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

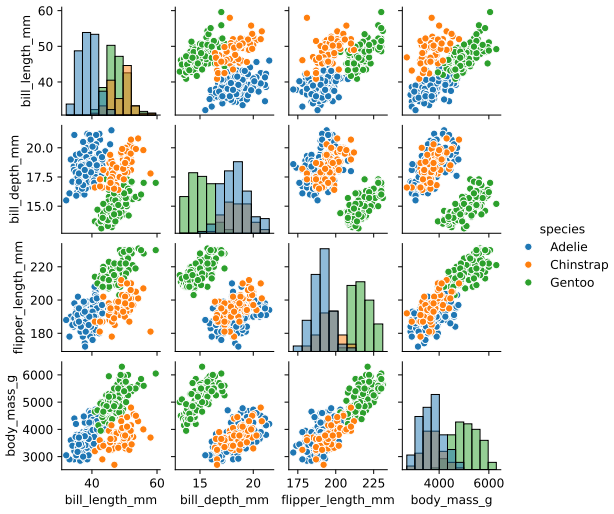
	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
0	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
1	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
2	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
4	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
5	Adelie	Torgersen	39.3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

Tabela 1.2: Sumário do Banco de Dados

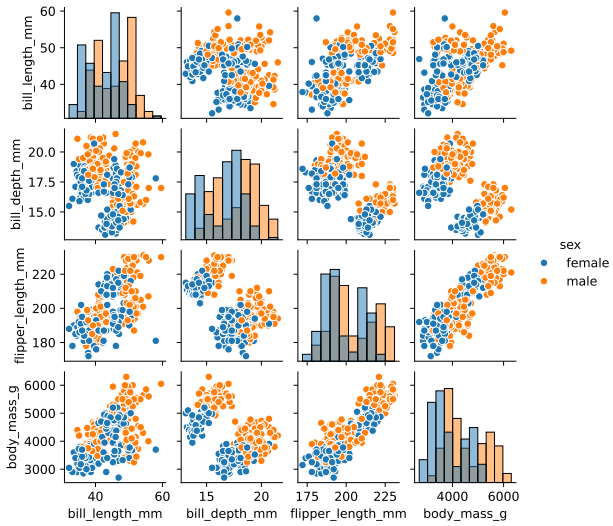
	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
count	333.000000	333.000000	333.00000	333.0000
mean	43.992793	17.164865	200.96697	4207.0571
std	5.468668	1.969235	14.01577	805.2158
min	32.100000	13.100000	172.00000	2700.0000
25%	39.500000	15.600000	190.00000	3550.0000
50%	44.500000	17.300000	197.00000	4050.0000
75%	48.600000	18.700000	213.00000	4775.0000
max	59.600000	21.500000	231.00000	6300.0000

