Regressão Linear em Python

Análise do Banco de Dados De Penguins do Arquipélago Palmer

Salvador Alves Ferreira Netto (2022040141)
Caique Izidoro Alvarenga (2021086814)
Nicolas Adam Berger Monteiro (2022039950)
Marcelo Pinheiro Filho (2020042686)

Índice

Capítulo 1

Introdução

Neste estudo, exploraremos a relação entre a massa corporal em gramas (`body_mass_g`) de pinguins e diversas variáveis específicas, utilizando análise de regressão linear em Python. As variáveis numéricas incluem o comprimento do bico em milímetros (`bill_length_mm`), o diâmetro do bico em milímetros (`bill_diameter_mm`) e o comprimento da nadadeira em milímetros (`flipper_length_mm`). Além disso, temos variáveis categóricas, como espécie (`species`), sexo (`sex`), ilha (`island`), e ano (`year`). O banco de dados contém 333 linhas e 8 colunas;

Tabela 1.1: Visualização das 5 Primeiras Linhas do Banco de Dados

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
0	Adelie	Torgersen	39.100000	18.700000	181.000000	3750.000000	male	2007
1	Adelie	Torgersen	39.500000	17.400000	186.000000	3800.000000	female	2007
2	Adelie	Torgersen	40.300000	18.000000	195.000000	3250.000000	female	2007
4	Adelie	Torgersen	36.700000	19.300000	193.000000	3450.000000	female	2007
5	Adelie	Torgersen	39.300000	20.600000	190.000000	3650.000000	male	2007

Tabela 1.2: Sumário do Banco de Dados

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
count	333.000000	333.000000	333.000000	333.000000
mean	43.992793	17.164865	200.966967	4207.057057
std	5.468668	1.969235	14.015765	805.215802
min	32.100000	13.100000	172.000000	2700.000000
25%	39.500000	15.600000	190.000000	3550.000000
50%	44.500000	17.300000	197.000000	4050.000000
75%	48.600000	18.700000	213.000000	4775.000000
max	59.600000	21.500000	231.000000	6300.000000

Capítulo 2

Seleção de Variáveis

O modelo de regressão linear múltipla inicial julgado como mais adequado foi:

body_mass_g ~ bill_depth_mm + flipper_length_mm + bill_length_mm + sex*species

Ao observarmos a Figura 2.1, é evidente que as três variáveis numéricas (`bill_depth_mm`, `flipper_length_mm` e `bill_length_mm`) apresentam uma correlação linear positiva com a variável de resposta `body_mass_g`. No entanto, as variáveis categóricas, como espécie e sexo, influenciam a forma como os coeficientes impactam a estimativa de y. Isso é evidenciado pelo fato de que, ao separar os dados por espécie ou sexo, observamos a formação de retas paralelas com interceptos diferentes.

Além disso, é intuitivo considerar que o sexo tem um efeito sobre o peso corporal do pinguim, dado que, na maioria das espécies, os machos tendem a ser mais pesados do que as fêmeas. Também, as diferenças físicas entre as espécies podem contribuir para variações no peso.

A variável `year` foi excluída do modelo devido à suposição de independência e distribuição igual entre os anos, corroborada pela observação gráfica de uma dispersão uniforme nos dados ao longo dos anos.

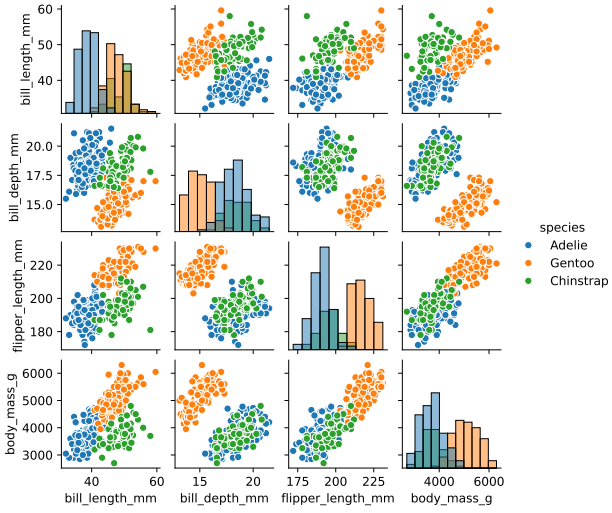
Tabela 2.1: Quantidade de Espécies por Ilha

	species	island	count
0	Adelie	Dream	55
1	Adelie	Torgersen	47
2	Adelie	Biscoe	44
3	Chinstrap	Dream	68
4	Gentoo	Biscoe	119

