



Análise de Variâncias

Disciplina: Delineamento Experimental

Mestrado: Inteligência Artificial e Ciências de Dados

Ano Lectivo:2025/2026

Docente: Professora Doutora Maria Manuela Melo Oliveira

Professor Doutor Nuno de Almeida Ribeiro

Relatório realizado por:

- Madalena Marques m68201

29 dezembro 2025

Índice

1. Introdução	4
2. Material e Métodos.....	4
3. Metodologia	4
4. Resultados/Discussão dos Resultados	5
4.1. ANOVA de 1 fator.....	5
4.1.1. Validações dos pressupostos	5
4.1.1.1. Independência dos Resíduos.....	5
4.1.1.2. Normalidade.....	6
4.1.1.3. Homogeneidade das Variâncias	6
4.1.2. Averiguar se existem diferenças de média de massa corporal por espécie 7	
4.1.3. Quais são os níveis do fator que diferem- Comparações múltiplas 7	
4.2. ANOVA de dois fatores	8
4.2.1. Validações dos pressupostos	8
4.2.1.1. Independência dos Resíduos.....	8
4.2.2.1 Normalidade.....	9
4.2.2.2 Homogeneidade das Variâncias	9
4.2.3 Averiguar se existem diferenças de média de massa corporal por espécie e sexo.....	10
4.2.4 Quais são os níveis do fator que diferem- Comparações múltiplas 10	
4.2.5 Intervalos de confiança a 95% para Massa Corporal por Espécie..	11
4.3. Regressão Linear simples.....	11
4.3.1. Validações dos Pressupostos.....	11
4.3.1.1. Independência dos Resíduos	12
4.3.1.2. Normalidade	12
4.3.1.3. Homogeneidade das Variâncias	13
4.3.2. Averiguar se existem diferenças de média de massa corporal por espécie 13	
4.3.3. Intervalos de Confiança a 95%	14
4.4. Regressão Linear múltipla	14
4.4.1. Validações dos Pressupostos	15
4.4.1.1. Independência de Resíduos.....	15

4.4.1.2.	Normalidade	15
4.4.1.3.	Homogeneidade das Variâncias	16
4.4.2.	Averiguar se existem diferenças de média de massa corporal por espécie e sexo.....	16
4.4.3.	Intervalos de Confiança a 95%	17
5.	Conclusão	17
6.	Referências Bibliográficas.....	18

1. Introdução

O presente relatório tem como objetivo aplicação de técnicas de análise de variâncias, tal como o modelo da anova e regressão linear na presente base de dados *penguins*.

O estudo da base de dados foi desenvolvido no âmbito da unidade curricular “Delineamento Experimental” na qual foram adquiridos conhecimentos do software *Rstudio*, que posteriormente foram aplicados para realização do presente relatório. Além disso, ao longo do relatório é possível encontrar diversos tipos de teste, nomeadamente o shapiro-wilk, levene e tukey, o último é usado para fazer comparações múltiplas.

O relatório encontra-se dividido em 4 partes, as quais correspondem a metodologia, os resultados obtidos através da anova de 1 fator e 2 fatores e da modelo regressão linear simples e múltipla. Após a análise dos resultados, procedeu-se à relação entre índice de massa corporal, sexo dos pinguins e a diversas espécies, terminando com a conclusão dos resultados obtidos.

2. Material e Métodos

A base de dados *penguins*, contém 344 observações e 8 variáveis, sendo que estas se classificam entre qualitativas e quantitativas. As variáveis *species* (espécie), *island* (ilha) e *sex* (sexo) enquadram-se nas variáveis qualitativas nominais. No entanto, as variáveis *year* (ano), *bill_lenght_mm* (comprimento do bico), *bill_depth_mm* (profundidade do bico), *flipper_length_mm* (comprimento da asa) e a *body_mass_g* (massa corporal) englobam as variáveis quantitativas. Além disso, foram identificados valores omissos (NA) em diversas variáveis cujos estes valores foram convertidos e alguns foram excluídos para garantir a consistências nos resultados. É importante referir que para este estudo foi só usado a espécie, massa corporal e o sexo.

