Relatório: Análise e Regressão Linear do Banco de Dados “Pinguins”

Igor Falcão, Beatriz Beckman e Natã Cavalcante

2024-10-07

```
# Setup para o relatório Quarto

knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
```

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)

# Caminho relativo para o arquivo .csv
dados <- read.csv("./terceiro_estagio.csv",
                  header = TRUE,
                  sep = ';',
                  dec = ',')

dados_clean <- na.omit(dados[,
                             c("especie", "ilha","comprimento_bico"
                            ,"profundidade_bico", "comprimento_nadadeira",
                            "massa_corporal", "sexo", "ano")])

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.

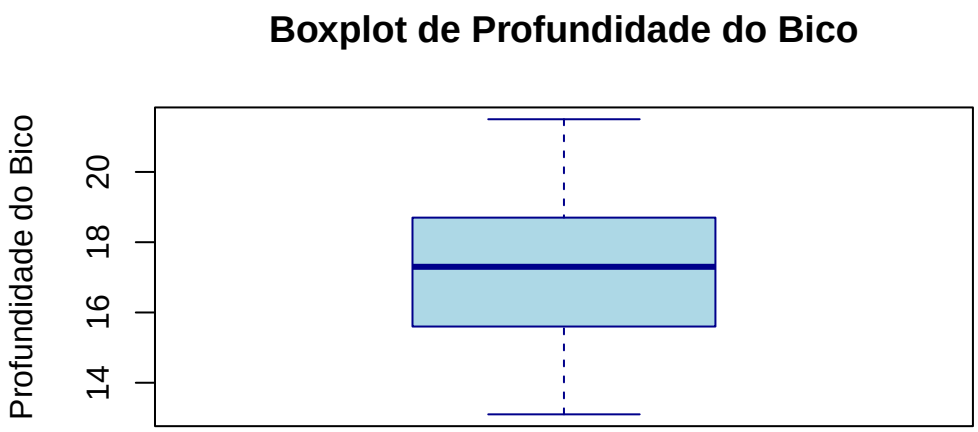
Entre as variáveis categóricas, a variável "espécie" contém dados sobre diferentes tipos de pinguins. A variável "ilha" apresenta três valores distintos, o que indica que as informações foram coletadas de três ilhas diferentes. A variável "sexo", também categórica, possui duas categorias principais,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

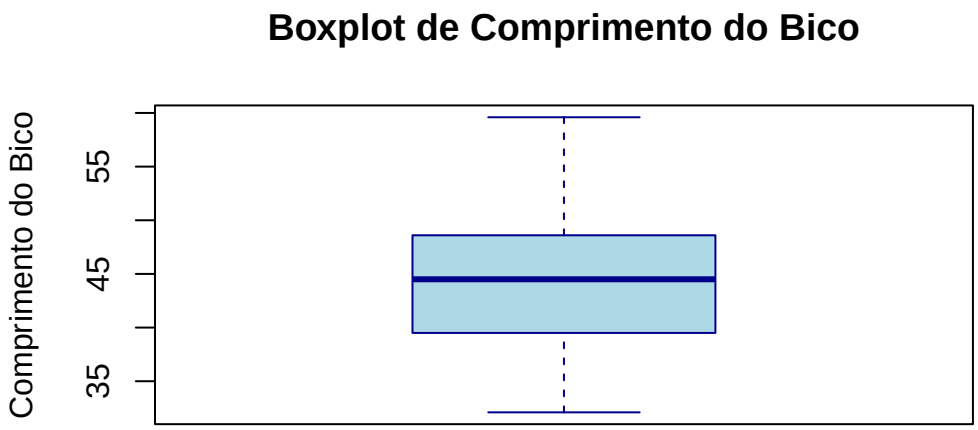
A análise dos percentis, como p25, p50 e p75, indica que a maioria dos valores está concentrada dentro de intervalos esperados, como no caso da **massa_corporal**, cuja maioria dos pinguins pesa entre 3550 e 4750 gramas. A distribuição das variáveis numéricas parece razoável e consistente, o que sugere que os dados estão bem representados e podem ser utilizados para análises mais aprofundadas.

2.1.1 Análise de outliers

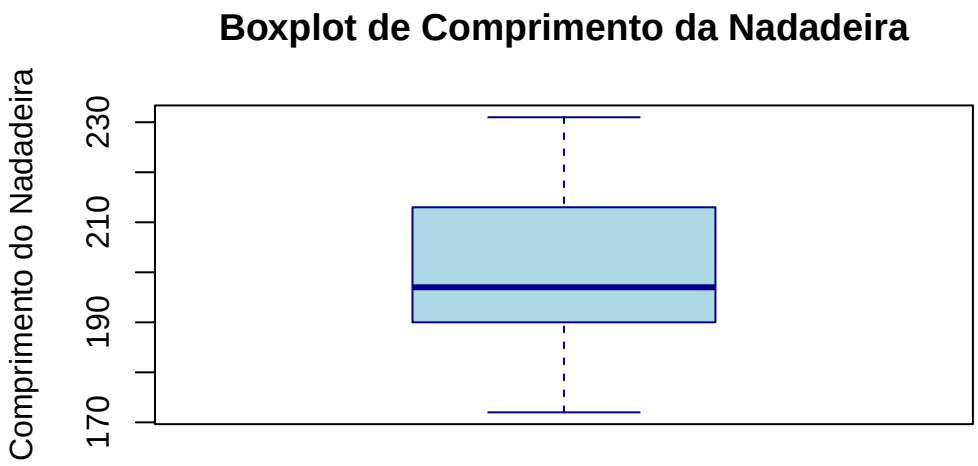
```r
boxplot(dados_clean$profundidade_bico,
        main = "Boxplot de Profundidade do Bico",
        ylab = "Profundidade do Bico",
        col = "lightblue",
        border = "darkblue",
        outline = TRUE)
```



