Data Analysis 2 - Assigment 2

Modelling Fish Weights based on Length, Height, Width and Species

Yuri Almeida Cunha

2021-01-01

Introduction

The objective of the analysis is to develop a multiple regression model predicting the **Weight** in grams **(dependent variable)** of a fish using its **physical dimensions**: Height, Width, Lengths (1,2,3) and its specie type. **(Explanatory Variables)**.

The **Lengths** are continuous variables measured in **centimeters (cm)**.

The **Height** is the maximum height as % of **Length3 (Height = Height/Length3 * 100)**

The **Width** is the maximum height as % of **Length3 (Width = Width/Length3 * 100)**

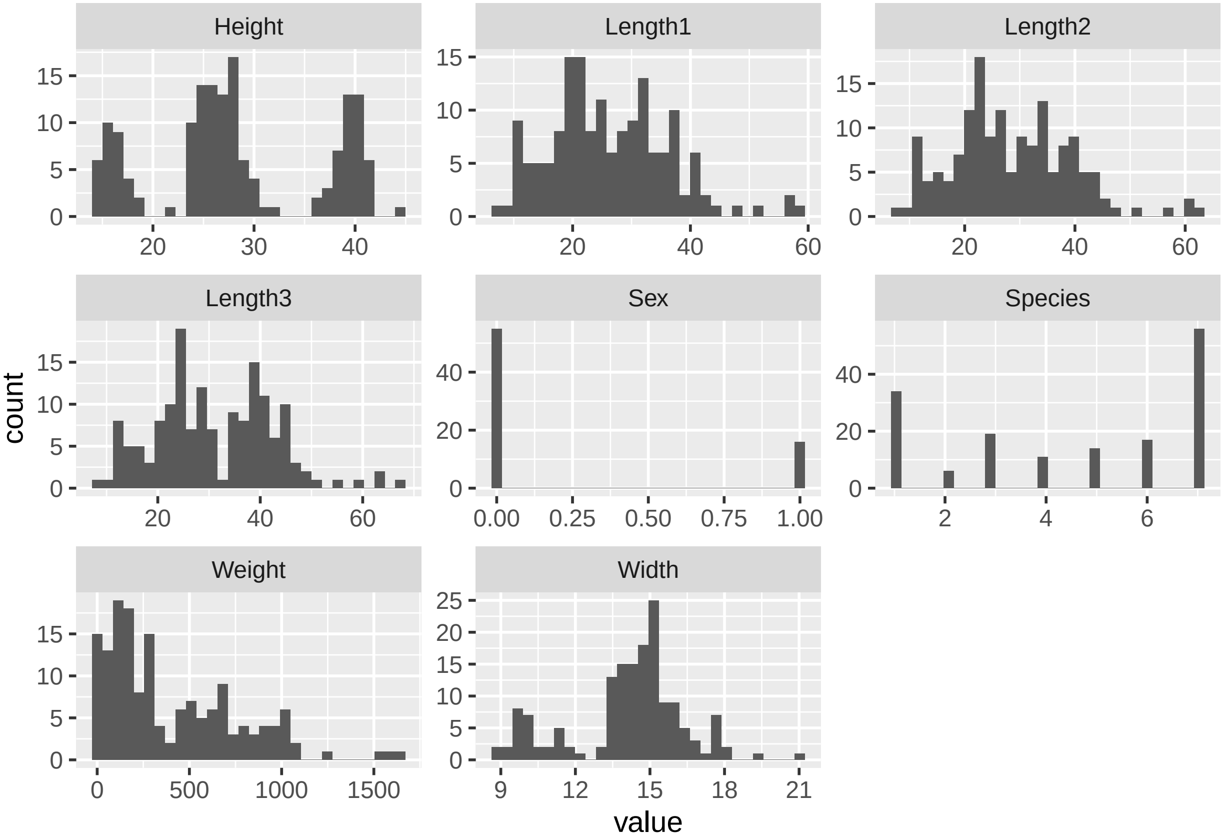
There are 7 species in the sample, with different numbers of observations of each.
The variable **Species** was transformed into dummy variables named from **(Specie1 to Specie7)**
The data set is sorted by species from 1 to 7 and then approximately by length within a species.

