Regressão Linear Múltipla - Exemplo Íris

Igor Falcão, Beatriz Beckman e Natã Cavalcante

2024-10-03

Relatório: Análise e Regressão Linear do Banco de Dados "Pinguins"

1.Introdução

Este relatório tem como objetivo ajustar um modelo de regressão linear múltipla para investigar a influência de determinadas características morfológicas e demográficas de pinguins sobre a variável de interesse "Profundidade do Bico". O estudo será conduzido com base no conjunto de dados "pinguins", que contém informações detalhadas de três espécies: Pinguim-de-adélia, Pinguim-de-barbicha e Pinguim-gentoo.

O dataset inclui medidas como massa corporal, comprimento da nadadeira e comprimento do bico, além de informações categóricas como o sexo dos pinguins e a ilha onde foram encontrados (Biscoe, Dream, Torgersen). Nosso objetivo é compreender como essas variáveis explicativas influenciam a profundidade do bico, uma característica importante para a alimentação e adaptação das espécies ao ambiente.

Este relatório apresentará a análise detalhada das variáveis, construção do modelo de regressão e interpretação dos resultados, com o intuito de identificar os fatores mais relevantes na determinação da profundidade do bico entre as diferentes espécies de pinguins.

2.Os dados

Para obter o dataset "pinguins (terceiro_estagio)", foi necessário entrar em contato com a professora, que nos enviou o arquivo. Em seguida, ele foi baixado para ser utilizado no R. Para que o dataset estivesse disponível e pudesse ser utilizado nos códigos, foi necessário carregá-lo no ambiente do R.

Observando a estrutura da base de dados, possuímos 344 observações e 8 variáveis:

- Sexo: Sexo do(a) pinguim, macho ou femea;
- Massa Corporal: Massa corporal, em gramas;
- Comprimento Nadadeira: Comprimento Nadadeira, em milímetros;
- Ano: Ano de estudo, (2007, 2008, 2009);
- Profundidade de Bico: Profundidade do bico, em milímetros;
- Comprimento de bico: Comprimento do Bico, em milímetros;
- Ilha: Ilha do arquipélago Palmer, na Antártida (Biscoe, Dream ,Torgersen)
- Espécie: Espécies de pinguim (Pinguim-de-adélia, Pinguim-de-barbicha, Pinguim-gentoo);

2.1 Análise exploratória dos dados

```
library(skimr)

#Colocar o caminho de onde está o arquivo .csv
dados <- read.csv("D:/Codigos/ESTATISCA/terceiro_estagio.csv", header = TRUE, sep=';', dec='

dados_clean <- na.omit(dados[, c("especie", "ilha","comprimento_bico" ,"profundidade_bico", '

skim(dados)
```

Tabela 1: Data summary

Name	dados
Number of rows	344
Number of columns	9
_____	
Column type frequency:	
character	3
numeric	6
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
especie	0	1.00	14	19	0	3	0
ilha	0	1.00	5	9	0	3	0
sexo	11	0.97	5	5	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
X	0	1.00	172.50	99.45	1.0	86.75	172.50	258.25	344.0	
comprimento_bico	2	0.99	43.92	5.46	32.1	39.23	44.45	48.50	59.6	
profundidade_bico	2	0.99	17.15	1.97	13.1	15.60	17.30	18.70	21.5	
comprimento_nadadeira	2	0.99	200.92	14.06	172.0	190.00	197.00	213.00	231.0	
massa_corporal	2	0.99	4201.75	801.95	2700.0	3550.00	4050.00	4750.00	6300.0	
ano	0	1.00	2008.03	0.82	2007.0	2007.00	2008.00	2009.00	2009.0	

A instrução library(skimr) carrega o pacote skimr, que serve para elaborar resumos estatísticos de conjuntos de dados de maneira minuciosa e estruturada. No exemplo apresentado, cria-se primeiramente um objeto denominado dados, que contém o conjunto de dados pinguins(terceiro_estagio). Após isso, utiliza-se o comando skim(dados) para produzir um resumo estatístico desse conjunto, apresentando informações como o número de valores ausentes, média, mediana, desvio padrão, valores mínimo e máximo, entre outros fatores. Essa função torna mais fácil a compreensão das características dos dados de modo claro e intuitivo.

