UNIVERSIDADE FEDERAL DE CAMPINA GRANDE - UFCG
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA - CEEI
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO - UASC

Análise do Banco de Dados “Pinguins”

Disciplina: Estatística Aplicada

Professor: Gilberto Silva Matos

Alunos:

Ana Beatriz Cavalcanti dos Santos - 121210095

Arthur Andrade de Miranda - 121210088

Raniel Miranda Dourado - 121110466

Campina Grande

Outubro de 2024

SUMÁRIO

1. INTRODUÇÃO 3

1.1 Objetivos 3

2. DESCRIÇÃO DA BASE DE DADOS 3

3. ANÁLISE EXPLORATÓRIA DOS DADOS 4

3.1 Identificação de Valores Ausentes e Outliers 4

3.2 Análise de Correlação e Verificação de Multicolinearidade 5

4. AJUSTE DO MODELO INICIAL 13

4.1 Análise da Qualidade do Modelo Ajustado de Regressão Linear Múltipla ….....13

4.2 Análise da Qualidade do Modelo Ajustado sem Massa Corporal …… 14

4.3 Análise da Qualidade do Modelo Ajustado Sem o Comprimento da Nadadeira.. 16

4.4 Seleção do Melhor Modelo……17

5. VERIFICAÇÃO DOS PRESSUPOSTOS DO MODELO 19

5.1 Normalidade dos Resíduos 22

5.2 Homocedasticidade 23

5.3 Independência dos Resíduos 23

6. INTERPRETAÇÃO DO MODELO SELECIONADO..... 23

7. ESTIMATIVA E PREVISÃO 25

8. CONCLUSÃO 27

1. INTRODUÇÃO

1.1 Objetivos

Este relatório tem como objetivo analisar o banco de dados "Pinguins", que contém informações sobre pinguins das ilhas da Antártica. Visa-se, por meio do treinamento de um modelo de Regressão Linear Múltipla (MRLM), identificar quais variáveis explicam as variações na profundidade do bico, nossa variável resposta. A profundidade do bico está relacionada a outras características físicas dos pinguins, como o comprimento do bico, massa corporal e comprimento da nadadeira.

2. DESCRIÇÃO DA BASE DE DADOS

O conjunto de dados **"Pinguins"** inclui informações sobre **344 pinguins adultos** (344 linhas) de 3 espécies diferentes de pinguins coletadas em 3 ilhas do Arquipélago Palmer, na Antártica: **Adélia**, **Chinstrap** e **Gentoo**. O objetivo da análise é explorar como diferentes variáveis explicativas (como comprimento do bico, comprimento da nadadeira e massa corporal) afetam a **profundidade do bico**, que será a variável resposta.

No total, temos 8 variáveis observadas (8 colunas), sendo elas:

- **especie**: três espécies distintas de pinguins, sendo elas: Pinguim-de-adélia, Pinguim-de-barbicha e Pinguim-gentoo;
- **ilha**: as três ilhas do Arquipélago Palmer, na Antártida: Biscoe, Dream, Togersen
- **comprimento_bico**: um número decimal que indica o comprimento do bico do pinguim, em milímetros
- **profundidade_bico**: um número decimal que indica a profundidade do bico do pinguim, em milímetros
- **comprimento_nadadeira**: número inteiro que indica o comprimento da nadadeira, em milímetros
- **massa_corporal**: um número inteiro que indica a massa corporal do pinguim, em gramas
- **sexo**: indicação do sexo do peguins, como fêmea ou macho
- **ano**: número inteiro que indica o ano que os dados foram coletados, sendo 2007, 2008 ou 2009.

3. ANÁLISE EXPLORATÓRIA DOS DADOS

3.1 Identificação de Valores Ausentes e Outliers

```
> summary(pinguins)
                especie          ilha     comprimento_bico profundidade_bico comprimento_nadadeira massa_corporal    sexo         ano
 Pinguim-de-adélia  :152   Biscoe   :168   Min.   :32.10    Min.   :13.10     Min.   :172.0        Min.   :2700   fêmea:165   Min.   :2007
 Pinguim-de-barbicha: 68   Dream    :124   1st Qu.:39.23    1st Qu.:15.60     1st Qu.:190.0        1st Qu.:3550   macho:168   1st Qu.:2007
 Pinguim-gentoo     :124   Torgersen: 52   Median :44.45    Median :17.30     Median :197.0        Median :4050   NA's : 11   Median :2008
                                           Mean   :43.92    Mean   :17.15     Mean   :200.9        Mean   :4202               Mean   :2008
                                           3rd Qu.:48.50    3rd Qu.:18.70     3rd Qu.:213.0        3rd Qu.:4750               3rd Qu.:2009
                                           Max.   :59.60    Max.   :21.50     Max.   :231.0        Max.   :6300               Max.   :2009
                                           NA's   :2        NA's   :2         NA's   :2            NA's   :2
```

Figura 1: Resultado da execução de summary(pinguins)

```
> summary(pinguins_limpos)
                especie          ilha     comprimento_bico profundidade_bico comprimento_nadadeira massa_corporal    sexo         ano
 Pinguim-de-adélia  :146   Biscoe   :163   Min.   :32.10    Min.   :13.10     Min.   :172          Min.   :2700   fêmea:165   Min.   :2007
 Pinguim-de-barbicha: 68   Dream    :123   1st Qu.:39.50    1st Qu.:15.60     1st Qu.:190          1st Qu.:3550   macho:168   1st Qu.:2007
 Pinguim-gentoo     :119   Torgersen: 47   Median :44.50    Median :17.30     Median :197          Median :4050               Median :2008
                                           Mean   :43.99    Mean   :17.16     Mean   :201          Mean   :4207               Mean   :2008
                                           3rd Qu.:48.60    3rd Qu.:18.70     3rd Qu.:213          3rd Qu.:4775               3rd Qu.:2009
                                           Max.   :59.60    Max.   :21.50     Max.   :231          Max.   :6300               Max.   :2009
```

Figura 2: Resultado da execução de summary(pinguins_limpos)

```
> glimpse(pinguins)
Rows: 344
Columns: 8
$ especie           <fct> Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de…
$ ilha              <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgerse…
$ comprimento_bico  <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, 42.0, 37.8, 37.8, 41.1, 38.6, 34.6, 36.6, 38.7, 42.5, 34.4, 46.0, 37.8, 37.7, 35.9, 38.2, 38.8, 35.3, 40.6, 40…
$ profundidade_bico <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, 20.2, 17.1, 17.3, 17.6, 21.2, 21.1, 17.8, 19.0, 20.7, 18.4, 21.5, 18.3, 18.7, 19.2, 18.1, 17.2, 18.9, 18.6, 17…
$ comprimento_nadadeira <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186, 180, 182, 191, 198, 185, 195, 197, 184, 194, 174, 180, 189, 185, 180, 187, 183, 187, 172, 180, 178, 178, 188…
$ massa_corporal    <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, 4250, 3300, 3700, 3200, 3800, 4400, 3700, 3450, 4500, 3325, 4200, 3400, 3600, 3800, 3950, 3800, 3800, 3550, 32…
$ sexo              <fct> macho, fêmea, fêmea, NA, fêmea, macho, fêmea, macho, NA, NA, NA, NA, fêmea, macho, macho, fêmea, fêmea, macho, fêmea, macho, fêmea, macho, fêmea, macho, macho, fê…
$ ano               <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, …
```

Figura 3: Resultado da execução de glimpse(pinguins)

```
> glimpse(pinguins_limpos)
Rows: 333
Columns: 8
$ especie           <fct> Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de-adélia, Pinguim-de…
$ ilha              <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Torgerse…
$ comprimento_bico  <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.1, 38.6, 34.6, 36.6, 38.7, 42.5, 34.4, 46.0, 37.8, 37.7, 35.9, 38.2, 38.8, 35.3, 40.6, 40.5, 37.9, 40.5, 39.5, 37.2, …
$ profundidade_bico <dbl> 18.7, 17.4, 18.0, 19.3, 20.6, 17.8, 19.6, 17.6, 21.2, 21.1, 17.8, 19.0, 20.7, 18.4, 21.5, 18.3, 18.7, 19.2, 18.1, 17.2, 18.9, 18.6, 17.9, 18.6, 18.9, 16.7, 18.1, …
$ comprimento_nadadeira <int> 181, 186, 195, 193, 190, 181, 195, 182, 191, 198, 185, 195, 197, 184, 194, 174, 180, 189, 185, 180, 187, 183, 187, 172, 180, 178, 178, 188, 184, 195, 196, 190, 18…
$ massa_corporal    <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 3200, 3800, 4400, 3700, 3450, 4500, 3325, 4200, 3400, 3600, 3800, 3950, 3800, 3800, 3550, 3200, 3150, 3950, 3250, 3900, …
$ sexo              <fct> macho, fêmea, fêmea, fêmea, macho, fêmea, macho, fêmea, macho, macho, fêmea, fêmea, macho, fêmea, macho, fêmea, macho, fêmea, macho, fêmea, macho, macho, fêmea, macho, fêmea, f…
$ ano               <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, …
```

Figura 4: Resultado da execução de glimpse(pinguins_limpos)

Percebe-se que das 344 observações do banco de dados original:

- Há 11 observações sem indicação do sexo do penguim;
- Há um total de 2 valores ausentes ("NA's") para as variáveis **comprimento_bico**, **profundidade_bico**, **comprimento_nadadeira** e **massa_corporal**.