Capítulo 2

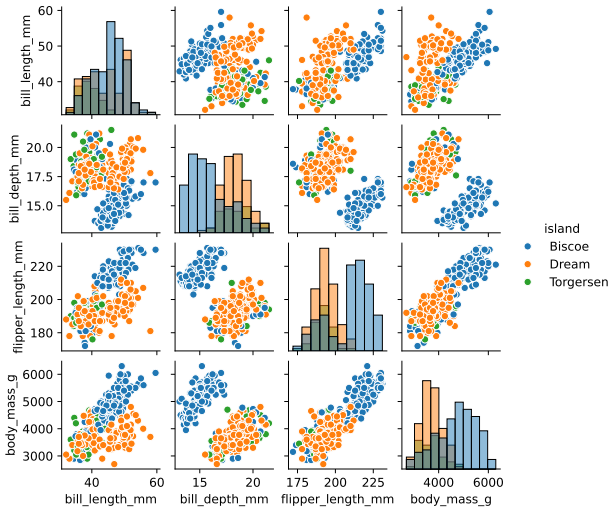
Seleção de Variáveis



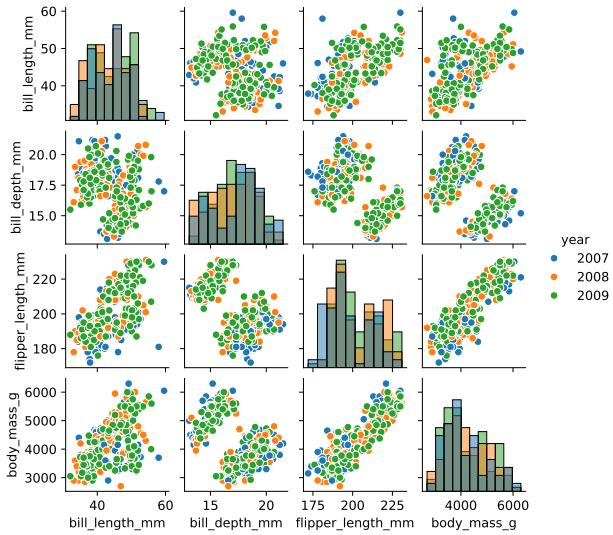
(a) Species



(b) Sex



(c) Island



(d) Year

Figura 2.1: Relações em Pares por Especies

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

```
(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,

(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,
```