3. Metodologia

A análise das variâncias foi feita através do programa de software estatístico *R studio* cujo output e os resultados serão apresentados e analisados ao longo do tópico 4 sobre gráficos e tabelas. Além disso foram usados diversos pacotes do R para facilitar a manipulação de dados e a criação dos gráficos entre os quais destaca-se os seguintes:

- *dplyr* para manipulação de dados;
- *car* para análise de regressão;

4. Resultados/Discussão dos Resultados

4.1. ANOVA de 1 fator

Anova de 1 fator é um modelo estatístico que permite comparar média de três ou mais grupos independentes com base em um único preditor categórico, que neste caso é espécie. Com este estudo pretende-se averiguar se se existe influência do índice de massa corporal dos pinguins entre espécies. Sendo que foi definido a seguinte hipótese:

Hipótese Nula (H_0): Não há diferenças nas médias de massa corporal entre as três espécies

$$H_0 = \mu_{adelie} = \mu_{chinstrap} = \mu_{gentoo}$$

Hipótese Alternativa: Pelo menos uma das médias de massa corporal das espécies é diferente

$$H_1 = \exists \text{ pelo menos um par } (i, j) \mu_i \neq \mu_j$$

4.1.1. Validações dos pressupostos

Para que os resultados da anova sejam considerados válidos e confiáveis é necessário proceder à validação dos seguintes pressupostos:

- 1) Independência
- 2) Normalidade
- 3) Homocedasticidade das variâncias

4.1.1.1. Independência dos Resíduos

O gráfico 1 mostra quatro gráfico de diagnóstico dos resíduos. Em que primeiro gráfico *residuals vs fitted*, verifica-se a homogeneidade das variâncias e a linearidade, em que se observa vários pequenos agrupamentos de pontos (um para combinação entre a espécies) e a linha vermelha de tendência é quase horizontal, indicando que a variâncias dos resíduos não muda drasticamente com o valor ajustado. Com isso pode-se concluir que o pressuposto de homocedasticidade está validado. O segundo gráfico verifica a normalidade dos resíduos (*Normal Q-Q*), a maioria dos pontos segue de perto a linha diagonal de referência, o que leva concluir que a normalidade dos resíduos está validada. O terceiro gráfico (*Scale-Location*), verifica a homocedasticidade tal como primeiro, verificando-se a linha vermelha de tendência plana e horizontal, o que leva concluir que homocedasticidade está validada. O último gráfico *Residuals vs Leverage* identifica observações influentes, após a sua análise, pode-se concluir que não há pontos de alta influência o que leva a crer que o modelo é robusto.

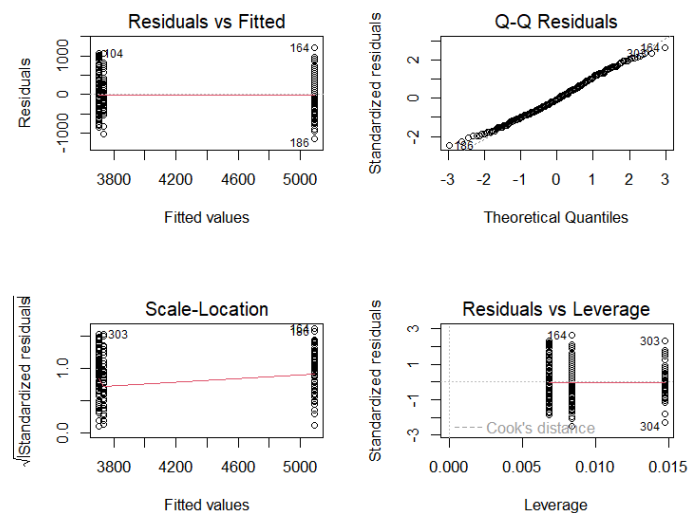


Gráfico 1-Gráfico de Diagnóstico

4.1.1.2. Normalidade

$$H_0 = y_i \sim N \text{ vs } H_1 = y_i \not\sim N$$

Para testar a normalidade, foi realizado o teste de shapiro-wilk, usado para testar normalidade dos resíduos do modelo ANOVA, cujo resultado está representado na figura 1. Após análise da figura, pode-se concluir que não se rejeita H_0 , pois p-value é superior ao nível de significância, que neste caso é 0.05, ou seja, não existe evidência estatística para se rejeitar H_0 , concluindo assim que o índice de massa corporal dos pinguins segue uma distribuição normal.

```
Shapiro-wilk normality test
data:  residuals(resultado_aov)
W = 0.9922, p-value = 0.07835
```

Figura 1- Teste de shapiro wilk para normalidade

4.1.1.3. Homogeneidade das Variâncias

$$H_0 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \text{ vs } H_1 \neq \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$$

Para testar a homocedasticidade das variâncias, foi realizado o teste de Bartlett, usado para homogeneidade das variâncias do modelo ANOVA, cujo resultado está representado na figura 2. Após análise da figura, pode-se concluir que não se rejeita H_0 , pois p-value é superior ao nível de significância, neste caso é 0.005, ou seja, não existe evidências estatísticas de que as variâncias são heterogêneas.

```

Bartlett test of homogeneity of variances

data:  body_mass_g by species
Bartlett's K-squared = 5.692, df = 2, p-value = 0.05808

```

Figura 2-Homogeneidade das Variâncias

Pode-se concluir que os pressupostos são válidos é possível avançar para a ANOVA de um fator.