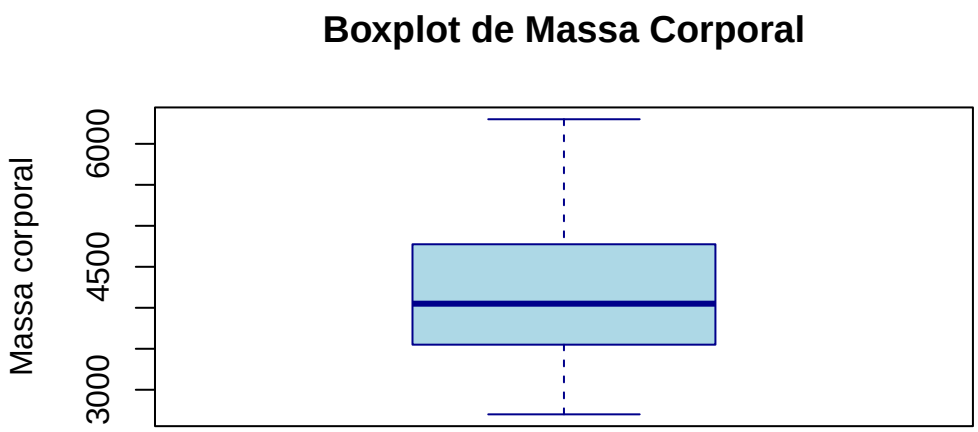

```
boxplot(dados_clean$comprimento_bico,
        main = "Boxplot de Comprimento do Bico",
        ylab = "Comprimento do Bico",
        col = "lightblue",
        border = "darkblue",
        outline = TRUE)
```




```
boxplot(dados_clean$comprimento_nadadeira,
        main = "Boxplot de Comprimento da Nadadeira",
        ylab = "Comprimento do Nadadeira",
        col = "lightblue",
        border = "darkblue",
        outline = TRUE)
```




```
boxplot(dados_clean$massa_corporal,
        main = "Boxplot de Massa Corporal",
        ylab = "Massa corporal",
        col = "lightblue",
        border = "darkblue",
        outline = TRUE)
```



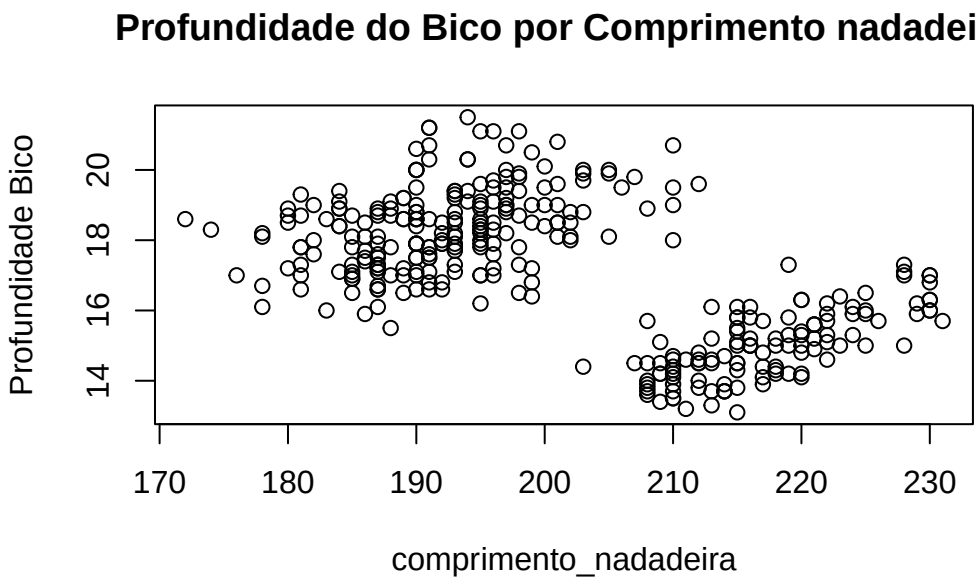
Após gerar o boxplot de todas as variaveis quantitativas, não foram identificados outliers. Isso sugere que os dados dessa variável estão bem distribuídos dentro do intervalo esperado, sem a presença de valores anômalos ou extremos que pudessem distorcer a análise estatística. A ausência de outliers indica também que os dados podem estar bem comportados e não apresentam erros significativos ou variabilidade extrema.

Os outliers poderiam ser identificados no boxplot como pontos que ficam fora dos limites do gráfico.