Variables Descriptive Analysis

Summary Statistics Table (without Species Dummy Variables)

Species	Weight	Length1	Length2	Length3	Height	Width	Sex
Min. :1.000	Min. : 5.9	Min. : 7.50	Min. : 8.40	Min. : 8.80	Min. :14.50	Min. : 8.70	Min. :0.0000
1st Qu.:2.000	1st Qu.: 120.0	1st Qu.:19.10	1st Qu.:21.00	1st Qu.:23.20	1st Qu.:24.20	1st Qu.:13.40	1st Qu.:0.0000
Median :5.000	Median : 273.0	Median :25.20	Median :27.30	Median :29.40	Median :26.90	Median :14.60	Median :0.0000
Mean :4.529	Mean : 401.2	Mean :26.27	Mean :28.44	Mean :31.24	Mean :28.26	Mean :14.12	Mean :0.2253
3rd Qu.:7.000	3rd Qu.: 650.0	3rd Qu.:32.70	3rd Qu.:36.00	3rd Qu.:39.70	3rd Qu.:37.80	3rd Qu.:15.30	3rd Qu.:0.0000
Max. :7.000	Max. :1650.0	Max. :59.00	Max. :63.40	Max. :68.00	Max. :44.50	Max. :20.90	Max. :1.0000
NA	NA	NA	NA	NA	NA	NA	NA's :86

Check the histograms distribution (without Species Dummy Variables)



Considering the analysis of the graphs and the summary statistics, it's possible to visualize that:

Whether **Height and Width** are closer to the regular **normal distribution**, the **Length(1,2,3) and Weight** variables are slightly skewed forming **Right-Tails** with the presence of some outliers.
Taking that into account, it's possible to apply some **log transformations** while the models are build, specially at the dependent variable **Weight** and the explanatory **Length**.
Variable **Sex** shouldn't be used for analysis because it has a good amount of **missing values** that may impact the model.

Variables Correlation Matrix

	Species	Weight	Length1	Length2	Length3	Height	Width
Species	1.0000000	-0.1327455	-0.0235951	-0.0355609	-0.1357486	-0.6619026	0.0891793
Weight	-0.1327455	1.0000000	0.9164552	0.9193671	0.9245350	0.1943756	0.1355157
Length1	-0.0235951	0.9164552	1.0000000	0.9995161	0.9921198	0.0351444	0.0317880
Length2	-0.0355609	0.9193671	0.9995161	1.0000000	0.9941898	0.0547546	0.0442597
Length3	-0.1357486	0.9245350	0.9921198	0.9941898	1.0000000	0.1326425	0.0377342
Height	-0.6619026	0.1943756	0.0351444	0.0547546	0.1326425	1.0000000	0.4560633
Width	0.0891793	0.1355157	0.0317880	0.0442597	0.0377342	0.4560633	1.0000000

As the correlation **results** between the Length's showed pretty much **closer to 1** with significant p-values:

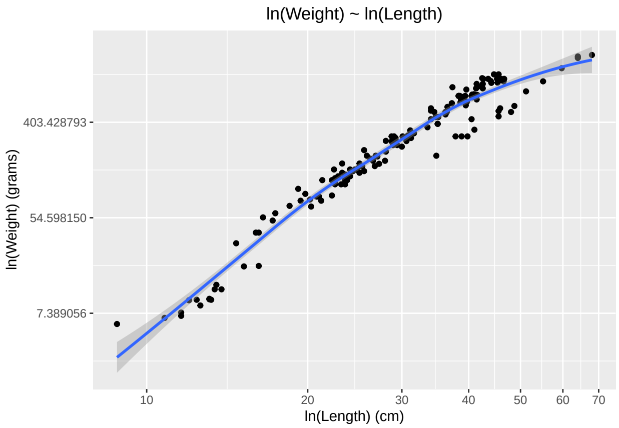
** Length1 ~ Length3: r = 0.99 and p = 0.000.

** Length2 ~ Length3: r = 0.99 and p = 0.000.
** Length1 ~ Length2: r = 1.00 and p = 0.000.

Having the three of them at the same regression model may lead to errors and won't produce significant better adjustments. With that, the variables **Length1 and Length2 won't be considered at any analysis**. Only the variable **Length3**, the total size of the fish, is going to be used as the **Final_Length measurement**.