4.1.2. Averiguar se existem diferenças de média de massa corporal por espécie

A tabela 1 mostra o resultado do teste para averiguar se existem diferenças de média massa corporal por espécie. Após análise da tabela, pode-se confirmar que o valor do teste F é 341.9 é superior ao valor tabelado $F(2,330)$ e além disso, o p-value ($2,2 \times 10^{-16}$) é inferior 0.05, rejeitando H_0 . O coeficiente de determinação (R^2) é 0.6745 indicando que aproximadamente 67.5% da variação observada no índice massa corporal pode ser explicada apenas pela diferença entre espécies. Com isso conclui-se que a espécie possui um efeito altamente significativo na massa corporal.

Tabela 1-Resultado da anova de 1 fator

```

      Df Sum Sq Mean Sq F value Pr(>F)
species    2 145190219 72595110   341.9 <2e-16 ***
Residuals 330  70069447  212332
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

4.1.3. Quais são os níveis do fator que diferem- Comparações múltiplas

Para saber quais as espécies que afetam a determinação da massa corporal, fez o teste Tukey para fazer comparação múltipla entre espécies, o resultado deste teste está representado na tabela 2. A comparação entre espécie *Chinstrap* vs *Adelie*, o p-value é superior a 0.05, com base nesse resultado, pode-se concluir que não existe diferença significativas, ou seja, a massa corporal de *Chinstrap* não é significativamente diferente da *Adelie*. Já a comparação entre espécies *Gentoo* vs *Adelie* a diferença é altamente significativa, ou seja, a espécie *Gentoo* é muito mais pesada que *Adelie*, já que o p-value é inferior a 0.05. E por último, a comparação entre espécies *Gentoo* vs *Chinstrap*, como a anterior, a diferença é altamente significativa, a espécie *Gentoo* é muito mais pesada que *Chinstrap* o p-value é inferior a 0.005.

Tabela 2-Teste de Tukey para comparações múltiplas

```

Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = body_mass ~ species, data = dados_limpos)