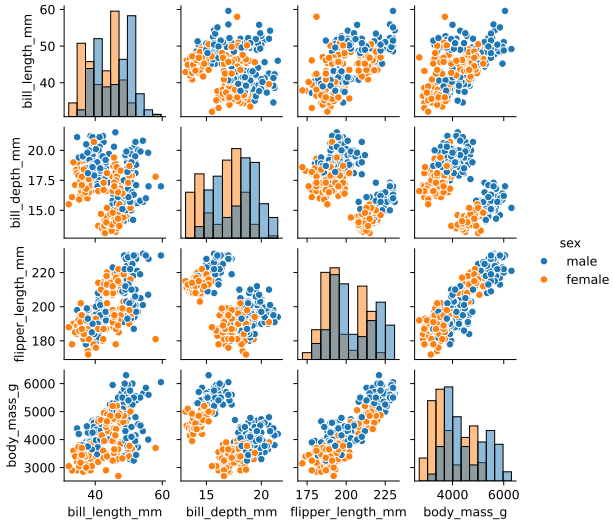
Quanto à variável categórica `island`, sua exclusão é justificada pela observação na tabela agrupada por espécie (Tabela 2.1). Nota-se que as espécies *Gentoo* e *Chinstrap* estão presentes exclusivamente em uma ilha cada, enquanto a espécie *Adelie* está praticamente distribuída de maneira equitativa nas três ilhas.

Figura 2.1: Relações em Pares por Especies

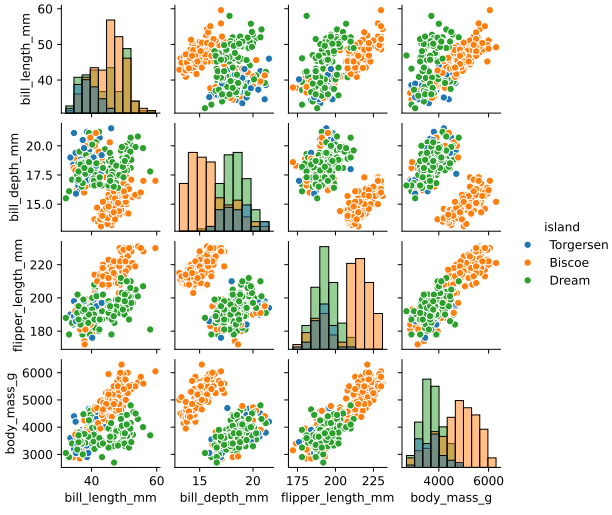
(a) Species



(b) Sex



(c) Island



(d) Year

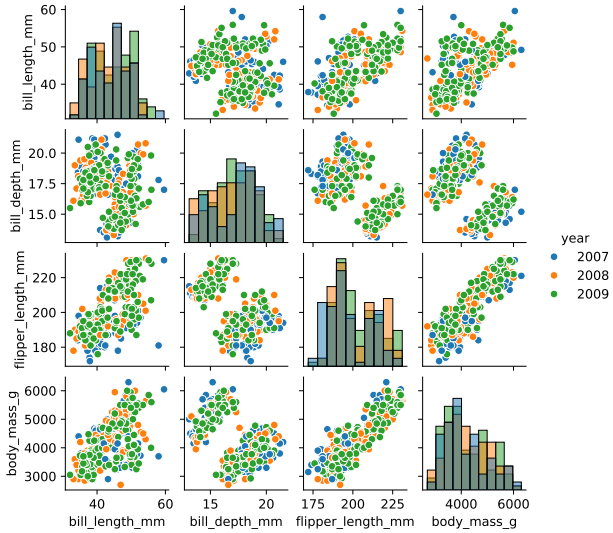
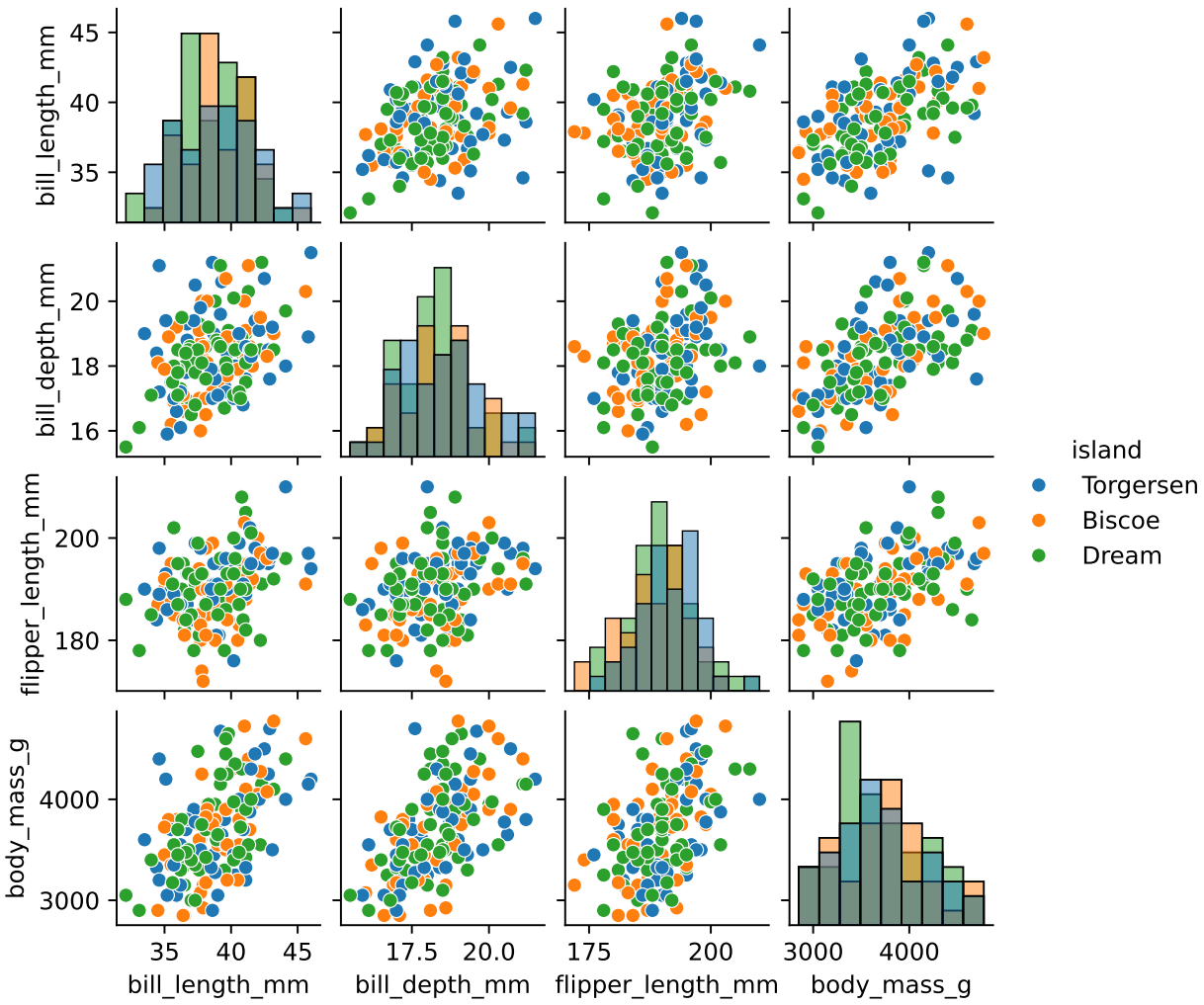