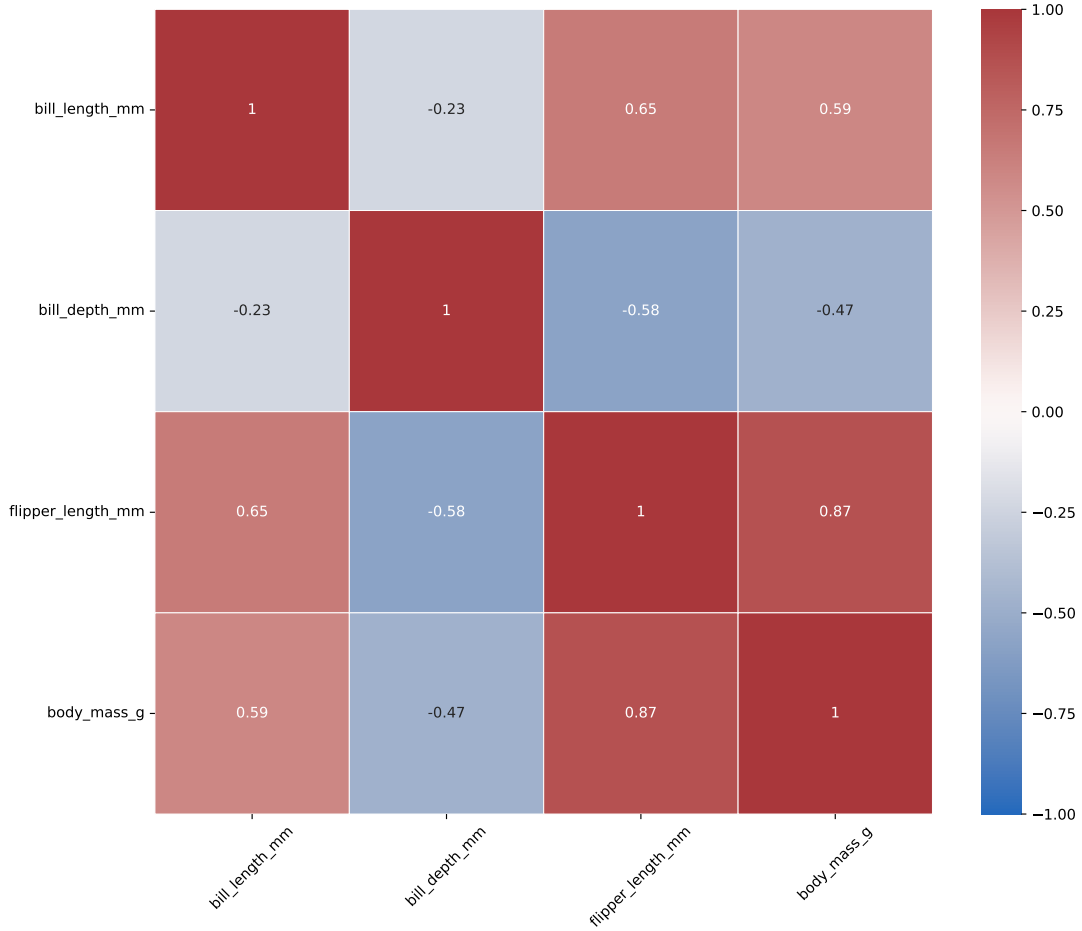


Figura 2.2: Correlações entre as Variáveis do Conjunto de Dados

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:                     381.3
Date:                sex, 24 nov 2023   Prob (F-statistic):          6.28e-122
Time:                        19:52:39   Log-Likelihood:                -2411.8
No. Observations:                 333   AIC:                             4834.
Df Residuals:                     328   BIC:                             4853.
Df Model:                           4
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept         -2288.4650    631.580     -3.623      0.000   -3530.924   -1046.006
sex[T.male]         541.0285     51.710     10.463      0.000     439.304     642.753
flipper_length_mm    38.8258      2.448     15.862      0.000      34.011      43.641
bill_depth_mm       -86.0882     15.570     -5.529      0.000    -116.718     -55.459
