

# Regressão Linear em Python

Análise do Banco de Dados De Penguins do Arquipélago Palmer

Salvador Alves Ferreira Netto (2022040141)      Caique Izidoro Alvarenga (2021086814)  
João Roberto Zuquim Filho (2022088831)

# Índice

1	Introdução	2
2	Seleção de Variáveis	3
3	Ajuste do Modelo e Multicolinearidade	7
4	Resíduos	9
5	Influência	11
6	Regressão Parcial	14
7	Conclusões	16

# 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

	<code>species</code>	<code>island</code>	<code>bill_length_mm</code>	<code>bill_depth_mm</code>	<code>flipper_length_mm</code>	<code>body_mass_g</code>	<code>sex</code>	<code>year</code>
0	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
1	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
2	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
5	Adelie	Torgersen	39.3	20.6	190	3650	male	2007

Tabela 1.2: Sumário do Banco de Dados

	<code>bill_length_mm</code>	<code>bill_depth_mm</code>	<code>flipper_length_mm</code>	<code>body_mass_g</code>
count	333.000000	333.000000	333.000000	333.0000
mean	43.992793	17.164865	200.96697	4207.0571
std	5.468668	1.969235	14.01577	805.2158
min	32.100000	13.100000	172.00000	2700.0000
25%	39.500000	15.600000	190.00000	3550.0000
50%	44.500000	17.300000	197.00000	4050.0000
75%	48.600000	18.700000	213.00000	4775.0000
max	59.600000	21.500000	231.00000	6300.0000

## Capítulo 2

# Seleção de Variáveis

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

$$\text{body\_mass\_g} \sim \text{bill\_depth\_mm} + \text{flipper\_length\_mm} + \text{bill\_length\_mm} + \text{sex} * \text{species}$$

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.

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.

Tabela 2.1: Quantidade de Espécies por Ilha

species	island	count
Adelie	Dream	55
	Torgersen	47
	Biscoe	44
Chinstrap	Dream	68
	Biscoe	0
	Torgersen	0
Gentoo	Biscoe	119
	Dream	0
	Torgersen	0