Dessa forma, para garantir a eficácia do modelo, estas 11 observações foram excluídas, obtendo-se como 333 observações aptas para a análise.

```
> skim(pinguins_limpos)
── Data Summary ────────────────────────
                           Values
Name                       pinguins_limpos
Number of rows             333
Number of columns          8
_______________________
Column type frequency:
  factor                   3
  numeric                  5
_______________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 especie               0             1 FALSE          3 Pin: 146, Pin: 119, Pin: 68
2 ilha                  0             1 FALSE          3 Bis: 163, Dre: 123, Tor: 47
3 sexo                  0             1 FALSE          2 mac: 168, fêm: 165

── Variable type: numeric ─────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable         n_missing complete_rate   mean      sd     p0    p25    p50    p75   p100 hist
1 comprimento_bico              0             1   44.0    5.47   32.1   39.5   44.5   48.6   59.6 ▃▇▇▆▁
2 profundidade_bico             0             1   17.2    1.97   13.1   15.6   17.3   18.7   21.5 ▅▇▇▆▅
3 comprimento_nadadeira         0             1  201.    14.0   172    190    197    213    231  ▂▇▃▅▂
4 massa_corporal                0             1 4207.   805.   2700   3550   4050   4775   6300 ▃▇▆▅▂
5 ano                           0             1 2008.    0.813 2007   2007   2008   2009   2009 ▇▁▇▁▇
```

Figura 5: Resultado da execução de skim(pinguins_limpos)

A figura 5 apresenta os fatores e as variáveis numéricas do dataset. Para os fatores, a tabela indica que não há valores faltantes, com cada variável tendo uma taxa de completude de 100%. Cada fator exibe as contagens das categorias.

No caso das variáveis numéricas, todas também possuem 100% de completude. A tabela resume estatísticas descritivas. Para a variável resposta **profundidade do bico**, a média é 17,2 mm, com um desvio padrão de 1,97 mm, o que indica uma variação moderada entre as observações. O valor mínimo registrado é 13,1 mm (p0), e o valor máximo é 21,5 mm (p100).

3.2 Análise de Correlação e Verificação de Multicolinearidade

Em uma análise inicial, foi utilizado um gráfico que relaciona todas as variáveis entre si, conforme mostrado na figura. Observa-se uma relação de decrescimento linear entre as variáveis explicativas comprimento do bico, massa corporal e comprimento da nadadeira com a variável resposta, que é a profundidade do bico. À medida que essas medidas aumentam, percebe-se o decrescimento da profundidade do bico. Além disso, é possível observar a relação entre a profundidade do bico e as demais variáveis: espécie, ilha, ano e sexo.

