Mergulho nas Misturas de Regressões: Uma Análise com Pinguins

Karina Kawashita Vicente de Azevedo (236174), Pedro Francisco Godoy Bernadinelli (251570), Rafaela da Silva Barril (204368) e Vitor Ribas Perrone (245040)

Introdução

Um pesquisador, ao estudar algumas espécies de pinguins, apagou acidentalmente os dados que identificavam as espécies em sua planilha, sem possibilidade de recuperação. Para lidar com esse problema, uma abordagem viável é o uso de modelos de mistura de regressão, que permitem identificar subgrupos em uma amostra e ajustar regressões separadas para cada um, sem a necessidade de informações explícitas sobre os grupos.

Sendo assim, caso de fato seja identificado um comportamento distinto para subgrupos de pinguins, ainda é possível prosseguir com seu estudo, a relação entre a massa de um pinguim e a profundidade de seu bico.

Metodologia

Para este trabalho, a fim de trabalhar com diversos grupos em uma mesma modelagem, foi utilizada a metodologia de Mistura de Regressão, com auxílio do algoritmo EM para determinar os grupos. Sendo assim, o método consiste em ajustar as regressões lineares ao mesmo tempo em que são determinados os grupos, para no fim obter em conjunto um ótimo ajuste a partir de um devido agrupamento. Com isso, o modelo é tal que, para m grupos:

$$y_i = \begin{cases} x_i^T\beta_1 + \varepsilon_{i1}, \text{se a } i\text{-ésima amostra pertence ao subgrupo 1,} \\ x_i^T\beta_2 + \varepsilon_{i2}, \text{se a } i\text{-ésima amostra pertence ao subgrupo 2,} \\ \vdots \\ x_i^T\beta_m + \varepsilon_{im}, \text{se a } i\text{-ésima amostra pertence ao subgrupo } m \end{cases}$$

A partir disso, o algoritmo consiste em 5 passos, sendo eles:

1. Distribua de forma aleatória os elementos da amostra entre os subgrupos;

2. Ajuste os modelos para cada um dos subgrupos de acordo com a suposta distribuição;

3. Calcule uma estimativa da log-verossimilhança dos modelos (esta é a etapa ”E”);

4. A etapa de maximização, redistribua as amostras de modo a maximizar a verossimilhança, ou seja, designando os elementos da amostra para os modelos que os descrevem melhor;

5. Retorne ao Passo 1 com a nova estimativa da distribuição dos exemplos e repita o processo.

Após o ajuste do modelo, para realizar inferências, isto é, calcular intervalos de confiança para os coeficientes, o Bootstrap é a alternativa mais utilizada, consistindo em, basicamente, realizar diversas amostras da variável aleatória de interesse. A partir disso, com uma distribuição empírica, o cálculo de um Intervalo de Confiança se torna mais viável, entretanto, erros de convergência do método podem prejudicar a estratégia.

Além disso, como na aplicação do método é preciso estabelecer uma distribuição, que no caso deste trabalho foi Normal com média 0 e variância constante dentro de cada grupo para os erros, é necessário verificar após o ajuste de a suposição foi válida. Para isso, é possível condicionar as retas para cada grupo e realizar análises gráficas.

Para todas as análises, foi utilizada a Linguagem de Programação R, com auxílio do pacote Tidyverse para gráficos e manipulações nos dados e do pacote Mixtools para o modelo.

Regressão Linear Simples

Inicialmente, para verificar o comportamento dos dados, aplicamos um modelo de Regressão Linear Simples para verificar se esse seria capaz de descrever adequadamente o comportamento dos dados.

