More Linear (and Logistic) Regression

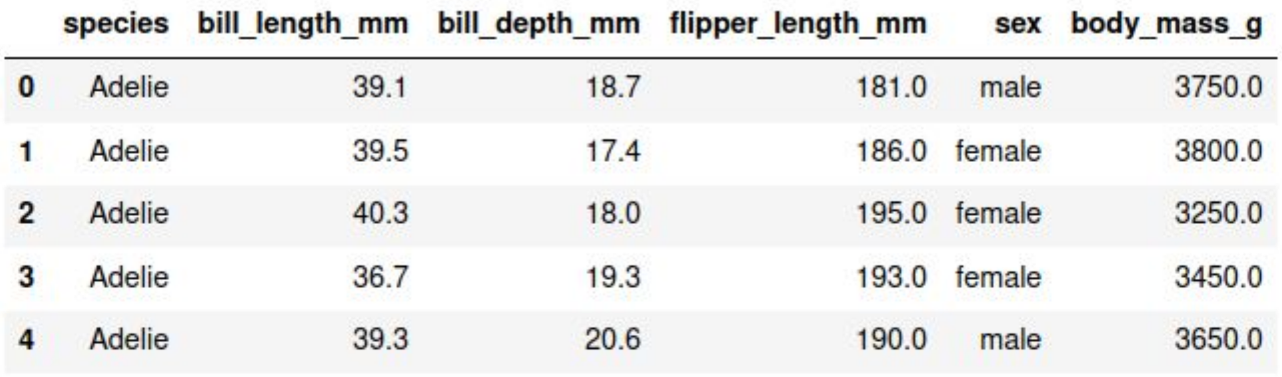
Linear Regression

Given *k* predictors $x^{(1)}$, $x^{(2)}$,...,$x^{(k)}$, linear regression uses the following equation to predict the target variable:

$$\hat{f}(\vec{x}) = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_k x^{(k)}$$

Here, β_0, β_1,...,β_k are constants that are determined by using the available training data.

Linear Regression - Example

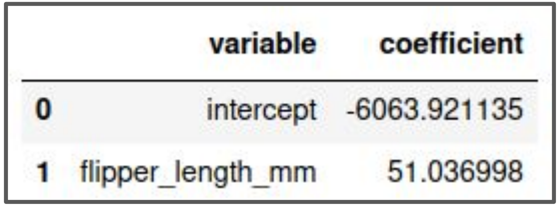


	species	bill_length_mm	bill_depth_mm	flipper_length_mm	sex	body_mass_g
0	Adelie	39.1	18.7	181.0	male	3750.0
1	Adelie	39.5	17.4	186.0	female	3800.0
2	Adelie	40.3	18.0	195.0	female	3250.0
3	Adelie	36.7	19.3	193.0	female	3450.0
4	Adelie	39.3	20.6	190.0	male	3650.0

Consider the penguins dataset.

Say we want to build a linear model to predict body_mass_g.