Figura 2.1: Relações em Pares por Espécies

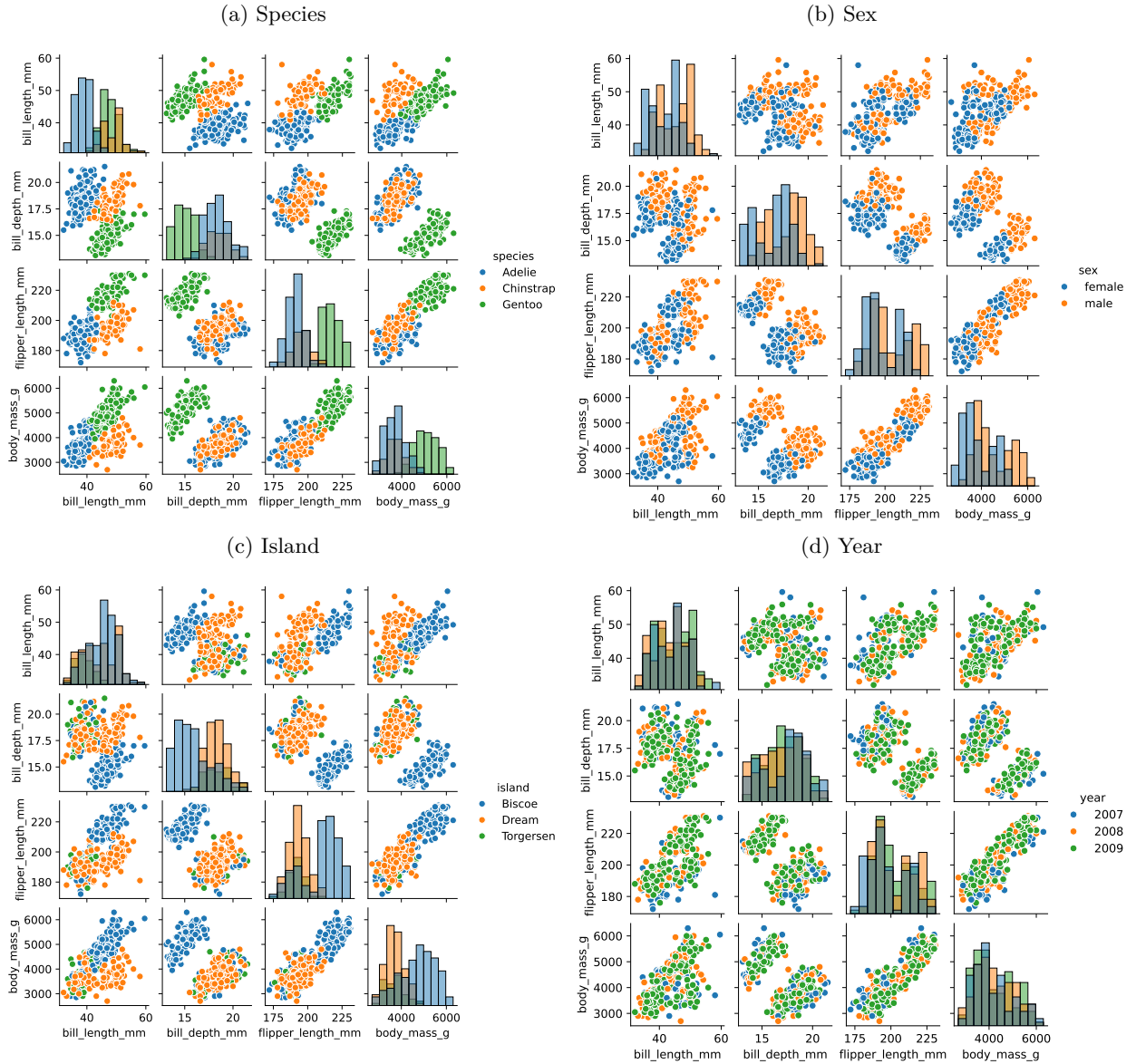
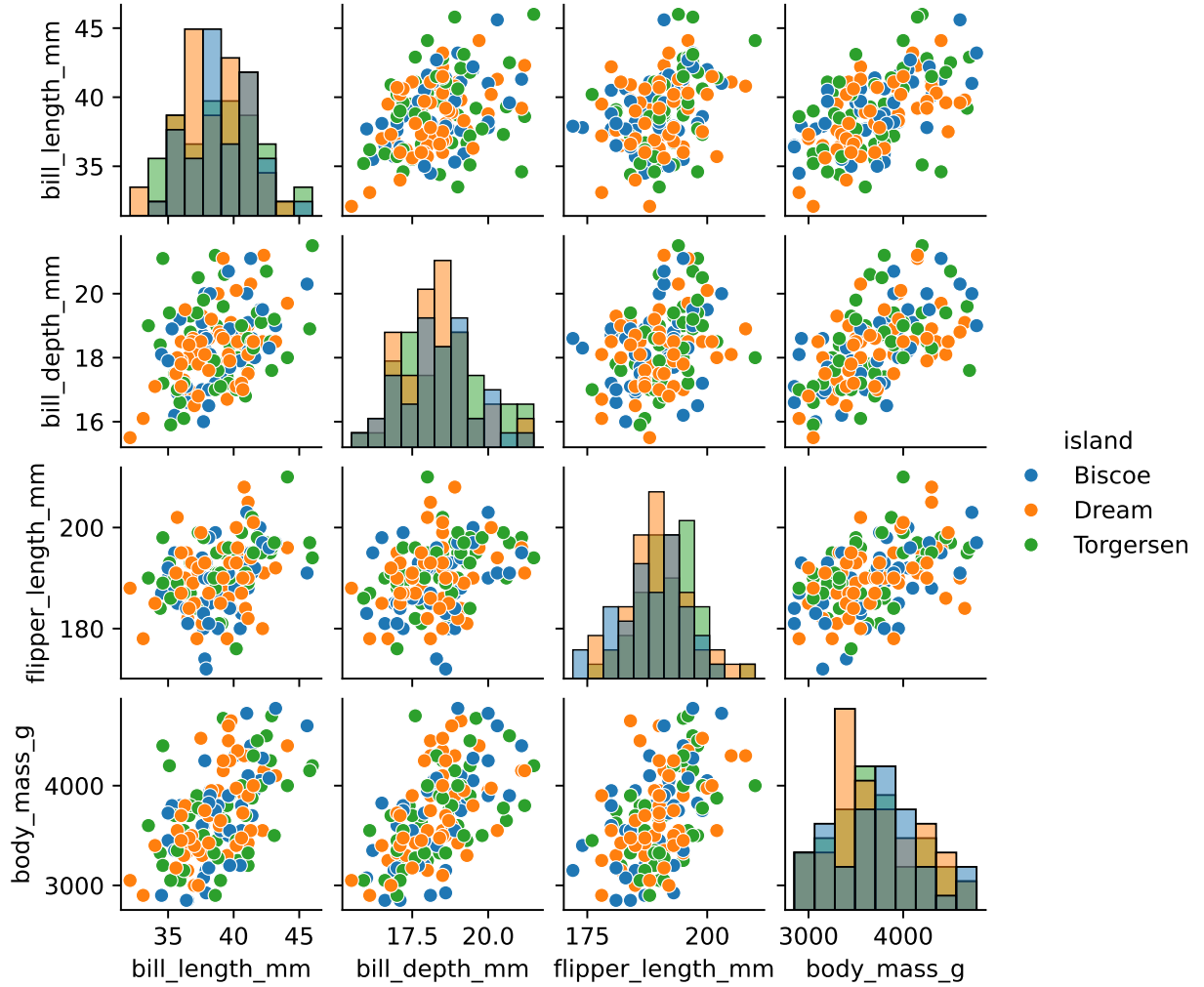