bill_length_mm       -2.3287      4.684     -0.497      0.619     -11.544       6.886
==============================================================================
Omnibus:                        2.598   Durbin-Watson:                   1.843
Prob(Omnibus):                  0.273   Jarque-Bera (JB):                2.125
Skew:                           0.062   Prob(JB):                        0.346
Kurtosis:                       2.629   Cond. No.                     7.01e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.01e+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])]
```

```
[1143.5578441010605, 1.9162481866596275, 3.364085891555125, 2.6869459960037685, 1.8756337894608972]
```

```
modelo.bse
```

```
Intercept            631.580155
sex[T.male]           51.709806
flipper_length_mm      2.447762
bill_depth_mm         15.569845
bill_length_mm         4.684302
dtype: float64
```

Capítulo 4

Resíduos

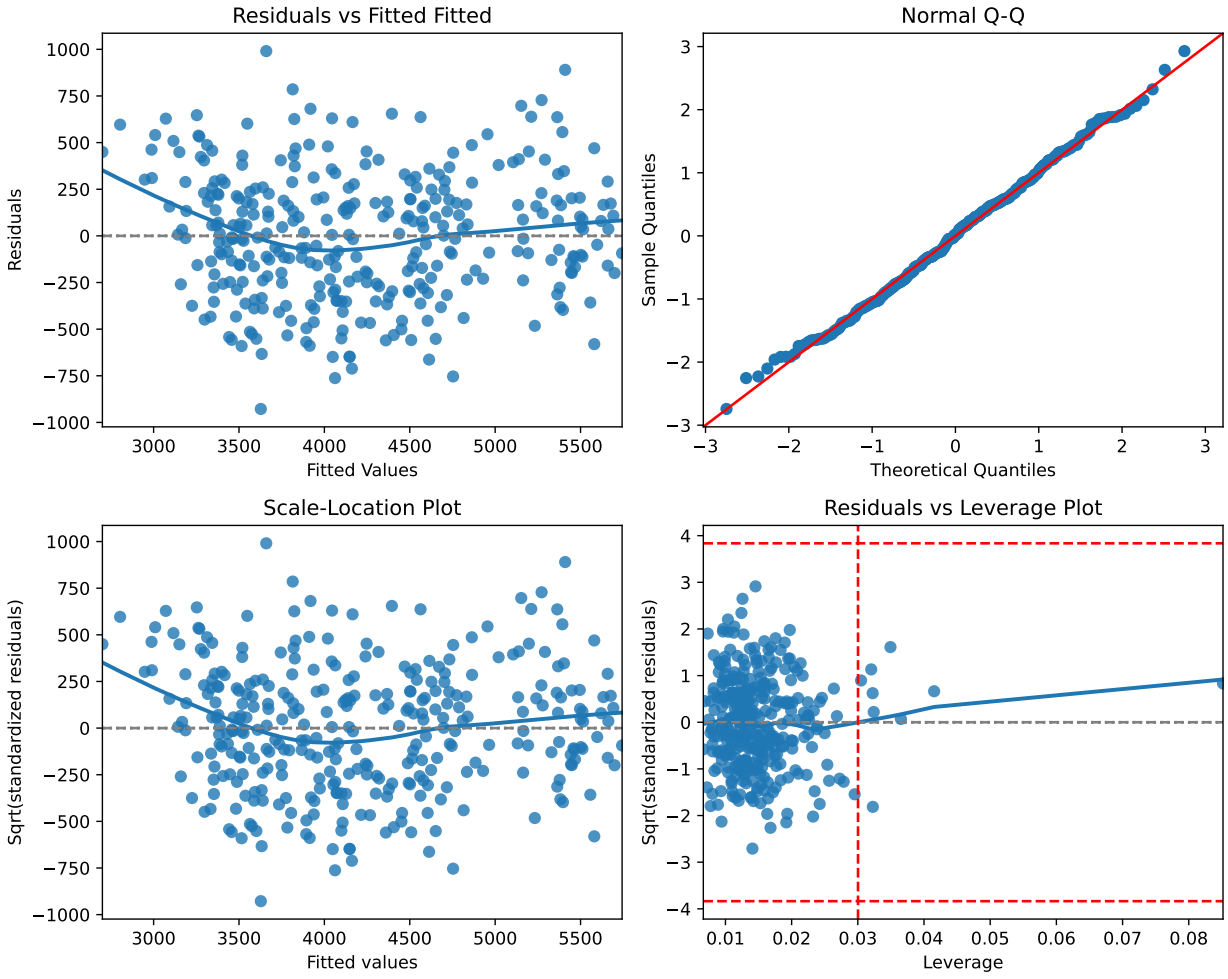


Figura 4.1: Análise Gráfica dos Resíduos