Linear Regression - Example

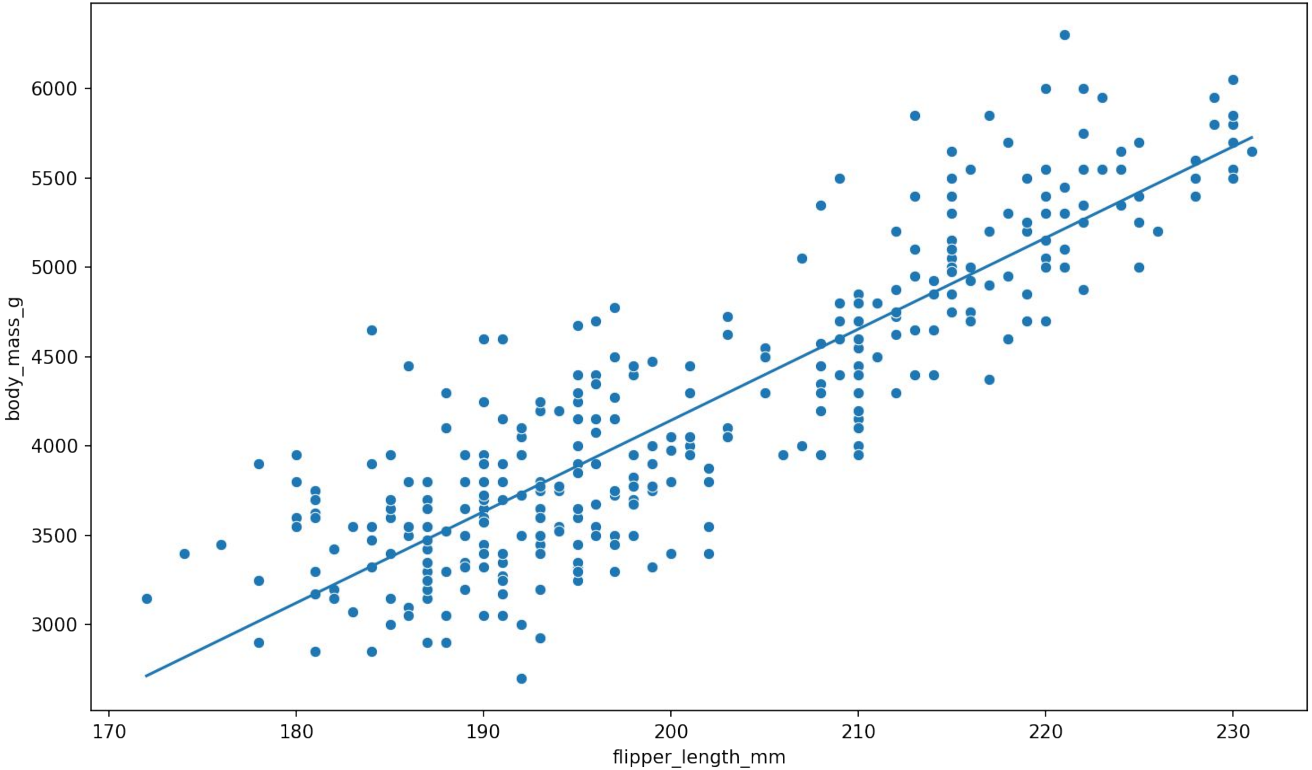


	variable	coefficient
0	intercept	-6063.921135
1	flipper_length_mm	51.036998

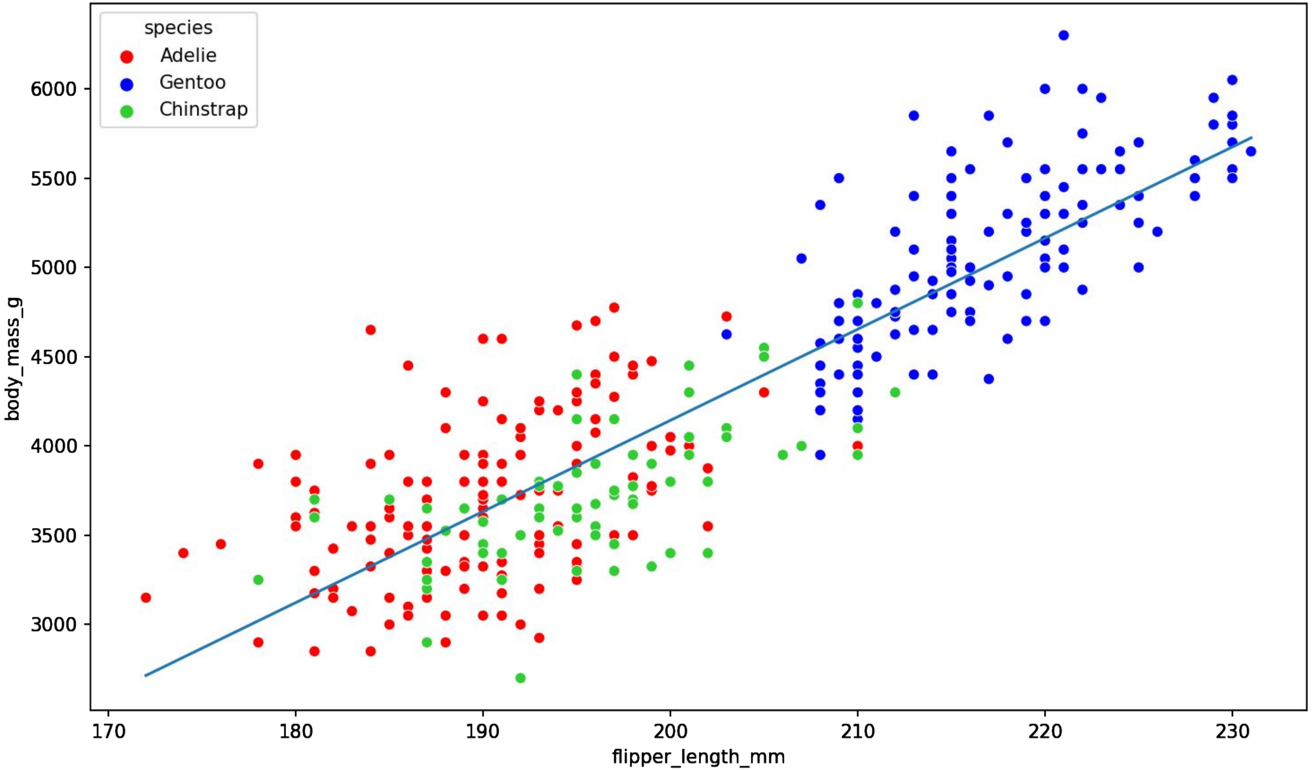
Using just flipper length gives these coefficients.

$$\text{predicted body\_mass} = -6064 + 51 \cdot \text{flipper\_length}$$

Linear Regression - Example



Linear Regression - Example



Linear Regression - Example

What if we want to include the species information?