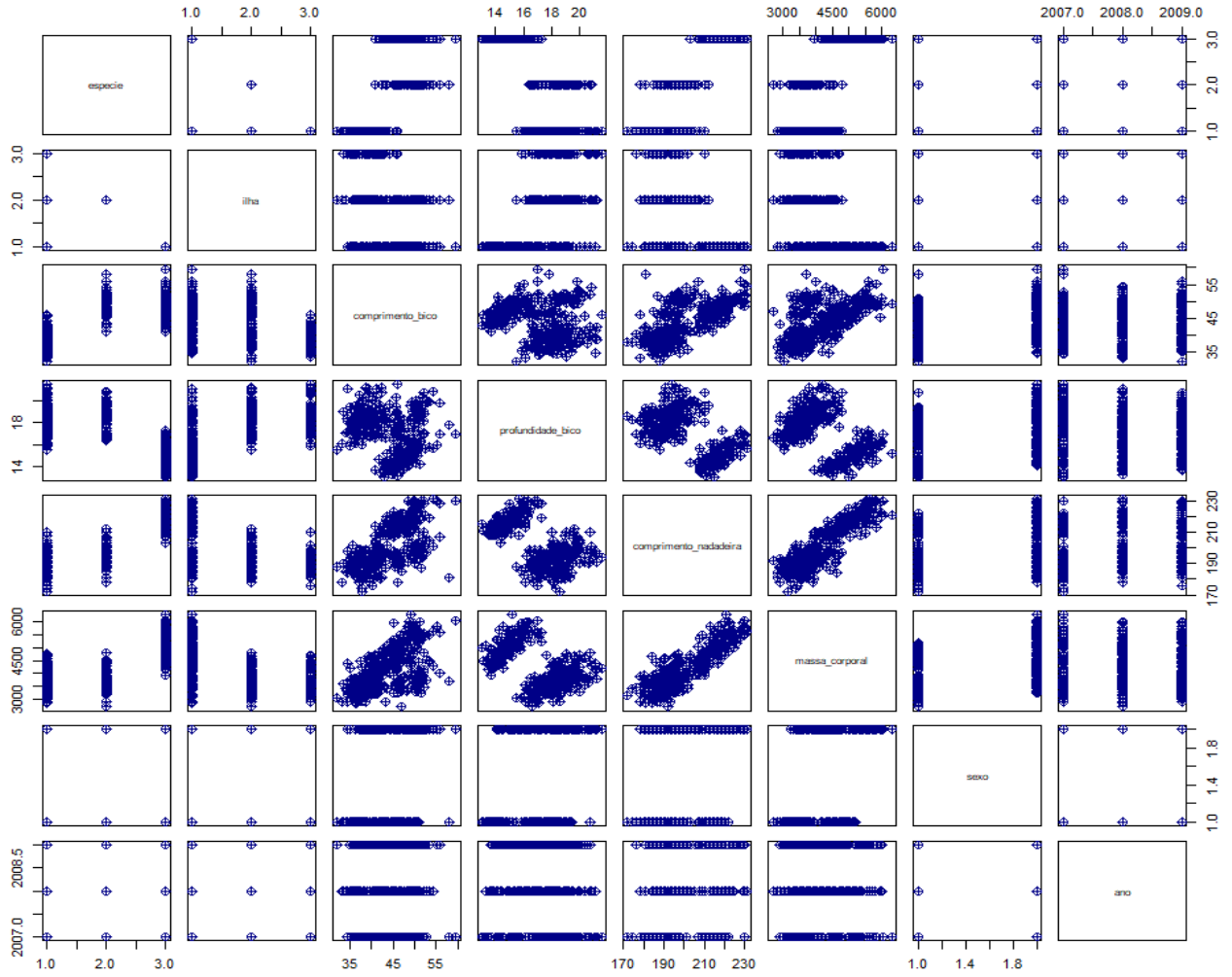

Figura 6: Plot da correlação entre todas as variáveis do banco

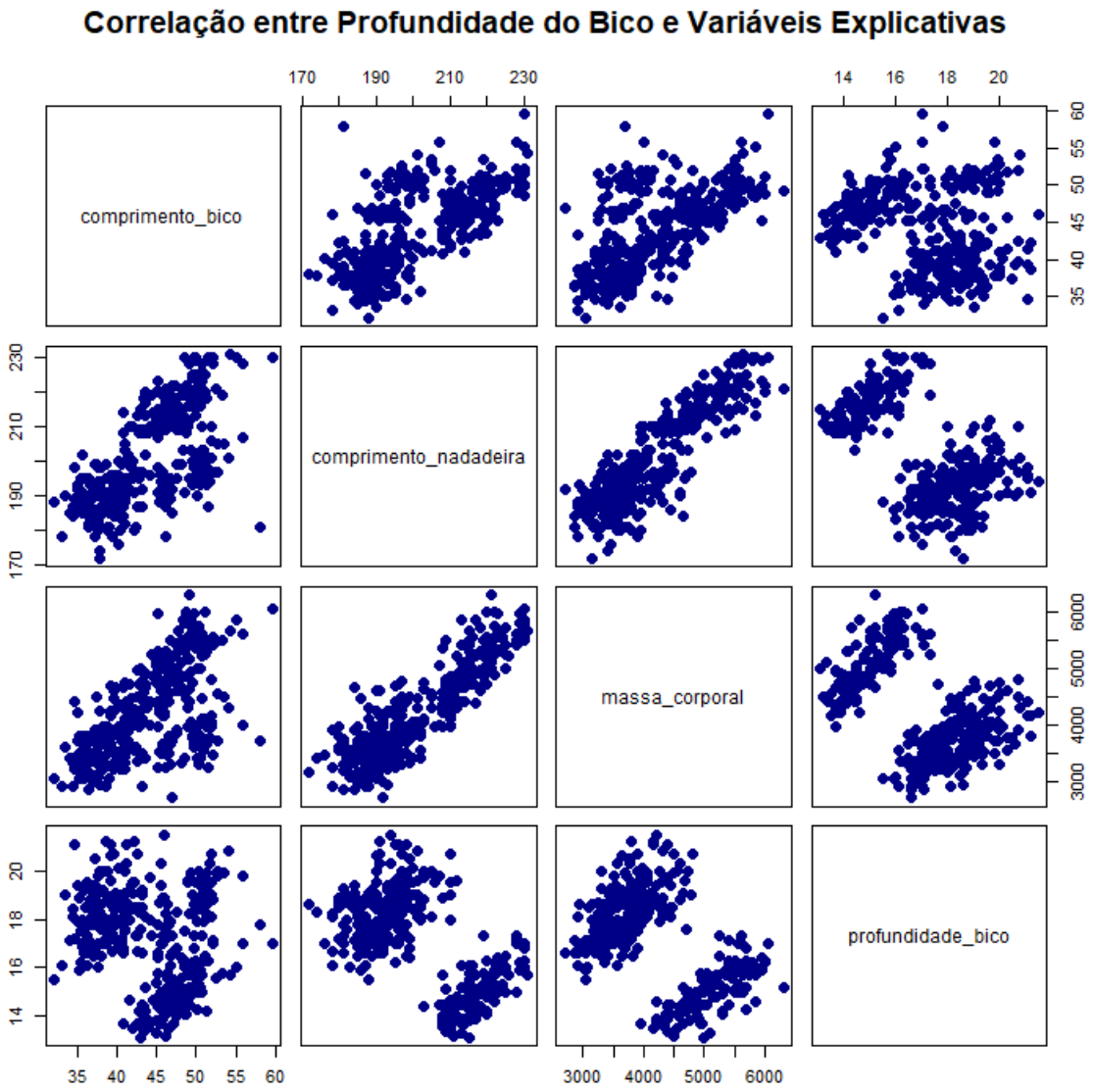

Figura 7: Plot da correlação entre a variável resposta “profundidade do bico" e variáveis explicativas

Olhando mais detalhadamente para a correlação da variável **profundidade_bico** com as outras variáveis, a variável **comprimento_bico** possui fraca correlação negativa com relação a profundidade do bico. Quanto maior o comprimento do bico do pinguim, menor a profundidade de seu bico.

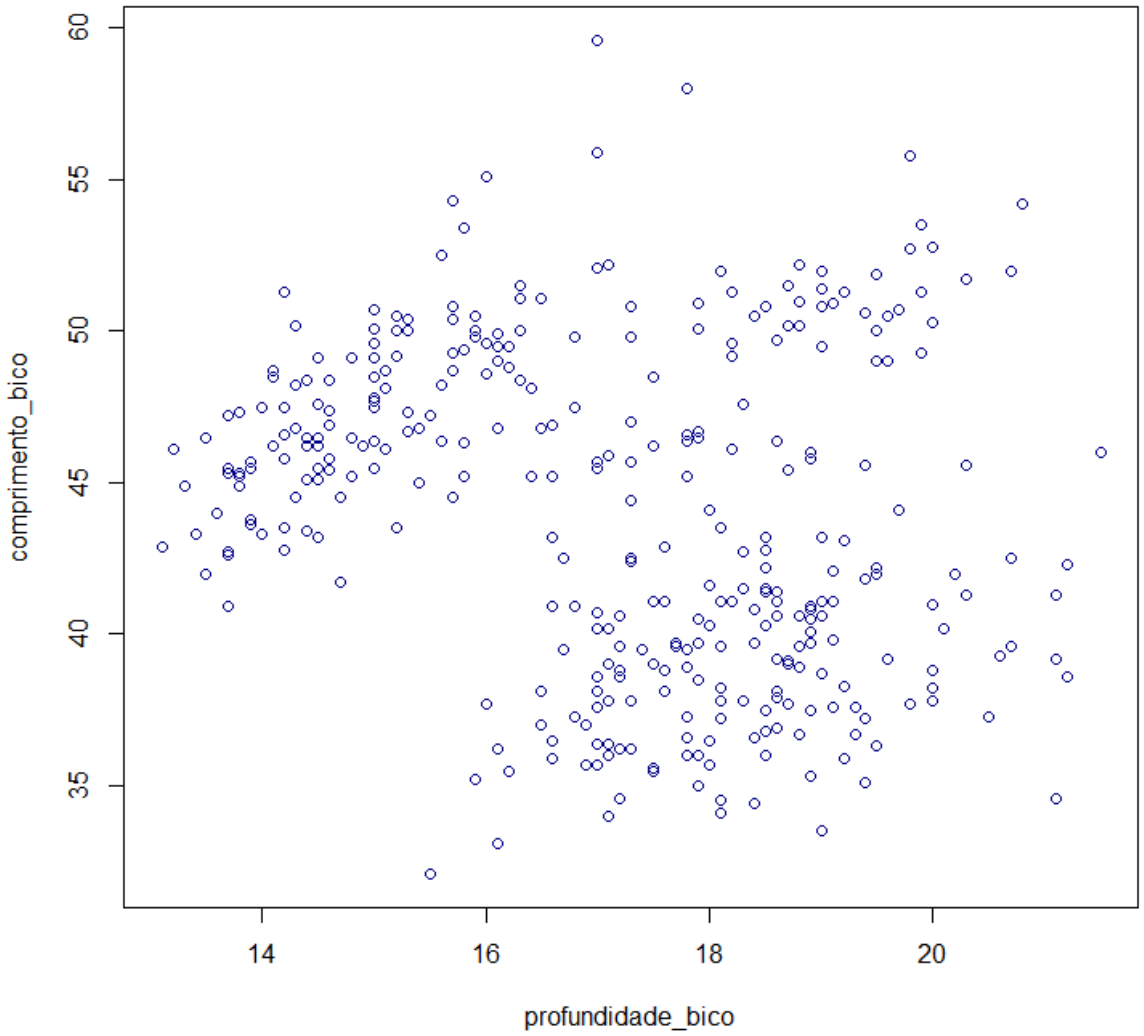

Figura 8: Plot da correlação entre profundidade do bico e comprimento do bico

O **comprimento_nadadeira** mostra uma relação moderada quanto à **profundidade_bico** bem dividida em dois grupos. Quanto mais comprida a nadadeira do pinguim, menor a profundidade do seu bico.

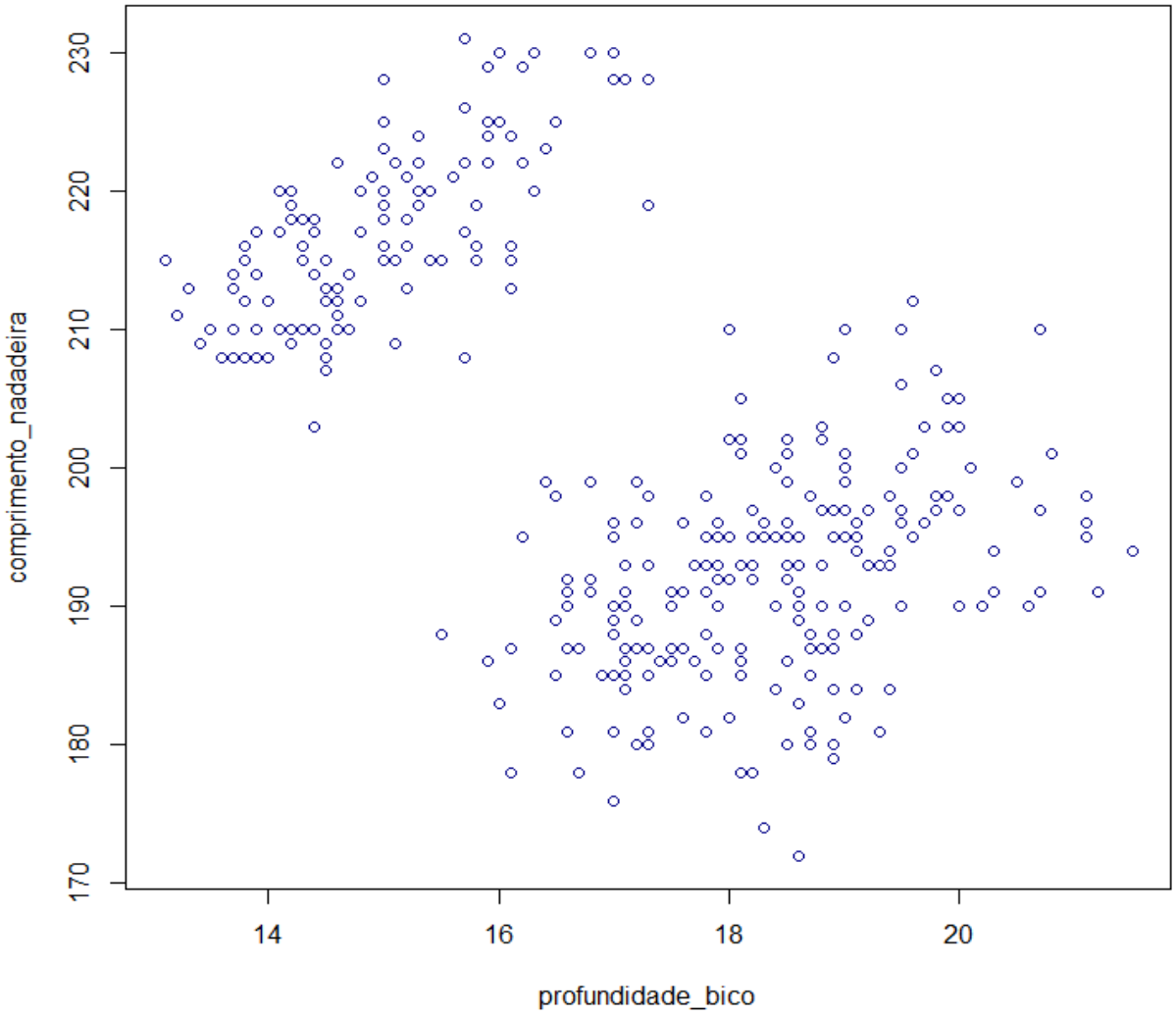

Figura 9: Plot da correlação entre profundidade do bico e comprimento da nadadeira

A variável **massa_corporal** também mostra uma relação moderada bem dividida em 2 grupos levemente dispersos. Podemos ver que, quão mais profundo o bico do pinguim, menor sua massa corporal. O contrário também é válido.