```
Shapiro Statistic:  0.996

Shapiro P-Value:  0.615

Durbin Watson Statistic: 1.8428622741934784
```

No gráfico de `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 relativamente aleatório dos resíduos em torno de zero, apesar de não ser perfeito.

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.

4.1 Influência

Tabela 4.1: Sumário das Observações Influentes

	dfb_Intercept	dfb_sex[T.male]	dfb_flipper_length_mm	dfb_bill_depth_mm	dfb_bill_length_mm	cooks_d	standard_resid	hat_diag	dffits_internal	student_resid	dffits
0	0.0528312	0.0536558	-0.0483769	-0.0356701	-0.0044319	0.0010554	0.5064642	0.0201585	0.0726440	0.5058894	0.0725616
1	0.0412107	-0.0354711	-0.0434127	-0.0109668	0.0027601	0.0030117	1.3458068	0.0082455	0.1227126	1.3474792	0.1228651
2	0.0574951	0.0866776	-0.0528788	-0.0726085	0.0274734	0.0027676	-1.1477585	0.0103953	-0.1176356	-1.1483158	-0.1176927
4	0.0023199	0.0024016	-0.0023511	-0.0028125	0.0017273	0.0000027	-0.0236217	0.0238131	-0.0036894	-0.0235857	-0.0036837
5	0.0107232	-0.0065326	-0.0097865	-0.0209595	0.0168132	0.0003496	-0.3386152	0.0150182	-0.0418121	-0.3381578	-0.0417556

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

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

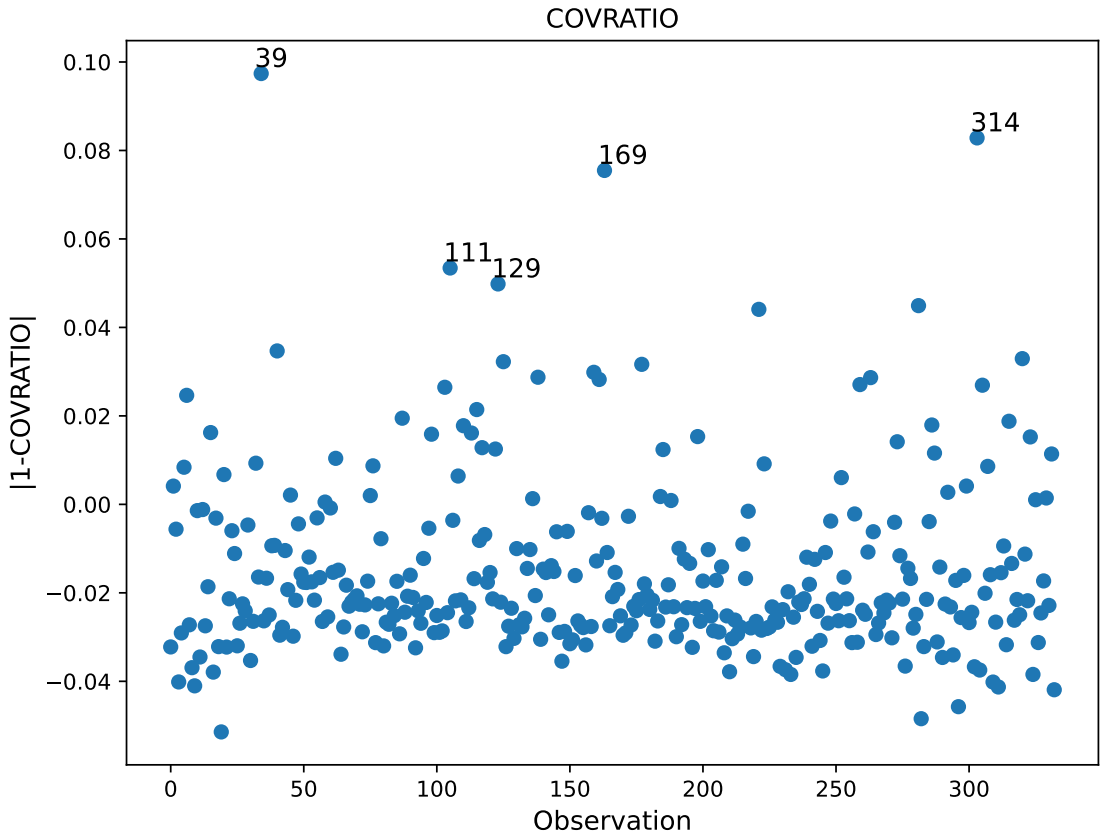
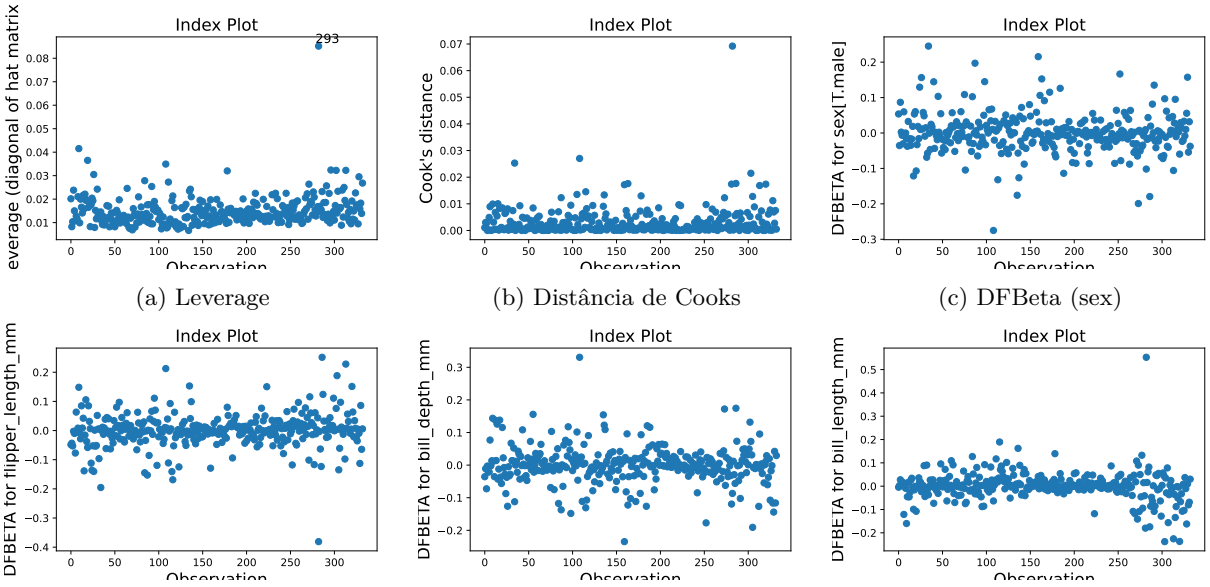


Figura 4.2: COVRATIO



(a) Leverage (b) Distância de Cooks (c) DFBeta (sex)

(d) DFBeta (flipper_length_mm) (e) DF Beta (bill_depth_mm) (f) DF Beta (bill_length_mm)