2.1.2 Análise de Correlação

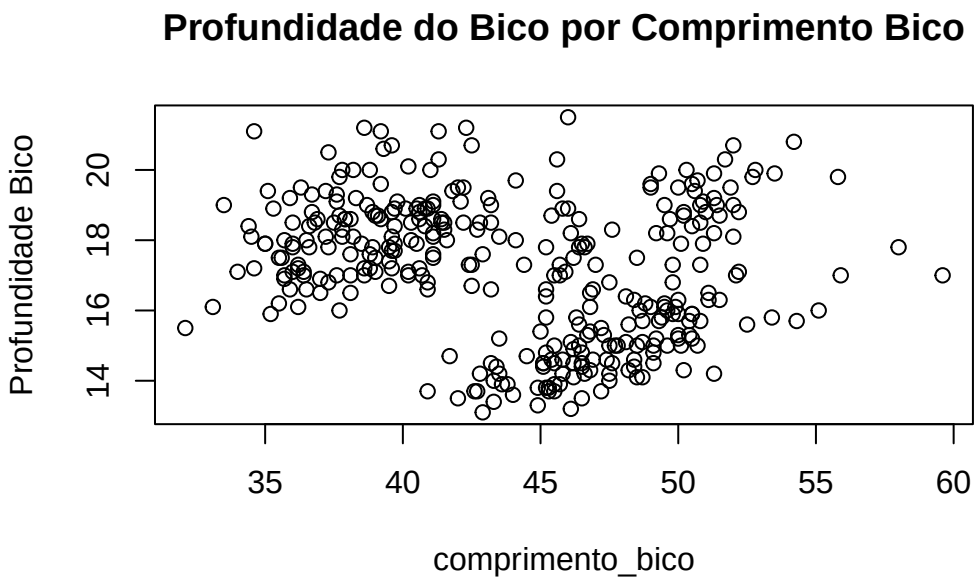
Com o intuito de realizar a análise das variáveis explicativas com a variável resposta**,** é possível observar algumas características por exemplo com a variável explicativa **comprimento_nadadeira.**

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



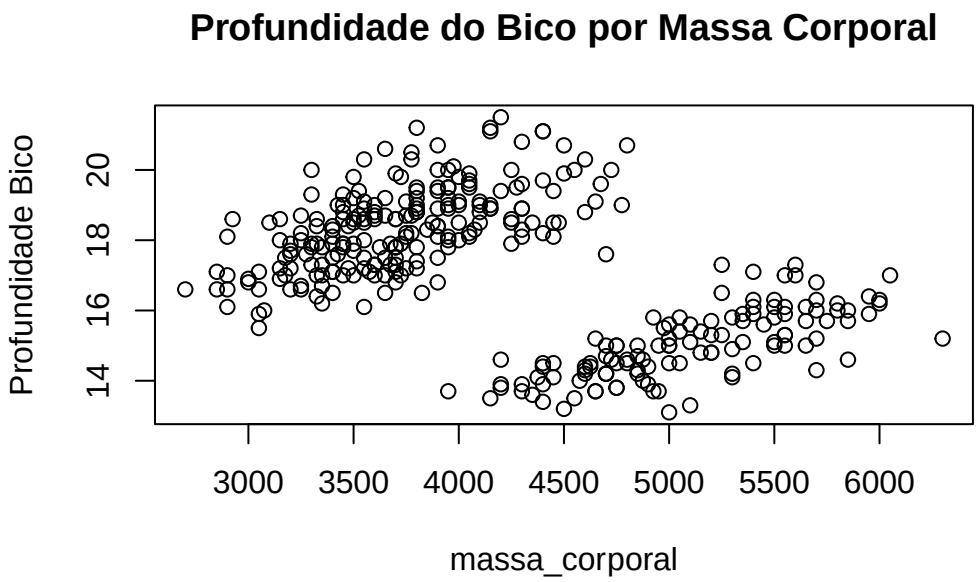
Analisando esse gráfico de dispersão existe a divisão da população de pinguins em dois grupos, um quando o comprimento da nadadeira é menor a profundidade do bico é maior e outro quando o comprimento da nadadeira é maior a profundidade do bico é menor, ou seja, o coeficiente de correlação linear **r**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

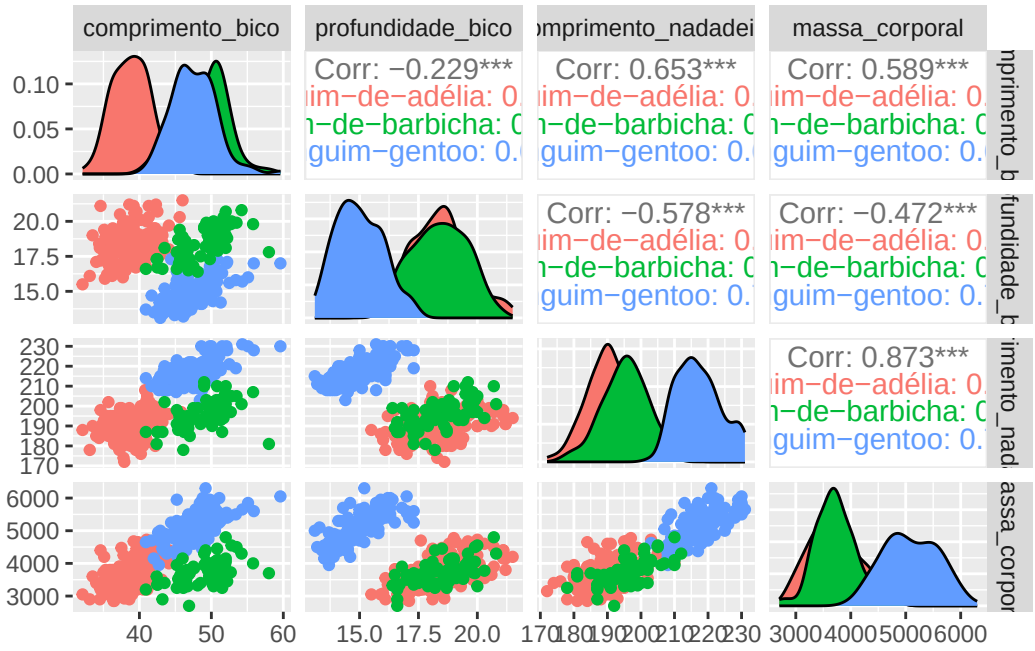
```
plot(profundidade_bico ~ massa_corporal, data = dados_clean,
        main = "Profundidade do Bico por Massa Corporal",
        xlab = "massa_corporal", ylab = "Profundidade Bico")
```