Figura 2.2: Pinguins da Espécie Adelie Disposto por Ilha



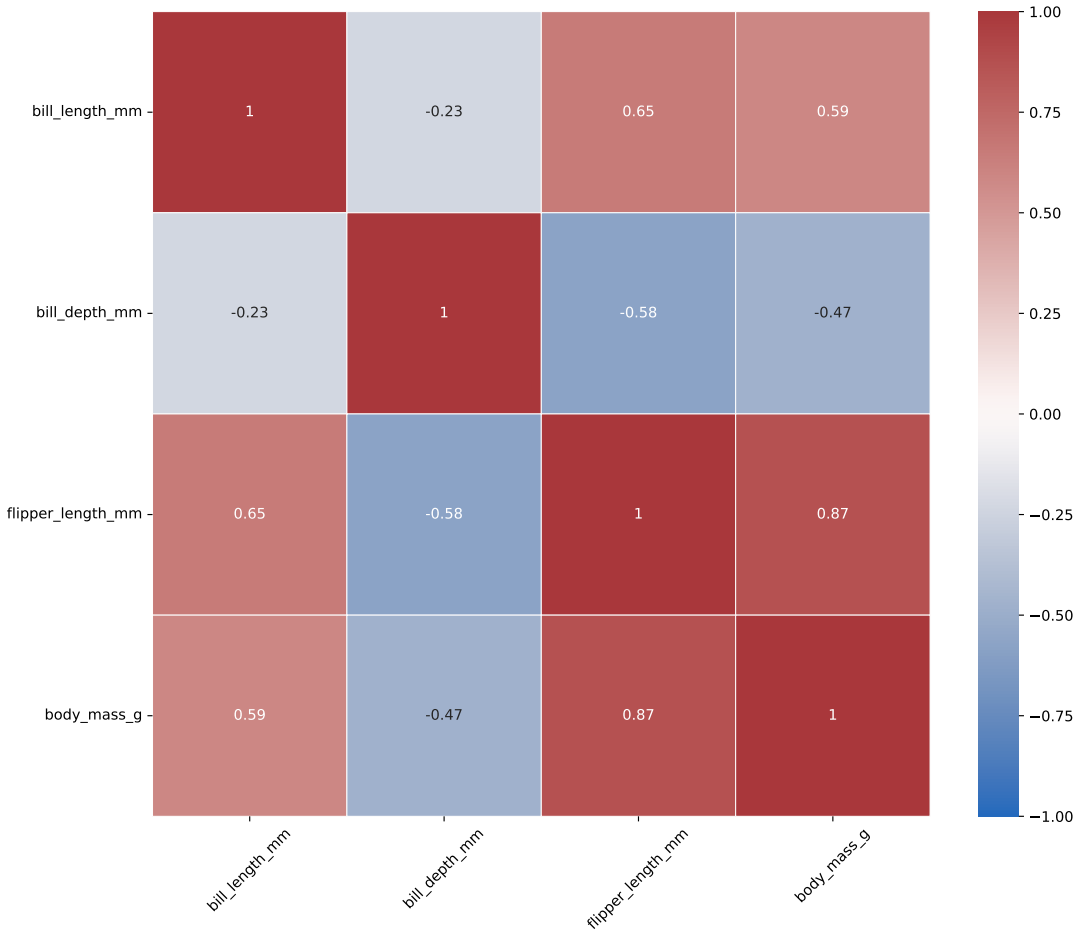
Ao analisarmos a Figura 2.2, podemos visualizar que o comportamento da espécie *Adelie* é consistente em todas as três ilhas. Concluímos, portanto, que o fator ilha não modifica o comportamento da espécie *Adelie*.