Figura 2.2: Pinguins da Espécie Adelie Disposto por Ilha



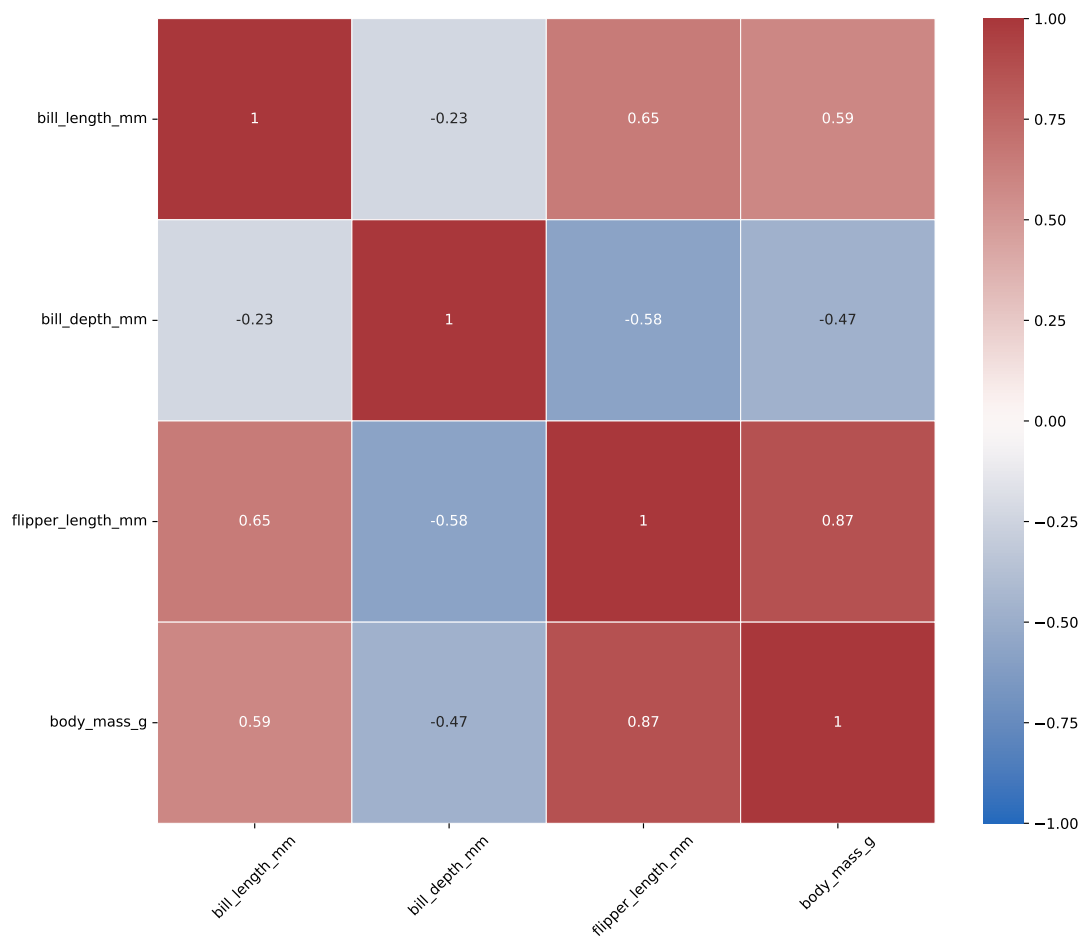
Ao analisarmos a Figura 2.2, podemos visualizar que o comportamento da espécie *Adelie* é consistente em todas as três ilhas. Concluimos, portanto, que o fator ilha não modifica o comportamento da espécie *Adelie*.