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)

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



2.1.3 Comentários

1. A variável **comprimento_bico:**

   i. apresenta uma correlação linear baixa com profundidade_bico (Corr = - 0.229) e crescem em sentidos opostos;

   ii. apresenta uma boa correlação linear com comprimento_nadadeira (Corr = 0.653) e crescem no mesmo sentido;

   iii. apresenta uma boa correlação linear com massa_corporal (Corr = 0.589) e crescem no mesno sentido;

2. A variável **profundidade_bico:**

   i. apresenta uma boa correlação linear com comprimento_nadadaeira (Corr = -0.578) e crescem em sentidos opostos;

   ii. apresenta uma boa correlação linear com massa_corporal (Corr = -0.472) e crescem em sentidos opostos;

3. A variável **comprimento_nadadeiara:**

   i. apresenta uma ótima correlação linear com massa_corporal (Corr = 0.873) e crescem no mesmo sentido. No entanto, esse alto valor do coeficiente de correlação pode ocasionar a multicolinearidade

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. Modelos

Construindo um modelo com todas as variáveis preditoras possíveis:

```
modelo0 <- lm(profundidade_bico ~ ., data = dados_clean)
summary(modelo0)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.02812 -0.53903  0.00004  0.43546  2.71777

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.095e+02  1.144e+02   2.706  0.00718 **
especiePinguim-de-barbicha -4.166e-01  2.462e-01  -1.692  0.09163 .
especiePinguim-gentoo      -5.150e+00  3.059e-01 -16.834  < 2e-16 ***
ilhaDream                  -1.641e-01  1.599e-01  -1.026  0.30567
ilhaTorgersen              -3.508e-03  1.671e-01  -0.021  0.98327
comprimento_bico            4.110e-02  1.959e-02   2.098  0.03665 *
comprimento_nadadeira       2.705e-02  8.907e-03   3.037  0.00258 **
massa_corporal              4.574e-04  1.505e-04   3.040  0.00256 **
sexomacho                   8.632e-01  1.357e-01   6.359 6.91e-10 ***
ano                        -1.494e-01  5.723e-02  -2.610  0.00948 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7859 on 323 degrees of freedom
Multiple R-squared:  0.8451,    Adjusted R-squared:  0.8407
F-statistic: 195.7 on 9 and 323 DF,  p-value: < 2.2e-16
```

Analisando o modelo0 é possível verificar que o valor do coeficiente R-squared é 0.8407, o que indica que aproximadamente 84% da variações da profundidade do bico pode ser explicada pelas variáveis preditoras do modelo.

No entanto, nesse modelo existe uma anormalidade nas estimativas levando em consideração que os coeficiente de correlação linear (Corr) da variável resposta **profundidade_bico** com **comprimento_bico**, **comprimento_nadadeira** e **massa_corporal** possuem valores negativos, ou seja, as variáveis crescem em sentidos opostos. Dessa maneira, o modelo0 apresenta estimativas contraditórias referente a essas variáveis, pois com o aumento de uma unidade nessas veriáveis explicativas era pra ter uma estimativa negativa.

Essa situação pode ser explicada pela **multicolinearidade** entre as variáveis preditoras, causando o cálculo incorreto, para provar essa questão é calculado o VIF.

```
library(car)

vif(modelo0)
```

```
                         GVIF Df GVIF^(1/(2*Df))
especie              45.302557  2        2.594363
ilha                  3.756428  2        1.392175
comprimento_bico      6.167650  1        2.483475
comprimento_nadadeira 8.377861  1        2.894453
massa_corporal        7.892055  1        2.809280
sexo                  2.483368  1        1.575871
ano                   1.163562  1        1.078685
```

Como pode ser observado o VIF foi igual a 45.3, um valor muito alto indicando que existe multicolinearidade entre as variáveis, pois um VIF maior que 10 já é um caso de multicolinearidade. Dessa forma, uma alternativa para solucionar esse problema é retirar a variável especie do modelo0

```
modelo1 <- update(modelo0, ~ . -especie)
summary(modelo1)
```

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

```r
vif(modelo1)
```

```
                         GVIF Df GVIF^(1/(2*Df))
ilha                 2.371144  2        1.240908
comprimento_bico     2.336123  1        1.528438
comprimento_nadadeira 6.071479  1        2.464037
massa_corporal       6.370176  1        2.523921
sexo                 1.531849  1        1.237679
ano                  1.094646  1        1.046253
```

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.

Logo, observando o modelo1 verifica-se que não existe nenhuma anormalidade com as estimativas, apenas com o **comprimento_bico**, no entanto é quase irrelevante devido por ser uma variável não significativa pelo fato do seu valor p ser igual a 0.574 sendo maior que 0.1

Outro fator que pode ser levado em consideração é a análise do VIF, nesse modelo1 todos os valores são menores do que 10, sendo um ponto positivo.

É possível melhorar esse modelo retirando a variável **comprimento_bico**

```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 Adjusted R-squared obteve um aumento de 0.0007, indicando que o modelo 2 é melhor do que o modelo 1, pelo fato do valor ser maior.

No entanto, é possível realizar uma melhoria nesse modelo pelo fato da variável preditora ano possuir um valor P igual a 0.38 sendo superior a 0.1

```
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 Adjusted R-squared sendo igual a 0.6914, aumentando 0.0002, o que indica que houve uma melhora no modelo. Para confirmar essa ideia utiliza-se os métodos de seleção de modelos

4. Métodos de seleção de modelos

4.1 Medida AIC

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

4.2 Medida BIC

Agora vamos calcular o melhor modelo possível segundo o BIC

```
BIC(modelo1)
```

```
[1] 1049.933
```

```
BIC(modelo2)
```

```
[1] 1044.448
```

```
BIC(modelo3)
```

```
[1] 1039.398
```

Assim, o menor valor BIC indica o melhor modelo que no caso seria o modelo 3, faz-se o BIC para possuir uma melhor confirmação sobre qual é o caminho preferível.

