UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO

DISCIPLINA: Estatística Aplicada
PERÍODO: 2023.2
DOCENTE: Amanda dos Santos Gomes
DISCENTE: Caroline de Oliveira Cordeiro
MATRÍCULA: 121111059

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

Campina Grande
Maio de 2024

Sumário

1. Introdução

Este trabalho visa analisar o banco de dados “Pinguins” que contém informações sobre pinguins das ilhas da Antártica, com o objetivo de, através do treinamento de um Modelo de Regressão Linear, mensurar quais variáveis explicam o crescimento ou decaimento da variável massa corporal, que diz sobre a massa corporal dos pinguins estudados. Na sessão 2, podemos ver com mais detalhes os dados encontrados nesse banco de dados, bem como as normalizações necessárias para que fosse possível trabalhar com eles. Na sessão 3, temos a análise descritiva dos dados, dando foco na relação entre a variável resposta (massa corporal) e as variáveis explicativas que se destacam. Na sessão 4 temos todo o processo de encontrar um modelo ajustável que fosse o mais significativo para nosso trabalho. Na sessão 5, temos o treino do modelo e o processo de análise da corretude do mesmo. Por fim, na sessão 6 temos as conclusões geradas pela análise do modelo ajustado.

2. Descrição dos Dados

O banco de dados pinguins contém informações sobre 344 pinguins adultos de 3 espécies diferentes, distribuídos em 3 ilhas da Antártida. No total, pode-se encontrar 8 variáveis observadas, 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.

A variável resposta que será observada aqui é **massa_corporal** e as demais variáveis serão explicativas.

2.1. Normalizações

O banco de dados original tem 344 observações e 8 colunas, porém 11 das observações tinham não continham indicação do sexo do penguim; além disso, 2 dessas 11 observações também não continham: **comprimento_bico**, **profundidade_bico**, **comprimento_nadadeira** e **massa_corporal**. Estas 11 observações foram excluídas para que a falta de valores não interferisse na eficácia do modelo, restando assim 333 observações hábeis para trabalho.

3. Análise Descritiva

Em uma primeira análise utilizando um plot que relaciona todas as variáveis entre si, apresentado na figura 1, podemos ver que existe uma relação de linear crescimento entre as variáveis explicativas **comprimento_bico**, **profundidade_bico** e **comprimento_nadadeira** com a variável resposta **massa_corporal**. À medida que essas medidas crescem, pode-se notar o crescimento da massa. Também é possível notar a relação entre entre **massa_corporal** e as outras variáveis: **especie**, **ilha, ano** e **sexo**.

