Análise de Regressão em Python

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

Índice

Capítulo 1

Introdução

O banco de dados possui 333 linhas não nulas 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	39.1	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

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
count	333.000000	333.000000	333.00000	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

```
(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.
```

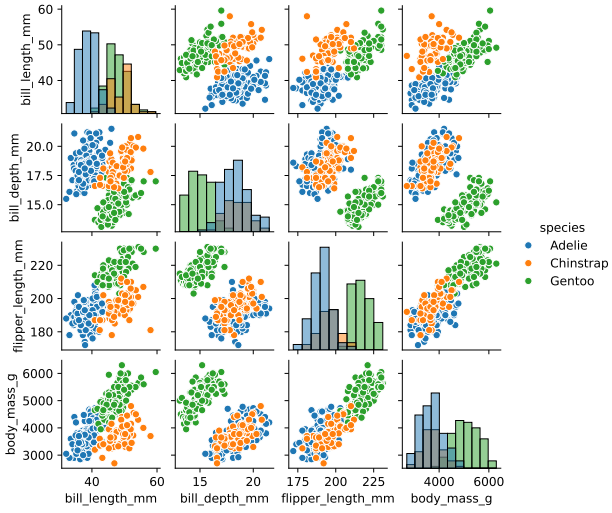
```
(array([0.5, 1.5, 2.5, 3.5]), [Text(0, 0.5, 'bill_length_mm'), Text(0, 1.5, 'bill_depth_mm'), Text(0,
```

Tabela 2.1: Quantidade de Espécies por Ilha

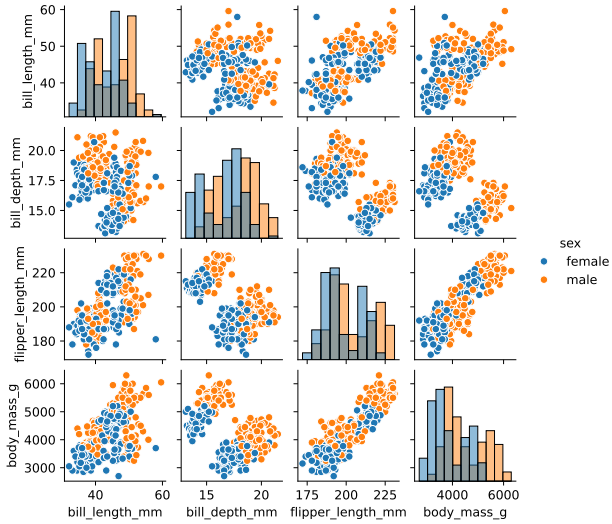
species	island	count
Adelie	Dream	55
	Torgersen	47
	Biscoe	44
Chinstrap	Dream	68
	Biscoe	0
	Torgersen	0
Gentoo	Biscoe	119
	Dream	0
	Torgersen	0