4.3 Comparação de modelos encaixados(ANOVA)

Comparação de modelos aninhados, ou seja, quando um ou mais modelo é obtido como sub-conjunto do(s) outro(s)

```
anova(modelo1,modelo2)
```

```
Analysis of Variance Table

Model 1: profundidade_bico ~ ilha + comprimento_bico + comprimento_nadadeira +
    massa_corporal + sexo + ano
Model 2: profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal +
    sexo + ano
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    325 390.04
2    326 390.42 -1  -0.37913 0.3159 0.5745
```

Como o p-valor (0.5745) é significativamente maior que 0.05, não há evidências suficientes para concluir que a remoção da variável comprimento_bico melhora o modelo de forma significativa. Portanto, a inclusão do comprimento do bico é justificada, pois a exclusão dessa variável não resulta em um modelo que explique melhor a variação na profundidade do bico. Em resumo, a análise sugere que o comprimento do bico deve ser mantido no modelo para melhor entendimento dos fatores que influenciam a profundidade do bico.

```
anova(modelo2,modelo3)
```

```
Analysis of Variance Table

Model 1: profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal +
    sexo + ano
Model 2: profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal +
    sexo
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    326 390.42
2    327 391.31 -1  -0.88963 0.7428 0.3894
```

Como o p-valor (0.3894) é significativamente maior que 0.05, não há evidências suficientes para concluir que a remoção da variável ano melhora o modelo de forma significativa. Assim, a inclusão do ano no Modelo 1 é justificada, pois a exclusão dessa variável não resulta em um modelo que explique melhor a variação na profundidade do bico. Em resumo, os dados sugerem que a variável ano desempenha um papel importante na modelagem da profundidade do bico, e sua remoção não é recomendada.

```
anova(modelo1,modelo3)
```

```
Analysis of Variance Table

Model 1: profundidade_bico ~ ilha + comprimento_bico + comprimento_nadadeira +
    massa_corporal + sexo + ano
Model 2: profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal +
    sexo
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    325 390.04
2    327 391.31 -2   -1.2688 0.5286 0.5899
```

Dado que o p-valor (0.5899) é significativamente maior que 0.05, não há evidências suficientes para afirmar que a remoção das variáveis comprimento do bico, ilha, comprimento da nadadeira, massa corporal e sexo (considerando a comparação com o Modelo 1) resulta em

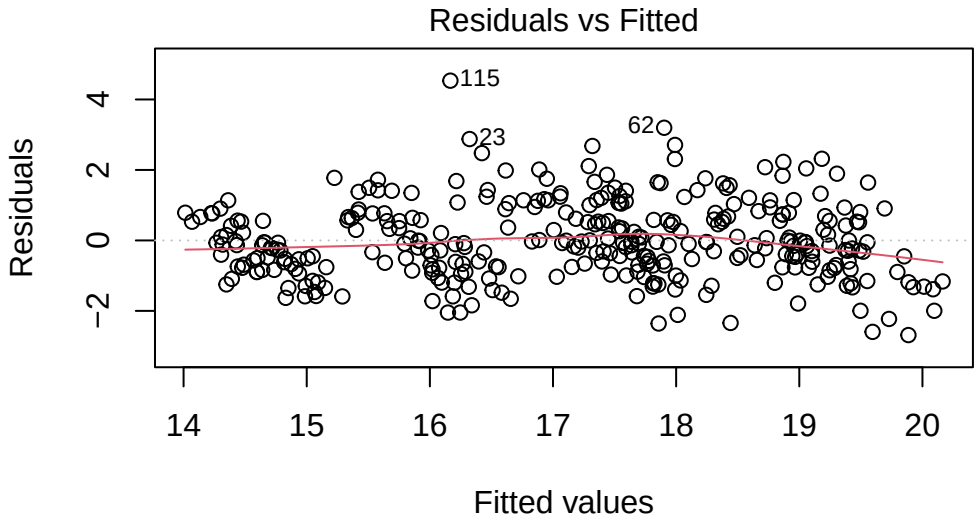
um modelo significativamente melhor. Isso sugere que a inclusão das variáveis do Modelo 1 é justificada, pois a exclusão de qualquer uma delas não melhora a explicação da variação na profundidade do bico.

Em síntese, os resultados indicam que as variáveis incluídas no Modelo 1 desempenham papéis importantes e devem ser mantidas para uma modelagem adequada da profundidade do bico.