Dado que as ilhas *Dream* e *Biscoe* contêm exclusivamente as espécies *Chinstrap* e *Gentoo*, respectivamente, enquanto a espécie *Adelie* está igualmente distribuída entre as três ilhas, concluímos que não há variação significativa nos pinguins *Adelie* entre as diferentes ilhas. Diante desse cenário, parece mais viável escolher somente a variável `species` para inclusão no modelo. Essa decisão é respaldada por motivos biológicos e também porque apenas uma espécie está presente em diversas ilhas.

```
(array([0.5, 1.5, 2.5, 3.5]), [Text(0.5, 0, 'bill_length_mm'), Text(1.5, 0, 'bill_depth_mm'), Text(2.
```

```
(array([0.5, 1.5, 2.5, 3.5]), [Text(0, 0.5, 'bill_length_mm'), Text(0, 1.5, 'bill_depth_mm'), Text(0,
```

Figura 2.3: Correlações entre as Variáveis do Conjunto de Dados



Temos correlações positivas entre as variáveis `flipper_length_mm` e `bill_length_mm`, bem como uma correlação negativa entre `bill_length_mm` e `bill_depth_mm`. Essa situação pode indicar a presença de alguns problemas de multicolinearidade.

Capítulo 3

Ajuste do Modelo e Multicolinearidade

Utilizamos o procedimento *stepwise* para verificar diferentes modelos que continham interação entre `species`, `sex`, e as variáveis numéricas. No entanto, sempre enfrentávamos problemas com multicolinearidade. Para simplificar o modelo, optamos pela fórmula: ***body_mass_g ~ flipper_length_mm + bill_depth_mm + bill_length_mm + sex + species***. Entre todos os modelos avaliados, aqueles que apresentavam resíduos bem comportados e nenhum problema de influência, este se mostrou o único com todas as variáveis significativas, conforme testado pelo teste de coeficiente de regressão individual.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            body_mass_g   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     380.2
Date:                sáb, 11 jan 2025   Prob (F-statistic):          6.82e-144
Time:                        12:10:20   Log-Likelihood:                -2354.0
No. Observations:                 333   AIC:                             4722.
Df Residuals:                     326   BIC:                             4749.
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                   -1460.9946    571.308     -2.557      0.011   -2584.911    -337.079
sex[T.male]                   389.8915     47.848      8.148      0.000     295.761     484.022
species[T.Chinstrap]         -251.4767     81.079     -3.102      0.002    -410.980     -91.973
species[T.Gentoo]            1014.6267    129.561      7.831      0.000     759.746    1269.507
flipper_length_mm              15.9502      2.910      5.482      0.000      10.226      21.674
bill_depth_mm                  67.2176     19.742      3.405      0.001      28.380     106.055
bill_length_mm                 18.2044      7.106      2.562      0.011       4.225      32.184
==============================================================================
Omnibus:                        0.879   Durbin-Watson:                   2.163
Prob(Omnibus):                  0.644   Jarque-Bera (JB):                0.871
Skew:                           0.124   Prob(JB):                        0.647
Kurtosis:                       2.959   Cond. No.                     7.55e+03
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.55e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
modelo = lm(body_mass_g ~ flipper_length_mm + bill_depth_mm + bill_length_mm + sex + species, data
vif(modelo)
```

```
                     GVIF Df GVIF^(1/(2*Df))
flipper_length_mm  6.687361  1        2.585993
bill_depth_mm      6.077457  1        2.465250
bill_length_mm     6.072898  1        2.464325
sex                2.308318  1        1.519315
species           41.074074  2        2.531582
```

