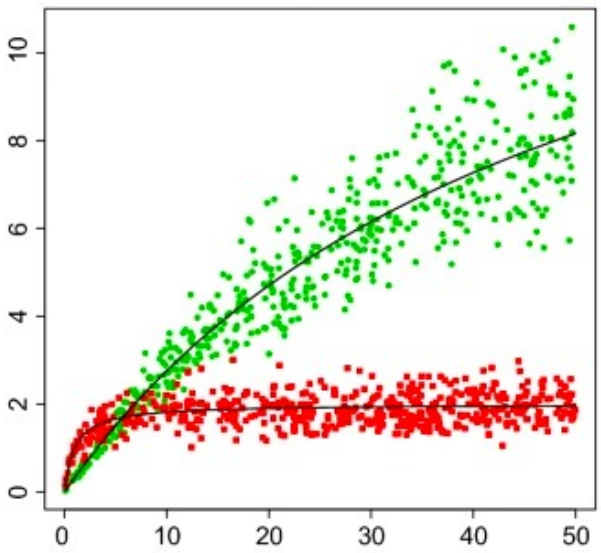
Para estimar tais parâmetros do modelo, é utilizado o algorítmo EM (Expectation - Maximization). Especificamente, o EM é aplicado porque os modelos de mistura envolvem componentes latentes (não observados), como a associação de uma observação a um determinado componente da mistura. O algoritmo resolve isso iterativamente, alternando entre dois passos principais:

- **Etapa E (Expectation)**: Nesta etapa, o algoritmo calcula as probabilidades a posteriori de que cada observação pertença a um componente específico da mistura, com base nos valores atuais dos parâmetros. Essas probabilidades servem como pesos que refletem a incerteza associada à alocação das observações aos componentes.

  Em termos de regressão, o algoritmo avalia a probabilidade de cada observação ter sido gerada por um dos modelos de regressão que compõem a mistura, dado o valor das variáveis independentes e os parâmetros estimados até o momento.
- **Etapa M (Maximization)**: Nesta etapa, o algoritmo ajusta os parâmetros do modelo (como os coeficientes de regressão, os parâmetros de dispersão e as proporções dos componentes) para maximizar a função de verossimilhança ponderada, onde os pesos são as probabilidades calculadas na etapa E.

  Para cada componente da mistura, são estimados separadamente os coeficientes de regressão, com os pesos refletindo o quanto cada observação contribui para aquele componente.

Um fator contribuinte para a suspeita de que podem existir subgrupos dentro dos dados sendo analisados é a assimetria dos gráficos de dispersão.



(a) Figura 1: Exemplo de Modelo de Mistura de Regressão, k = 2

O exemplo apresentado na Figura 1 representa a modelagem de dados com 2 subpopulações, tais que:

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

$$\mu_1 = \frac{x}{3 + \frac{x}{16}}; \mu_2 = \frac{x}{0.5 + \frac{x}{2}}; x \sim U(0, 50); (\theta_0, \theta_1) = (\frac{1}{50}, \frac{1}{30}) \tag{2}$$

2.2 Banco de dados

2.2.1 Preparação dos Dados

Os dados do banco utilizado foram simulados utilizando funções do R. Foram criadas as seguintes variáveis:

- **HorasEstudo**: Gerada como um valor numérico e contínuo, aleatório entre 1 e 10 horas;
- **NotaProva**: Definida com diferentes relações lineares para cada programa, incluindo um ruído gaussiano, criando uma versão do Paradoxo de Simpson;
- **Programa**: Gerada como uma variável categórica de dois níveis, "Programa A"e "Programa B". No entanto, para o propósito desta análise, a variável Programa foi removida e não foi considerada diretamente nos modelos.