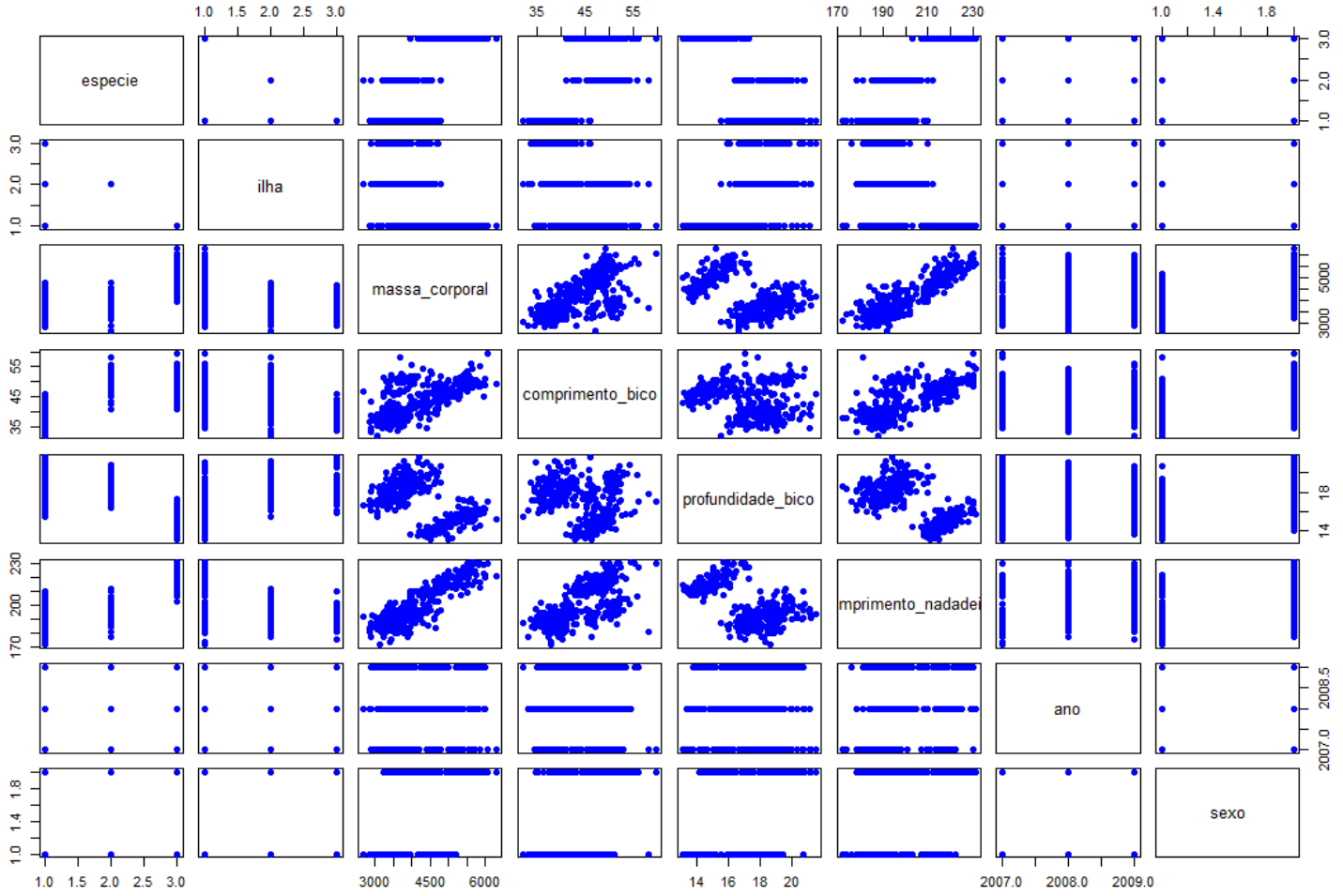

Figura 1: plot relacionando todas as variáveis do banco.

Olhando mais detalhadamente para a correlação entre as outras variáveis, a variável **profundidade_bico** mostra uma relação bem dividida em 2 grupos levemente dispersos, como pode-se notar na figura 2. Podemos ver que, quão mais profundo o bico do pinguim, menor sua massa corporal. O contrário também é válido.