Linear Regression - Example

What if we want to include the species information?

Species is a categorical variable, but we can include it if we make dummy columns. This is also known as one-hot encoding.

Linear Regression - Example

What if we want to include the species information?

Species is a categorical variable, but we can include it if we make dummy columns. This is also known as one-hot encoding.

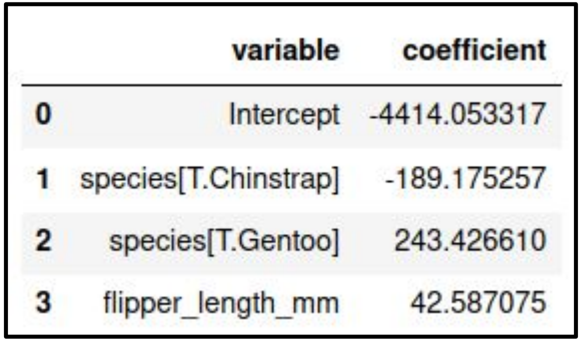
We have 3 species (Adelie, Chinstrap, and Gentoo), so we’ll create two new 0/1 columns.

Linear Regression - Example

species
Adelie
Chinstrap
Gentoo

species_Chinstrap	**species_Gentoo**
0	0
1	0
0	1

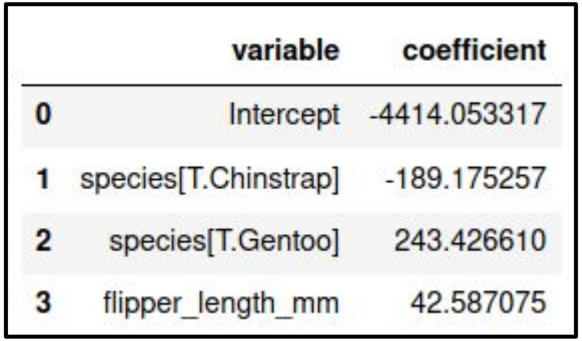
Linear Regression - Example



	variable	coefficient
0	Intercept	-4414.053317
1	species[T.Chinstrap]	-189.175257
2	species[T.Gentoo]	243.426610
3	flipper_length_mm	42.587075

$$y = -4414 - 189(\text{Chinstrap}) + 243 \cdot (\text{Gentoo}) + 43 \cdot \text{flipper\_length}$$

Linear Regression - Example

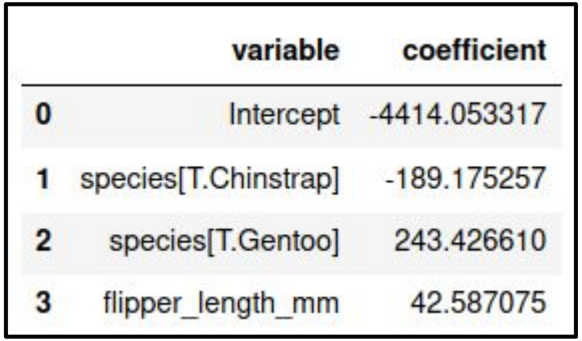


	variable	coefficient
0	Intercept	-4414.053317
1	species[T.Chinstrap]	-189.175257
2	species[T.Gentoo]	243.426610
3	flipper_length_mm	42.587075

$$y = -4414 - 189(\text{Chinstrap}) + 243 \cdot (\text{Gentoo}) + 43 \cdot \text{flipper\_length}$$

We have three different parallel lines, one per species.

Linear Regression - Example



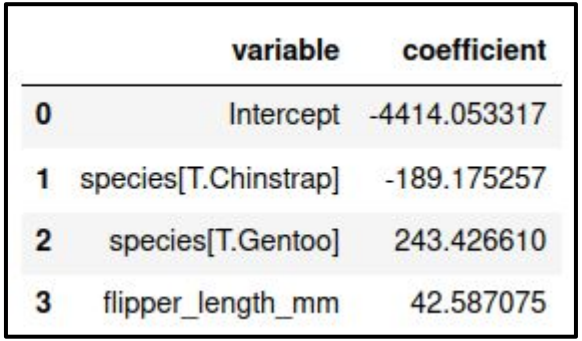
	variable	coefficient
0	Intercept	-4414.053317
1	species[T.Chinstrap]	-189.175257
2	species[T.Gentoo]	243.426610
3	flipper_length_mm	42.587075

$$y = -4414 - 189(\text{Chinstrap}) + 243 \cdot (\text{Gentoo}) + 43 \cdot \text{flipper\_length}$$

We have three different parallel lines, one per species.