Figura 4.3: Medidas de 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.

4.2 Regressão Parcial

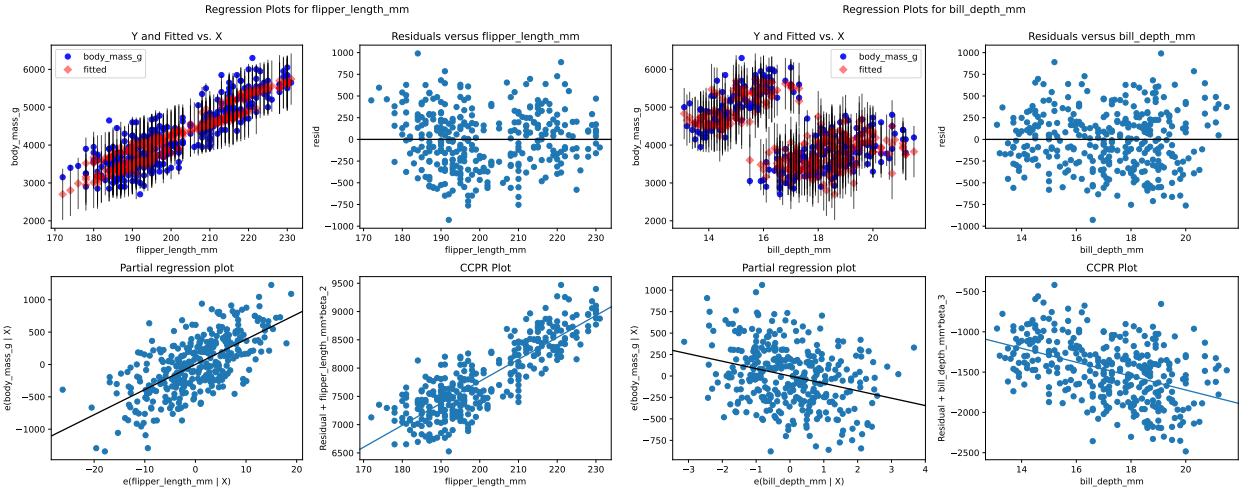
Ao analisar o Figura 4.4c, notamos um padrão nulo na variável `bill_length_mm`, sugerindo que essa variável explicativa pode não ser necessária no modelo. Para confirmar essa suspeita, realizamos o teste t, e a variável `bill_length_mm` apresenta um valor p de 0.619. Com base nesse resultado, temos evidências suficientes para considerar a remoção dessa variável do modelo.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            body_mass_g   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.821
Method:                 Least Squares   F-statistic:                     381.3
Date:                sex, 24 nov 2023   Prob (F-statistic):          6.28e-122
Time:                        19:52:47   Log-Likelihood:                -2411.8
No. Observations:                 333   AIC:                             4834.
Df Residuals:                     328   BIC:                             4853.
Df Model:                           4
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept         -2288.4650    631.580     -3.623      0.000   -3530.924   -1046.006
sex[T.male]         541.0285     51.710     10.463      0.000     439.304     642.753
flipper_length_mm    38.8258      2.448     15.862      0.000      34.011      43.641