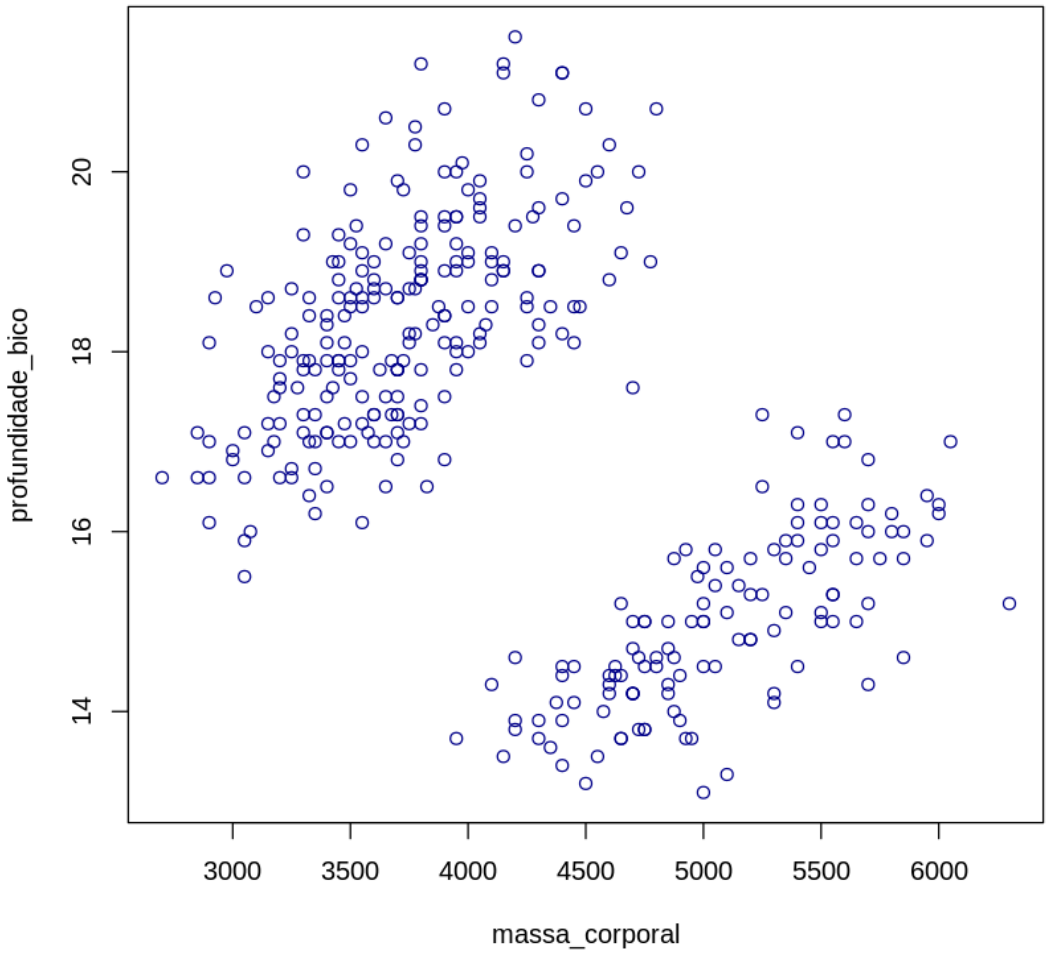

Figura 2: Diagrama de dispersão entre as variáveis profundidade_bico e massa_corporal.

Na figura 2, vemos que o **comprimento_bico** mostra um comportamento mais linear com relação a massa. Quanto maior o comprimento do bico do pinguim, maior sua massa corporal.

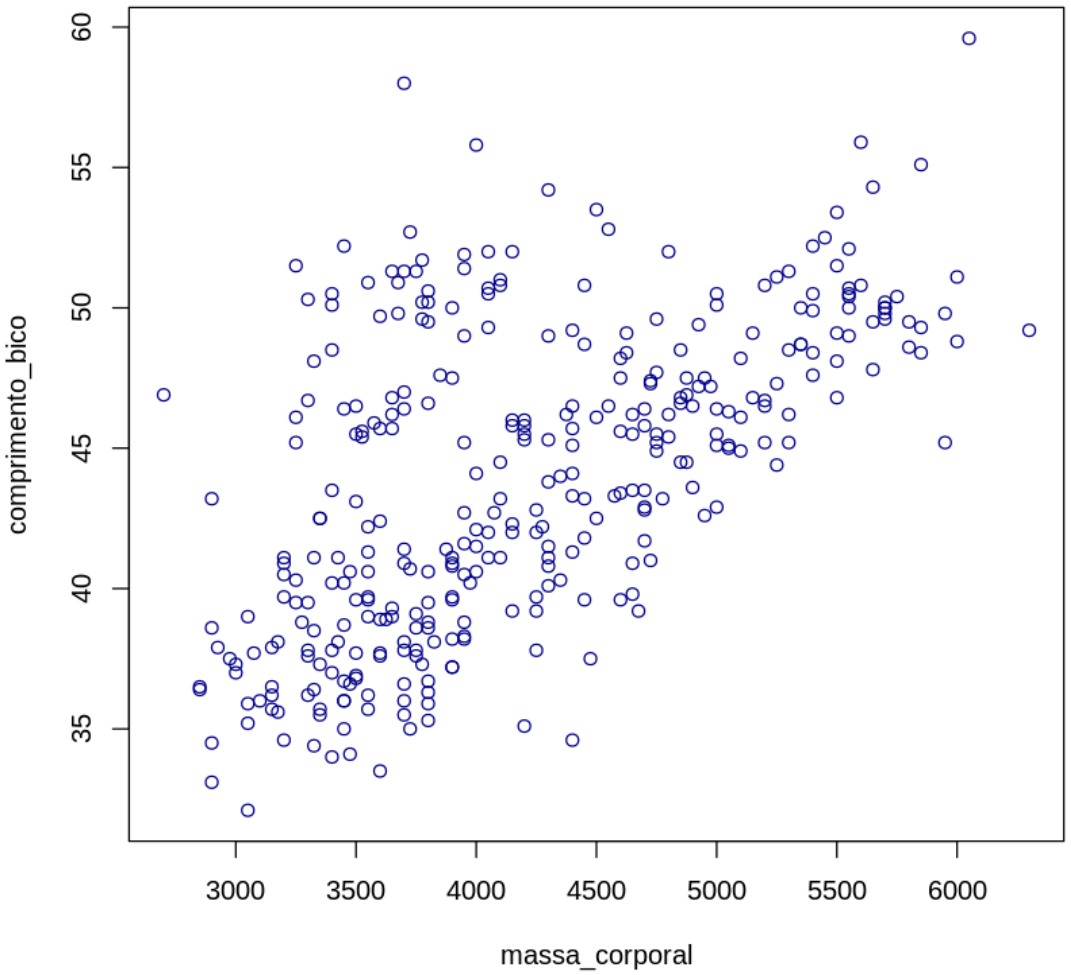

Figura 3: Diagrama de dispersão entre as variáveis comprimento_bico e massa_corporal.