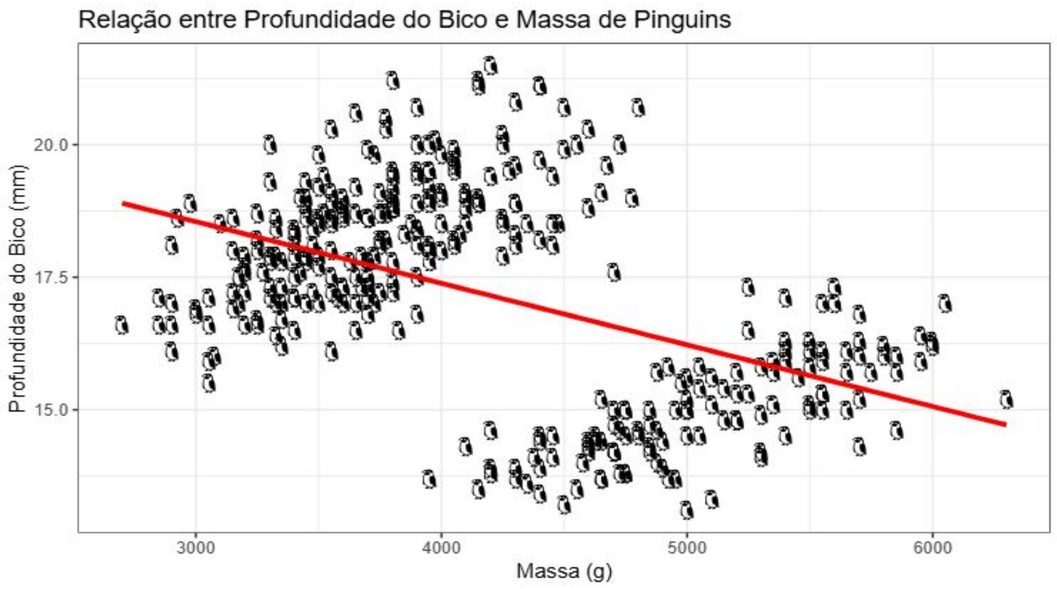

Figura 1: Regressão Linear Simples para Profundidade do Bico e Massa dos Pinguins

Dessa maneira, a partir da Figura 1, ficou visualmente evidente que a Regressão Linear Simples não se trata de um modelo adequado para explicar efetivamente os dados em questão. Entretanto, foi possível observar que os dados indicam tendências comportamentais que podem ser divididas em dois grupos possibilitando, portanto, a aplicação de uma Mistura de Regressão.

Mistura de Regressão

Supondo inicialmente k = 2 (quantidade de grupos) e betas iguais a 22, 0.002, -15 e 0.002, pelo algoritmo EM obtemos os resultados expressos na Figura 2:

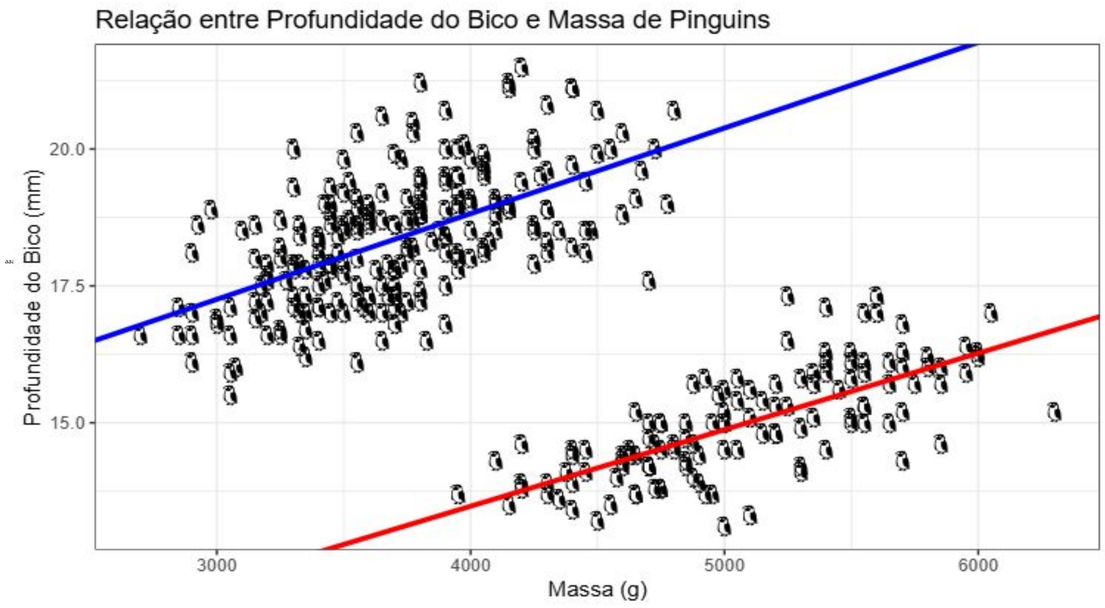

Figura 2: Mistura de Regressão para Profundidade do Bico e Massa dos Pinguins.

Considerando o Grupo 1 como o que está associado a reta azul e o Grupo 2 com a vermelha, temos os respectivos coeficiente indicados na Tabela 1.