$species
      diff      lwr      upr    p adj
Chinstrap-Adelie  26.92385 -132.3528  186.2005  0.916431
Gentoo-Adelie    1386.27259 1252.2897 1520.2554  0.000000
Gentoo-Chinstrap 1359.34874 1194.4304 1524.2671  0.000000

```

4.2. ANOVA de dois fatores

Anova de dois fatores permite verificar se dois fatores, neste caso espécies e o sexo, interagem entre si de maneira a influencia o índice massa corporal. Em que iremos três conjuntos de hipóteses:

1. Efeito principal da espécie: O índice massa corporal médio difere entre as espécies independentemente do sexo.

$$H_0 = \mu_1 = \mu_2 = \mu_3$$

$$H_1 = \exists \text{ pelo menos um par } (i, j) \mu_i \neq \mu_j$$

2. Efeito principal do sexo: Existe dimorfismo sexual (os pinguins do sexo masculino são mais pesados que as fêmeas).
3. Efeito de Interação (Espécie vs Sexo): A diferença de peso entre machos e fêmeas é para todas as espécies, ou o dimorfismo sexual é mais acentuado em espécie do que em outras?

4.2.1. Validações dos pressupostos

Para que os resultados da anova sejam considerados válidos e confiáveis é necessário proceder à validação dos seguintes pressupostos:

- 1) Independência
- 2) Normalidade
- 3) Homocedasticidade das variâncias

4.2.1.1. Independência dos Resíduos

O gráfico 3 mostra quatro gráfico de diagnóstico dos resíduos. Em que primeiro gráfico *residuals vs fitted*, verifica-se a homogeneidade das variâncias e a linearidade, em que se observa vários pequenos agrupamentos de pontos (um para combinação entre a espécie e a ilha) e a linha vermelha de tendência é quase horizontal, indicando que a variâncias dos resíduos não muda drasticamente com o valor ajustado. Com isso pode-se concluir que o pressuposto de homocedasticidade está validado. O segundo gráfico

verifica a normalidade dos resíduos (*Normal Q-Q*), a maioria dos pontos segue de perto a linha diagonal de referência, o que leva concluir que a normalidade dos resíduos está validada. O terceiro gráfico (*Scale-Location*), verifica a homocedasticidade tal como primeiro, verificando-se a linha vermelha de tendência plana e horizontal, o que leva concluir que homocedasticidade está validada. O último gráfico *Residuals vs Leverage* identifica observações influentes, após a sua análise, pode-se concluir que não há pontos de alta influência o que leva a crer que o modelo é robusto.

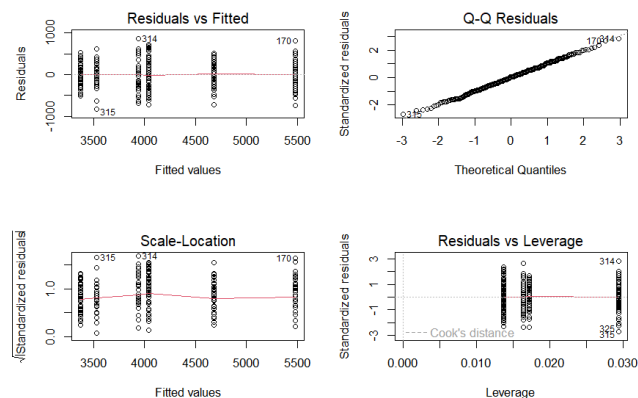


Gráfico 2- Gráfico de Diagnóstico de Resíduos

4.2.2.1 Normalidade

$$H_0 = y_i \sim N \text{ vs } H_1 = y_i \not\sim N$$

Para testar a normalidade, foi realizado o teste de shapiro-wilk, usado para testar normalidade dos resíduos do modelo ANOVA, cujo resultado está representado na figura 3. Após análise da figura, pode-se concluir que não se rejeita H_0 , pois p-value é superior ao nível de significância, que neste caso é 0.05, ou seja, não existe evidência estatística para se rejeitar H_0 , concluindo assim que o índice de massa corporal dos pinguins segue uma distribuição normal.

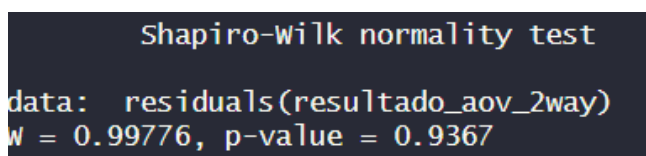


Figura 3-Teste de shapiro-wilk para testar a normalidade

4.2.2.2 Homogeneidade das Variâncias

$$H_0 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \text{ vs } H_1 = \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$$

Para testar a homocedasticidade das variâncias, foi realizado o teste de Bartlett, usado para homogeneidade das variâncias do modelo ANOVA, cujo resultado está representado na tabela 3. Após análise da figura, pode-se concluir que não se rejeita H_0 , pois p-value é superior ao nível de significância, neste caso é 0.005, ou seja, não existe evidências estatística de que as variâncias são heterogêneas.

Tabela 3- Teste de Levene para testar homogeneidade das variâncias

Levene's Test for Homogeneity of Variance (center = median)			
	Df	F value	Pr(>F)
group	5	1.3908	0.2272
	327		

Pode-se concluir que os pressupostos são válidos é possível avançar para a ANOVA de um fator.

4.2.3 Averiguar se existem diferenças de média de massa corporal por espécie e sexo

A tabela 4 mostra o resultado do teste para averiguar se existem diferenças de média massa corporal por espécie e sexo. Após a análise da tabela, consegue-se observar que existe diferença de peso entre os machos e fêmeas, já que o p-value da interação entre a espécie e o sexo dá 0.000197 inferior ao nível de significância (0.001), concluindo assim que o valor p é altamente significativo.

Tabela 4- Resultado da anova de 2 fatores

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
species	2	145190219	72595110	758.358	< 2e-16 ***
sex	1	37090262	37090262	387.460	< 2e-16 ***
species:sex	2	1676557	838278	8.757	0.000197 ***
Residuals	327	31302628	95727		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

4.2.4 Quais são os níveis do fator que diferem- Comparações múltiplas

Para saber quais as espécies que são afetadas pelo dimorfismo sexual, fez um teste de comparação múltipla entre espécies, o resultado deste teste está representado na tabela 5. Na primeira comparação entre espécies *Chinstrap-Adelie*, o p-value é superior a 0.05, com base nesse resultado, pode-se concluir que não existe diferença significativas no dimorfismo sexual, ou seja, não é altamente significativo na massa corporal. Já na comparação entre espécies *Gentoo-Adelie*, o p-value é inferior a 0.05, o que leva a concluir que existe diferença significativas no dimorfismo sexual, ou seja, é altamente significativo na massa corporal e os machos são significativamente mais pesados que as fêmeas nestas duas espécies. Na última comparação, entre espécies *Gentoo-Chinstrap*, o p-value é inferior a 0.05, tal que como a comparação anterior, existe

diferenças significativas no dimorfismo sexual, ou seja, é altamente significativo na massa corporal e os machos são significativos mais pesados que as fêmeas nestas duas espécies.

Tabela 5- Teste de Tukey para anova de dois fatores