Já o **comprimento_nadadeira** mostra um comportamento mais linear com relação à **massa_corporal**, mostrando um crescimento linearmente positivo apresentado na figura 4. Quanto mais comprida a nadadeira do pinguim, maior sua massa corporal.

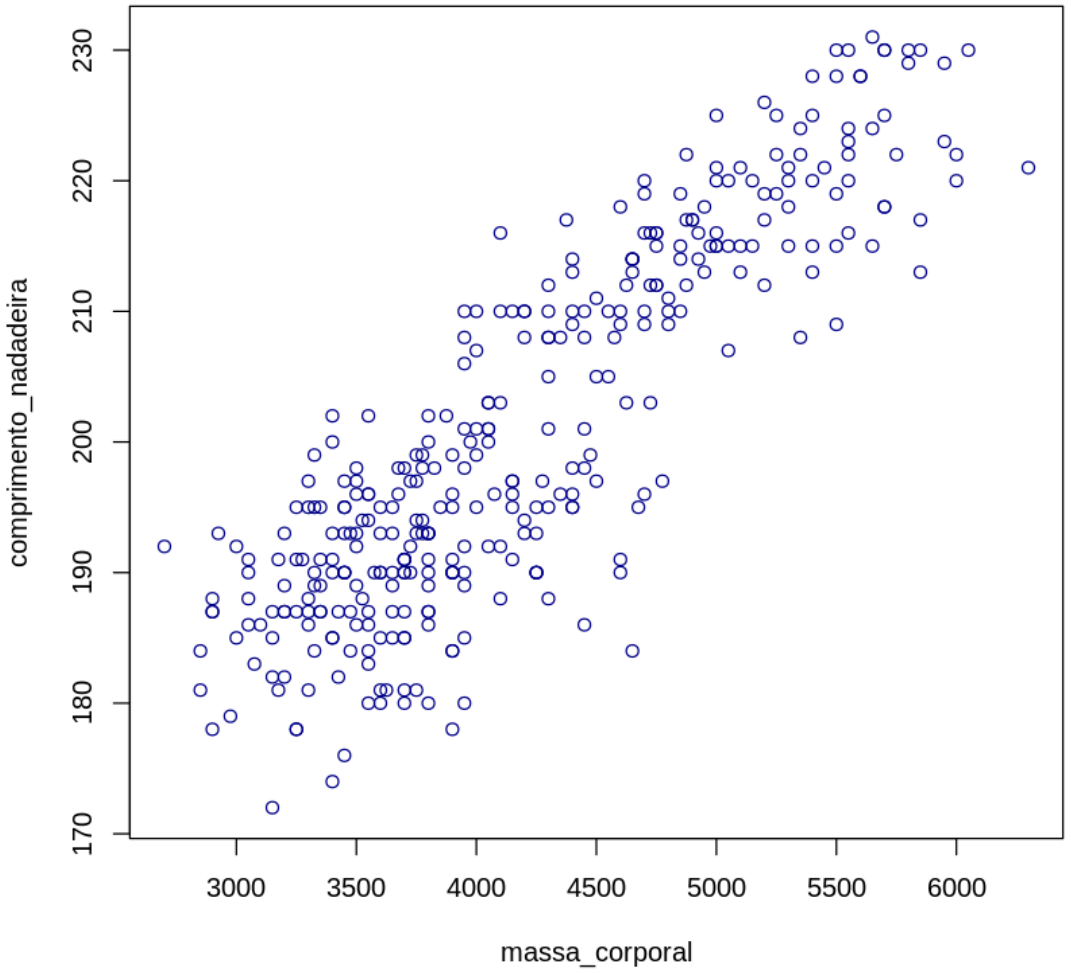

Figura 4: Diagrama de dispersão entre as variáveis comprimento_nadadeira e massa_corporal.

- Agora, observando a Matriz de Correlação da figura 5 entre as variáveis quantitativas, podemos observar que **massa_corporal** e **comprimento_nadadeira** tem uma forte correlação positiva, sendo próxima de 0.9. Há também uma relação média entre **massa_corporal** e **comprimento_bico**, porém não chega a ser tão alta sendo aproximadamente 0.6. Já com **profundidade_bico** e **ano**, não há tanta relação.