```
modelo.bse
```

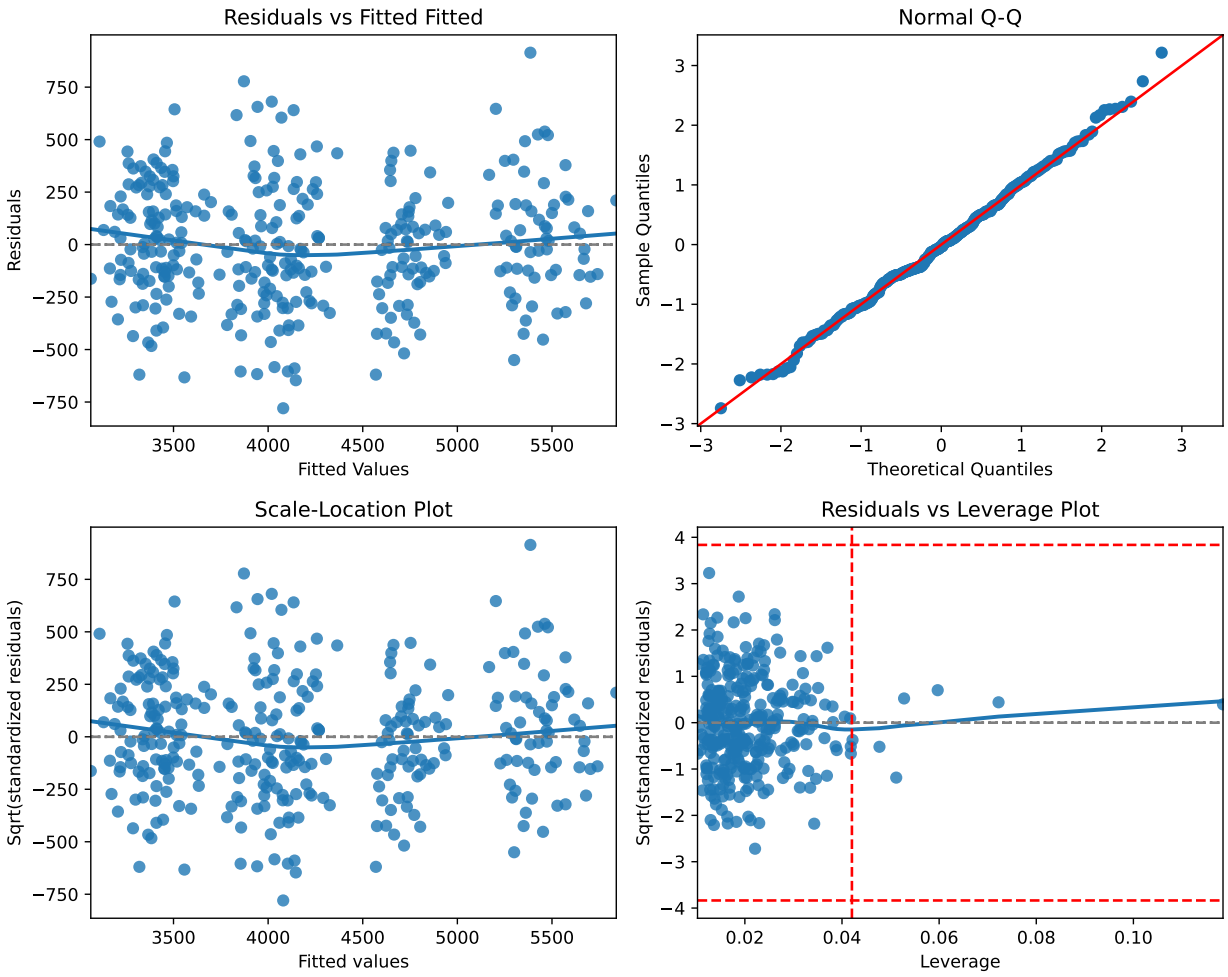
```
Intercept               571.308144
sex[T.male]              47.848346
species[T.Chinstrap]     81.078824
species[T.Gentoo]       129.560586
flipper_length_mm         2.909612
bill_depth_mm            19.741850
bill_length_mm            7.106258
dtype: float64
```

Não identificamos indícios de multicolinearidade, visto que os valores de *VIF* estão abaixo de 3.

Capítulo 4

Resíduos

Figura 4.1: Análise Gráfica dos Resíduos




```
Shapiro Statistic:  0.997

Shapiro P-Value:  0.735
```

```
Durbin Watson Statistic: 2.163212829533633
```

Na Figura 4.1, nos gráficos `Resíduos versus valores ajustados` e `Gráfico escala-locação`, podemos observar que a validade da suposição de linearidade existe no modelo, assim como a validade da suposição de homocedasticidade das variâncias. Isso é evidenciado pelo padrão aleatório dos resíduos em torno de zero.

O teste de *Durbin-Watson* para autocorrelação dos erros não mostra indícios de autocorrelação. Além disso, na figura `Normal Q-Q`, a verificação da suposição de normalidade dos erros é confirmada, e o teste de *Shapiro-Wilk* confirma esse resultado.

No gráfico `Resíduos versus Alavancagem`, observamos a presença de pontos de alavancagem, mas não identificamos pontos inconsistentes. Portanto, não atribuiremos atenção excessiva a esses pontos.

Capítulo 5

Influência

Em resumo, tanto as inspeções visuais quanto as análises estatísticas indicam que as observações não apresentam problemas significativos ou influências prejudiciais para a validade do nosso modelo.