```

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = body_mass_g ~ species * island, data = dados_limpos)

$species
      diff      lwr      upr    p adj
Chinstrap-Adelie  26.92385 -132.8397  186.6874 0.9169133
Gentoo-Adelie    1386.27259 1251.8801 1520.6650 0.0000000
Gentoo-Chinstrap 1359.34874 1193.9262 1524.7713 0.0000000

$island
      diff      lwr      upr    p adj
Dream-Biscoe     -3.090032 -133.0586  126.8786 0.9982740
Torgersen-Biscoe  1.402898 -178.7614  181.5672 0.9998147
Torgersen-Dream   4.492930 -182.1127  191.0986 0.9982299

```

4.2.5 Intervalos de confiança a 95% para Massa Corporal por Espécie

A tabela 6 mostra os limites do Intervalo de Confiança a 95% para cada das três espécies. Após a análise da tabela, pode-se concluir que as médias populacionais são significativamente diferentes.

Tabela 6- Intervalo de confiança a 97.5%

	2.5 %	97.5 %
(Intercept)	3572.5873	3846.7309
speciesChinstrap	-133.1643	196.6135
speciesGentoo	1222.3542	1543.2016
islandDream	-192.1966	175.6057
islandTorgersen	-191.8789	189.5820

4.3. Regressão Linear simples

O modelo de regressão linear simples tem como base entender a relação entre uma variável dependente, neste caso será a massa corporal e uma única variável independente neste caso seria a espécie.

A equação do Modelo:

$$Y = \beta_0 + \beta_1 X$$

Onde:

B₀= intercepção;

B₁= inclinação: o quanto o peso dos pinguins aumenta por espécie;

4.3.1. Validações dos Pressupostos

Para que os resultados da regressão linear simples sejam considerados válidos e confiáveis é necessário proceder à validação dos seguintes pressupostos:

- 1) Independência

- 2) Normalidade
- 3) Homocedasticidade das variâncias

4.3.1.1. Independência dos Resíduos

O gráfico 5 mostra quatro gráfico de diagnóstico dos resíduos. Em que primeiro gráfico *residuals vs fitted*, verifica-se a homogeneidade das variâncias e a linearidade, em que se observa vários pequenos agrupamentos de pontos (um para combinação entre a espécie e a ilha) e a linha vermelha de tendência é quase horizontal, indicando que a variâncias dos resíduos não muda drasticamente com o valor ajustado. Com isso pode-se concluir que o pressuposto de homocedasticidade está validado. O segundo gráfico verifica a normalidade dos resíduos (*Normal Q-Q*), a maioria dos pontos segue de perto a linha diagonal de referência, o que leva concluir que a normalidade dos resíduos está validada. O terceiro gráfico (*Scale-Location*), verifica a homocedasticidade tal como primeiro, verificando-se a linha vermelha de tendência plana e horizontal, o que leva concluir que homocedasticidade está validada. O último gráfico *Residuals vs Leverage* identifica observações influentes, após a sua análise, pode-se concluir que não há pontos de alta influência o que leva a crer que o modelo é robusto.

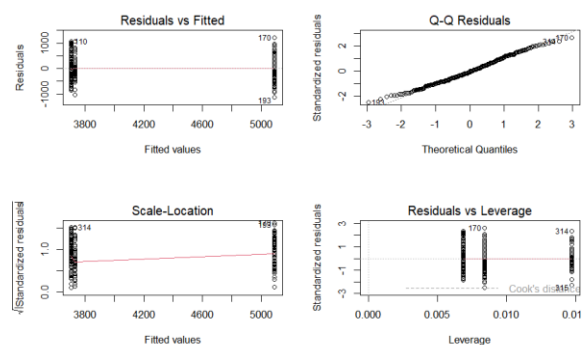


Gráfico 3- Gráfico de Diagnóstico de Resíduos

4.3.1.2. Normalidade

$$H_0 = y_i \sim N \text{ vs } H_1 = y_i \not\sim N$$

Para testar a normalidade, foi realizado o teste de shapiro-wilk, usado para testar normalidade dos resíduos do modelo da regressão linear, cujo resultado está representado na tabela 7. Após análise da tabela, pode-se concluir que não se rejeita H_0 , pois p-value é superior ao nível de significância, que neste caso é 0.05, ou seja, não existe evidência estatística para se rejeitar H_0 , concluindo assim que índice de massa corporal dos pinguins segue uma distribuição normal.

Tabela 7- Teste de shapiro wilk para normalidade