Sobre a saída do codigo, o conjunto de dados possui 344 linhas e 8 colunas, sendo que 3 delas são variáveis categóricas (do tipo character) e as outras 5 são numéricas. Essa configuração nos proporciona uma visão sobre a estrutura do conjunto, que abrange informações tanto qualitativas quanto quantitativas.

Entre as variáveis categóricas, a variável “espécie” contém dados sobre diferentes tipos de pinguins. A variável “ilha” apresenta três valores distintos, o que indica que as informações foram coletadas de três ilhas diferentes. A variável “sexo”, também categórica, possui duas categorias principais, que provavelmente correspondem a “macho” e “fêmea”.

No que diz respeito às variáveis numéricas, as estatísticas descritivas incluem média, desvio padrão e percentis, facilitando uma compreensão mais aprofundada da distribuição dos dados. A média do comprimento do bico é de 43,9 mm, enquanto a profundidade média do bico é de 17,2 mm. A média da massa corporal é de 4202 gramas, sugerindo uma ligação entre essa variável e o peso dos pinguins.

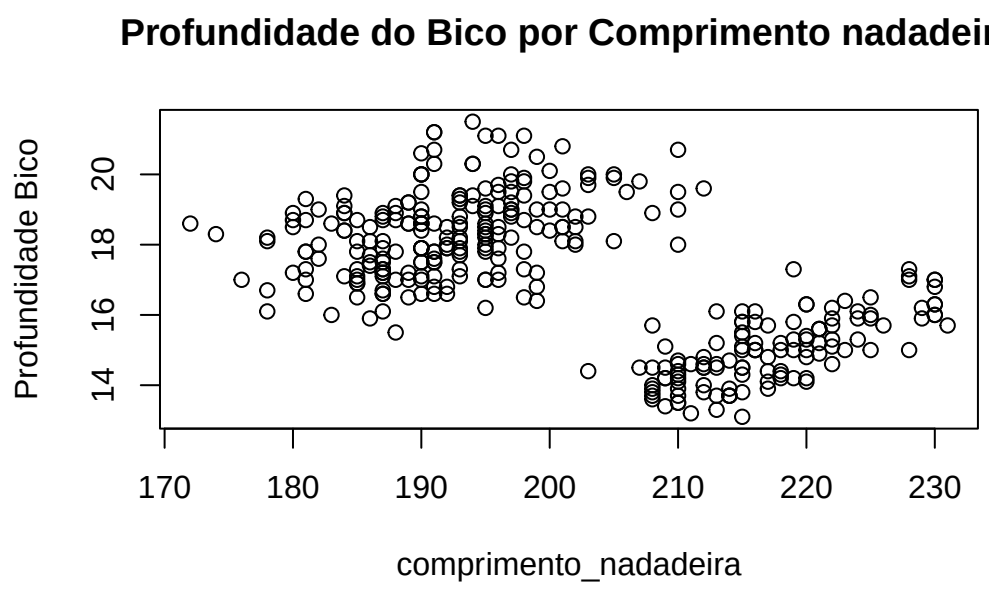
A análise dos percentis, como p25, p50 e p75, indica que a maioria dos valores está concentrada dentro de intervalos esperados, como no caso da **massa_corporal**, cuja maioria dos pinguins

pesa entre 3550 e 4750 gramas. A distribuição das variáveis numéricas parece razoável e consistente, o que sugere que os dados estão bem representados e podem ser utilizados para análises mais aprofundadas.