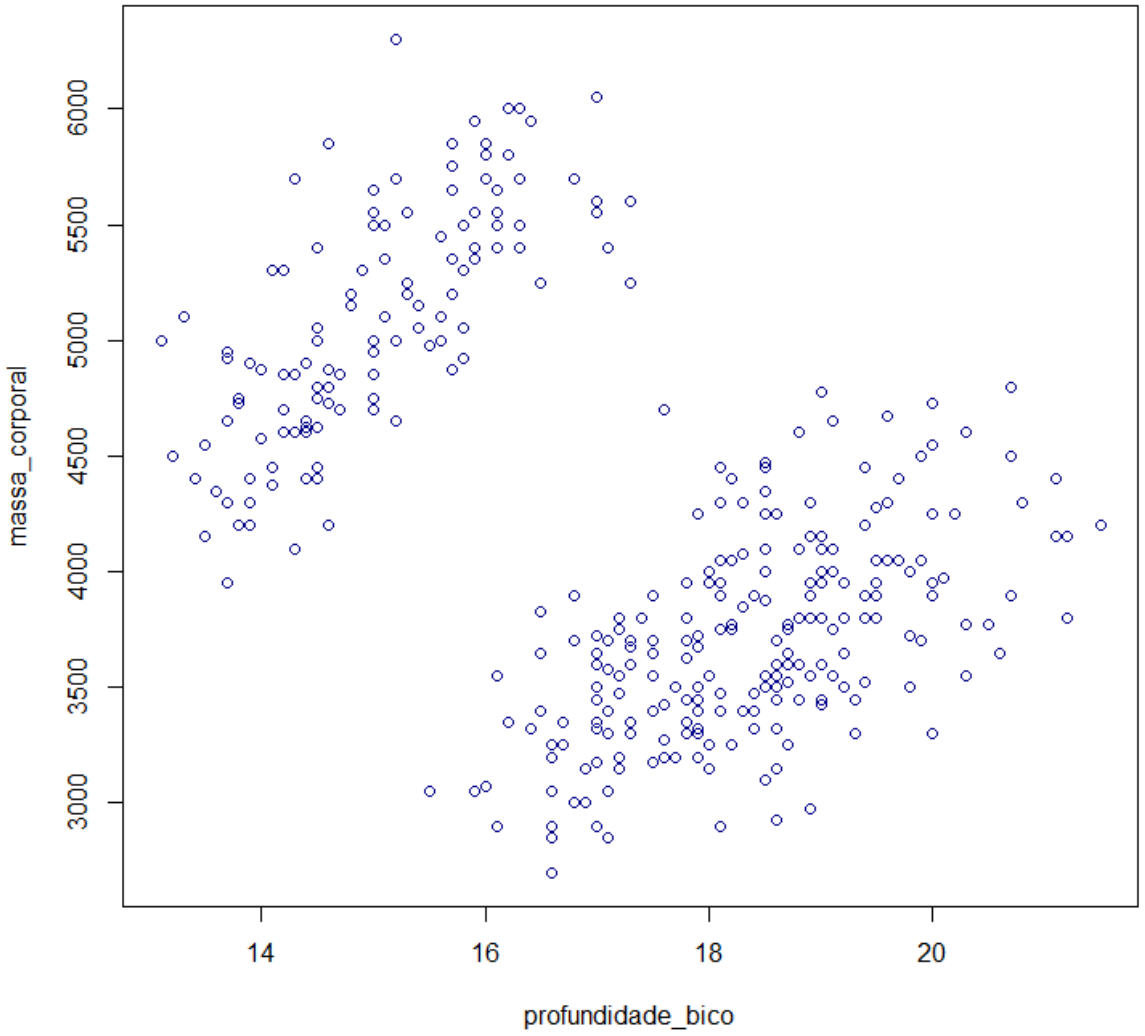

Figura 10: Plot da correlação entre profundidade do bico e massa corporal

Observando a Matriz de Correlação da figura 12 entre as variáveis quantitativas, podemos observar que **profundidade_bico** tem uma moderada correlação negativa com **massa_corporal** e **comprimento_nadadeira**. Já com **comprimento_bico** e ano, não há tanta relação.

```
> correlacao
                   profundidade_bico comprimento_bico comprimento_nadadeira massa_corporal         ano
profundidade_bico          1.0000000       -0.2286256            -0.5777917    -0.47201566 -0.04818160
comprimento_bico          -0.2286256        1.0000000             0.6530956     0.58945111  0.03265690
comprimento_nadadeira     -0.5777917        0.6530956             1.0000000     0.87297890  0.15106792
massa_corporal            -0.4720157        0.5894511             0.8729789     1.00000000  0.02186213
ano                       -0.0481816        0.0326569             0.1510679     0.02186213  1.00000000
```

Figura 11: Tabela de correlação entre as variáveis numéricas do banco de dados.

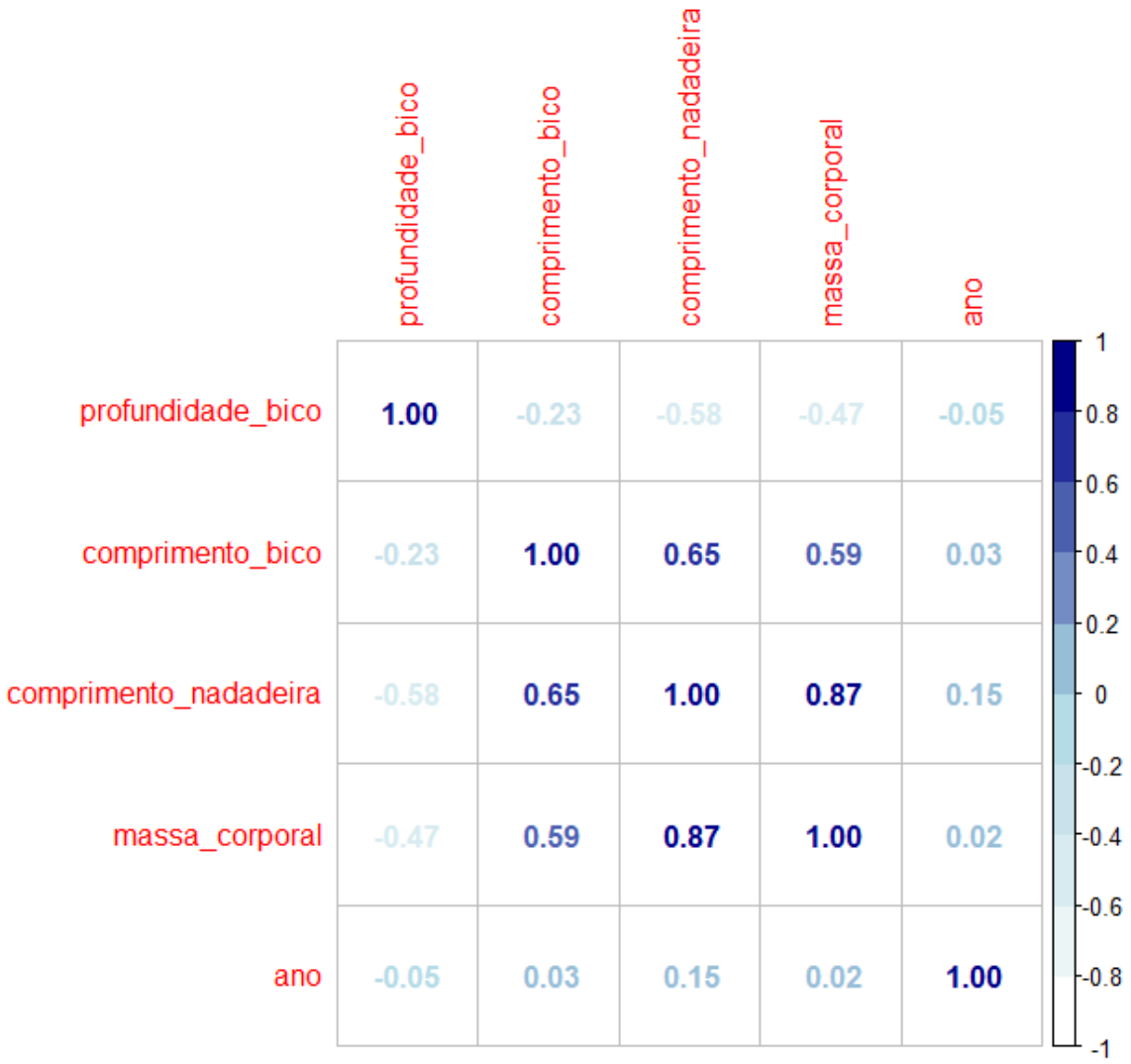

Figura 12: Matriz de correlação entre as variáveis numéricas do banco de dados.