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, concluimos 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(2.5, 0, 'flipper_length_mm'), Text(3.5, 0, 'body_mass_g')])
```

```
(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, 2.5, 'flipper_length_mm'), Text(0, 3.5, 'body_mass_g')])
```

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:	qua, 29 nov 2023	Prob (F-statistic):	6.82e-144			
Time:	20:53:57	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 <sup>1/(2*Df)</sup>
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

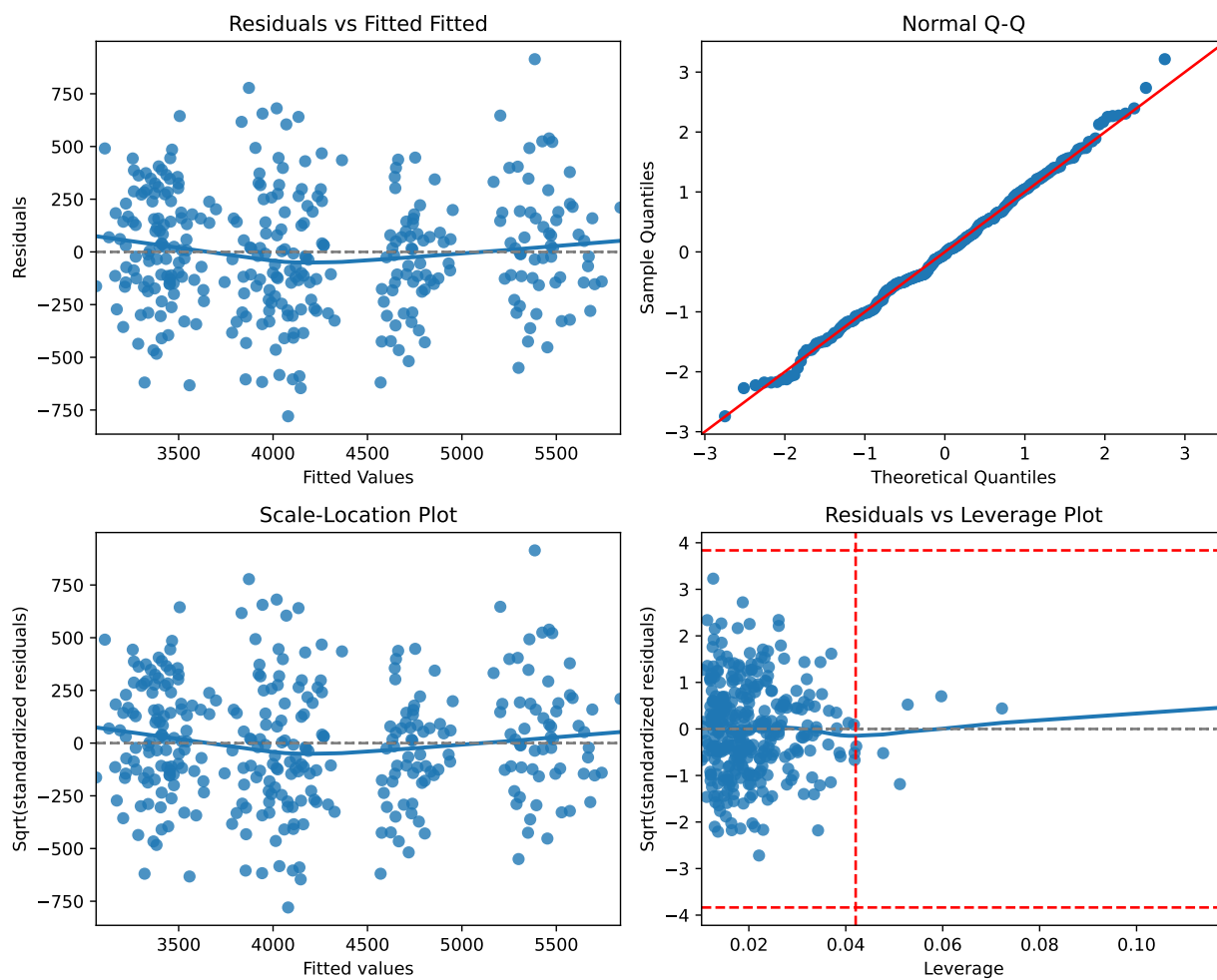
dtype: float64

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: 2.1632128295336326