Weight and Length

ln(Weight) ~ ln(Final_Length)

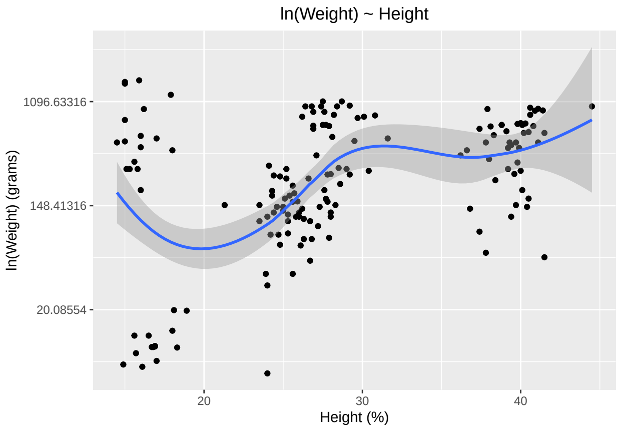


According to the results obtained from the graphs and model comparison (please review the following **Appendix section** for more detailed information), the **log-log regression** shows not only a **better fit**, but also an **extraordinary R-squared** adjusted with a really **significant p-value**. This is supported by the **Right-Tailed Distributions Patterns** encountered both in **Weight** and **Length**.
The variable **Length itself** could be used **alone** for the **predictions**.

Considering that, the first regression model for analysis is:
Reg1: ln(Weight) ~ ln(Final_Length)

Weight and Height

ln(Weight) ~ Height



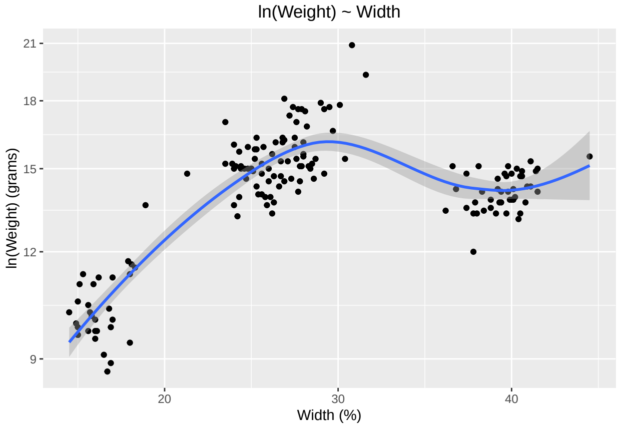
The analysis of the graphs and the comparisons between the **models with and without** the log-transformations (please review the following **Appendix section** for more detailed information), didn't show **any significant statistic difference**. The **pattern** expressed in both graphs are quite **similar**, and the results obtained in terms of **R-squared** are pretty **close** as well. With that, due to the **ease and more tangible interpretation** of the results, the model **chosen** was the **log-lvl** one.

Another important aspect to highlight is the fact that the pattern observed can possibly fit a **quadratic or cubic polynomial models**. Although the cubic and quadratic models showed a slightly better fit than the simple log-lvl, **the interpretation is much easier under the simple model**. The differences aren't that significant enough in terms R-Squared as well.

Considering that, the second regression model for analysis is:
Reg2: ln(Weight) ~ ln(Final_Length) + Height

Weight and Width

ln(Weight) ~ Width



The analysis of the graphs and the comparisons between the **models with and without** the log-transformations (please review the following **Appendix section** for more detailed information), didn't show **any significant statistic difference**. The **pattern** expressed in both graphs are quite **similar**, and the results obtained in terms of **R-squared** are pretty **close** as well. With that, due to the **ease and more tangible interpretation** of the results, the model **chosen** was the **log-lvl** one.

Another important aspect to highlight is the fact that the pattern observed can possibly fit a **quadratic or cubic polynomial models**. Although the cubic and quadratic models showed a slightly better fit than the simple log-lvl, **the interpretation is much easier under the simple model**. The differences aren't that significant enough in terms R-Squared as well.