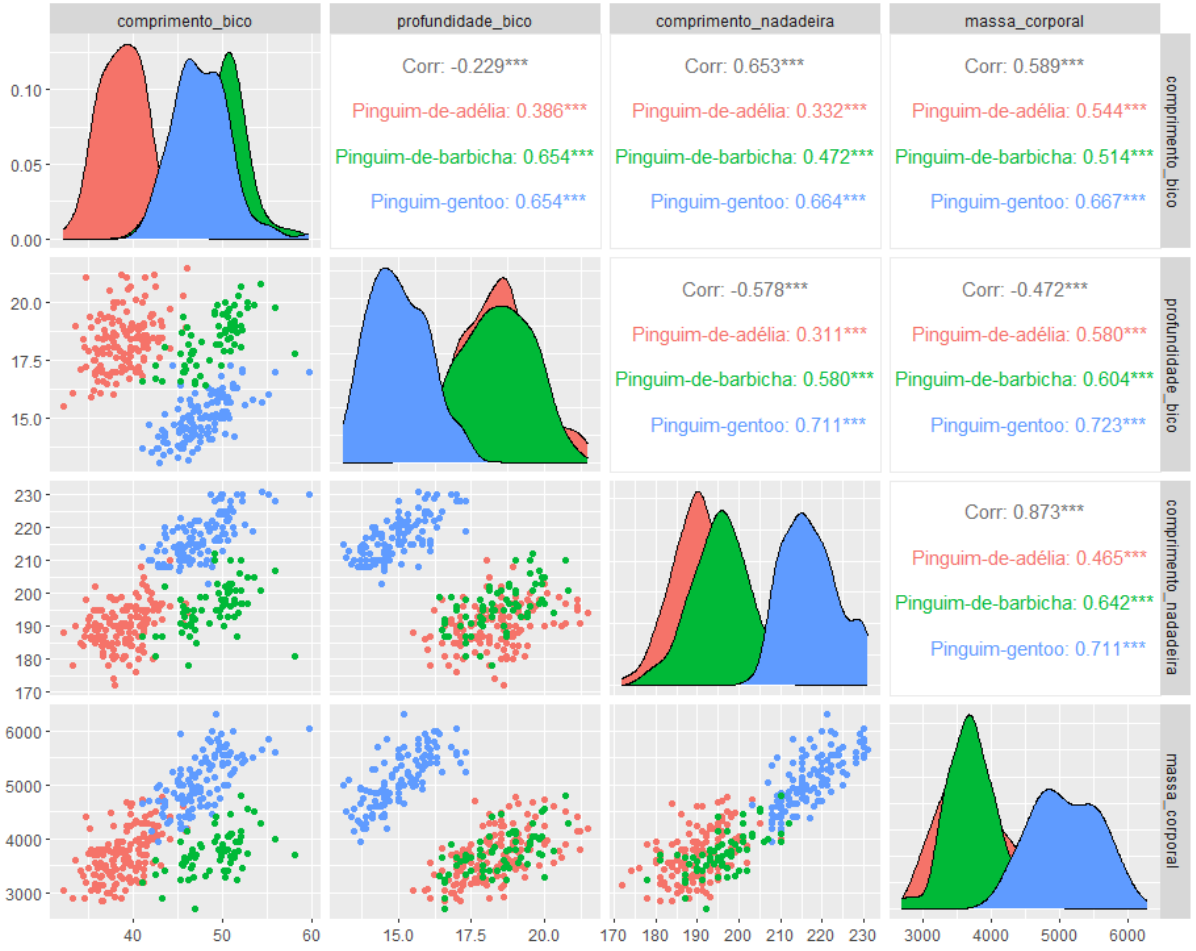

Figura 13: Matriz de gráficos de pares

Com relação à análise de correlação, altas correlações entre as variáveis independentes sugerem indícios de problemas de multicolinearidade ao ajustar um modelo de regressão linear múltipla (MRLM). Como regra geral, isso ocorre quando há correlações superiores a 0.8 ou 0.9 entre as variáveis preditoras.

Dito isto, é possível observar na figura 13 que:

A variável independente comprimento_bico:

- Apresenta correlação positiva com comprimento_nadadeira ($r = 0.653$, $p < 0.001$), sem ser superior a 0.8, não havendo indicativos de multicolinearidade.
- Apresenta correlação significativa com massa_corporal ($r = 0.589$, $p < 0.001$), sem ultrapassar os limites para preocupação com multicolinearidade.

A variável independente comprimento_nadadeira:

- Apresenta forte correlação positiva com massa_corporal ($r = 0.873$, $p < 0.001$), o que sugere que conforme o comprimento da nadadeira aumenta, a massa corporal também aumenta. Como essa correlação é superior a 0.8, há um possível indício de multicolinearidade.

Em síntese, as correlações mostram que as variáveis comprimento_nadadeira e massa_corporal têm uma relação forte, sugerindo potencial multicolinearidade se usadas juntas em um modelo de regressão para prever a profundidade do bico. No entanto, as demais variáveis não apresentam relações problemáticas com a variável resposta.

4. AJUSTE DO MODELO INICIAL

4.1 Análise da Qualidade do Modelo Ajustado de Regressão Linear Múltipla

```
Call:
lm(formula = profundidade_bico ~ . - especie - ilha - sexo -
    ano, data = pinguins_limpos)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0497 -1.1274 -0.0975  1.0387  4.3923

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          35.8005205  1.7811822  20.099  < 2e-16 ***
comprimento_bico      0.0919907  0.0207443   4.435 1.26e-05 ***
comprimento_nadadeira -0.1187744  0.0134050  -8.860  < 2e-16 ***
massa_corporal        0.0002822  0.0002187   1.290    0.198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.563 on 329 degrees of freedom
Multiple R-squared:  0.3756,    Adjusted R-squared:  0.3699
F-statistic: 65.96 on 3 and 329 DF,  p-value: < 2.2e-16
```

Figura 14: Modelo com ajuste de regressão linear múltipla

O modelo de regressão linear (figura 14) foi ajustado para prever a profundidade do bico de pinguins com base em variáveis como espécie, ilha, sexo e ano.

- Resíduos:
    - Mínimo: -3.0497 Máximo: 4.3923
- Coeficientes:
    - Intercepto: 35.80 ($p < 2e\text{-}16$)
    - Comprimento do bico: 0.09199 ($p < 1e\text{-}05$)
    - Comprimento da nadadeira: -0.11877 ($p < 2e\text{-}16$)
    - Massa corporal: 0.00028 ($p = 0.198$, não significativo)
- Estatísticas do Modelo:
    - Erro padrão residual: 1.563
    - R-quadrado: 0.3756 (37.56% da variação explicada)
    - R-quadrado ajustado: 0.3699
    - Estatística F: 65.96 ($p < 2.2e\text{-}16$, modelo significativo)

Deste modo, o modelo na figura 14 apresenta um ajuste razoável, mas com um R-quadrado abaixo de 0.4, indicando que outras variáveis podem ser relevantes. A massa corporal não contribui significativamente, e a análise de resíduos é recomendada para melhorar o modelo.

4.2 Análise da Qualidade do Modelo Ajustado sem Massa Corporal

```
Call:
lm(formula = profundidade_bico ~ comprimento_bico + comprimento_nadadeira,
    data = pinguins_limpos)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1084 -1.1321 -0.0776  1.0896  4.5009

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          34.153619   1.243384  27.468  < 2e-16 ***
comprimento_bico      0.093390   0.020737   4.504 9.28e-06 ***
comprimento_nadadeira -0.104979   0.008091 -12.975  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.565 on 330 degrees of freedom
Multiple R-squared:  0.3724,    Adjusted R-squared:  0.3686
F-statistic: 97.91 on 2 and 330 DF,  p-value: < 2.2e-16
```

Figura 15: Modelo com ajuste sem massa corporal

O modelo de regressão linear (figura 15) foi ajustado para prever a profundidade do bico de pinguins com base no comprimento do bico e no comprimento da nadadeira.

- Resíduos
    - Mínimo: -3.1084 Máximo: 4.5009
- Interpretação: A variação dos resíduos sugere que o modelo pode ter previsões que se afastam consideravelmente dos valores reais, indicando a presença de outliers.
- Coeficientes
    - Intercepto: 34.153619 ($p < 2e-16$), significativo.
    - Comprimento do bico: 0.093390 ($p < 9.28e-06$), significativo e com efeito positivo.
    - Comprimento da nadadeira: -0.104979 ($p < 2e-16$), significativo e com efeito negativo.
- Estatísticas do Modelo
    - Erro padrão residual: 1.565, indicando a média dos erros de previsão.
    - R-quadrado: 0.3724, sugerindo que aproximadamente 37.24% da variação na profundidade do bico é explicada pelas variáveis independentes.
    - R-quadrado ajustado: 0.3686, que ajusta o R-quadrado para o número de preditores.
    - Estatística F: 97.91 ($p < 2.2e-16$), indicando que o modelo é estatisticamente significativo.