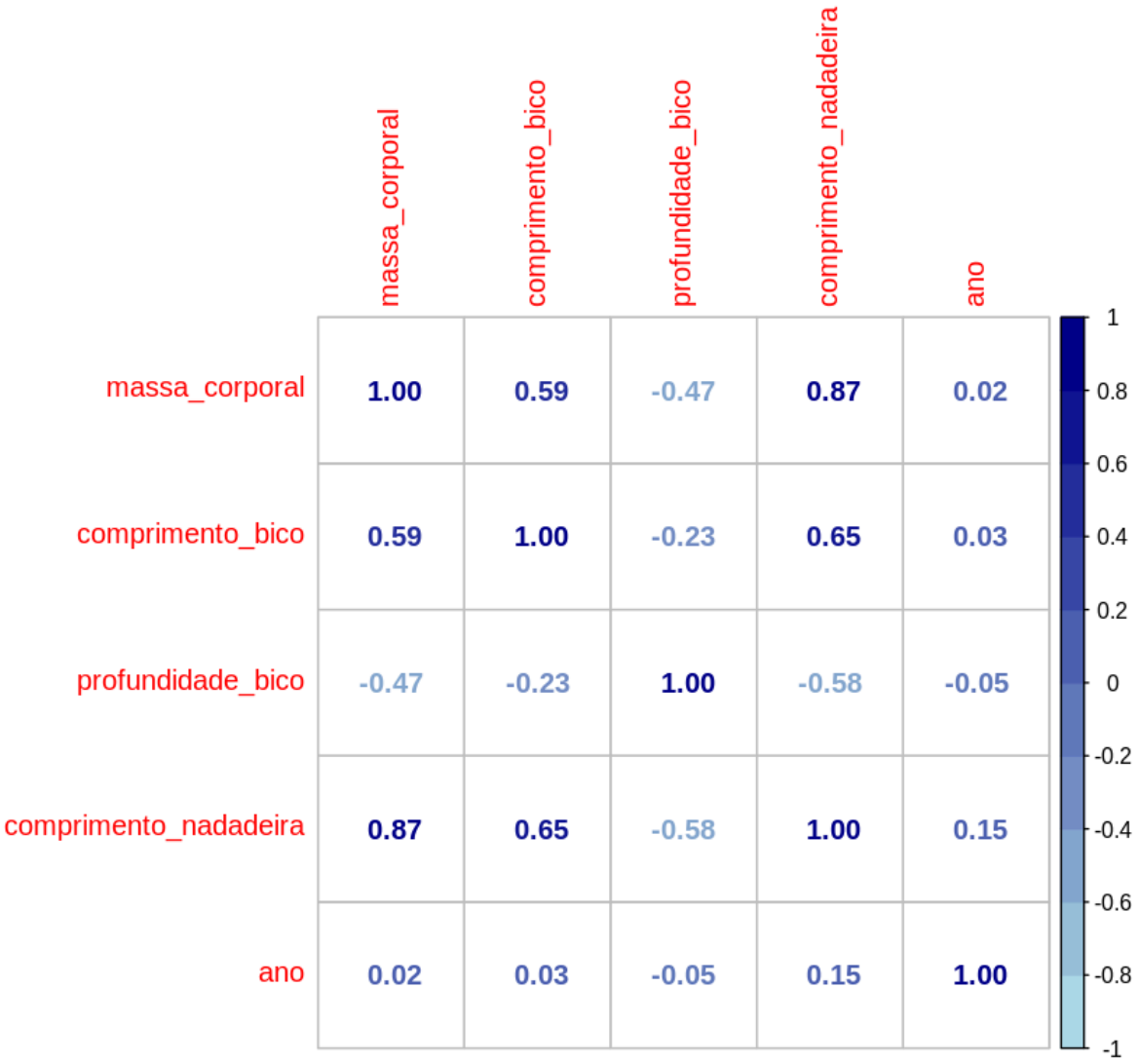

Figura 5: Matriz de correlação entre as variáveis numéricas do banco de dados.

4. O modelo

4.1. Modelo Inicial

Inicialmente, foi testada a construção do modelo utilizando todas as variáveis com o intercepto e sem o intercepto. Já o modelo com intercepto, mostrado na figura 6, mostrou R^2 mais baixo, de 87,68%. Ainda assim, o intercepto se mostra significativo.

```
Residuals:
    Min      1Q  Median      3Q     Max
-809.70 -180.87   -6.25  176.76  864.22

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                84087.945  41912.019   2.006  0.04566 *
especiePinguim-de-barbicha  -282.539     88.790  -3.182  0.00160 **
especiePinguim-gentoo        890.958    144.563   6.163 2.12e-09 ***
ilhaDream                    -21.180     58.390  -0.363  0.71704
ilhaTorgersen                -58.777     60.852  -0.966  0.33482
comprimento_bico              18.964      7.112   2.667  0.00805 **
profundidade_bico             60.798     20.002   3.040  0.00256 **
comprimento_nadadeira         18.504      3.128   5.915 8.46e-09 ***
ano                          -42.785     20.949  -2.042  0.04194 *
sexomacho                    378.977     48.074   7.883 4.95e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 286.5 on 323 degrees of freedom
Multiple R-squared:  0.8768, Adjusted R-squared:  0.8734
F-statistic: 255.4 on 9 and 323 DF,  p-value: < 2.2e-16
```