Adelie: $y = -4414 + 43 \cdot \text{flipper\_length}$

Linear Regression - Example



	variable	coefficient
0	Intercept	-4414.053317
1	species[T.Chinstrap]	-189.175257
2	species[T.Gentoo]	243.426610
3	flipper_length_mm	42.587075

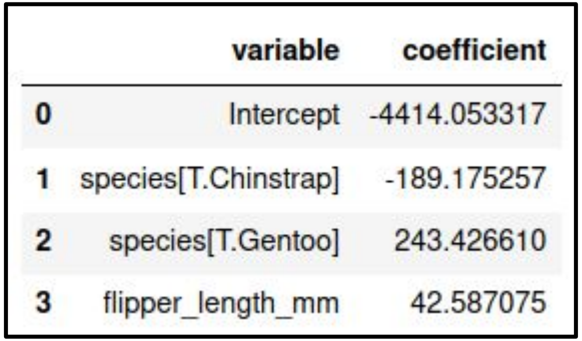
$$y = -4414 - 189(\text{Chinstrap}) + 243 \cdot (\text{Gentoo}) + 43 \cdot \text{flipper\_length}$$

We have three different parallel lines, one per species.

Adelie: $y = -4414 + 43 \cdot \text{flipper\_length}$

Chinstrap: $y = -4603 + 43 \cdot \text{flipper\_length}$

Linear Regression - Example



	variable	coefficient
0	Intercept	-4414.053317
1	species[T.Chinstrap]	-189.175257
2	species[T.Gentoo]	243.426610
3	flipper_length_mm	42.587075

$$y = -4414 - 189(\text{Chinstrap}) + 243 \cdot (\text{Gentoo}) + 43 \cdot \text{flipper\_length}$$

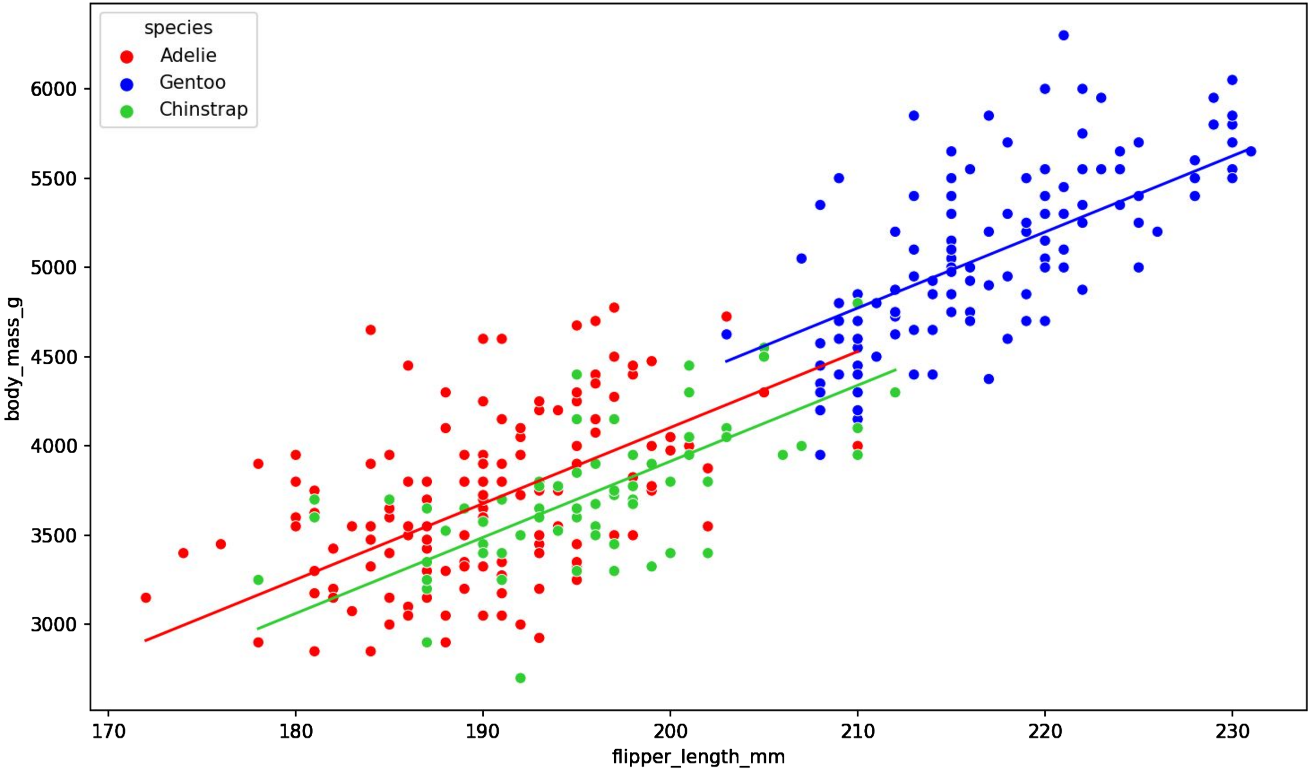
We have three different parallel lines, one per species.