Considering that, the second regression model for analysis is:
Reg5: ln(Weight) ~ ln(Final_Length) + Height + Width

Weight and Species

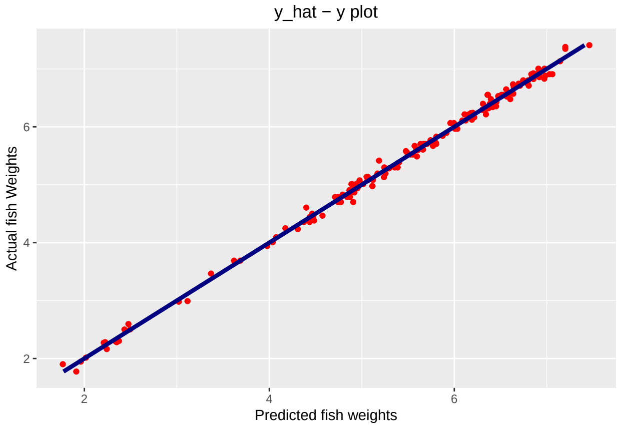
Considering what was done so far, the best possible model so far is represented by reg5. However, it's still missing the impact of the species variables. Another regression model can actually be created for that situation:

Reg8: ln(Weight) ~ ln(Final_Length) + Height + Width + as.factor(Species)

Adding the dummy variables created for species, the final model, **reg8**, shows itself really representative with **higher fit** and **prediction capabilities**. Although the R-squared **didn't present any substantive statistical** difference from the previous models, the **p-values were highly significant**. It means that the model explains well the dataset chosen, using a good number of really representative variables.

For all of that, the **reg8** is going to be the model **used for predictions**. (Please check the **final table resume** at page number 15 - **Appendix**)

Predictions



As expected the model **predicted very well** the values for **Weight**. The **residuals** are pretty low and the graph shows small discrepancy between the actual and the predicted values. It's important to mention that that the **residuals were evaluated in terms of the log transformations, as well the values predicted**. The **conversion** from **logarithmic Weight to absolute values should be done carefully**, considering some possible error bias. The **variables** presented themselves significant **(lower p-values)** and important for the model description. The **R-squared value** is really **representative** and generates a great understanding of how the dependent variable **Weight is changing in the dataset**.

Conclusion

The **main goal** of this project, create a reasonable model that could predict the fish Weight based on its physical characteristics, was achieved **successfully**. Considering the **high R-Squared** and **lower p-values**, the model was build with a **strong comprehend of the dataset** structure and patterns generating on that way really **accurate predictions**.

Although it's **impossible** to infer any **causality**, this model seems a **good income** for **further studies** in the area.