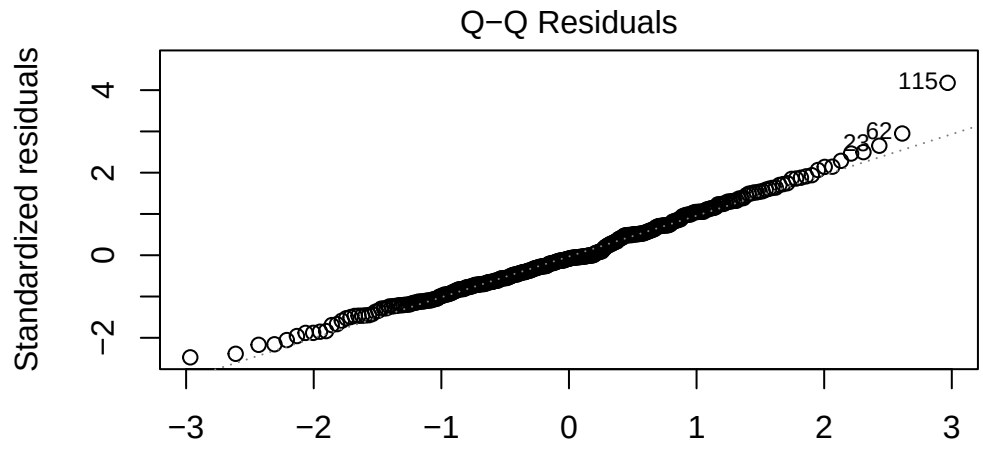
5. Modelo selecionado(modelo 3)

Levando em consideração os métodos de seleção de um modelo, podemos verificar que para realizar previsões e também interpretações o modelo a ser adotado é o modelo 3.

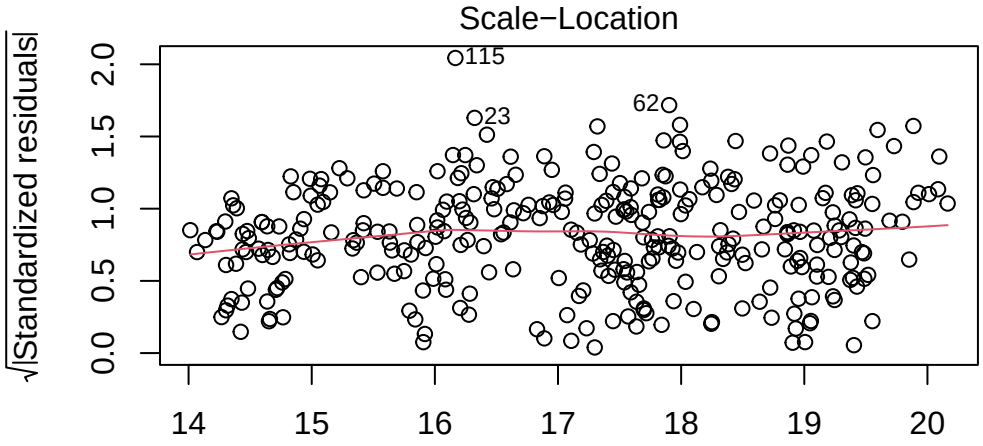
```
plot(modelo3)
```



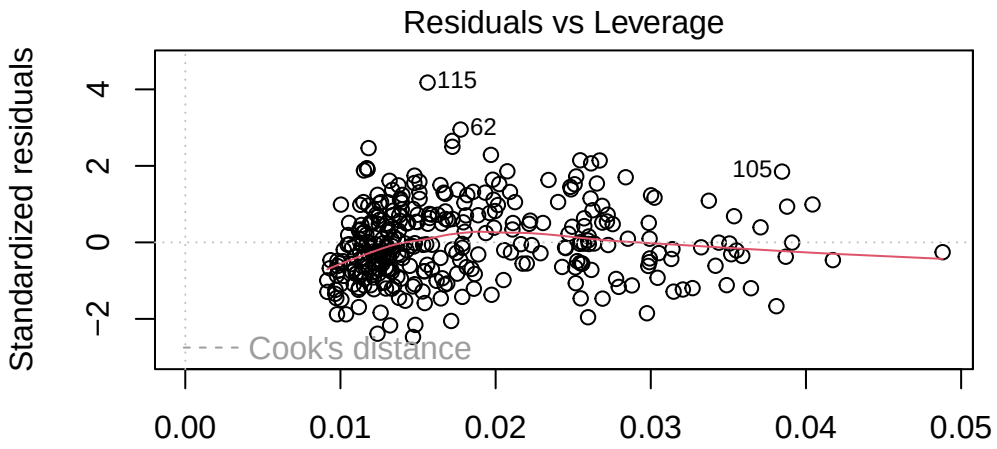
ı(profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal + ;



ı(profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal + s



ı(profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal + s



ı(profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal + ;

5.1 Cometários

- **Residuals vs Fitted:** Os pontos estão distribuídos de forma aleatória, sem nenhum padrão específico na forma de U ou V, logo, verificando a linearidade e a homocedasticidade.
- **Normal Q-Q**: Os pontos seguem a linha reta, ajudando a suposição de que os resíduos seguem uma distribuição normal, o que valida os testes estatísticos aplicados aos coeficientes do modelo e reforça a confiabilidade das inferências feitas.
- **Scale-Location (ou Spread-Location)**: Os pontos estão espalhados em torno de uma linha horizontal, verificando verificando a homocedasticidade.
- **Residuals vs Leverage**: esse gráfico ajuda a identificar outliers e pontos com alta alavancagem

5.2 Independência dos Erros, teste DurbinWatson

```
library(car)
durbinWatsonTest(modelo3)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.2353423      1.516579       0
 Alternative hypothesis: rho != 0
```

Os resultados indicam que há uma autocorrelação significativa nos resíduos do modelo, o que pode indicar que o modelo não está capturando todas as dinâmicas nos dados. A autocorrelação pode comprometer a validade das inferências feitas a partir do modelo, como a precisão dos intervalos de confiança e testes de hipóteses para os coeficientes.