2.1.1 Análise de Correlação

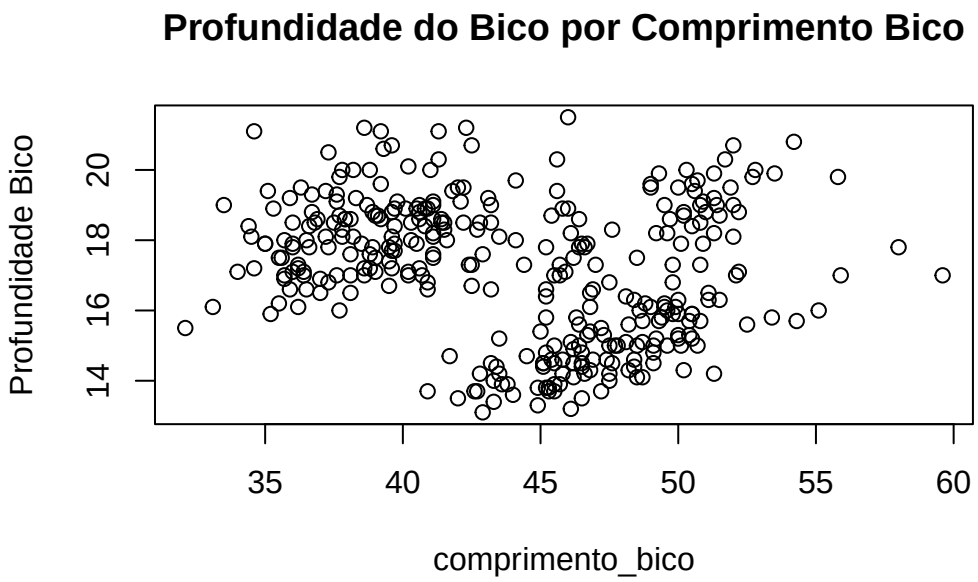
Com o intuito de realizar a análise das variáveis explicativas com a variável resposta**,** é possível observar algumas características por exemplo com a variável explicativa **comprimento_nadadeira.**

```
plot(profundidade_bico ~ comprimento_nadadeira, data = dados_clean,
        main = "Profundidade do Bico por Comprimento nadadeira",
        xlab = "comprimento_nadadeira", ylab = "Profundidade Bico")
```



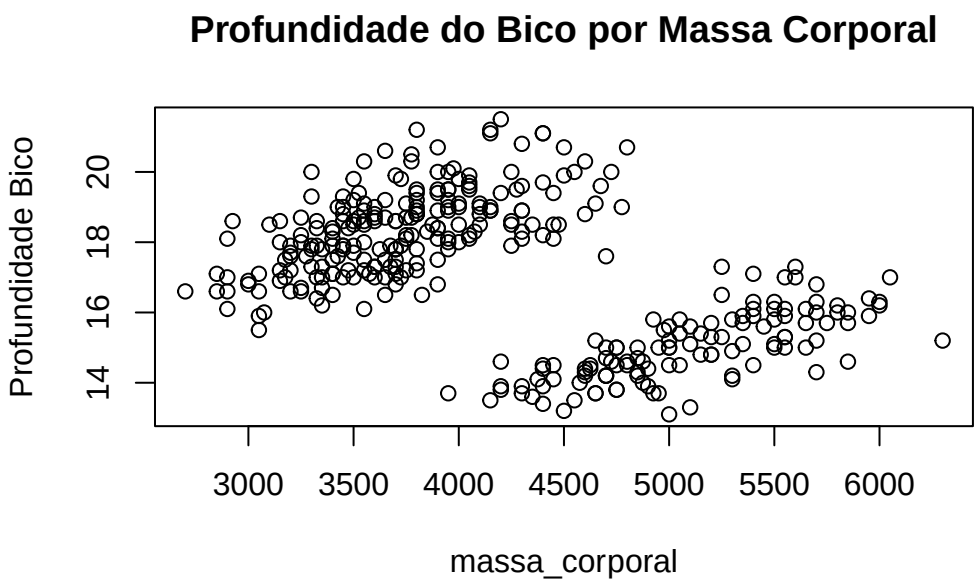
Analisando esse gráfico de dispersão existe a divisão da população de pinguins em dois grupos, um quando o comprimento da nadadeira é menor a profundidade do bico é maior e outro quando o comprimento da nadadeira é maior a profundidade do bico é menor, ou seja, o coeficiente de correlação linear **r** será negativo

```
plot(profundidade_bico ~ comprimento_bico, data = dados_clean,
        main = "Profundidade do Bico por Comprimento Bico",
        xlab = "comprimento_bico", ylab = "Profundidade Bico")
```