O modelo apresenta um ajuste moderado, com um R-quadrado de 0.3724, o que indica que há espaço para melhorias. Os coeficientes são significativos, mas a variação explicada é relativamente baixa, sugerindo que outras variáveis podem ser relevantes. A análise de resíduos e a verificação de suposições do modelo são recomendadas para aprimorar a qualidade do ajuste.

4.3 Análise da Qualidade do Modelo Ajustado Sem o Comprimento da Nadadeira

```
Call:
lm(formula = profundidade_bico ~ comprimento_bico + massa_corporal,
    data = pinguins_limpos)

Residuals:
   Min     1Q Median     3Q    Max
-3.754 -1.204 -0.003  1.214  4.436

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      21.2781272  0.7749058  27.459  < 2e-16 ***
comprimento_bico  0.0273732  0.0215806   1.268    0.206
massa_corporal   -0.0012639  0.0001466  -8.624 2.76e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.737 on 330 degrees of freedom
Multiple R-squared:  0.2266,    Adjusted R-squared:  0.2219
F-statistic: 48.34 on 2 and 330 DF,  p-value: < 2.2e-16
```

Figura 16: Modelo com ajuste sem o comprimento da nadadeira

O modelo de regressão linear foi ajustado para prever a profundidade do bico de pinguins com base no comprimento do bico e no comprimento da nadadeira.

- Resíduos:
    - Mínimo: -3.1084 Máximo: 4.5009
- Interpretação: A variação dos resíduos sugere que o modelo pode ter previsões que se afastam consideravelmente dos valores reais, indicando a presença de outliers.
- Coeficientes:
    - Intercepto: 34.153619 ($p < 2e\text{-}16$), significativo.
    - Comprimento do bico: 0.093390 ($p < 9.28e\text{-}06$), significativo e com efeito positivo.
    - Comprimento da nadadeira: -0.104979 ($p < 2e\text{-}16$), significativo e com efeito negativo.
- Estatísticas do Modelo
    - Erro padrão residual: 1.565, indicando a média dos erros de previsão.
    - R-quadrado: 0.3724, sugerindo que aproximadamente 37.24% da variação na profundidade do bico é explicada pelas variáveis independentes.
    - R-quadrado ajustado: 0.3686, que ajusta o R-quadrado para o número de preditores.

- Estatística F: 97.91 ($p < 2.2e-16$), indicando que o modelo é estatisticamente significativo.

O modelo apresenta um ajuste moderado, com um R-quadrado de 0.3724, o que indica que há espaço para melhorias. Os coeficientes são significativos, mas a variação explicada é relativamente baixa, sugerindo que outras variáveis podem ser relevantes. A análise de resíduos e a verificação de suposições do modelo são recomendadas para aprimorar a qualidade do ajuste.

4.4 Seleção do Melhor Modelo

O modelo ajustado de regressão linear lúltipla com todas as variáveis (incluindo massa corporal) é o melhor dado que possui um R-quadrado um pouco mais alto e um erro padrão ligeiramente inferior. No entanto, a presença de uma variável não significativa (massa corporal) sugere que talvez o modelo possa ser otimizado mais ainda.

Portanto, idealmente, o modelo sem a massa corporal pode ser potencialmente mais útil no contexto de previsão, mesmo que tenha um R-quadrado um pouco mais baixo, já que mantém a simplicidade do modelo e elimina uma variável não significativa.

Assim, uma análise de validação cruzada ou um teste com dados independentes poderia ajudar a decidir qual modelo realmente se comporta melhor em um cenário prático.

```
> #AIC - MODELO 2
> paste("AIC do modelo 2: ", AIC(modelo_2))
[1] "AIC do modelo 2:  1248.18633653088"
> #AIC - MODELO 3
> paste("AIC do modelo 3: ", AIC(modelo_3))
[1] "AIC do modelo 3:  1317.76917408576"
> #BIC - MODELO 2
> paste("BIC do modelo 2: ", BIC(modelo_2))
[1] "BIC do modelo 2:  1263.41890649081"
> #BIC - MODELO 3
> paste("BIC do modelo 3: ", BIC(modelo_3))
[1] "BIC do modelo 3:  1333.00174404568"
```

Figura 17: AIC e BIC dos modelos.

Ao comparar os dois modelos após a remoção individual das variáveis que apresentam multicolinearidade, nota-se que o modelo 2 exibe o valor mais baixo de AIC, tornando-o a opção mais adequada. É importante ressaltar que, como regra geral, uma diferença superior a 10 entre os valores de AIC é considerada um indicativo relevante de uma distinção significativa entre os modelos.

De forma similar à interpretação do AIC, a análise do BIC também aponta que o modelo 2 possui o menor valor, reforçando sua preferência. Novamente, aplica-se a regra prática de que uma diferença maior que 10 nos valores de BIC sugere uma distinção significativa entre os modelos avaliados.

Portanto, tanto o critério AIC quanto o BIC convergem para a escolha do modelo 2 como o mais apropriado para a análise em questão, demonstrando consistência na seleção do modelo preferível.

5. VERIFICAÇÃO DOS PRESSUPOSTOS DO MODELO

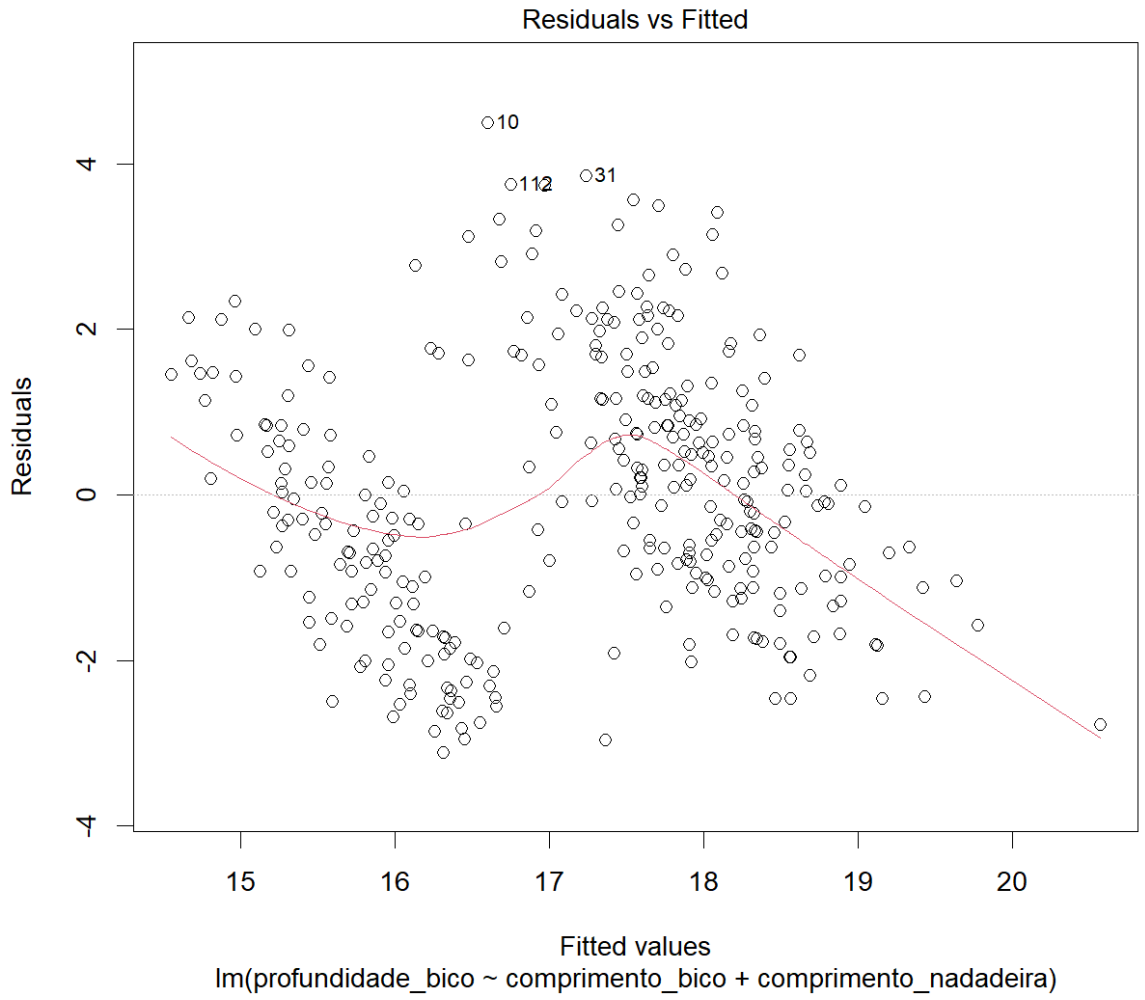

Figura 18: Resíduos vs Valores Ajustados: Mostra a relação entre resíduos e valores previstos pelo modelo.