Figura 6: modelo com intercepto e com todas as variáveis.

Como se pode observar na figura 7, o modelo aplicado sem o intercepto teve um R^2 de 99,57%, com uma diferença de apenas 0,02% no ajuste.

```
Residuals:
    Min      1Q  Median      3Q     Max
-809.70 -180.87   -6.25  176.76  864.22

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
especiePinguim-de-adélia   84087.945  41912.019   2.006  0.04566 *
especiePinguim-de-barbicha 83805.405  41901.594   2.000  0.04633 *
especiePinguim-gentoo      84978.903  41865.444   2.030  0.04320 *
ilhaDream                    -21.180     58.390  -0.363  0.71704
ilhaTorgersen                -58.777     60.852  -0.966  0.33482
comprimento_bico              18.964      7.112   2.667  0.00805 **
profundidade_bico             60.798     20.002   3.040  0.00256 **
comprimento_nadadeira         18.504      3.128   5.915 8.46e-09 ***
ano                          -42.785     20.949  -2.042  0.04194 *
sexomacho                    378.977     48.074   7.883 4.95e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 286.5 on 323 degrees of freedom
Multiple R-squared:  0.9957, Adjusted R-squared:  0.9955
F-statistic:  7409 on 10 and 323 DF,  p-value: < 2.2e-16
```

Figura 7: modelo sem intercepto e com todas as variáveis.

4.2. Modelo Ajustado

Após todos os testes de ajuste, foram encontrados dois modelos que são equivalentes. Ambos utilizam as variáveis **comprimento_bico**, **comprimento_nadadeira**, **profundidade_bico, espécie** e **ano**, tendo como principal diferença o intercepto, que aparece em uma e não na outra. O modelo com intercepto, na figura 8, demonstra um R^2 de 85,21% e seu ajuste 84,94%. Já o modelo sem o intercepto, na figura 9, tem R^2 de 99,48% e ajuste de 99,47%. Como ambos os modelos estavam bons, foi verificado o Critério de Informação de Akaide - AIC - para verificar a acurácia deles, e foi notada uma diferença de apenas 0,00000000002 em ambos os modelos. Sendo assim, ambos os modelos passaram para a fase de treino e teste para encontrarmos os melhores a partir de métricas dos erros dos modelos.

```
Residuals:
    Min      1Q  Median      3Q     Max
-791.83 -195.71  -29.15  198.59  970.50

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                105356.726  45399.715   2.321   0.0209 *
comprimento_bico              40.207      7.177   5.602 4.51e-08 ***
comprimento_nadadeira         22.928      3.307   6.933 2.22e-11 ***
profundidade_bico            131.193     19.521   6.721 8.09e-11 ***
especiePinguim-de-barbicha  -520.935     82.476  -6.316 8.78e-10 ***
especiePinguim-gentoo        852.874    148.220   5.754 2.01e-08 ***
ano                          -54.768     22.677  -2.415   0.0163 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 312.5 on 326 degrees of freedom
Multiple R-squared:  0.8521, Adjusted R-squared:  0.8494
F-statistic: 313.1 on 6 and 326 DF,  p-value: < 2.2e-16

'AIC do modelo com intercepto:  4779.76412091895'
```

Figura 8: modelo ajustado com o intercepto.

```
Residuals:
    Min      1Q  Median      3Q     Max
-791.83 -195.71  -29.15  198.59  970.50

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
comprimento_bico              40.207      7.177   5.602 4.51e-08 ***
comprimento_nadadeira         22.928      3.307   6.933 2.22e-11 ***
profundidade_bico            131.193     19.521   6.721 8.09e-11 ***
especiePinguim-de-adélia    105356.726  45399.715   2.321   0.0209 *
especiePinguim-de-barbicha  104835.791  45389.863   2.310   0.0215 *
especiePinguim-gentoo       106209.601  45354.020   2.342   0.0198 *
ano                             -54.768     22.677  -2.415   0.0163 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 312.5 on 326 degrees of freedom
Multiple R-squared:  0.9948, Adjusted R-squared:  0.9947
F-statistic:  8892 on 7 and 326 DF,  p-value: < 2.2e-16

'AIC do modelo sem o intercepto:  4779.76412091893'
```

Figura 9: modelo ajustado sem o intercepto.