```
# DFFitS
summ_df[summ_df['dffits'] > 3*np.sqrt(p/(n-p))]['dffits']
```

```
Series([], Name: dffits, dtype: float64)
```

Tabela 5.1: **?(caption)**

	dfb_Intercept	dfb_sex[T.male]	...	student_resid	dffits
0	-0.013805	-0.010631	...	-0.121923	-0.019308
1	0.006024	-0.074324	...	1.424014	0.171866
2	0.098837	0.115395	...	-1.206164	-0.172198
4	0.039437	0.047064	...	-0.467584	-0.075284
5	0.004672	-0.010751	...	-1.445773	-0.230489

[5 rows x 13 columns]

Figura 5.1: COVRATIO

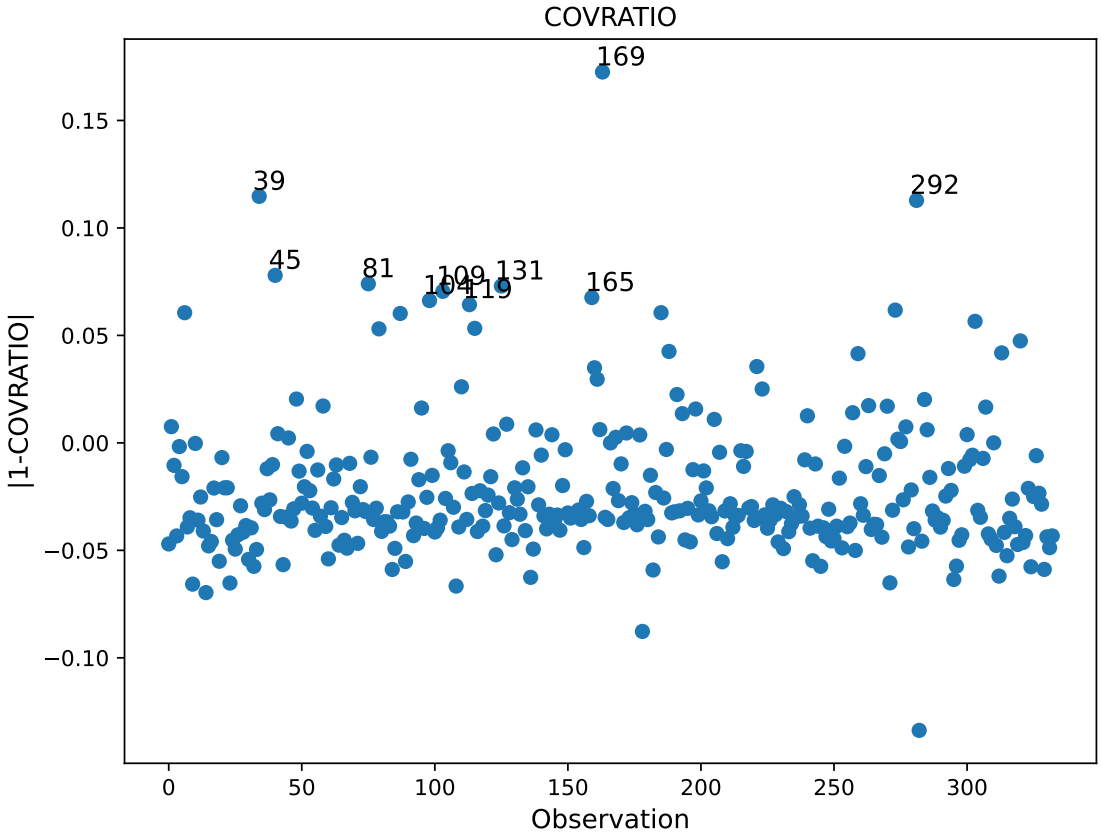


Figura 5.2: Medidas de Influência Leverage e Cook's

(a) Leverage

(b) Distância de Cooks

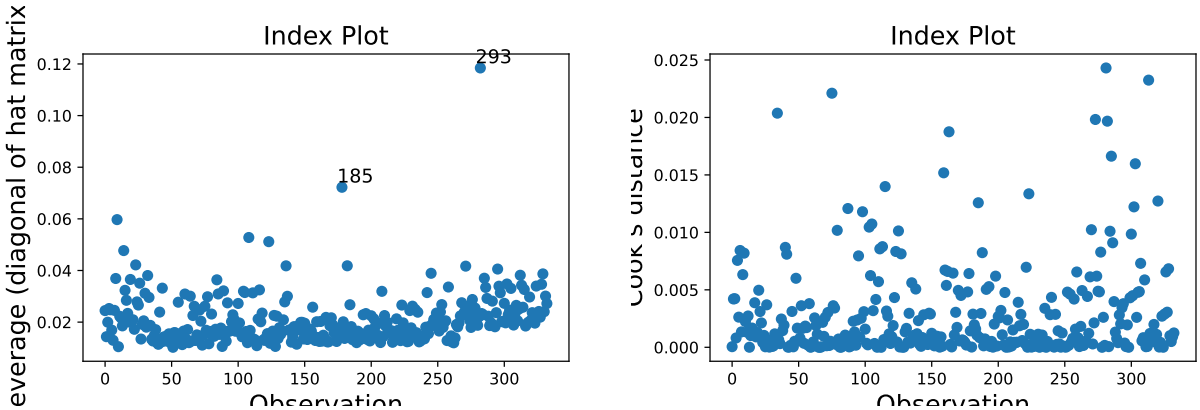
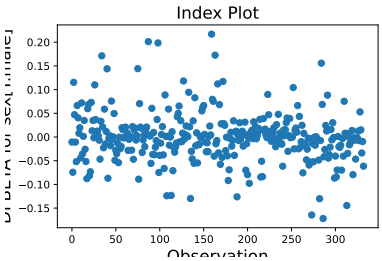
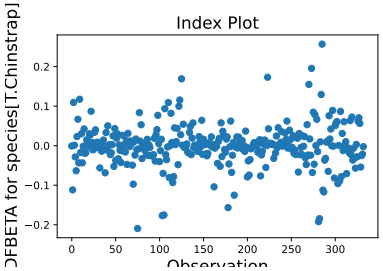


Figura 5.3: Medidas de Influência DFBeta

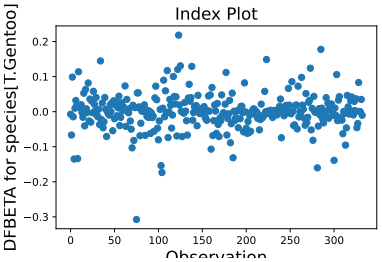
(a) DFBeta (species)



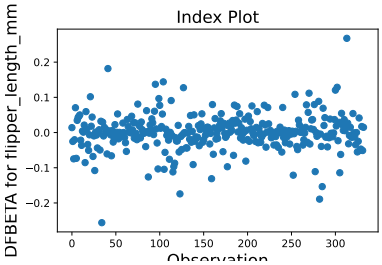
(b) DFBeta (flipper_length_mm)



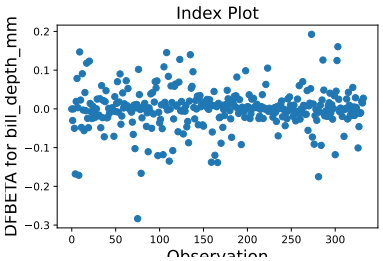
(c) DFBeta (bill_depth_mm)



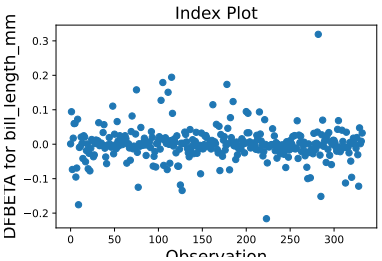
(d) DFBeta (bill_length_mm)



(e) DFBeta (sex)



(f) DFBeta (species)



Capítulo 6

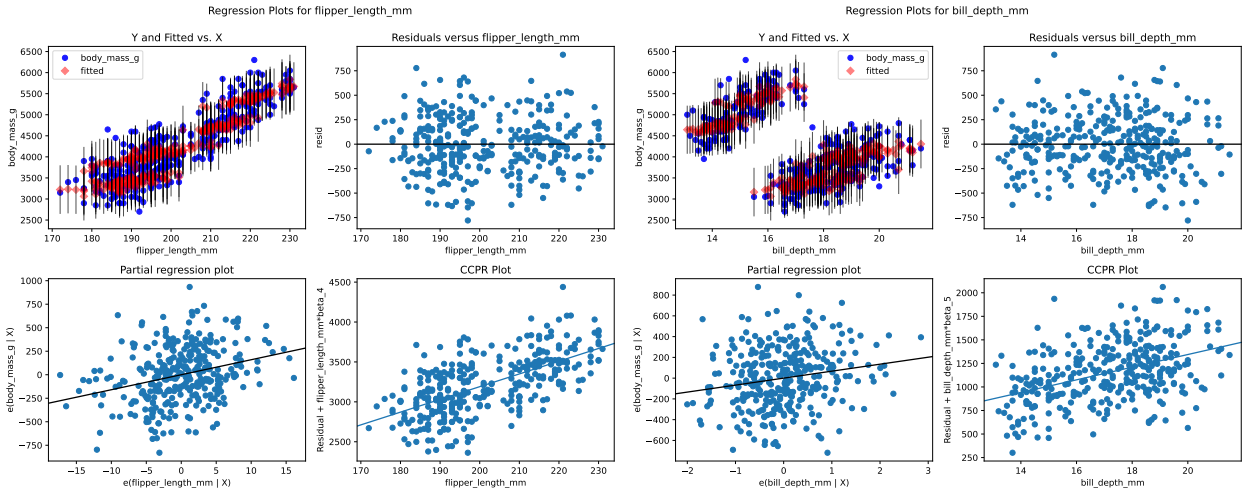
Regressão Parcial

Na Figura 6.1 podemos observar um padrão linear passando pela origem, mostrando que as variáveis explicativa estão linearmente relacionadas com a variável resposta.

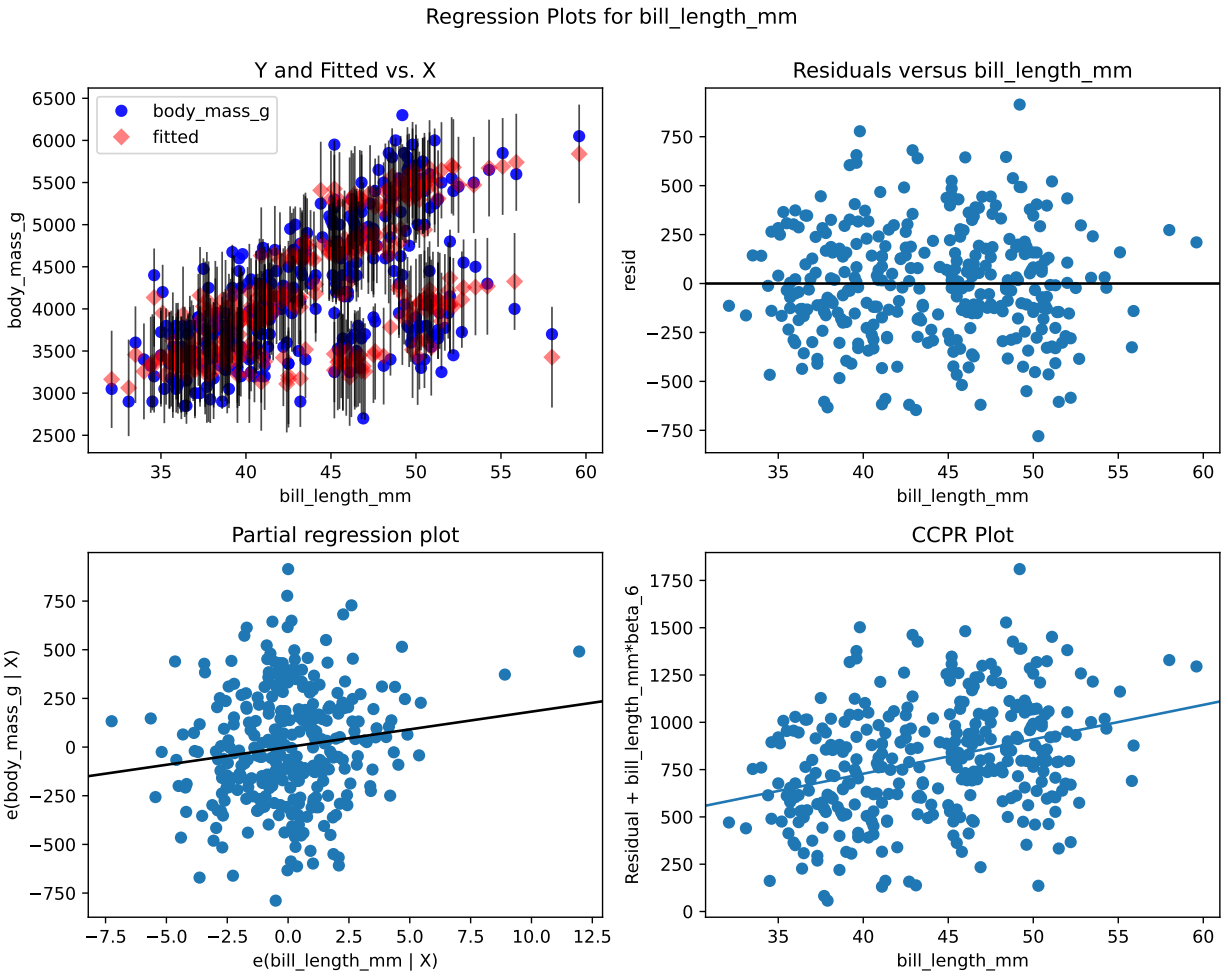
Figura 6.1: Regressão Parcial

(a) Flipper Length