Adelie: $y = -4414 + 43 \cdot \text{flipper\_length}$

Chinstrap: $y = -4603 + 43 \cdot \text{flipper\_length}$

Gentoo: $y = -4171 + 43 \cdot \text{flipper\_length}$

Linear Regression - Example



Linear Regression - Example

Just adding dummy columns limits us to just changing the intercept but not the slope per species.

Linear Regression - Example

Just adding dummy columns limits us to just changing the intercept but not the slope per species.

If we think that perhaps the effect of flipper length should be different per species, we can add **interaction terms**.

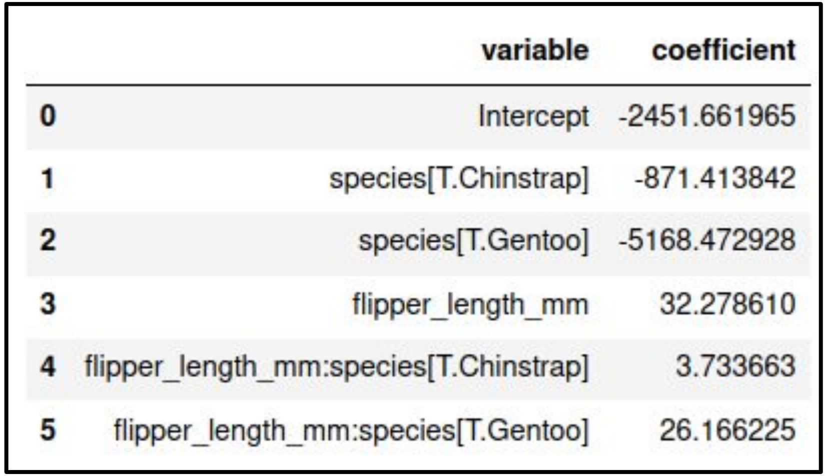
Linear Regression - Example

Just adding dummy columns limits us to just changing the intercept but not the slope per species.

If we think that perhaps the effect of flipper length should be different per species, we can add **interaction terms**.

We get these by multiplying the value across two variables.

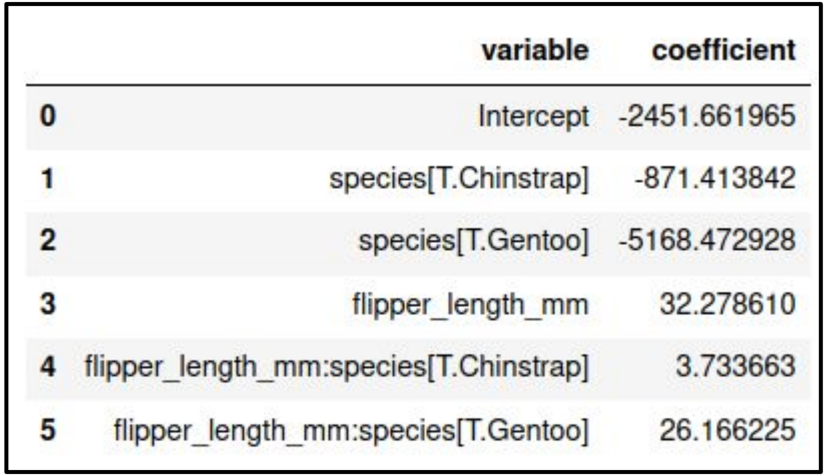
Linear Regression - Example



	variable	coefficient
0	Intercept	-2451.661965
1	species[T.Chinstrap]	-871.413842
2	species[T.Gentoo]	-5168.472928
3	flipper_length_mm	32.278610
4	flipper_length_mm:species[T.Chinstrap]	3.733663
5	flipper_length_mm:species[T.Gentoo]	26.166225

Now, we have 3 different lines, one per species:

Linear Regression - Example

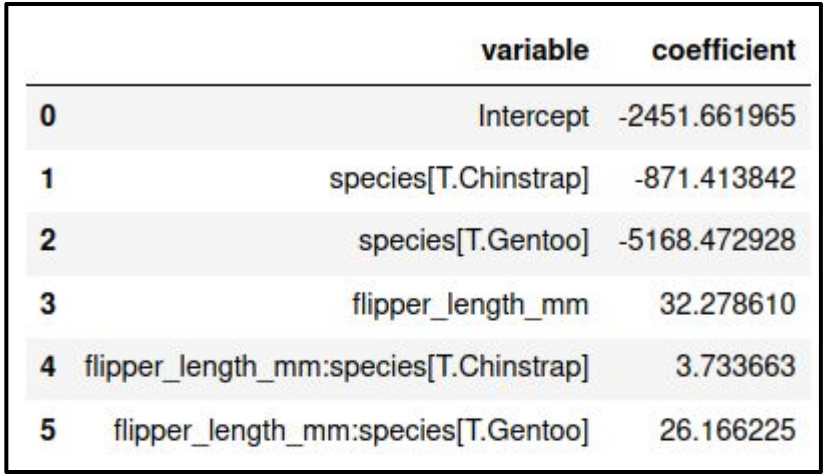


	variable	coefficient
0	Intercept	-2451.661965
1	species[T.Chinstrap]	-871.413842
2	species[T.Gentoo]	-5168.472928
3	flipper_length_mm	32.278610
4	flipper_length_mm:species[T.Chinstrap]	3.733663
5	flipper_length_mm:species[T.Gentoo]	26.166225

Now, we have 3 different lines, one per species:

Adelie: $y = -2452 + 32 \cdot \text{flipper\_length}$

Linear Regression - Example



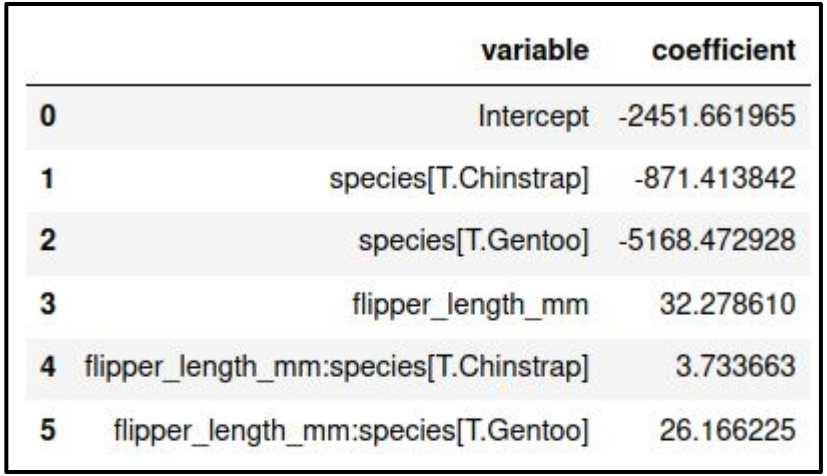
	variable	coefficient
0	Intercept	-2451.661965
1	species[T.Chinstrap]	-871.413842
2	species[T.Gentoo]	-5168.472928
3	flipper_length_mm	32.278610
4	flipper_length_mm:species[T.Chinstrap]	3.733663
5	flipper_length_mm:species[T.Gentoo]	26.166225

Now, we have 3 different lines, one per species:

Adelie: $y = -2452 + 32 \cdot \text{flipper\_length}$

Chinstrap: $y = -3323 + 36 \cdot \text{flipper\_length}$

Linear Regression - Example



	variable	coefficient
0	Intercept	-2451.661965
1	species[T.Chinstrap]	-871.413842
2	species[T.Gentoo]	-5168.472928
3	flipper_length_mm	32.278610
4	flipper_length_mm:species[T.Chinstrap]	3.733663
5	flipper_length_mm:species[T.Gentoo]	26.166225