Appendix

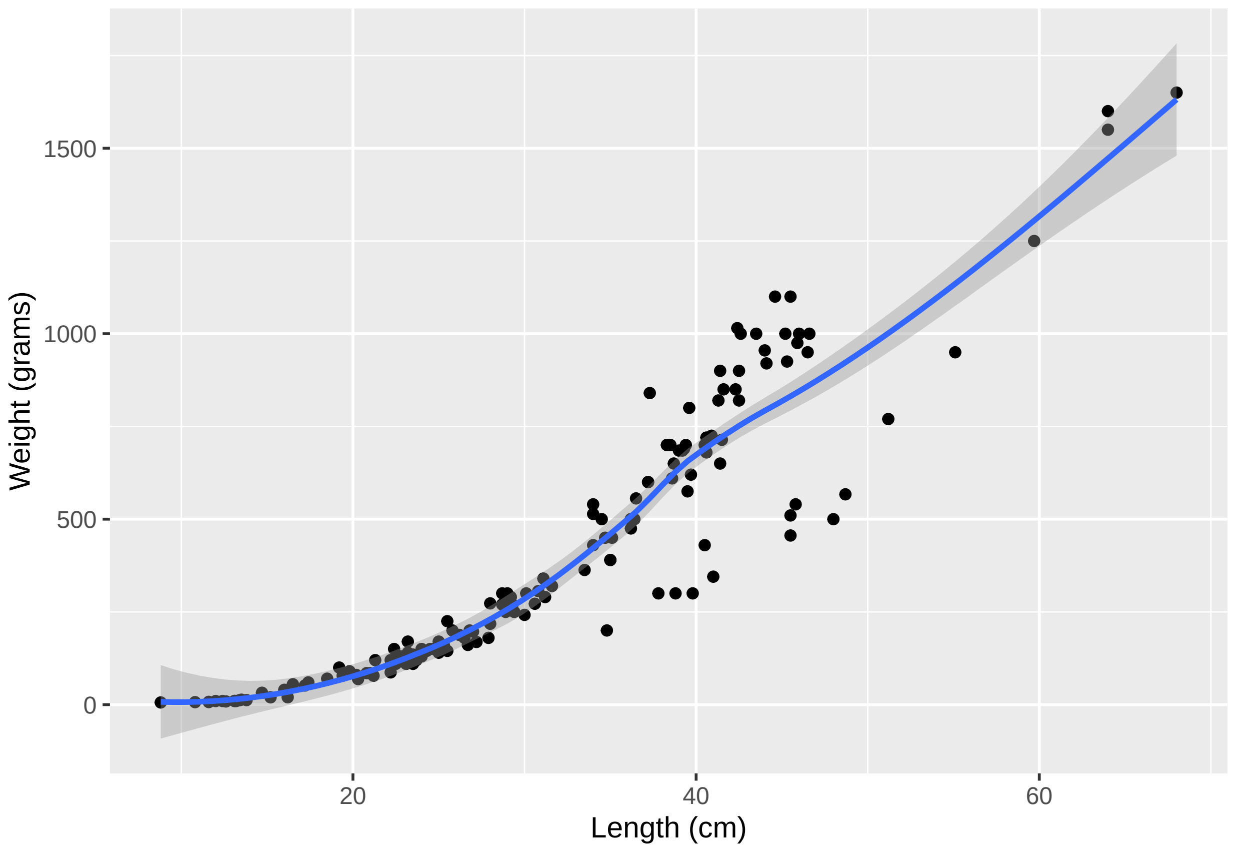
Correlation Matrix

```
##             Species     Weight     Length1     Length2     Length3      Height
## Species  1.00000000 -0.1327455 -0.02359505 -0.03556088 -0.13574860 -0.66190261
## Weight  -0.13274545  1.0000000  0.91645518  0.91936709  0.92453500  0.19437562
## Length1 -0.02359505  0.9164552  1.00000000  0.99951615  0.99211983  0.03514437
## Length2 -0.03556088  0.9193671  0.99951615  1.00000000  0.99418981  0.05475455
## Length3 -0.13574860  0.9245350  0.99211983  0.99418981  1.00000000  0.13264248
## Height  -0.66190261  0.1943756  0.03514437  0.05475455  0.13264248  1.00000000
## Width    0.08917930  0.1355157  0.03178795  0.04425967  0.03773417  0.45606333
##              Width
## Species 0.08917930
## Weight  0.13551573
## Length1 0.03178795
## Length2 0.04425967
## Length3 0.03773417
## Height  0.45606333
## Width   1.00000000
```