(b) Bill Depth



(c) Bill Length



Capítulo 7

Conclusões

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            body_mass_g   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     380.2
Date:                sáb, 11 jan 2025   Prob (F-statistic):          6.82e-144
Time:                        12:10:28   Log-Likelihood:                -2354.0
No. Observations:                 333   AIC:                             4722.
Df Residuals:                     326   BIC:                             4749.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept            -1460.9946    571.308     -2.557      0.011   -2584.911    -337.079
sex[T.male]            389.8915     47.848      8.148      0.000     295.761     484.022
species[T.Chinstrap]  -251.4767     81.079     -3.102      0.002    -410.980     -91.973
species[T.Gentoo]     1014.6267    129.561      7.831      0.000     759.746    1269.507
flipper_length_mm       15.9502      2.910      5.482      0.000      10.226      21.674
bill_depth_mm           67.2176     19.742      3.405      0.001      28.380     106.055
bill_length_mm          18.2044      7.106      2.562      0.011       4.225      32.184
==============================================================================
Omnibus:                        0.879   Durbin-Watson:                   2.163
Prob(Omnibus):                  0.644   Jarque-Bera (JB):                0.871
Skew:                           0.124   Prob(JB):                        0.647
Kurtosis:                       2.959   Cond. No.                     7.55e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.55e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Com um R^2 *ajustado* de aproximadamente 87.3%, o modelo demonstra um bom ajuste aos dados, indicando que grande parte da variação no peso corporal pode ser explicada pelas variáveis consideradas.

Todas as variáveis são estatisticamente significativas, e de acordo com o VIF, não temos multicolinearidade. Além disso, a `F-statistic` sugere que o modelo como um todo é estatisticamente significativo.

Observamos que pinguins do sexo masculino tendem a ter um peso corporal médio aproximadamente 389.89 gramas maior do que pinguins do sexo feminino, mantendo outras variáveis constantes. Quanto à espécie, pinguins *Chinstrap* apresentam um peso médio cerca de 251.48 gramas menor em comparação com a espécie de referência (*Adelie*, *Female*), enquanto pinguins *Gentoo* têm um peso médio aproximadamente 1014.63 gramas maior.

As variáveis físicas também desempenham um papel significativo. Para cada unidade adicional no comprimento da nadadeira, observamos um aumento médio de 15.95 gramas no peso corporal, mantendo outras variáveis constantes. Similarmente, aumentos na profundidade e comprimento do bico estão associados a acréscimos médios de 67.22 e 18.20 gramas no peso corporal, respectivamente.