Figura 2.1: Relações em Pares por Especies

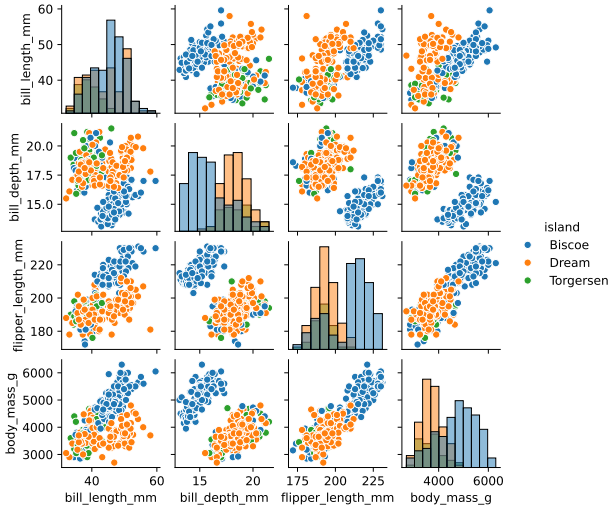
(a) Species



(b) Sex



(c) Island



(d) Year

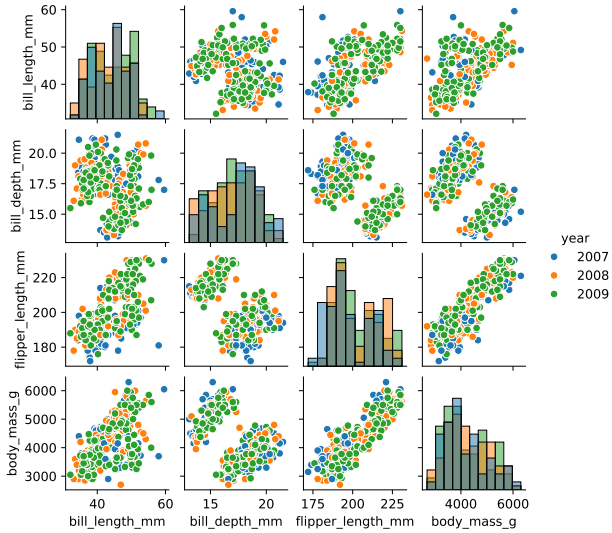
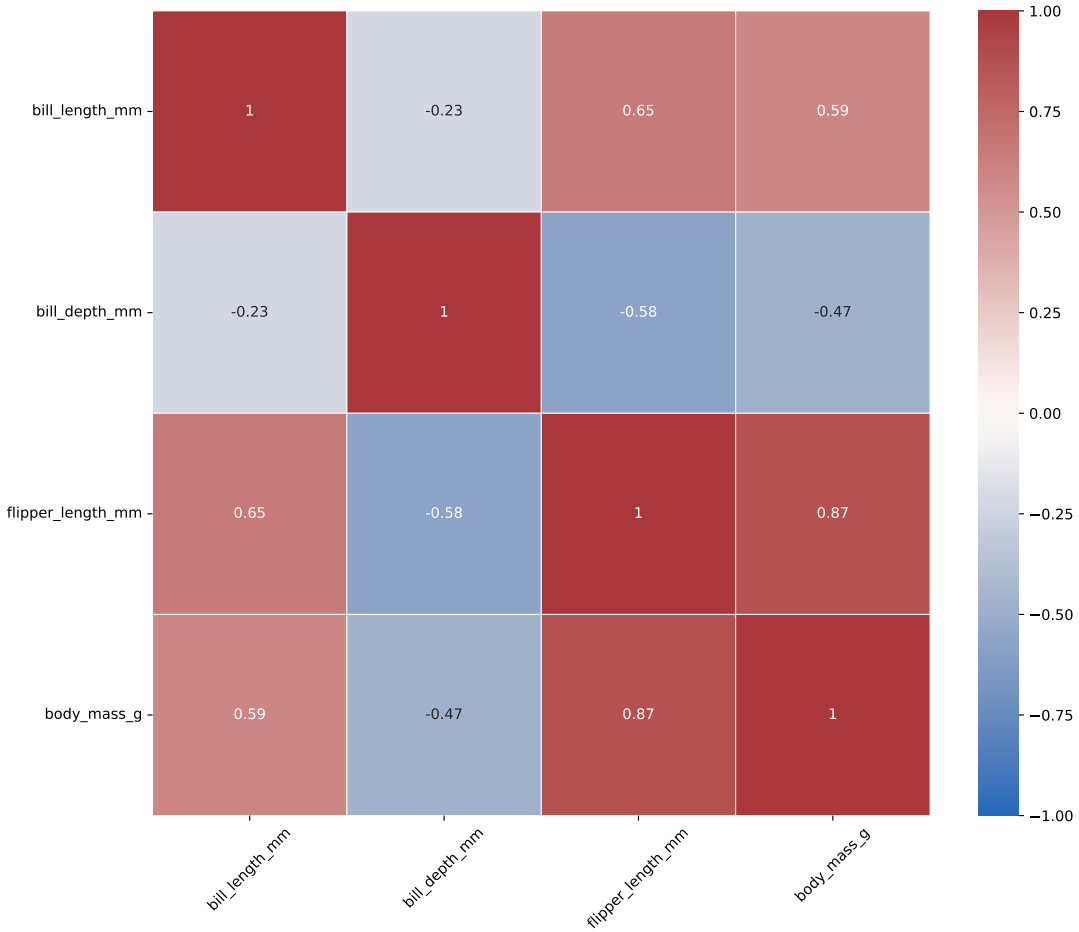