Na Figura 4.1, nos gráficos **Resíduos versus valores ajustados** e **Gráfico escala-locção**, 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.

Tabela 5.1: Sumário das Observações Influentes

	dfb_intercept	dfb_sex[T.male]	dfb_species[T.Chinstrap]	dfb_species[T.Gentoo]	dfb_flipper_length_mm	dfb_bill_depth_mm	dfb_bill_length_mm	cooks_d	standard_resid	hat_diag	dffits_internal	student_resid	dffits
0	-0.0138047	-0.0106306	-0.0007172	-0.0072067	0.0139850	-0.0002869	0.0009662	0.0000534	-0.1221080	0.0244651	-0.0193373	-0.1219233	-0.0193981
1	0.0060244	-0.0743242	-0.1118377	-0.0665715	-0.0253104	-0.0299943	0.0943237	0.0042064	1.4217742	0.0143572	0.1715956	1.4240138	0.1718659
2	0.0988371	0.1153950	0.1092888	0.0986741	-0.0747684	-0.0000788	-0.0733625	0.0042301	-1.2053238	0.0199747	-0.1720779	-1.2061644	-0.1721979
4	0.0394374	0.0470639	0.0011448	-0.0146354	-0.0203938	-0.0501422	0.0175913	0.0008116	-0.4681456	0.0252680	-0.0753744	-0.4675842	-0.0752840
5	0.0046725	-0.0107510	-0.0284445	-0.1351557	0.0705987	-0.1679537	0.0598328	0.0075640	-1.4433612	0.0247857	-0.2301049	-1.4457727	-0.2304894

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

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

Figura 5.1: COVRATIO

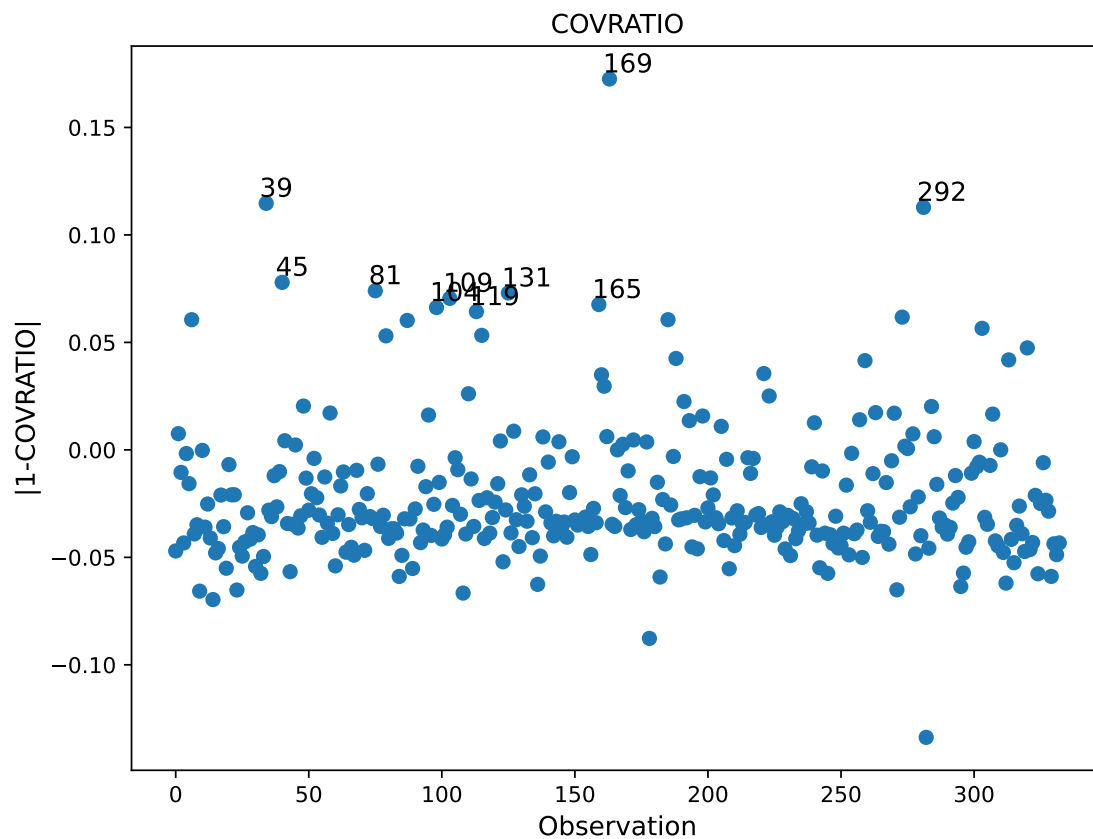


Figura 5.2: Medidas de Influência Leverage e Cook's

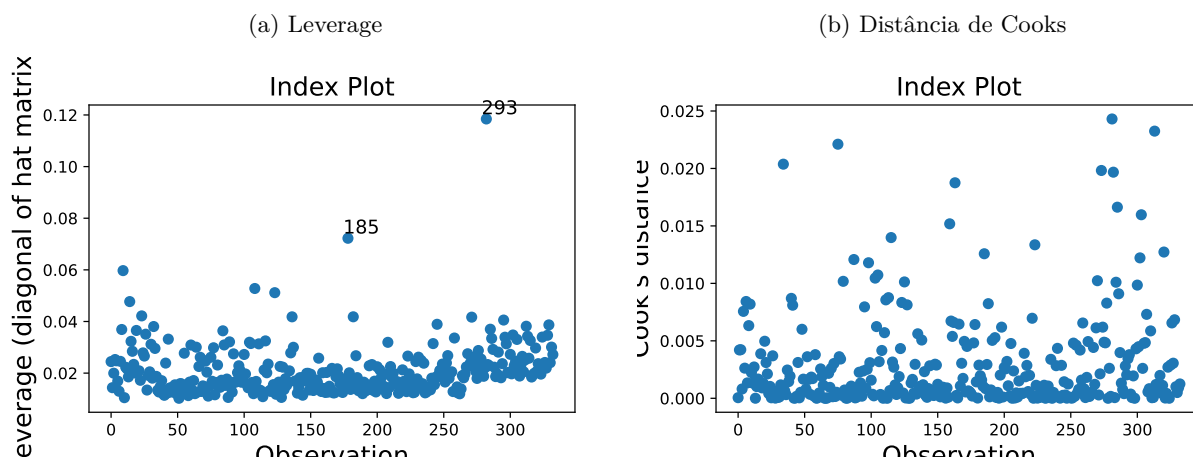
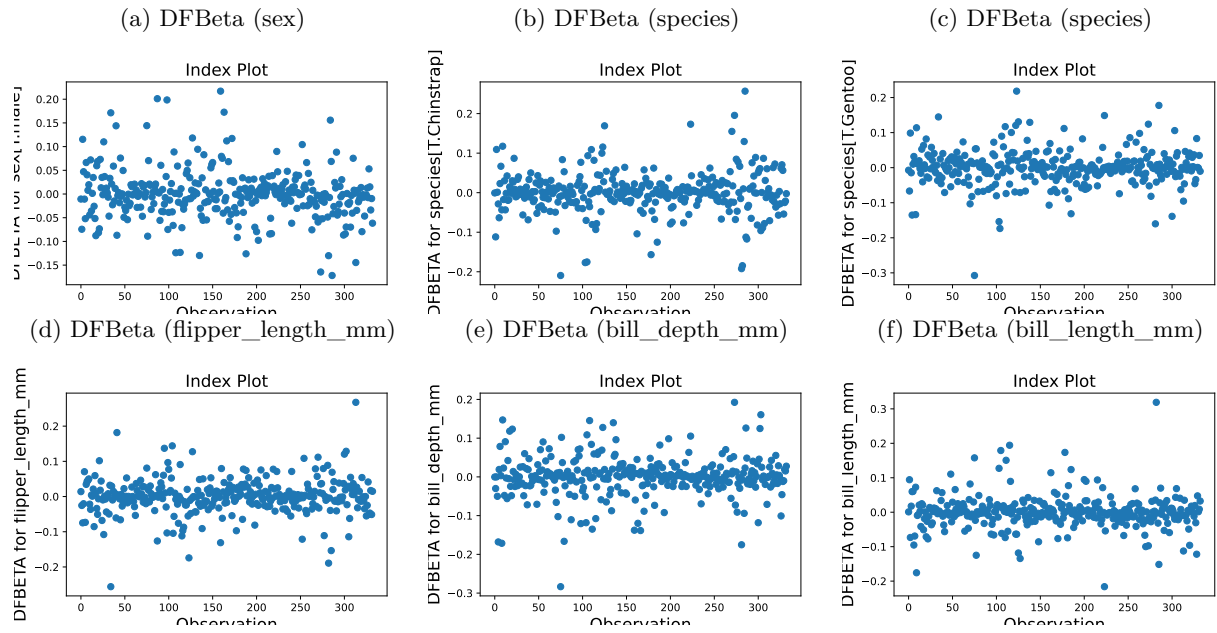