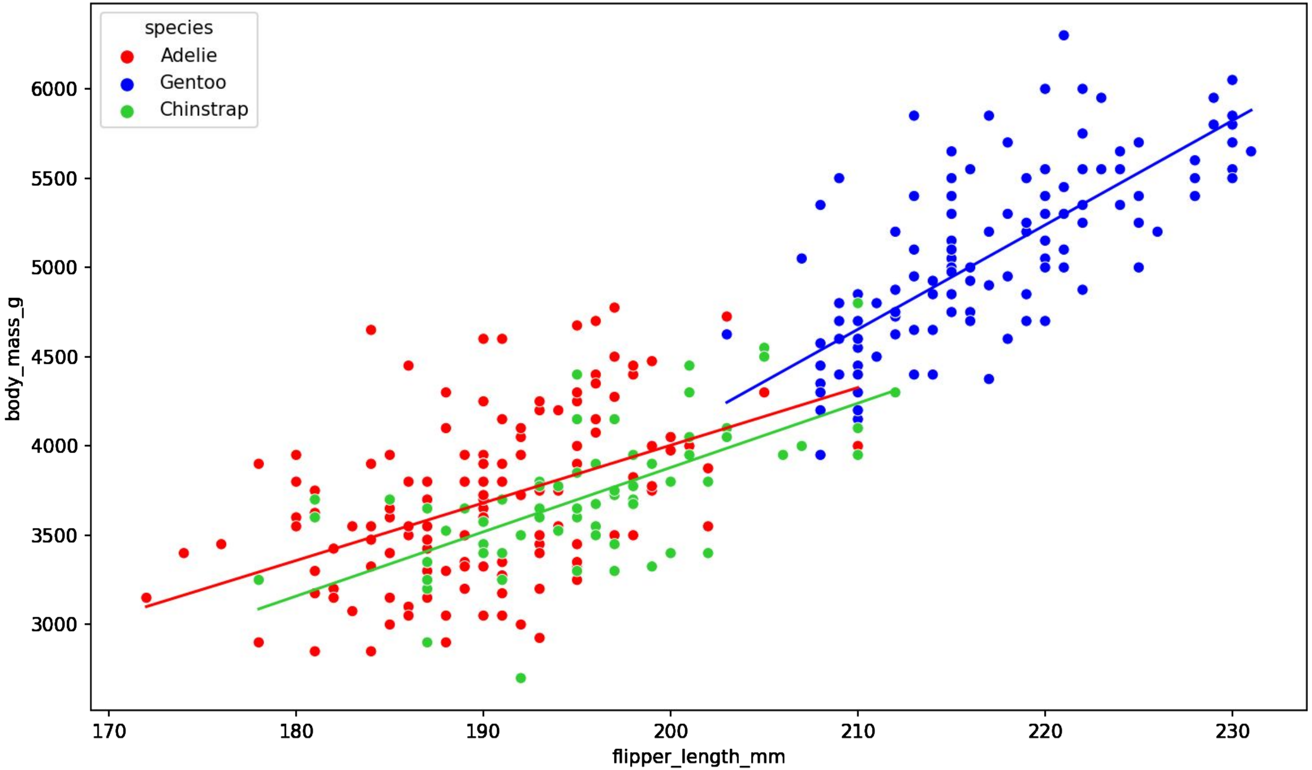
Now, we have 3 different lines, one per species:

Adelie: $y = -2452 + 32 \cdot \text{flipper\_length}$

Chinstrap: $y = -3323 + 36 \cdot \text{flipper\_length}$

Gentoo: $y = -7620 + 58 \cdot \text{flipper\_length}$

Linear Regression - Example

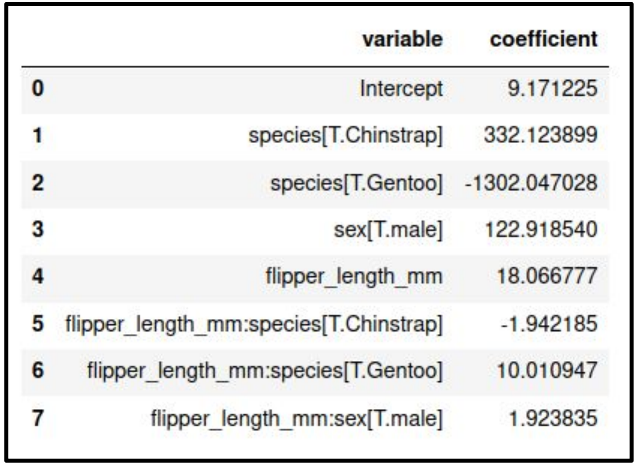


Linear Regression - Example

What if we also include the sex variable?

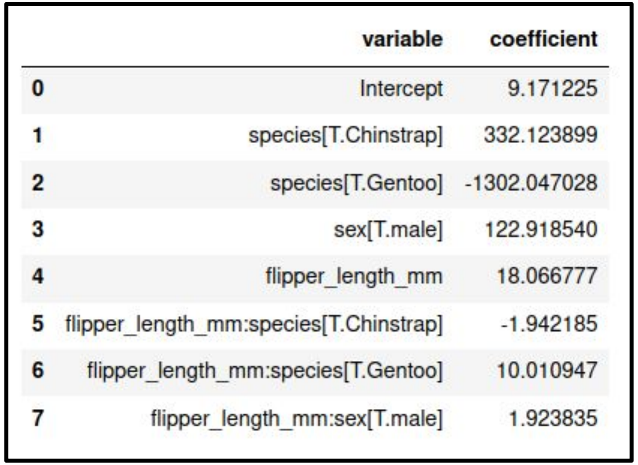
We’ll add it and the interactions with the flipper length.