O gráfico de dispersão com a variável **comprimento_bico**, não é possível encontrar alguma relação entre as variáveis.

```
plot(profundidade_bico ~ massa_corporal, data = dados_clean,
        main = "Profundidade do Bico por Massa Corporal",
        xlab = "massa_corporal", ylab = "Profundidade Bico")
```



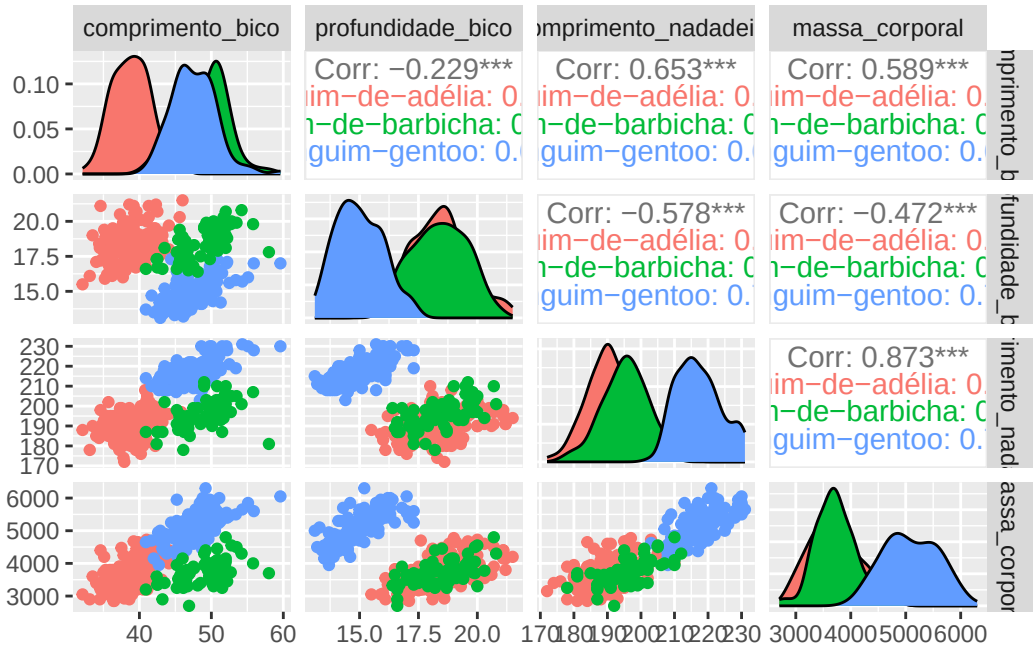
Com relação à **massa_corporal** é possível observar que quanto maior a massa corporal menor a profundidade do bico e quando menor a massa corporal maior será o bico, logo, o coeficiente de correlação linear também será negativo

```
library(GGally)
```

```
Carregando pacotes exigidos: ggplot2
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```
graf1 <- ggpairs(dados_clean, columns = 3:6, ggplot2::aes(colour=especie))
graf1
```



Importante ressaltar que as variáveis preditoras comprimento nadadeira e massa corporal podem ocasionar uma multicolinearidade, pois o coeficiente de correlação linear (Corr) é maior do que 0.8. Dessa forma, esse fato tem que ser corrigido no nosso modelo a ser escolhido.

3. Modelo

Inicialmente temos o primeiro modelo

```
modelo1 <- lm(profundidade_bico ~ . -especie, data = dados_clean)
summary(modelo1)
```

```
Call:
lm(formula = profundidade_bico ~ . - especie, data = dados_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7329 -0.7450 -0.1286  0.6913  4.4479

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)           -1.124e+02  1.550e+02  -0.725  0.46888
ilhaDream              1.032e+00  1.849e-01   5.584 4.97e-08 ***
ilhaTorgersen          1.127e+00  2.150e-01   5.239 2.91e-07 ***
comprimento_bico       9.445e-03  1.680e-02   0.562  0.57446
comprimento_nadadeira -5.480e-02  1.057e-02  -5.184 3.82e-07 ***
massa_corporal        -5.304e-04  1.885e-04  -2.814  0.00518 **
sexomacho              2.194e+00  1.486e-01  14.766  < 2e-16 ***
ano                    7.010e-02  7.738e-02   0.906  0.36567
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.095 on 325 degrees of freedom
Multiple R-squared:  0.697, Adjusted R-squared:  0.6905
F-statistic: 106.8 on 7 and 325 DF,  p-value: < 2.2e-16
```