Figura 5.3: Medidas de Influência DFBeta



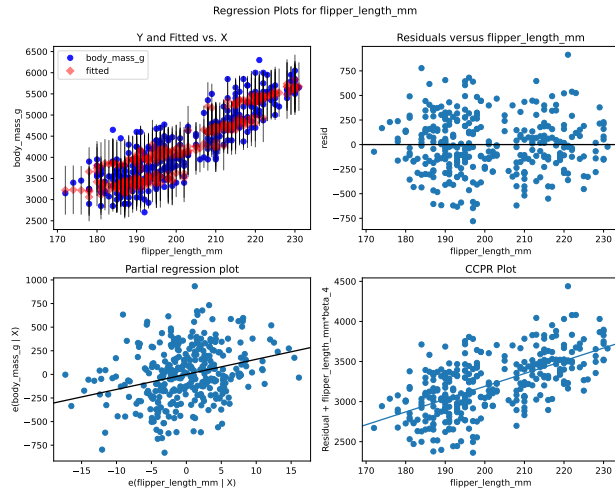
## Capítulo 6

# Regressão Parcial

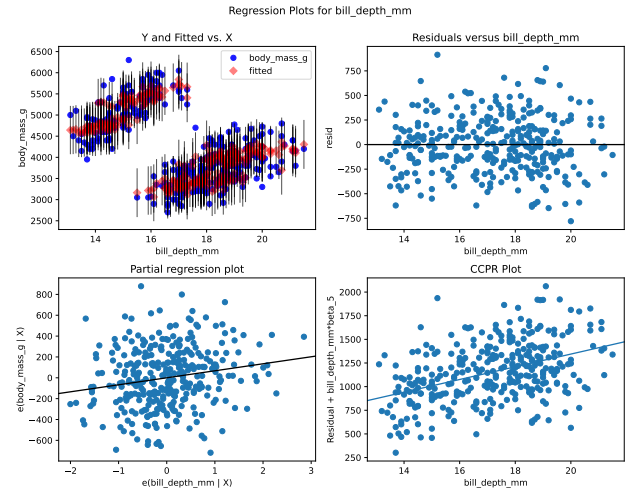
Na Figura [6.1](#) podemos observar um padrão linear passando pela origem, mostrando que as variáveis explicativas estão linearmente relacionadas com a variável resposta.

Figura 6.1: Regressão Parcial

(a) Bill Depth

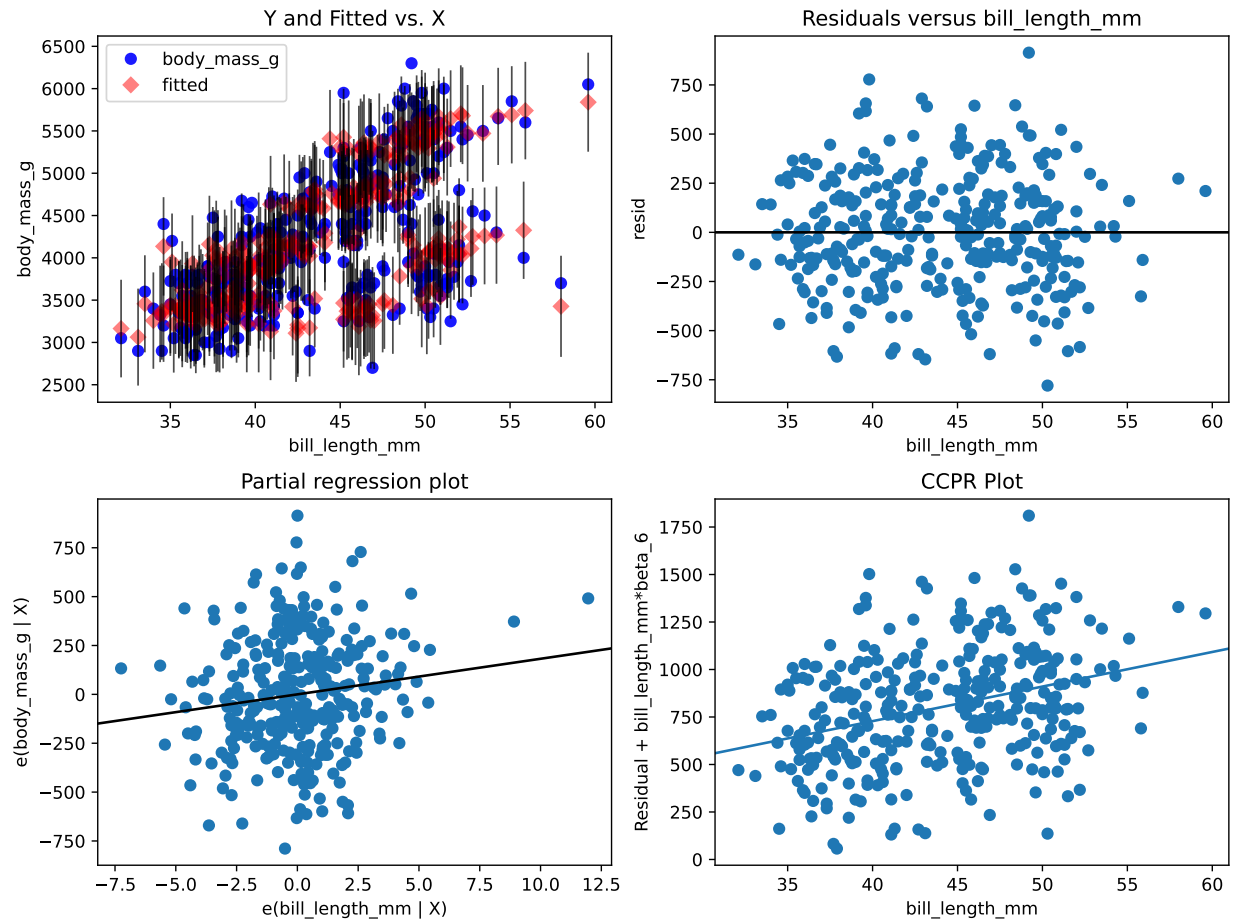


(b) Bill Length



(c) Flipper Length

Regression Plots for bill\_length\_mm





## 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:	qua, 29 nov 2023	Prob (F-statistic):	6.82e-144			
Time:	20:54:06	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.