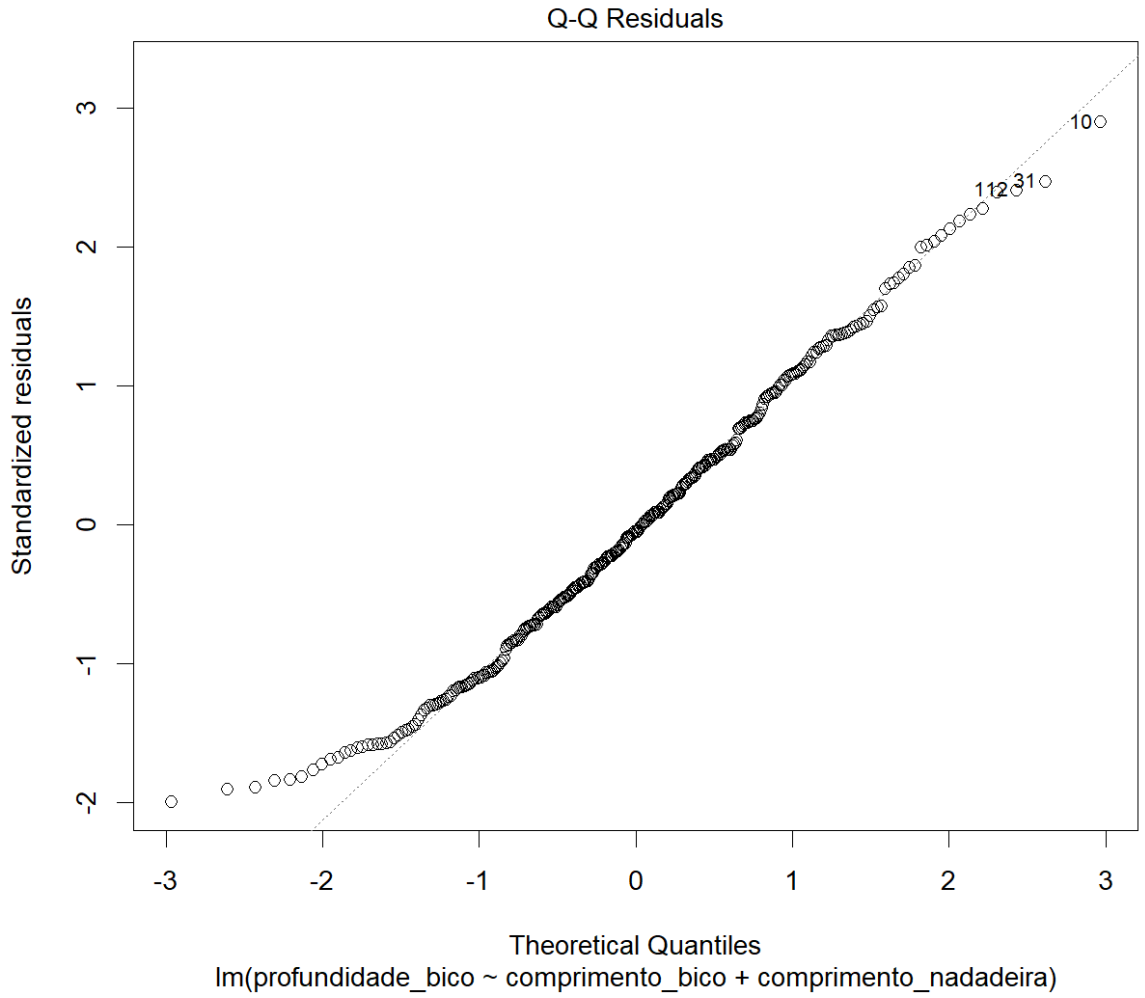

Figura 19: Q-Q de Resíduos: Compara a distribuição dos resíduos padronizados com a distribuição normal teórica.

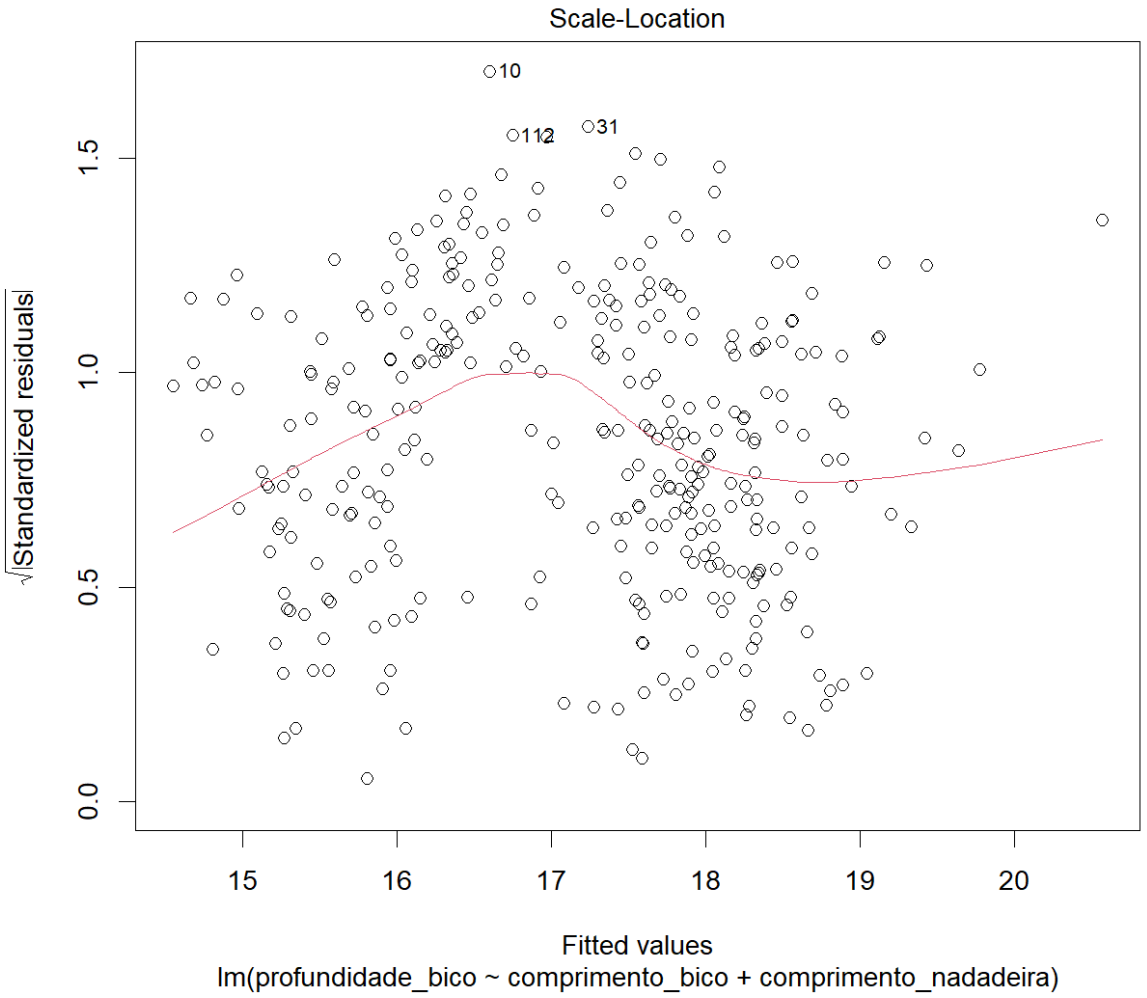

Figura 20: Scale-Location: Avalia a homocedasticidade dos resíduos padronizados.