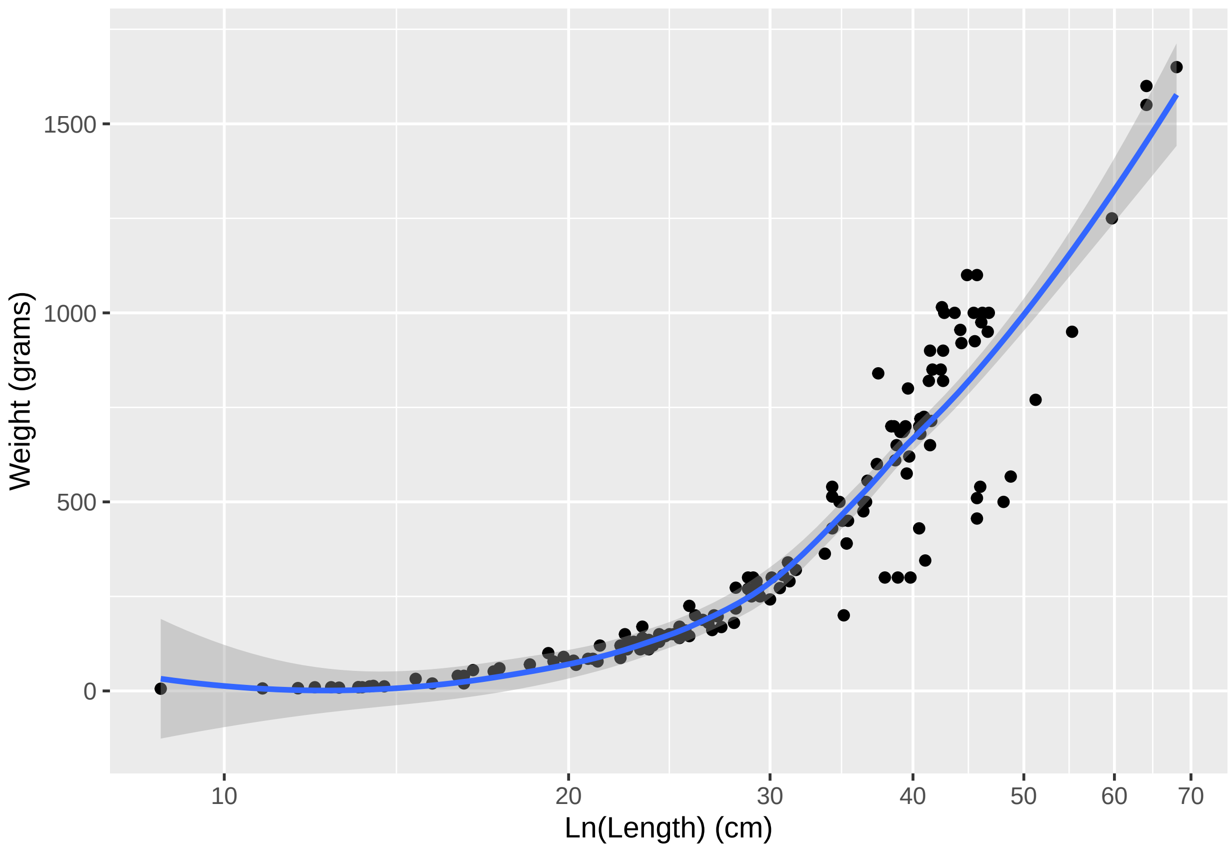
```
##            Species     Weight   Length1   Length2    Length3       Height
## Species         NA 0.09744931 0.7692749 0.6583765 0.09003959 0.000000e+00
## Weight  0.09744931         NA 0.0000000 0.0000000 0.00000000 1.471291e-02
## Length1 0.76927491 0.00000000        NA 0.0000000 0.00000000 6.621304e-01
## Length2 0.65837646 0.00000000 0.0000000        NA 0.00000000 4.958074e-01
## Length3 0.09003959 0.00000000 0.0000000 0.0000000         NA 9.771168e-02
## Height  0.00000000 0.01471291 0.6621304 0.4958074 0.09771168           NA
## Width   0.26669809 0.09059758 0.6926813 0.5820380 0.63892903 1.943337e-09
##                Width
## Species 2.666981e-01
## Weight  9.059758e-02
## Length1 6.926813e-01
## Length2 5.820380e-01
## Length3 6.389290e-01
## Height  1.943337e-09
## Width             NA
```

Weight and Length

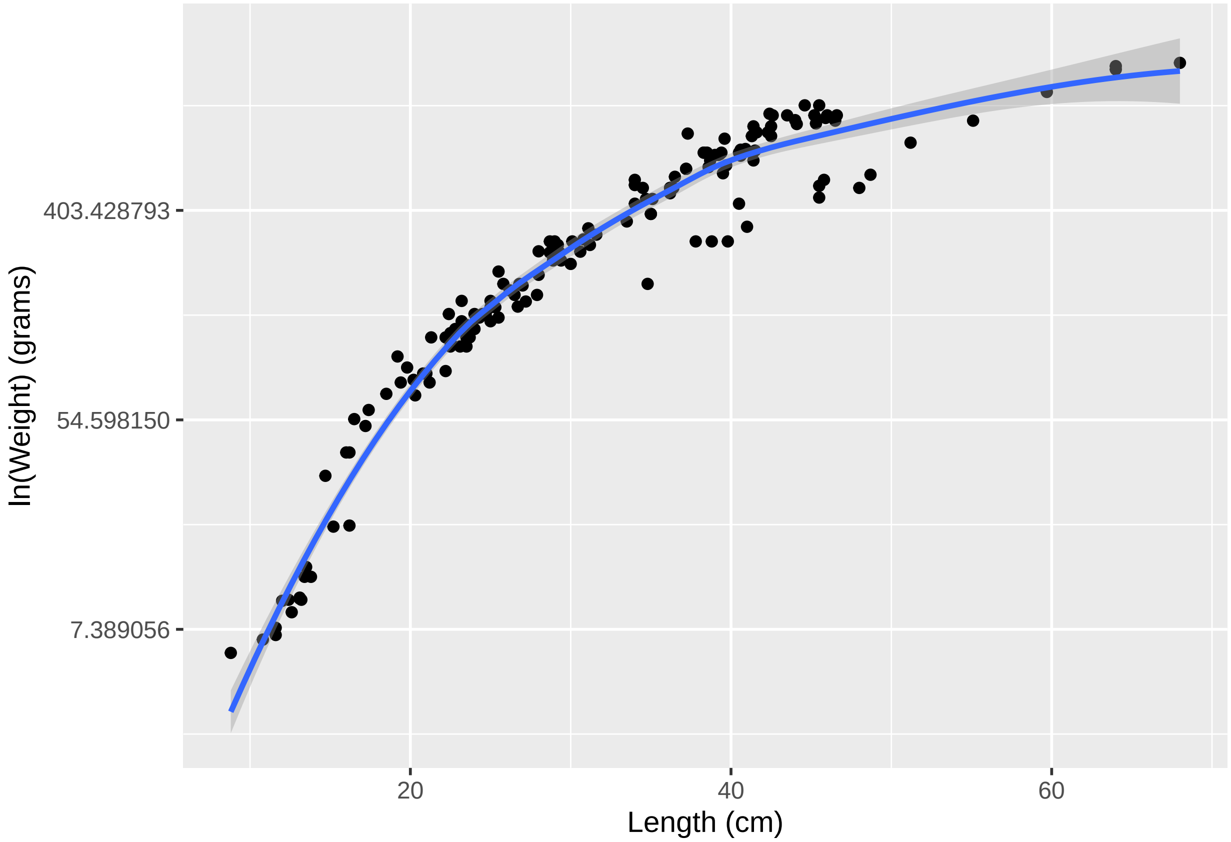
```
## `geom_smooth()` using formula 'y ~ x'
```