Figura 2.2: Correlações entre as Variáveis do Conjunto de Dados



Capítulo 3

Ajuste do Modelo e Multicolinearidade

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            body_mass_g   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.821
Method:                 Least Squares   F-statistic:                     509.5
Date:                qua, 22 nov 2023   Prob (F-statistic):          2.90e-123
Time:                        19:02:17   Log-Likelihood:                -2412.0
No. Observations:                 333   AIC:                             4832.
Df Residuals:                     329   BIC:                             4847.
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept         -2246.8293    625.286     -3.593      0.000   -3476.892   -1016.767
sex[T.male]         538.0800     51.310     10.487      0.000     437.144     639.017
flipper_length_mm    38.1896      2.084     18.324      0.000      34.090      42.290
bill_depth_mm       -86.9467     15.456     -5.625      0.000    -117.352     -56.541
==============================================================================
Omnibus:                        2.262   Durbin-Watson:                   1.829
Prob(Omnibus):                  0.323   Jarque-Bera (JB):                1.901
Skew:                           0.051   Prob(JB):                        0.387
Kurtosis:                       2.644   Cond. No.                     6.79e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.79e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
X = modelo.model.exog
[variance_inflation_factor(X, i) for i in range(X.shape[1])]
```

```
[1123.4483623889216, 1.8910378402256625, 2.4444701909365647, 2.653895148514561]
```

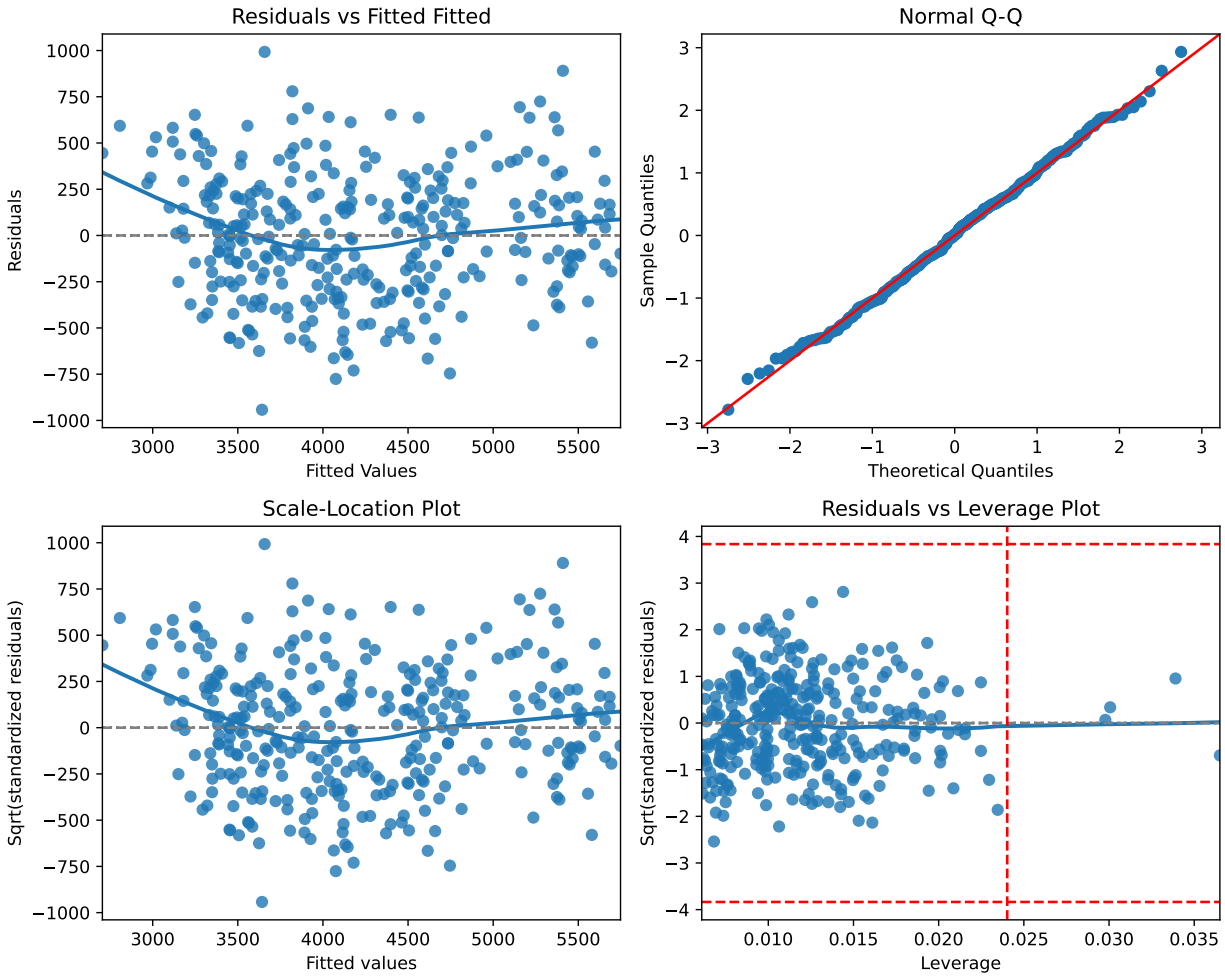
```
modelo.bse
```