```
Shapiro-Wilk normality test
data: residuals(modelo_regressao_fator)
W = 0.9922, p-value = 0.07835
```

4.3.1.3. Homogeneidade das Variâncias

$$H_0 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \text{ vs } H_1 \neq \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$$

Para testar a homogeneidade das variâncias, foi realizado o teste de levene, usado para homogeneidade das variâncias do modelo regressão linear, cujo resultado está representado na tabela 8. Após análise da figura, pode-se concluir que não se rejeita H_0 , pois $p\text{-value} > 0.05$, concluindo assim não existe evidências estatística para dizer que as variâncias são heterogêneas.

Tabela 8- Teste de Levene para homogeneidade das variâncias

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  2  5.1349 0.006367 **
      330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.3.2. Averiguar se existem diferenças de média de massa corporal por espécie

Para averiguar se existem diferenças de média de massa corporal por espécie, foi usado o modelo de regressão linear simples com as seguintes hipóteses:

1. Para β_0 (interceção)

Para este modelo foi usado uma espécie de referência (*Adelie*)

$$H_0 = \mu_1 = 0 \text{ vs } H_1 = \mu_1 \neq 0$$

2. Para outras espécies (β_i)

$$H_0 = \beta_i = 0 \text{ vs } H_1 = \beta_i \neq 0 \text{ } i = 1, 2, 3$$

A tabela 9 mostra o resultado do modelo para averiguar se existem diferenças de média de massa corporal por espécie. Observa-se que a espécie influencia o índice de massa corporal, o valor do teste F(2,330) é 341.9 e valor do p-value do β_0 é 2×10^{-16} , sendo inferior ao nível de significância, rejeitando o H_0 , ou seja, não existe evidência estatística suficiente que β_0 seja zero, concluindo assim que a média da espécie *adelie* é estatisticamente sólida. Já espécie *chinstrap* o valor de p-value é 0.691, como é superior ao nível de significância, não se rejeitando H_0 , ou seja, não existe evidência de que a espécie *chinstraps* tenha um peso diferente da *adelies*. Por um lado, a espécie *gentoo*

o valor de p-value é 2×10^{-16} , sendo inferior ao nível de significância, rejeitando H_0 , ou seja, existe evidência estatística de que a espécie *gentoos* têm um peso significativamente diferente que a espécie *adelies*. O modelo de regressão linear simples para teste é o seguinte:

$$\text{body mass} = 97.184 + 0.398X + 24.359X$$

Tabela 9-O modelo de regressão linear simples

```
Call:
lm(formula = body_mass_g ~ species, data = dados_limpos)

Residuals:
    Min       1Q   Median       3Q      Max
-1142.44  -342.44   -33.09   307.56  1207.56