Segue abaixo o código comentado que gerou o conjunto simulado:

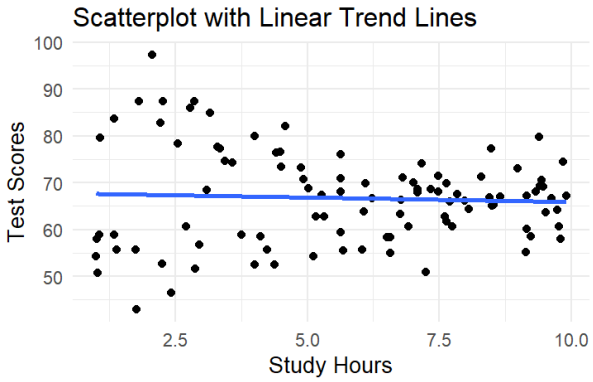
```r
library(dplyr)

set.seed(42)
n <- 100
data <- data.frame(
  HorasEstudo = runif(n, 1, 10),
  Programa = sample(c("Programa A", "Programa B"), n, replace = TRUE)
)

# Gerando NotaProva com Paradoxo de Simpson
data <- data %>%
  mutate(
    NotaProva = ifelse(
      Programa == "Programa A",
      90 - 3 * HorasEstudo + rnorm(n, 0, 5), #Relacao neg no 'Programa A'
      50 + 2 * HorasEstudo + rnorm(n, 0, 5)  #Relacao pos no 'Programa B'
    )
  )
```

2.3 Aplicação ao Banco de Dados

2.3.1 Modelagem Inicial

A primeira abordagem consistiu na aplicação de um modelo de regressão linear simples, que foi ajustado considerando apenas a variável `HorasEstudo` como preditora para `NotaProva`. O modelo foi ajustado da seguinte forma:

$$NotaProva = \beta_0 + \beta_1 \cdot HorasEstudo + \epsilon$$



(a) Figura 2: Gráfico de Regressão Linear Simples

Esse modelo foi ajustado utilizando a função `lm()` do R. A partir dos resultados obtidos, foi possível verificar a significância do coeficiente de HorasEstudo e a qualidade do ajuste do modelo, por meio de métricas como R^2 e teste F.

Porém, somente pela análise gráfica da Figura 2, é possível notar um formato de sino ao redor da reta de regressão, denotando uma forte assimetria dos dados modelados, o que é indício de que este pode ser um caso em que a modelagem por meio de mistura de regressão pode ser o caminho ideal

2.3.2 Modelo Mistura de Regressão

A segunda abordagem foi conduzida utilizando o pacote `flexmix` para identificar agrupamentos latentes.

Usando o gráfico de dispersão para direcionar a suspeita de haver dois ou mais subgrupos latentes, foram ajustados modelos com 1 a 5 componentes utilizando a função `stepFlexmix`, repetindo a inicialização 10 vezes para garantir robustez.

```
library(flexmix)

#Especificar numero de componentes (ex: k=2)
fittedModel_1_5_c <-  stepFlexmix( NotaProva~HorasEstudo, k = c(1,2,3,4,5),
                                   nrep = 10, data = data)
fittedModel_1_5_c
plot(fittedModel_1_5_c)
#Melhor modelo: 2 grupos, pelo BIC e ICL
best_model <- getModel(fittedModel_1_5_c, which=2)
summary(best_model)
summary(refit(best_model))
parameters(best_model)

plot(best_model)
```

O número ideal de componentes foi escolhido com base no Critério de Informação Bayesiano (BIC) e no Critério de Informação Integrada Completa (ICL). A partir destes critérios, foi selecionado o grupo com 2 componentes.

Também foi separado as atribuições de cada *cluster* às observações, para as análises que serão apresentadas posteriormente. Para tal foi utilizado as funções `cluster()`, que informa qual o grupo designado para as observações, e o `posterior()`, que fornece a probabilidade daquela

observação pertencer ao cluster 1 ou 2.

```
#Qual obs vai em qual grupo
posterior(best_model)
clusters(best_model)
#Adicionar os clusters designados ao df
data_cluster <- data
data_cluster$cluster <- factor(clusters(best_model))
```

3 Análise e Resultados

3.1 Análise Descritiva

Aqui, apresenta-se uma breve análise descritiva do conjunto de dados

Tabela 1: Estatísticas descritivas das Horas de Estudo

Média	Mediana	Mínima	Máxima
5.72	5.86	1.002	9.90

A Tabela 1 apresenta as estatísticas descritivas das horas de estudo. A média das horas de estudo é de 5.72, com uma mediana de 5.86, indicando uma leve inclinação negativa. A maior quantidade de horas dedicadas pelos alunos é de 9.90, enquanto a menor dedicação é de 1.002.