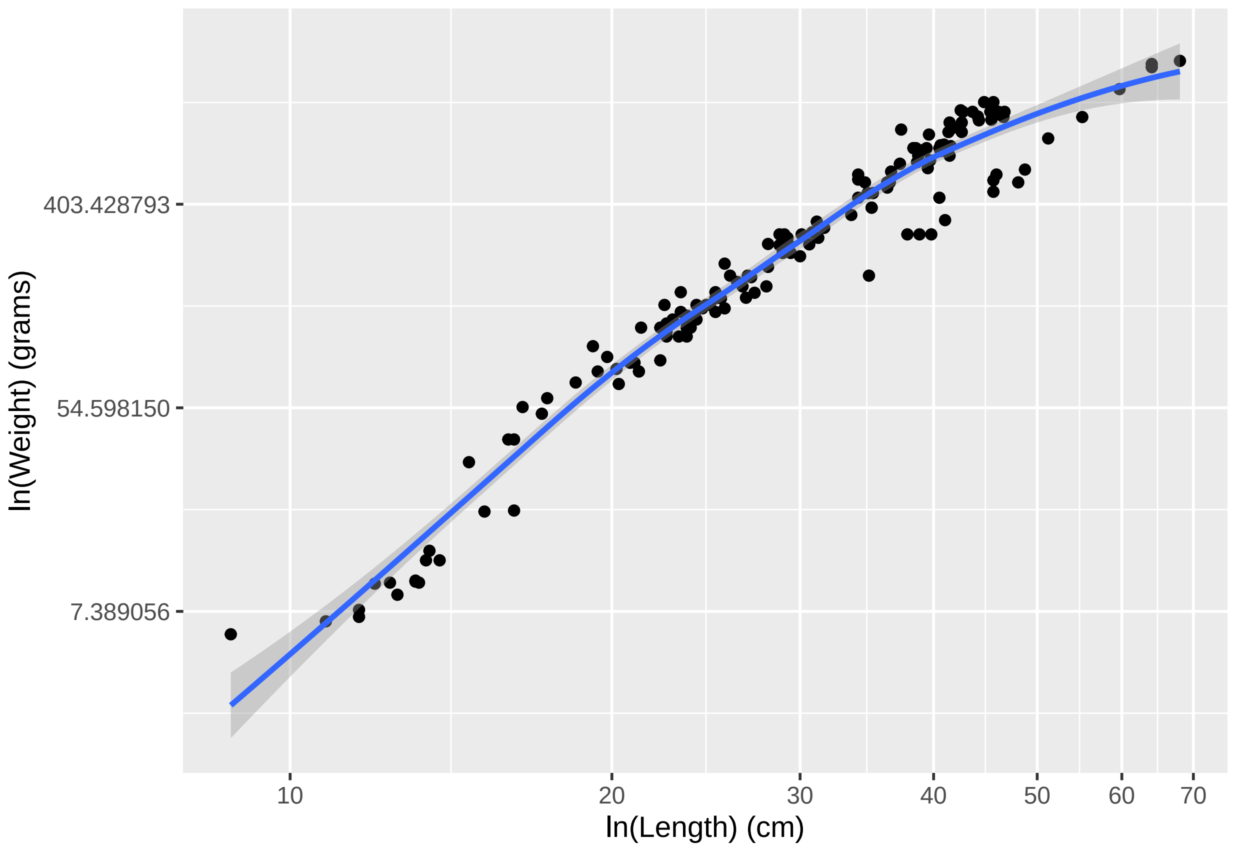
```
## `geom_smooth()` using formula 'y ~ x'
```



```
## 'geom_smooth()' using formula 'y ~ x'
```