Devido a existência de uma correlação da variável resposta **profundidade_bico** com as variáveis explicativas **comprimento_nadadeira** e **massa_corporal** como já foi observado anteriormente elas variam em sentidos opostos, no entanto essas variáveis preditoras podem ocasionar a multicolinearidade. Assim, analisando o modelo 1 não existe nenhuma anormalidade com as estimativas, pois a medida que aumenta uma unidade no comprimento da nadadeira, tem uma estimativa de diminuir a profundidade do bico em -5.480e-02 o mesmo acontece com a massa corporal, pois quando aumenta uma unidade na massa corporal tem a estimativa de diminuir a profundidade do bico em -5.304e-04.

No entanto, a variável preditora **comprimento_bico** possui o valor P igual a 0.57446 o que é maior do que 10%, sendo uma boa alternativa retirá-la do modelo 1.

```r
modelo2 <- update(modelo1, ~ . -comprimento_bico)
summary(modelo2)
```

```
Call:
lm(formula = profundidade_bico ~ ilha + comprimento_nadadeira +
    massa_corporal + sexo + ano, data = dados_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7533 -0.7624 -0.1015  0.7184  4.4564

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.051e+02  1.543e+02  -0.681  0.49635
```

```
ilhaDream              1.069e+00  1.728e-01   6.185 1.86e-09 ***
ilhaTorgersen          1.119e+00  2.144e-01   5.219 3.20e-07 ***
comprimento_nadadeira -5.223e-02  9.519e-03  -5.486 8.25e-08 ***
massa_corporal        -5.259e-04  1.881e-04  -2.796  0.00548 **
sexomacho              2.208e+00  1.463e-01  15.090  < 2e-16 ***
ano                    6.638e-02  7.701e-02   0.862  0.38938
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.094 on 326 degrees of freedom
Multiple R-squared:  0.6968,    Adjusted R-squared:  0.6912
F-statistic: 124.8 on 6 and 326 DF,  p-value: < 2.2e-16
```

Com o novo modelo o valor R-squared obteve um aumento de 0.0007, indicando que o modelo 2 é melhor do que o modelo 1, pelo fato do valor R^2 ser maior. No entanto, é possível realizar uma melhoria nesse modelo pelo fato da variável preditora ano possuir um valor P superior a 10%

```r
modelo3 <- update(modelo2, ~ . -ano)
summary(modelo3)
```

```
Call:
lm(formula = profundidade_bico ~ ilha + comprimento_nadadeira +
    massa_corporal + sexo, data = dados_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6861 -0.7580 -0.0936  0.6901  4.5334

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          27.9043913  1.3521204  20.638  < 2e-16 ***
ilhaDream             1.0626246  0.1726214   6.156 2.19e-09 ***
ilhaTorgersen         1.1179289  0.2142790   5.217 3.23e-07 ***
comprimento_nadadeira -0.0499521  0.0091429  -5.464 9.26e-08 ***
massa_corporal        -0.0005633  0.0001829  -3.079  0.00225 **
sexomacho             2.2176123  0.1458956  15.200  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.094 on 327 degrees of freedom
```

```
Multiple R-squared:  0.6961,    Adjusted R-squared:  0.6914
F-statistic: 149.8 on 5 and 327 DF,  p-value: < 2.2e-16
```

Logo, podemos perceber uma melhoria nos valores tanto na estimativa, como também no valor R^2, 0.6914, aumentando 0.0002, o que indica que houve uma melhora no modelo.Para confirmar essa ideia vamos calcular a medida AIC dos modelos.

4. Métodos de seleção de modelos

Tendo em vista que temos três possíveis modelos calculamos o AIC

```
AIC(modelo1)
```

```
[1] 1015.66
```

```
AIC(modelo2)
```

```
[1] 1013.983
```

```
AIC(modelo3)
```

```
[1] 1012.741
```

Dessa forma, o menor valor AIC indica o melhor modelo que no caso seria o modelo 3