Tabela 1: Coeficientes das Retas de Regressão para Cada Grupo

Grupo	β_0	β_1
1	12.562	0.002
2	7.887	0.001

Ou seja, temos que para cada uma unidade de aumento de massa (em gramas) , a profundidade do bico aumenta em média 0.002 mm para o Grupo 1 e 0.001 mm para o Grupo 2.

Análise de Resíduos

Para que o mistura de regressões represente um bom ajuste para os dados é necessário verificar se as suposições sobre os erros são atendidas, para tal realizaremos análises gráficas:

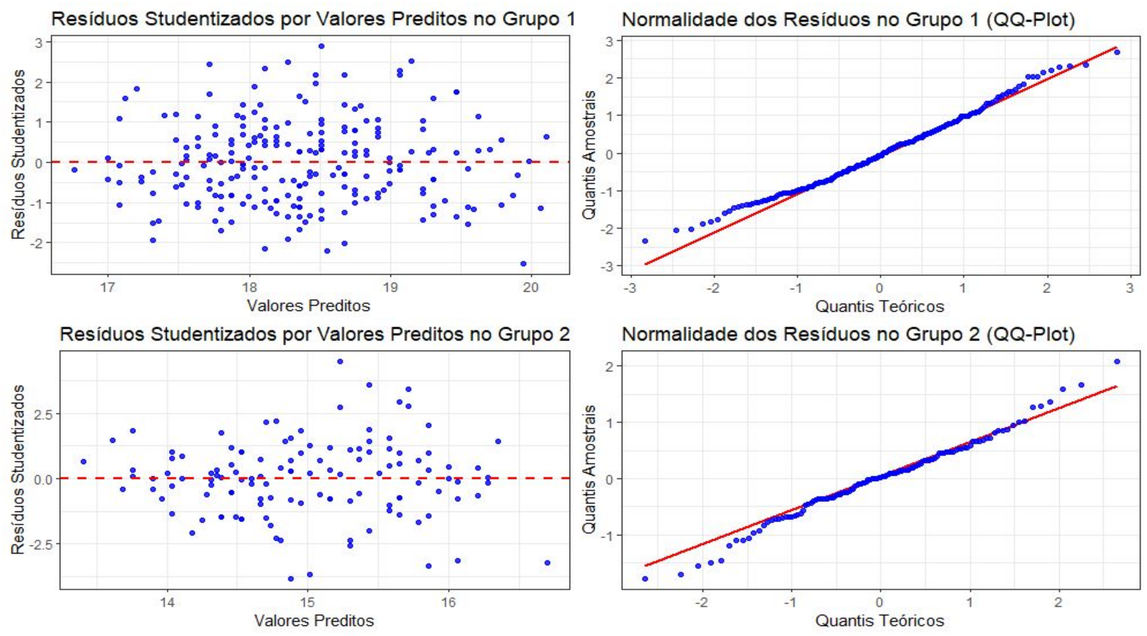

Figura 3: Conjunto de Gráficos para Análise de Resíduos.

Assim, observamos para ambos os grupos, que os resíduos não apresentam tendências específicas indicando, então, independência e homocedasticidade, em acréscimo, vemos que apesar de indícios de cauda pesada, é possível afirmar a normalidade dos dados.

Conclusão

Portanto, foi possível avaliar que de fato existe uma relação crescente entre a massa de um pinguim e a profundidade de seu bico.

Embora a mistura de regressão resultou em um modelo melhor ajustado para os dados, no o caso em questão, observamos que o algoritmo EM dependia fortemente da escolha de bons valores iniciais. Sem esses chutes iniciais adequados, o algoritmo frequentemente não convergia de forma satisfatória, devido à presença de máximos e mínimos locais nos dados. Isso decorre do fato de que o algoritmo EM não garante a convergência global.

Dessa forma, foi necessário recorrer a ferramentas de visualização para verificar a convergência, o que limitou a possibilidade de incluir mais parâmetros no modelo.

Referências Bibliográficas

[1] Shane MUELLER. Mixture Modeling: Mixture of Regressions.

url: https ://pages.mtu.edu/˜shanem/psy5220/daily/Day19/Mixture_of_regressions.html.

[2] Susana FARIA e Gilda SOROMENHO. “Fitting mixtures of linear regressions”. Em: Journal of Statistical Computation and Simulation (2010).

[3] BENAGLIA, T. et al. mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, v. 32, n. 6, 1 jan. 2009.