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3706.16     38.14   97.184  <2e-16 ***
speciesChinstrap    26.92     67.65    0.398    0.691
speciesGentoo   1386.27     56.91   24.359  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.8 on 330 degrees of freedom
Multiple R-squared:  0.6745,    Adjusted R-squared:  0.6725
F-statistic: 341.9 on 2 and 330 DF,  p-value: < 2.2e-16
```

4.3.3. Intervalos de Confiança a 95%

A tabela 10 mostra o intervalo de confiança a 95% da regressão linear entre espécies, com os dois propósitos principais, analisar a significância dos preditores e medir a precisão das previsões. Na espécie *chinstrap* é estatisticamente não significativo e na espécie *gentoo* é estatisticamente significativo.

Tabela 10-Intervalo de confiança a 95%

	2.5 %	97.5 %
(Intercept)	3631.1448	3781.1840
speciesChinstrap	-106.1606	160.0083
speciesGentoo	1274.3226	1498.2226

4.4. Regressão Linear múltipla

O modelo de regressão linear múltipla procura explicar a variação do peso através das diferentes espécies e o sexo.

A equação do modelo:

$$Y = \beta_0 + \beta_1(\text{espécie}) + \beta_2(\text{Sexo}) + \beta_3(\text{Espécie} \times \text{sexo})$$

Onde:

B_0 (interceção): Representa a média do grupo de referência;

B_1 e β_2 : Representa os efeitos principais de cada fator;

B₃: Indica se o efeito do sexo na massa corporal muda depende da espécie.

4.4.1. Validações dos Pressupostos

Para que os resultados da regressão linear simples sejam considerados válidos e confiáveis é necessário proceder à validação dos seguintes pressupostos:

- 1) Independência
- 2) Normalidade
- 3) Homocedasticidade das variâncias

4.4.1.1. Independência de Resíduos

O gráfico 6 mostra quatro gráfico de diagnóstico dos resíduos. Em que primeiro gráfico residuals vs fitted, verifica-se a homogeneidade das variâncias e a linearidade, em que se observa vários pequenos agrupamentos de pontos (um para combinação entre a espécie e a ilha) e a linha vermelha de tendência é quase horizontal, indicando que a variâncias dos resíduos não muda drasticamente com o valor ajustado. Com isso pode-se concluir que o pressuposto de homocedasticidade está validado. O segundo gráfico verifica a normalidade dos resíduos (Normal Q-Q), a maioria dos pontos segue de perto a linha diagonal de referência, o que leva concluir que a normalidade dos resíduos está validada. O terceiro gráfico (Scale-Location), verifica a homocedasticidade tal como primeiro, verificando-se a linha vermelha de tendência plana e horizontal, o que leva concluir que homocedasticidade está validada. O último gráfico Residuals vs Leverage identifica observações influentes, após a sua análise, pode-se concluir que não há pontos de alta influência o que leva a crer que o modelo é robusto.

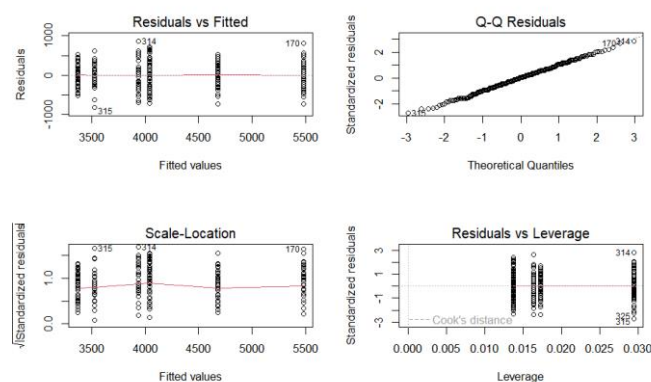


Gráfico 4-Gráfico de Diagnóstico de Resíduos

4.4.1.2. Normalidade

$$H_0 = y_i \sim N \text{ vs } H_1 = y_i \not\sim N$$

Para testar a normalidade, foi realizado o teste de shapiro-wilk, usado para testar normalidade dos resíduos do modelo da regressão linear, cujo resultado está representado na tabela 10. Após análise da tabela, pode-se concluir que não se rejeita H_0 , pois p-value é superior ao nível de significância, que neste caso é 0.05, ou seja, não existe evidência estatística para se rejeitar H_0 , concluindo assim que índice de massa corporal dos pinguins segue uma distribuição normal.

Tabela 10- Teste do Shapiro-wilk para normalidade

```
Shapiro-wilk normality test
data: residuals(modelo_regressao_2f)
W = 0.99776, p-value = 0.9367
```

4.4.1.3. Homogeneidade das Variâncias

$$H_0 = \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \text{ vs } H_1 \neq \sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$$

Para testar a homogeneidade das variâncias, foi realizado o teste de levene, usado para homogeneidade das variâncias do modelo regressão linear, cujo resultado está representado na tabela 11. Após análise da figura, pode-se concluir que não se rejeita H_0 , pois p-value>0.05, concluindo assim não existe evidências estatística para dizer que as variâncias são heterogêneas.

Tabela 11- Teste de Levene para homogeneidade das variâncias