Linear Regression - Example



	variable	coefficient
0	Intercept	9.171225
1	species[T.Chinstrap]	332.123899
2	species[T.Gentoo]	-1302.047028
3	sex[T.male]	122.918540
4	flipper_length_mm	18.066777
5	flipper_length_mm:species[T.Chinstrap]	-1.942185
6	flipper_length_mm:species[T.Gentoo]	10.010947
7	flipper_length_mm:sex[T.male]	1.923835

Linear Regression - Example

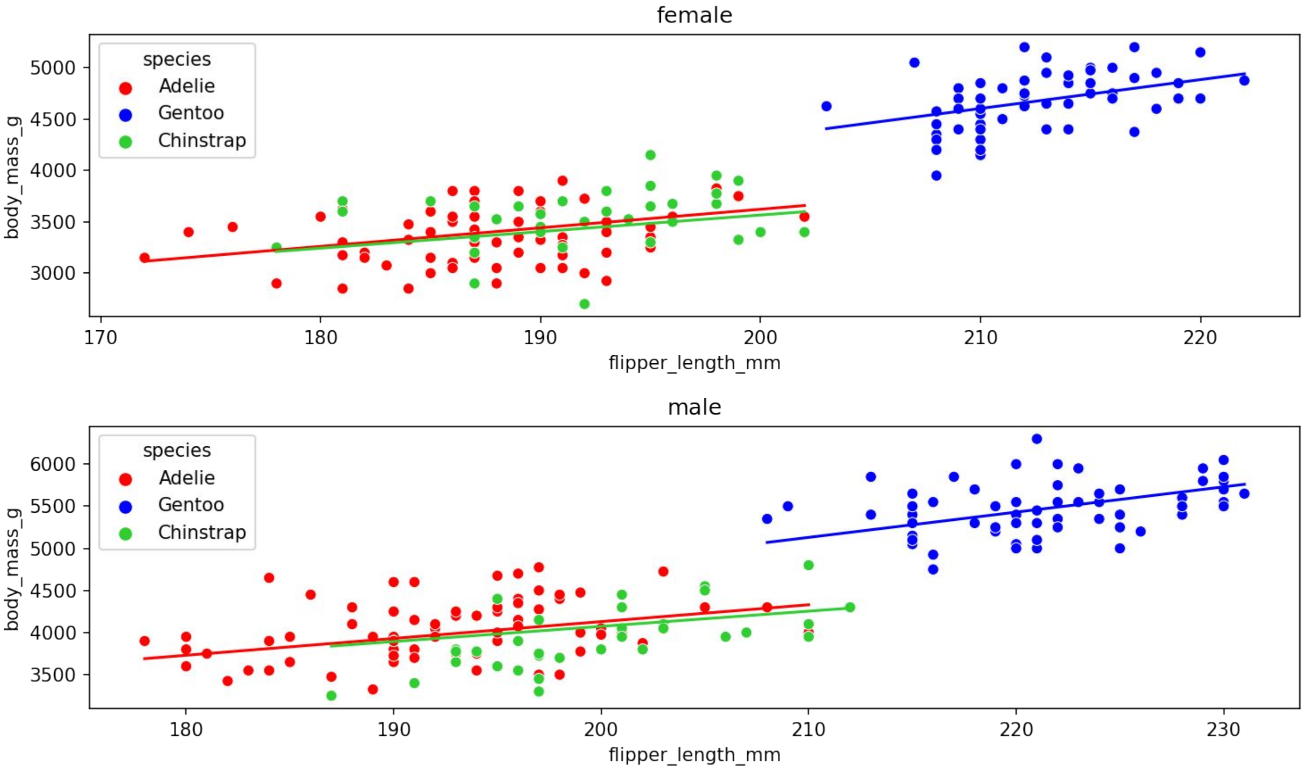


	variable	coefficient
0	Intercept	9.171225
1	species[T.Chinstrap]	332.123899
2	species[T.Gentoo]	-1302.047028
3	sex[T.male]	122.918540
4	flipper_length_mm	18.066777
5	flipper_length_mm:species[T.Chinstrap]	-1.942185
6	flipper_length_mm:species[T.Gentoo]	10.010947
7	flipper_length_mm:sex[T.male]	1.923835

Now, we have 6 different lines, one per species/sex combination:

	female	male
Adelie	y = 9 + 18 · (flipper length)	y = 132 + 20 · (flipper length)
Chinstrap	y = 341 + 16 · (flipper length)	y = 464 + 18 · (flipper length)
Gentoo	y = -1293 + 28 · (flipper length)	y = -1170 + 30 · (flipper length)

Linear Regression - Example



Linear Regression - Example

Question: Would we ever not want to do this? What are the potential downsides?