Tabela 2: Estatísticas descritivas das Horas de Estudo

Média das Notas	Mediana das Notas	Nota Mínima	Nota Máxima
66.74	66.97	42.98	97.34

Na Tabela 2 temos as estatísticas descritivas das notas de prova. A média de notas é de 66.74, com uma mediana de 66.97, indicando uma sutil inclinação negativa. A maior nota observada é de 97.34, enquanto a menor é de 42.98.

3.2 Modelo de regressão simples

Na abordagem inicial, foi aplicado um modelo de regressão simples. O modelo ficou definido na seguinte equação:

$$NotaProva = 67.8595 - 0.1960 \cdot HorasEstudo$$

Ademais, o sumário avaliando a qualidade do modelo ajustado se apresenta abaixo:

$$R^2 = 0.002786 \quad \text{Estatística F} = 0.2738$$

Tem-se aqui o coeficiente de determinação R^2 de 0,002%, o que já é suficiente para comprovar que o modelo consegue explicar quase nenhuma variabilidade do conjunto de dados nessa

situação. Um dos fatores principais para um modelo com poder de predição tão baixo se dá pela Figura 2, o gráfico de dispersão na próxima seção. Nele, é possível observar a ausência de variância constante, um dos pressupostos necessários para o desempenho do modelo de regressão linear.

3.3 Resultados

Nesta seção, é realizada uma análise detalhada dos resultados obtidos pela aplicação de um modelo de mistura de regressão aos dados em estudo. O objetivo é examinar como diferentes subpopulações presentes nos dados contribuem para a relação geral entre as variáveis dependente e independentes.

Com base na segmentação fornecida pelo modelo de mistura, procura-se identificar padrões distintos de comportamento que podem não ser evidenciados por meio de uma abordagem tradicional de regressão linear. Essa análise desempenha um papel fundamental na validação das hipóteses de heterogeneidade dos dados, bem como na avaliação da robustez e utilidade do modelo ajustado. São discutidos os parâmetros estimados para cada componente do modelo, as probabilidades associadas a essas componentes e o desempenho geral do modelo, considerando critérios de informação e análise de resíduos.

As tabelas a seguir fornecem os detalhes dos coeficientes de regressão ajustados para cada componente identificado pelo modelo de mistura.

Componente 1					
	Estimativa	Erro Padrão	Estatística z	Pr(>|z|)	
(Intercepto)	50.7311	1.9056	26.622	< 2.2e-16	***
HorasEstudo	1.8856	0.3054	6.174	6.657e-10	***
Componente 2					
	Estimativa	Erro Padrão	Estatística z	Pr(>|z|)	
(Intercepto)	89.58505	1.91881	46.6878	< 2.2e-16	***
HorasEstudo	-2.94476	0.34838	-8.4528	< 2.2e-16	***

Tabela 3

Primeiramente, aqui estão definidos cada item da Tabela 1:

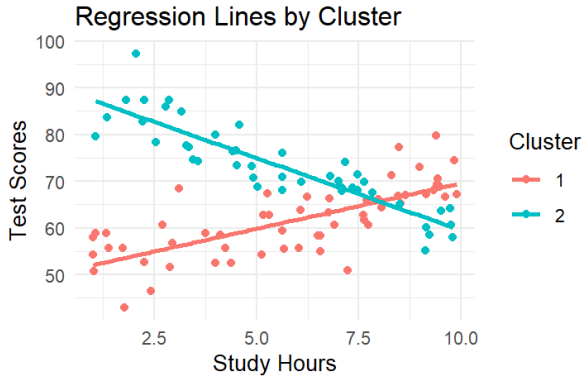
- **Erro Padrão**: Um Erro Padrão menor indica que o coeficiente foi estimado com maior precisão.
- **Estatística z**: Um valor da estatística z alto (em magnitude) indica que o coeficiente estimado está muito longe de zero, ou seja, é estatisticamente significativo (Estatística z = Estimativa do Coeficiente / Erro Padrão).
- **Pr(>|z|)** Para um nível de significância $\alpha = 0.05$, |z|< 1.96 indica que o coeficiente é significante.
- **Significância**: Os asteriscos apresentados ao final da tabela são uma forma mais direta de se entender o nível de significância da estatística analisada, de forma que:
  - '***' < 0.001; '**' < 0.01 ; '*' < 0.05 ; '.' < 0.1