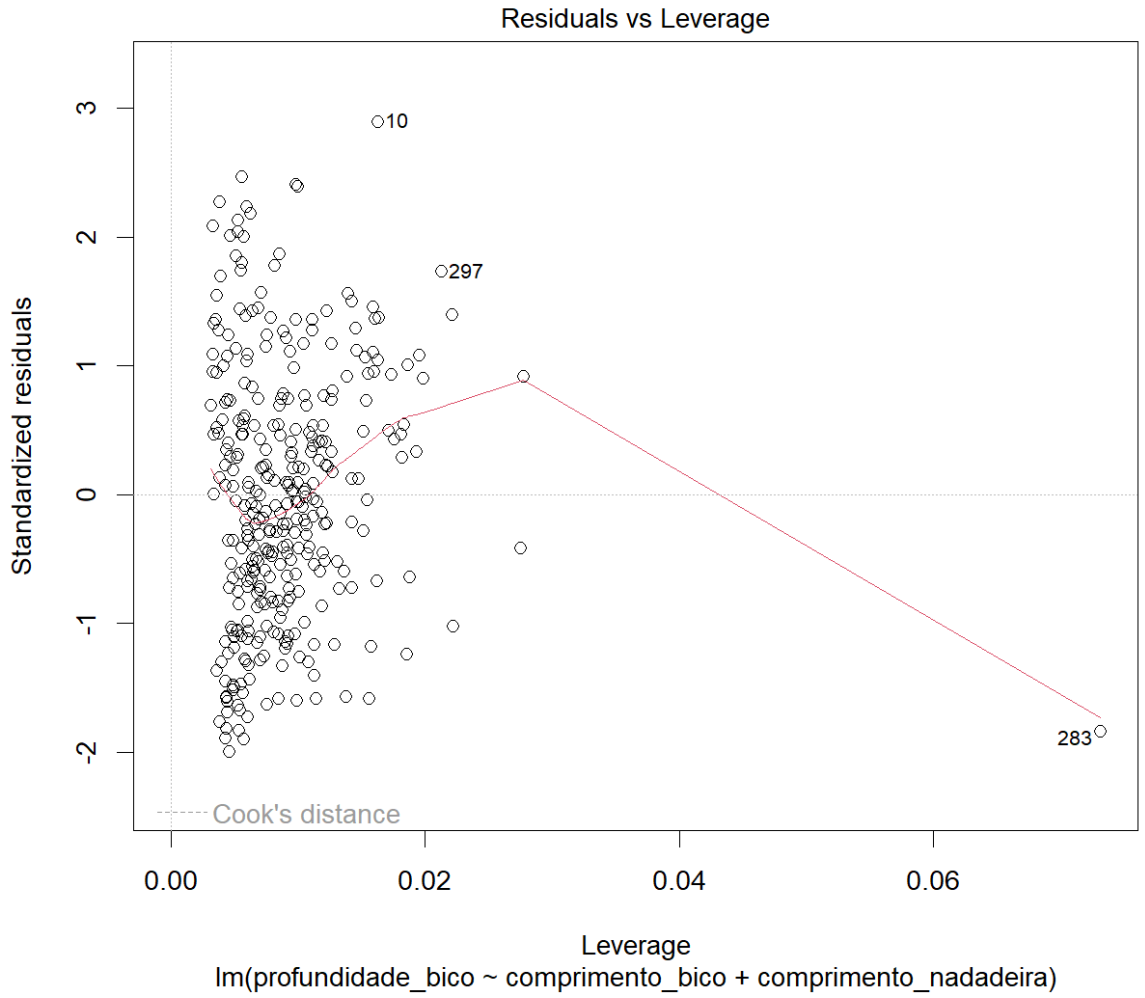

Figura 21: Resíduos vs Leverage: Identifica pontos com alta influência e outliers potenciais.

5.1 Normalidade dos resíduos:

O gráfico Q-Q de resíduos mostra que os pontos seguem aproximadamente uma linha reta, com alguns desvios nas extremidades. Isso sugere que os resíduos têm uma distribuição próxima da normal, embora haja alguns indícios de caudas mais pesadas (pontos 10, 31 e outros nas extremidades se afastam da linha).

5.2 Homocedasticidade:

O gráfico de Resíduos vs Valores Ajustados não mostra um padrão claro de aumento ou diminuição da variância dos resíduos ao longo dos valores ajustados. Isso sugere que a homocedasticidade é razoavelmente satisfeita, embora haja uma leve tendência curvilínea na linha vermelha.

5.3 Interpretação dos resíduos:

Não há um padrão claro nos resíduos que indique não-linearidade significativa.

Alguns pontos (10, 31, 297, 283) parecem ser outliers potenciais, com resíduos maiores.

O gráfico de Scale-Location mostra uma distribuição relativamente uniforme dos resíduos padronizados, mas com uma leve tendência ascendente.

O gráfico de Resíduos vs Leverage indica que o ponto 283 tem alta alavancagem, podendo ser influente no modelo.

6. INTERPRETAÇÕES DO MODELO SELECIONADO

Ajustamos um modelo linear (estimado usando OLS) para prever a profundidade_bico com comprimento_bico e comprimento_nadadeira (fórmula: profundidade_bico ~ comprimento_bico + comprimento_nadadeira). O modelo explica uma proporção de variância estatisticamente significativa e substancial ($R^2 = 0,37$, $F(2, 330) = 97,91$, $p < 0,001$, R^2 ajustado = 0,37). O intercepto do modelo, correspondente a comprimento_bico = 0 e comprimento_nadadeira = 0, é de 34,15 (IC 95% [31,71, 36,60], $t(330) = 27,47$, $p < 0,001$). Dentro deste modelo:

- O efeito do comprimento_bico é estatisticamente significativo e positivo (beta = 0,09, IC 95% [0,05, 0,13], $t(330) = 4,50$, $p < 0,001$; beta padrão = 0,26, IC 95% [0,15, 0,37])

- O efeito do comprimento_nadadeira é estatisticamente significativo e negativo (beta = -0,10, IC 95% [-0,12, -0,09], t(330) = -12,97, p < 0,001; beta padrão = -0,75, IC 95% [-0,86, -0,63])

Os parâmetros padronizados foram obtidos ajustando o modelo em uma versão padronizada do conjunto de dados. Os intervalos de confiança de 95% (ICs) e os valores de p foram calculados usando uma aproximação da distribuição t de Wald.

7. ESTIMATIVA E PREVISÃO

Os intervalos de confiança fornecem uma estimativa do valor médio da variável resposta para as duas combinações de preditores especificados (comprimento do bico e da nadadeira). Para a **primeira combinação**, com **comprimento_bico = 45** e **comprimento_nadadeira = 190**, a estimativa pontual (média) foi estimada em **18.41**, e o intervalo de confiança a 95% foi **[18.14, 18.67]**.

Isso significa que, para pinguins que possuem um comprimento de bico de 45 mm e um comprimento de nadadeira de 190 mm, a **média esperada da variável resposta** é **18.41**. O **intervalo de confiança** indica que, com 95% de certeza, a verdadeira média da variável resposta para pinguins com essas características está entre **18.14 e 18.67**.

Para a **segunda combinação**, com **comprimento_bico = 50** e **comprimento_nadadeira = 210**, a estimativa pontual (média) foi **16.78**, com intervalo de confiança **[16.53, 17.03]**.

```
> novas.preditoras <- data.frame(comprimento_bico = c(45, 50),comprimento_nadadeira = c(190, 210))
> previsoes <- predict(modelo_2, newdata = novas.preditoras, interval = "confidence")
> previsoes
       fit      lwr      upr
1 18.41023 18.14569 18.67476
2 16.77760 16.52630 17.02891
```

Figura 22: Resultado do Intervalo de Confiança

Os intervalos de predição são usados para prever uma nova observação individual. Esses intervalos são mais amplos porque incorporam a variabilidade entre indivíduos. No caso da **primeira combinação**, com **comprimento_bico = 45mm** e **comprimento_nadadeira = 190mm**, a predição pontual foi **18.41**, com um intervalo de predição de **[13.32, 21.50]**. A predição pontual indica que, para um pinguim com um comprimento de bico de 45 mm e comprimento de nadadeira de 190 mm, a profundidade do bico esperada é de **18.41 mm**. Contudo, devido à variabilidade entre indivíduos, a profundidade real do bico de um novo pinguim com essas mesmas medidas pode variar. O intervalo de predição nos informa que, com **95% de confiança**, a profundidade do bico estará entre **13.32 mm** e **21.50 mm**.

Para a **segunda combinação**, com **comprimento_bico = 50mm** e **comprimento_nadadeira = 210mm**, a predição pontual foi **16.78**, e o intervalo de predição foi **[13.69, 19.87]**. Para um pinguim com um comprimento de bico de 50 mm e nadadeira de 210 mm, a predição pontual da profundidade do bico foi de **16.78 mm**. Esse valor representa a melhor estimativa fornecida pelo modelo. O intervalo de predição de **[13.69 mm, 19.87 mm]** indica que, com **95% de confiança**, a profundidade real do bico de um novo pinguim com essas medidas estará dentro desse intervalo. Isso reflete a incerteza natural de novas observações individuais, que pode ser influenciada por outros fatores não capturados pelo modelo.

```
> novas.preditoras <- data.frame(comprimento_bico = c(45, 50),comprimento_nadadeira = c(190, 210))
> previsoes <- predict(modelo_2, newdata = novas.preditoras, interval = "prediction")
> previsoes
       fit      lwr      upr
1 18.41023 15.32073 21.49972
2 16.77760 13.68921 19.86599
```

Figura 23: Resultado do Intervalo de Predição

8. CONCLUSÃO

Nesta análise abrangente do banco de dados "Pinguins", com o objetivo de identificar as variáveis que influenciam a profundidade do bico dos pinguins das ilhas da Antártica. Através da aplicação de modelos de Regressão Linear Múltipla (MRLM), foram exploradas as relações entre a profundidade do bico e variáveis como comprimento do bico, comprimento da nadadeira e massa corporal.

Os resultados indicaram que, embora o modelo inicial, que incluía todas as variáveis, apresentasse um R-quadrado de 0.3756, a massa corporal não se mostrou uma variável significativa. A análise subsequente, excluindo a massa corporal, resultou em um modelo com R-quadrado de 0.3724, que ainda explicava uma proporção considerável da variação na profundidade do bico, mas com um ajuste mais simples e eficaz.

A verificação dos pressupostos do modelo revelou que os resíduos apresentavam uma distribuição próxima da normalidade e que a homocedasticidade estava razoavelmente satisfeita, embora alguns outliers e pontos de alta alavancagem tenham sido identificados. Isso sugere que, embora o modelo seja estatisticamente significativo, há espaço para melhorias, possivelmente através da inclusão de outras variáveis relevantes ou da aplicação de técnicas de ajuste mais robustas.

Além disso, a análise de estimativas e previsões forneceu intervalos de confiança e predição que podem ser úteis para futuras observações, permitindo uma melhor compreensão das características dos pinguins em relação à profundidade do bico.

Em suma, a pesquisa não apenas contribuiu para o entendimento das variáveis que afetam a profundidade do bico dos pinguins, mas também destacou a importância de uma análise cuidadosa e da validação de modelos em estudos estatísticos.