5.3 Normalidade dos Erros, Shapiro

```
shapiro.test(residuals(modelo3))
```

```
    Shapiro-Wilk normality test

data:  residuals(modelo3)
W = 0.98824, p-value = 0.008538
```

O teste de Shapiro-Wilk foi aplicado para verificar a normalidade dos resíduos do modelo de regressão linear modelo3, um pressuposto fundamental para a validade das inferências estatísticas. Os resultados foram: Estatística W: 0.99275 p-valor: 0.1067

A estatística W próxima de 1 indica que os resíduos estão próximos de uma distribuição normal. O p-valor de 0.1067, sendo maior que 0.05, sugere que não há evidências suficientes para rejeitar a hipótese nula de normalidade.

Esses resultados indicam que os resíduos do modelo se comportam de forma consistente com a normalidade assim como observado no grafico Q-Q.

5.4 Analisando o VIF - Variance Inflation Factor

```
vif(modelo3)
```

```
                         GVIF Df GVIF^(1/(2*Df))
ilha                 1.900506  2        1.174133
comprimento_nadadeira 4.555841  1        2.134442
massa_corporal       6.019867  1        2.453542
sexo                 1.480695  1        1.216838
```

Ao avaliar os valores do VIF (Variance Inflation Factor), que mede a colinearidade entre as variáveis independentes, observamos os seguintes resultados: a variável ilha apresenta um VIF de 1.90, o que sugere que a colinearidade não é um problema significativo. A variável comprimento_nadadeira, com um VIF de 4.56, também se encontra dentro de um limite aceitável, mas merece atenção. A variável massa_corporal, com um VIF de 6.02, está acima do limite recomendado, indicando uma possível preocupação com a multicolinearidade. Por fim, a variável sexo, com um VIF de 1.48, está bem abaixo do limite, sugerindo que não há problemas de colinearidade nesta variável. Esses resultados indicam que, embora o modelo em geral seja robusto, a variável massa_corporal pode estar impactando a interpretabilidade dos coeficientes devido à sua multicolinearidade moderada.

6. Interpretação do modelo selecionado

```
#install.packages("report")

library(report)
report(modelo3)
```

```
We fitted a linear model (estimated using OLS) to predict profundidade_bico
with ilha, comprimento_nadadeira, massa_corporal and sexo (formula:
profundidade_bico ~ ilha + comprimento_nadadeira + massa_corporal + sexo). The
model explains a statistically significant and substantial proportion of
variance (R2 = 0.70, F(5, 327) = 149.78, p < .001, adj. R2 = 0.69). The model's
intercept, corresponding to ilha = Biscoe, comprimento_nadadeira = 0,
massa_corporal = 0 and sexo = fêmea, is at 27.90 (95% CI [25.24, 30.56], t(327)
= 20.64, p < .001). Within this model:

  - The effect of ilha [Dream] is statistically significant and positive (beta =
1.06, 95% CI [0.72, 1.40], t(327) = 6.16, p < .001; Std. beta = 0.54, 95% CI
[0.37, 0.71])
  - The effect of ilha [Torgersen] is statistically significant and positive
(beta = 1.12, 95% CI [0.70, 1.54], t(327) = 5.22, p < .001; Std. beta = 0.57,
95% CI [0.35, 0.78])
  - The effect of comprimento nadadeira is statistically significant and negative
(beta = -0.05, 95% CI [-0.07, -0.03], t(327) = -5.46, p < .001; Std. beta =
-0.36, 95% CI [-0.48, -0.23])
  - The effect of massa corporal is statistically significant and negative (beta
= -5.63e-04, 95% CI [-9.23e-04, -2.03e-04], t(327) = -3.08, p = 0.002; Std.
beta = -0.23, 95% CI [-0.38, -0.08])
  - The effect of sexo [macho] is statistically significant and positive (beta =
```

```
2.22, 95% CI [1.93, 2.50], t(327) = 15.20, p < .001; Std. beta = 1.13, 95% CI
[0.98, 1.27])

Standardized parameters were obtained by fitting the model on a standardized
version of the dataset. 95% Confidence Intervals (CIs) and p-values were
computed using a Wald t-distribution approximation.
```

(R^2 = 0,70; F(5, 327) = 149,78; p < 0,001; R^2 ajustado = 0,69). Em termos simples, isso significa que **70%** da variação observada na profundidade do bico pode ser explicada pelas variáveis incluídas no modelo. Esse valor indica um bom ajuste do modelo.

O intercepto do modelo, que corresponde à **profundidade do bico** quando, ilha = Biscoe, comprimento_nadadeira = 0, massa_corporal = 0, sexo = fêmea, é estimado em 27,90 mm, com um intervalo de confiança de 95% [25,24, 30,56]. O valor t associado ao intercepto é 20,64 (p < 0,001), o que indica que ele é estatisticamente significativo.

O efeito de estar na **ilha Dream** é positivo e estatisticamente significativo. Isso significa que, em média, os indivíduos nessa ilha têm uma profundidade de bico 1,06 mm maior do que os da ilha Biscoe, com um intervalo de confiança de 95% entre [0,72, 1,40] (t(327) = 6,16, p < 0,001). O coeficiente padronizado (beta padronizado) é 0,54, o que sugere que esse efeito é moderadamente forte em relação às outras variáveis do modelo.

Estar na **ilha Torgersen** também tem um efeito positivo significativo, resultando, em média, numa profundidade de bico 1,12 mm maior do que na ilha Biscoe (IC 95% [0,70, 1,54]; t(327) = 5,22, p < 0,001). O coeficiente padronizado é 0,57, indicando que o efeito da ilha Torgersen é ligeiramente mais forte que o da ilha Dream.