Assim, agora é possível realizar uma análise de cada componente, como está apresentado a seguir:

Para o **Componente 1**, valor do intercepto (50,7311) representa a predição média de NotaProva quando o número de horas de estudo (HorasEstudo) é igual a zero. Isso indica que os alunos deste grupo que não dedicam tempo ao estudo têm, em média, uma nota inicial de 50,73. Já o coeficiente associado a HorasEstudo (1,8856) sugere que, para cada hora adicional de estudo, a nota na prova aumenta, em média, em 1,89 pontos. Esse coeficiente positivo reflete a relação direta entre o aumento das horas de estudo e a melhora no desempenho dos alunos nesse grupo. Além disso, os valores de significância (Pr(>|z|)) para ambos os coeficientes são extremamente baixos (< 0,001), evidenciando que os coeficientes são altamente significativos estatisticamente.

Em contrapartida, para o **Componente 2**, o intercepto (89,58505) indica a predição média de NotaProva quando o número de horas de estudo (HorasEstudo) é igual a zero. Para o componente 2, observa-se que os alunos sem horas de estudo possuem uma nota inicial significativamente mais alta, de 89,59, em comparação ao componente 1. Por outro lado, o coeficiente de HorasEstudo (-2,94476) revela que, para cada hora adicional de estudo, a nota na prova diminui, em média, 2,94 pontos. Esse coeficiente negativo sugere que, neste grupo, o aumento nas horas de estudo está associado a uma redução no desempenho, possivelmente refletindo práticas de estudo ineficazes ou uma sobrecarga dos alunos. Além disso, os valores de significância (Pr(>|z|)) permanecem extremamente baixos (< 0,001), confirmando que os coeficientes são altamente significativos do ponto de vista estatístico.

Esses coeficientes resultaram na mistura de regressão ajustada que pode ser observado na Figura 3.



(a) Figura 3: Gráfico de mistura de regressão com k = 2

A partir da análise da Figura 3, é possível perceber que o modelo de mistura de regressão fez um trabalho ótimo na modelagem dos dados, conseguindo capturar o comportamento de cada um dos componentes.

As probabilidades estimadas de cada componente estão apresentadas na tabela 5

Proporção de cada componente

	Anterior	Tamanho	Posterior	Razão
Comp. 1	0.556	56	91	0.615
Comp.2	0.444	44	84	0.524

Tabela 4

A partir da observação da tabela, é possível perceber que, por mais que 91 observações tenham

probabilidade maior que 0 de pertencer ao subgrupo 1, e 84 ao grupo 2, a partir do algorítmo EM do modelo, foi selecionado 56 e 44 dados, respectivamente, a cada componente.

4 Conclusões

Neste estudo, foram investigados os Modelos de Mistura de Regressão, explorando seus conceitos fundamentais, aplicações práticas e um caso específico em um conjunto de dados. A análise destacou a capacidade desses modelos de lidar com a heterogeneidade implícita em populações, permitindo a identificação de subgrupos distintos e a modelagem detalhada das relações entre variáveis em cada segmento.

A aplicação prática resultou em um modelo bem ajustado, composto por dois subgrupos, evidenciando a superioridade da abordagem de misturas em relação à regressão linear tradicional para o conjunto de dados estudado. Essa diferenciação tornou-se clara ao comparar os gráficos gerados, reforçando a relevância dessa técnica para cenários onde a segmentação dos dados é essencial para compreender padrões e comportamentos complexos.

Assim, este trabalho reafirma a utilidade dos Modelos de Mistura de Regressão em análises modernas, onde a personalização e a identificação de padrões são fundamentais. Nos contextos onde são aplicados, estes modelos oferecem uma ferramenta robusta para interpretar e aproveitar a riqueza de informações contida em populações heterogêneas, contribuindo para avanços em diversas áreas.

5 Referências

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).