Como é possível observar, os resíduos de ambos os modelos também são muito parecidos. Observando os gráficos de dispersão na figura 10, podemos observar que há de fato uma grande dispersão demonstrando que há evidências de homocedasticidade.

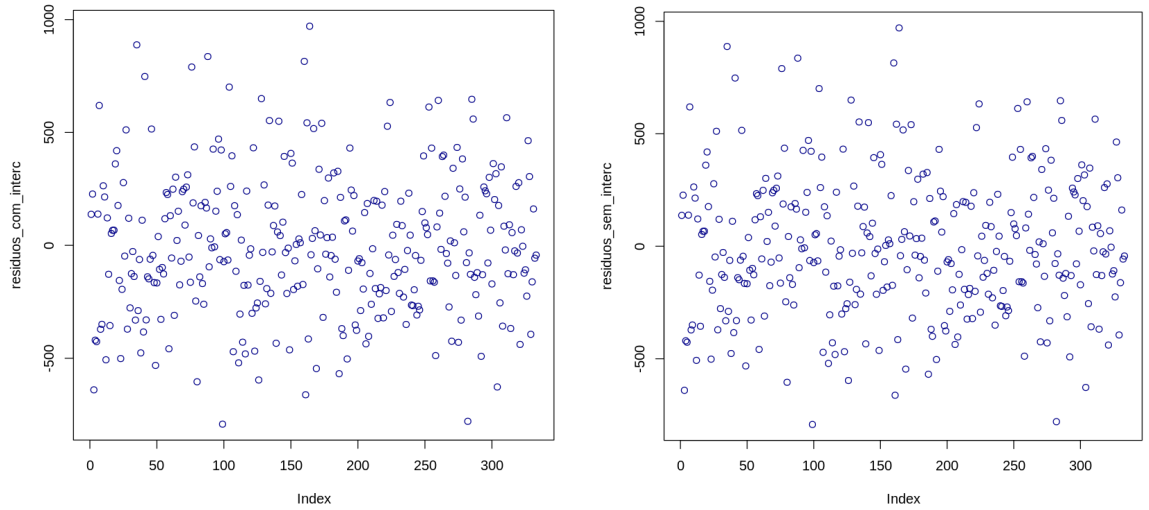

Figura 10: Diagramas de dispersão dos resíduos dos modelos ajustados.

Na figura 11, vemos que os histogramas dos resíduos tem uma distribuição aproximadamente normal.

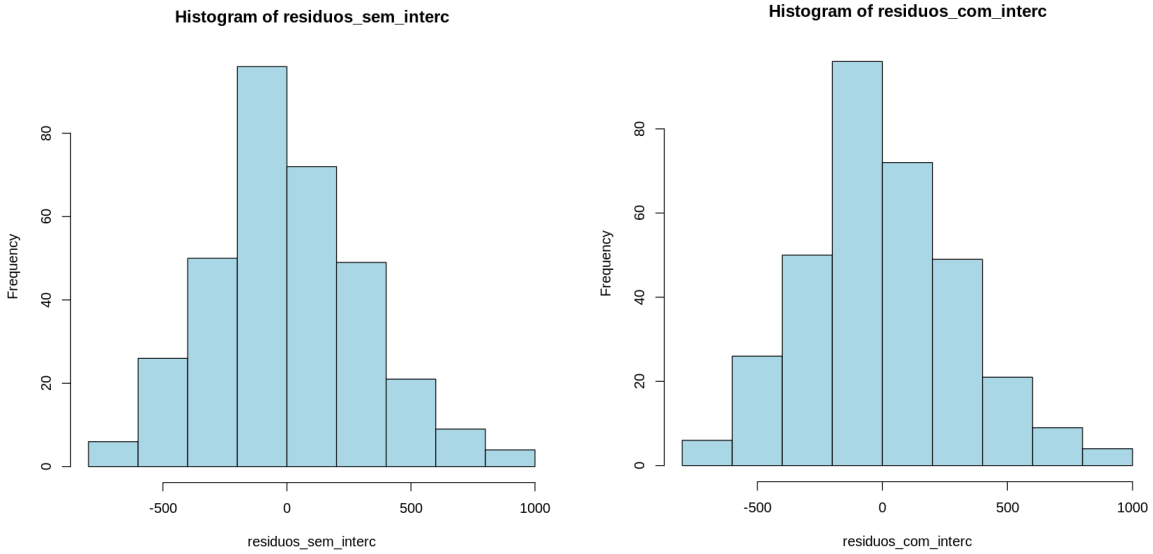

Figura 11: histogramas dos resíduos dos modelos ajustados.