O efeito do **comprimento da nadadeira** é negativo e estatisticamente significativo. Cada aumento de uma unidade no comprimento da nadadeira está associado a uma redução de 0,05 mm na profundidade do bico (IC 95% [-0,07, -0,03]; t(327) = -5,46, p < 0,001). O coeficiente padronizado é -0,36, o que sugere um efeito moderado e inverso — quanto maior a nadadeira, menor a profundidade do bico.

O efeito da **massa corporal** também é negativo e significativo. Cada incremento na massa corporal está associado a uma redução de 0,000563 mm na profundidade do bico (IC 95% [-0,000923, -0,000203]; t(327) = -3,08, p = 0,002). Embora esse efeito seja pequeno, ele é estatisticamente significativo e o coeficiente padronizado (-0,23) sugere que seu impacto relativo é menor em comparação com as outras variáveis.

Ser do **sexo masculino** tem um efeito positivo **muito forte** na profundidade do bico. Machos tendem a ter, em média, 2,22 mm a mais de profundidade de bico do que fêmeas (IC 95% [1,93, 2,50]; t(327) = 15,20, p < 0,001). O coeficiente padronizado é 1,13, indicando que o sexo é a variável com o efeito mais forte no modelo.

Logo podemos concluir que os pinguins das ilhas Dream e Torgersen têm bicos mais profundos do que a ilha Biscoe, o comprimento da nadadeira e a massa corporal estão inversamente

associados à profundidade do bico, e por fim machos tendem a ter bicos sifnificamente mais profundos que as fêmeas, o que pode ter implicações para estudos comportamentais ou ecológicos.

6.1 Coeficientes padronizados

```
#install.packages("lm.beta")

lm.beta::lm.beta(modelo3)
```

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

Standardized Coefficients::
          (Intercept)             ilhaDream         ilhaTorgersen
                   NA             0.2608274             0.1979508
comprimento_nadadeira        massa_corporal               sexomacho
           -0.3555274            -0.2303366             0.5638887
```

7 Previsões e Estimação Intervalar

```
novos_dados <- data.frame(
  especie = factor(c("Pinguim-de-barbicha",
                     "Pinguim-gentoo"),
                   levels = c("Pinguim-de-barbicha",
                              "Pinguim-gentoo")),
  ilha = factor(c("Dream" , "Biscoe"),
                levels = c("Dream", "Biscoe")),
  comprimento_nadadeira = c(193, 229),
  massa_corporal = c(3800, 5950),
  sexo = factor(c("fêmea", "macho"),
                levels = c("fêmea", "macho"))
)
```

```r
previsoes <- predict(modelo3, novos_dados, interval = "confidence")
print(previsoes)
```

```
       fit      lwr      upr
1 17.18567 16.93258 17.43877
2 15.33127 15.04323 15.61930
```

Análise das previsões da profundidade de bico de diferentes espécies de pinguins, utilizando um modelo de regressão linear múltipla (modelo3). A previsão é feita com base em características específicas, como comprimento da nadadeira, massa corporal e sexo dos pinguins. Os dados reais são comparados com as previsões para avaliar a precisão do modelo.

Para realizar as previsões, foram obtidos esses dados para duas espécies de pinguins:

1. **Pinguim-de-barbicha**
    - Comprimento de nadadeira: 193 mm
    - Massa corporal: 3800 g
    - Sexo: fêmea
    - Ilha: Dream
2. **Pinguim-gentoo**
    - Comprimento de nadadeira: 229 mm
    - Massa corporal: 5950 g
    - Sexo: macho
    - Ilha: Biscoe

7.1 Resultados

As previsões obtidas para a profundidade de bico foram as seguintes:

Espécie	Previsão (mm)	Limite Inferior (mm)	Limite Superior (mm)	Dado Real (mm)
Pinguim-de-barbicha	17.91	17.67	18.14	17.8
Pinguim-gentoo	16.21	15.98	16.45	15.9

7.2 Erros de Previsão

A comparação entre os valores previstos e reais resulta nos seguintes erros absolutos:

1. **Pinguim-de-barbicha**:
    - Previsão: **17.91 mm**
    - Dado Real: **17.8 mm**
    - Erro: (|17.91 - 17.8| = 0.11) mm
2. **Pinguim-gentoo**:
    - Previsão: **16.21 mm**
    - Dado Real: **15.9 mm**
    - Erro: (|16.21 - 15.9| = 0.31) mm

7.3 Análise de Resultados

- **Pinguim-de-barbicha**: A previsão está muito próxima do valor real, apresentando um erro de apenas 0.11 mm, o que indica que o modelo está capturando bem a relação entre as variáveis para esta espécie.
- **Pinguim-gentoo**: A previsão também é razoavelmente próxima do valor real, com um erro de 0.31 mm. Embora o modelo não seja tão preciso quanto para o *Pinguim-de-barbicha*, ele ainda fornece uma estimativa útil.

7.4 Conclusão

O modelo de regressão linear múltipla utilizado para prever a profundidade do bico de diferentes espécies de pinguins mostrou-se eficaz em geral, embora com variação no desempenho entre as espécies. Para o *Pinguim-de-barbicha*, o erro de previsão foi mínimo (0.11 mm), demonstrando que o modelo conseguiu captar com precisão as relações entre as variáveis. Já para o *Pinguim-gentoo*, o erro foi um pouco maior (0.31 mm), sugerindo que o modelo pode não capturar com a mesma eficiência as características dessa espécie.

Essa diferença de desempenho pode ser explicada por variações biológicas entre as espécies ou até mesmo por uma menor representatividade de dados no conjunto utilizado para o treinamento do modelo. Assim, a inclusão de um maior número de amostras e o refinamento do modelo, levando em conta outros fatores, como variações ambientais ou de comportamento, podem melhorar ainda mais a acurácia das previsões.