bill_depth_mm       -86.0882     15.570     -5.529      0.000    -116.718     -55.459
bill_length_mm       -2.3287      4.684     -0.497      0.619     -11.544       6.886
==============================================================================
Omnibus:                        2.598   Durbin-Watson:                   1.843
Prob(Omnibus):                  0.273   Jarque-Bera (JB):                2.125
Skew:                           0.062   Prob(JB):                        0.346
Kurtosis:                       2.629   Cond. No.                     7.01e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.01e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

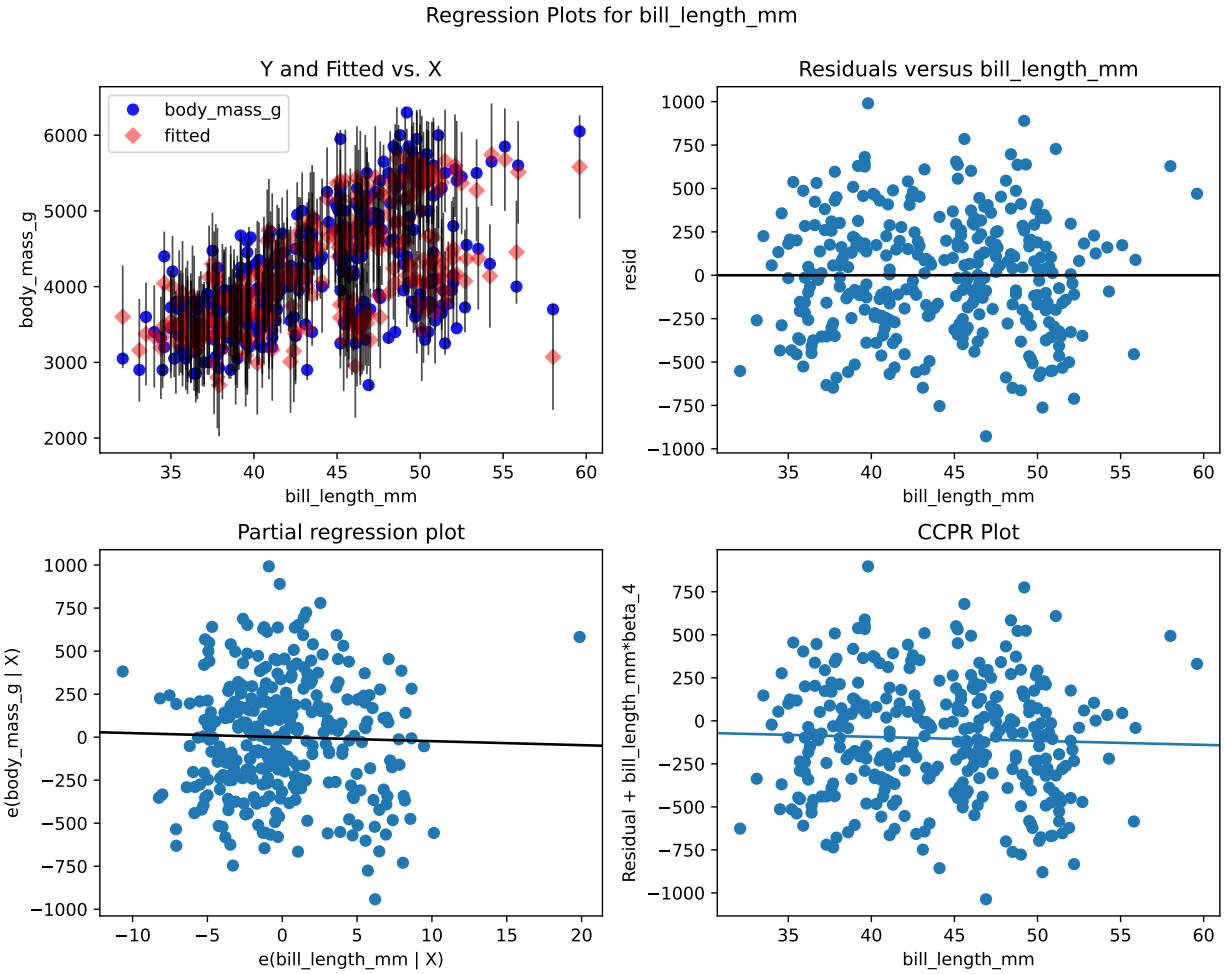
Removendo 'bill__length__mm'

```
                            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:                sex, 24 nov 2023   Prob (F-statistic):          2.90e-123
Time:                        19:52:47   Log-Likelihood:                -2412.0
No. Observations:                 333   AIC:                             4832.
Df Residuals:                     329   BIC:                             4847.
Df Model:                           3
Covariance Type:            nonrobust
```



(a) Flipper Length

(b) Bill Depth



(c) Bill Length

Figura 4.4: Regressão Parcial

```
==============================================================================
                 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 5

Conclusões