```
Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
group 5 1.3908 0.2272
327
```

4.4.2. Averiguar se existem diferenças de média de massa corporal por espécie e sexo

A tabela 12 mostra o resultado do modelo para averiguar se existem diferenças de média de massa corporal por espécie e sexo. Observa-se que o sexo e a espécie têm influência no índice de massa corporal dos pinguins, já que o valor do teste F(5,327) é 348.3 e o valor do p-value do β_0 (interceção) é 2×10^{-16} , sendo inferior ao nível de significância, rejeitando o H_0 , ou seja, não existe evidência estatística suficiente que β_0 seja zero. Já na espécie *chinstrap* o valor de p-value é 0.1420, como é superior ao nível de significância, não se rejeita H_0 , ou seja, não existe evidência de que a espécie *chinstrap* tenha um peso diferente da *adelies*. Por um lado, a espécie *gentoo* o valor de p-value é 2×10^{-16} , sendo inferior ao nível de significância, rejeitando H_0 , ou seja, existe evidência estatística de que a espécie *gentoos* têm um peso significativamente diferente que a espécie *adelies*. No sexo masculino, o valor de p-value é 2×10^{-16} , rejeitando H_0 , ou seja, existe dimorfismo sexual, concluindo assim os machos são mais pesados que

fêmeas. Nos pinguins machos e espécie *chinstrap* o valor de p-value é 0.00406, rejeitando o H_0 , ou seja, os pinguins machos desta espécie são mais pesados que fêmeas. E por último, nos pinguins machos da espécie *gentoo* o valor p-value é 0.08886, não se rejeitando H_0 , ou seja, os pinguins machos e fêmeas têm um peso semelhante nesta espécie. O modelo de regressão linear múltipla para teste é o seguinte:

$$body\ mass = 93.030 + 2.465X + 24.088X + 13.174 X - 2.894X + 1.706X$$

Tabela 12- Modelo de Regressão Linear múltipla

```
Call:
lm(formula = body_mass ~ species * sex, data = dados_limpos)

Residuals:
    Min       1Q   Median       3Q      Max
-827.21 -213.97   11.03   206.51  861.03

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3368.84     36.21   93.030 < 2e-16 ***
speciesChinstrap  158.37     64.24    2.465  0.01420 *
speciesGentoo    1310.91     54.42   24.088 < 2e-16 ***
sexmale         674.66     51.21   13.174 < 2e-16 ***
speciesChinstrap:sexmale -262.89    90.85   -2.894  0.00406 **
speciesGentoo:sexmale   130.44     76.44    1.706  0.08886 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 309.4 on 327 degrees of freedom
Multiple R-squared:  0.8546,    Adjusted R-squared:  0.8524
F-statistic: 384.3 on 5 and 327 DF,  p-value: < 2.2e-16
```

4.4.3. Intervalos de Confiança a 95%

A tabela 13 mostra o intervalo de confiança a 95% da regressão linear entre espécies e sexo, com os dois propósitos principais, analisar a significância dos preditores e medir a precisão das previsões. Em ambas espécies não estaticamente significativo.

Tabela 13-Intervalos de Confiança a 95%

	2.5 %	97.5 %
(Intercept)	3297.59731	3440.0739
speciesChinstrap	31.99386	284.7467
speciesGentoo	1203.84380	1417.9677
sexmale	573.91136	775.4037
speciesChinstrap:sexmale	-441.61606	-84.1696
speciesGentoo:sexmale	-19.93039	280.8047

5. Conclusão

A análise das variâncias da base dados *penguins*, com especial destaque para variáveis massa corporal, sexo e espécie, permite concluir que o peso dos pinguins é diferente de espécie para espécie, mas onde há maior diferença é na espécie *gentoo* em relação há outra duas espécies *adelie* e *chinstrap*.

Além disso também existe dimorfismo sexual, ou seja, o sexo também influencia o peso de todas as espécies deste estudo, com especial destaque para macho que são relativamente mais pesados que as fêmeas. Com base nos resultados anterior consegue-se observar que dimorfismo sexual não é constante, ou seja, na espécie *gentoo* é mais acentuado.

6. Referências Bibliográficas

Fox, J (2002). An R and S-Plus Companion to Applied Regression. Sage Publications.
Faraway, J. Practical Regression and Anova using R
(<http://cran.rproject.org/doc/contrib/Faraway-PRA.pdf>). Dados e scripts disponíveis em
<http://www.maths.bath.ac.uk/~jjf23/LMR/>