```
Intercept            625.285686
sex[T.male]           51.309722
flipper_length_mm      2.084158
bill_depth_mm         15.456076
dtype: float64
```

Capítulo 4

Resíduos

Figura 4.1: Análise Gráfica dos Resíduos

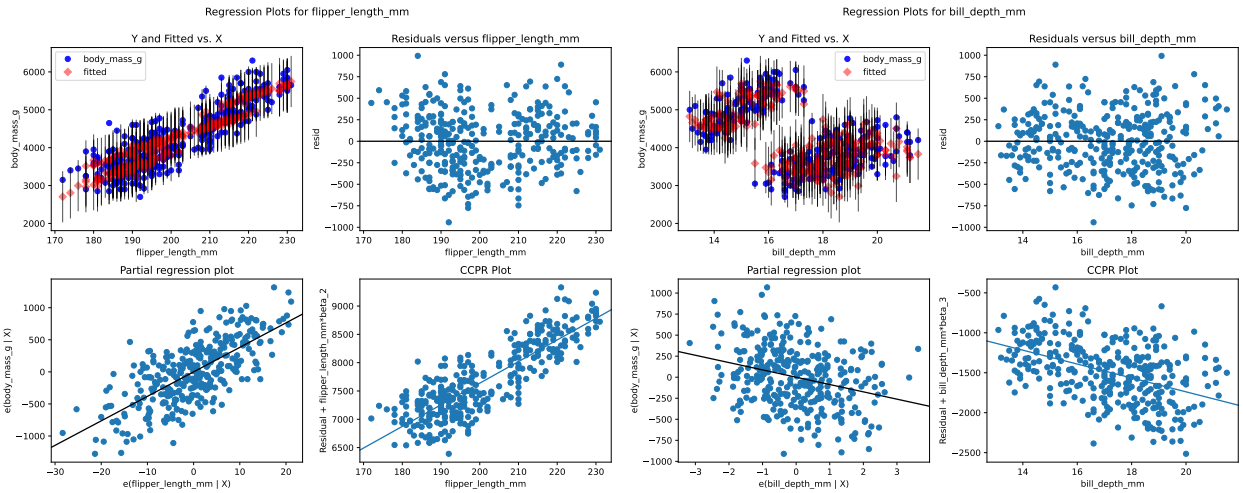



```
print('Shapiro Statistic: ', stats.shapiro(residuals)[0])
```

Figura 4.2: Regressão Parcial

(a) Flipper Length

(b) Bill Depth