A distribuição Normal também pode ser notada no Normal Q-Q Plot de ambos, onde há grande concentração dos pontos com a reta, havendo apenas uma leve dispersão nas pontas, como mostra a figura 12.

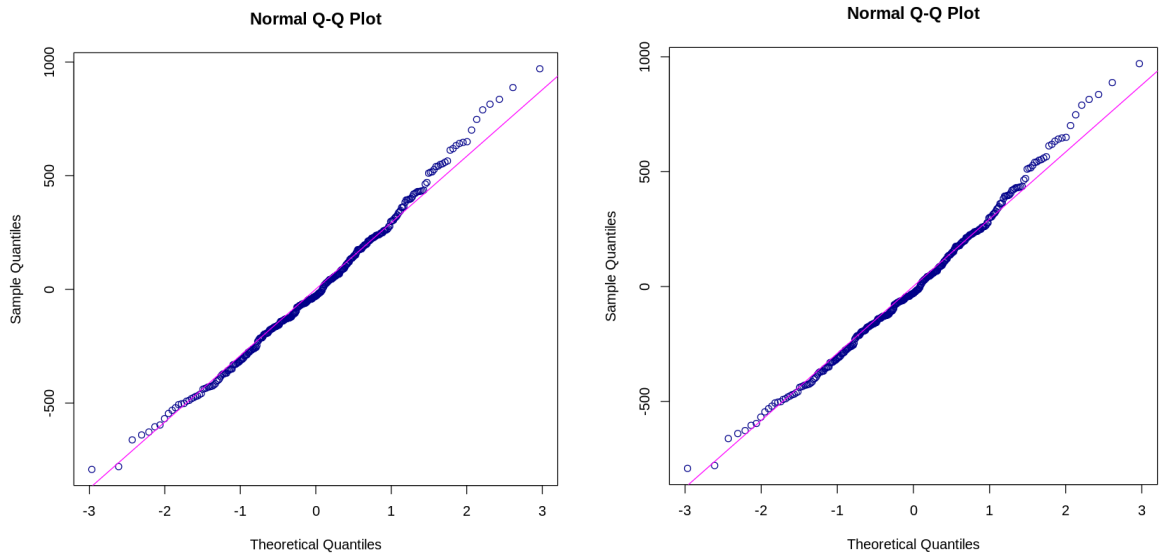

Figura 12: Normal Q-Q Plot dos resíduos dos modelos ajustados.

5. Acurácia do Modelo

Primeiramente, o banco de dados foi dividido entre teste e treino, sendo 75% dos dados para treino, resultando em 252 dados, e 25% para teste, resultando em 81 dados. Assim, foi feito o treinamento e teste para ambos os modelos, e encontradas as medidas Erro Quadrático Médio (MSE) e Erro Absoluto Médio (MAE), que seriam comparadas em ambos os modelos a fim de descobrirmos qual modelo é mais eficiente.

Para o modelo com o intercepto, foram retornados os dados presentes na figura 13 após o fim do treino.

```
[1] "MSE Final: 102026.805917758"
[1] "MAE Final: 256.786076813107"
```

Figura 13: resultado das métricas de erro para o modelo com o intercepto.

Para o modelo sem o intercepto, foram retornados os dados presentes na figura 14 após o fim do treino.

```
[1] "MSE Final: 102026.805917765"
[1] "MAE Final: 256.786076813109"
```

Figura 14: resultado das métricas de erro para o modelo sem o intercepto.

Percebe-se que novamente, a diferença entre entre os valores é mínima, podendo-se assim assumir que não há diferença entre os resultados.

6. Considerações Finais

Por fim, podemos concluir que as variáveis **comprimento_bico**, **comprimento_nadadeira**, **profundidade_bico,** **espécie** e **ano** impactam significativamente a variável **massa_corporal**, ou seja, a massa corporal dos pinguins desse arquipélago está relacionada ao comprimento do seu bico e nadadeira, a profundidade do bico, a espécie a qual o animal pertence e ao ano em que foi feita a pesquisa. Dado que o R^2 do modelo com intercepto é menor que o do modelo sem o intercepto, e que o modelo sem o intercepto mostra uma maior facilidade de explicação para um possível cliente, conclui-se que o modelo sem intercepto seria a minha escolha.