```
Shapiro Statistic:  0.9966068863868713
```

```
print('Shapiro P-Value: ', stats.shapiro(residuals)[1])
```

```
Shapiro P-Value:  0.7077585458755493
```

```
print('\nDurbin Watson Statistic:', durbin_watson(residuals))
```

```
Durbin Watson Statistic: 1.8294478334581534
```

```
print('\nOutlier Test:', modelo.outlier_test(cutoff= 0.05))
```

```
Outlier Test: Empty DataFrame
Columns: [student_resid, unadj_p, bonf(p)]
Index: []
```

Capítulo 5

Influência

Tabela 5.1: Sumário das Observações Influentes

	dfb_Intercept	dfb_sex[T.male]	dfb_flipper_length_mm	dfb_bill_depth_mm	cooks_d	standard_resid	hat_diag	dffits_internal	student_resid	dffits
0	0.0544170	0.0540189	-0.0600462	-0.0367387	0.0013399	0.5113851	0.0200833	0.0732101	0.5108104	0.0731278
1	0.0412244	-0.0354016	-0.0492541	-0.0107309	0.0037656	1.3463271	0.0082413	0.1227288	1.3479979	0.1228811
2	0.0537974	0.0895363	-0.0447377	-0.0693049	0.0032073	-1.1368413	0.0098289	-0.1132653	-1.1373484	-0.1133158
4	-0.0011369	-0.0014105	0.0009155	0.0014213	0.0000008	0.0127650	0.0185777	0.0017563	0.0127456	0.0017536
5	0.0079331	-0.0042907	-0.0010815	-0.0177869	0.0003139	-0.3138702	0.0125833	-0.0354321	-0.3134398	-0.0353835

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

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

Figura 5.1: COVRATIO

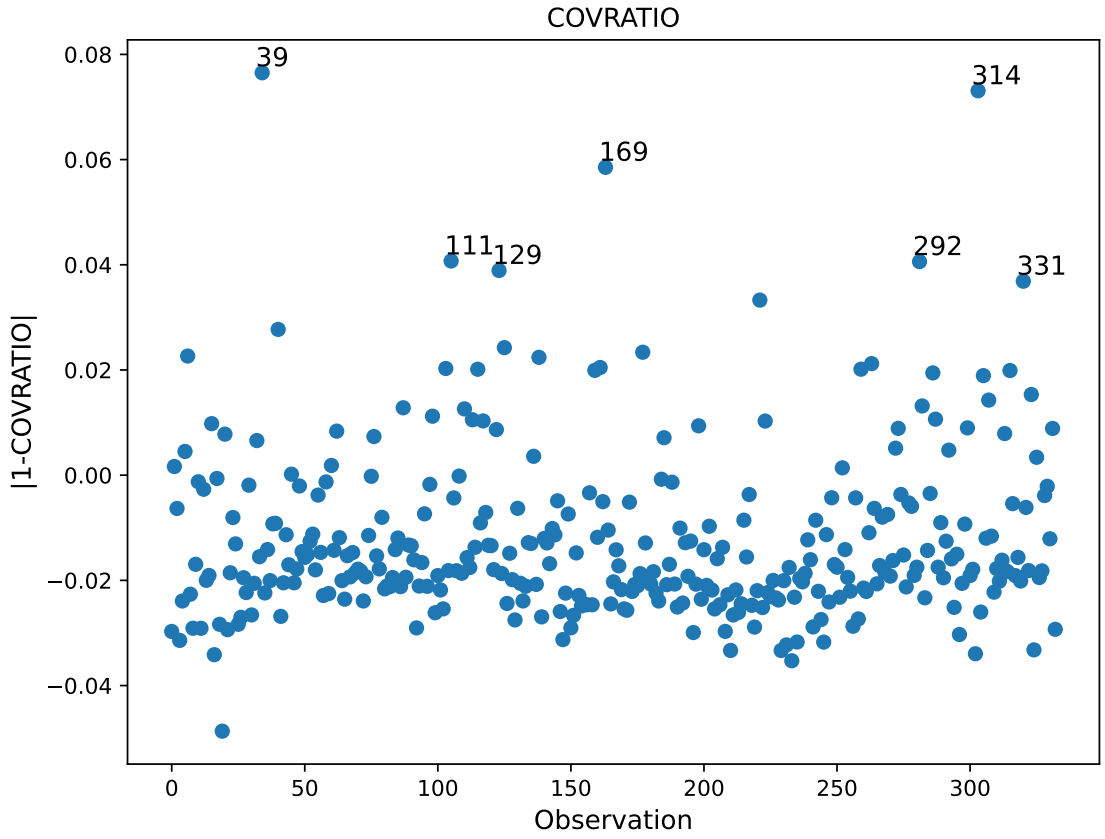
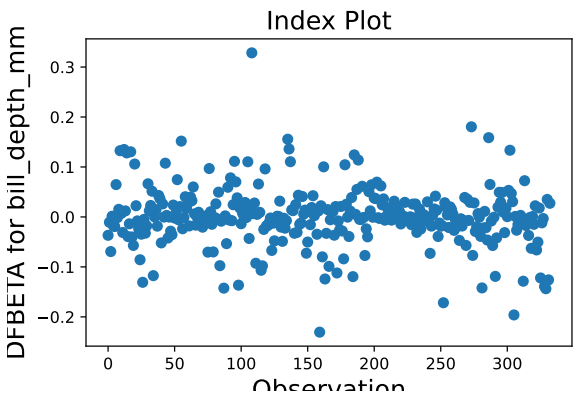


Figura 5.2: Medidas de Influência

(a) Distância de Cooks



(b) DFBetaS