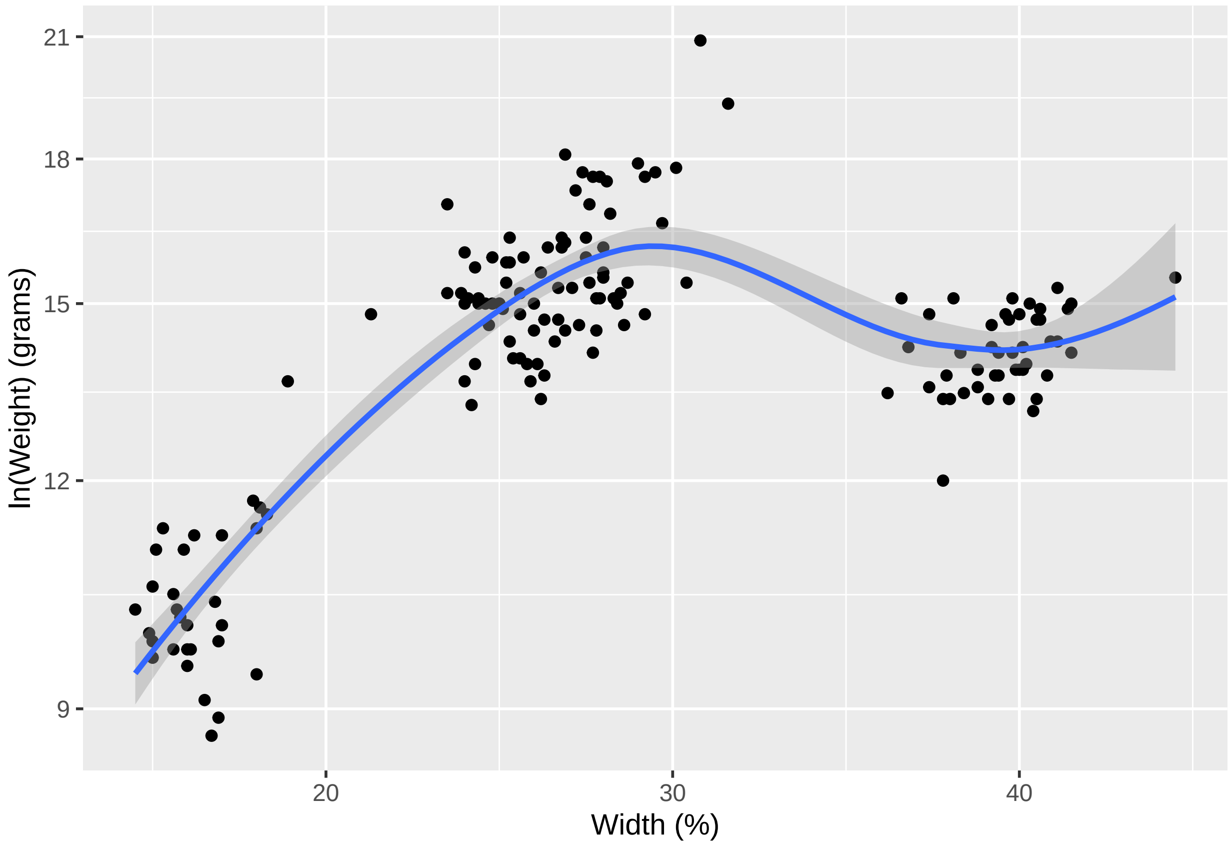
```
## `geom_smooth()` using formula 'y ~ x'
```



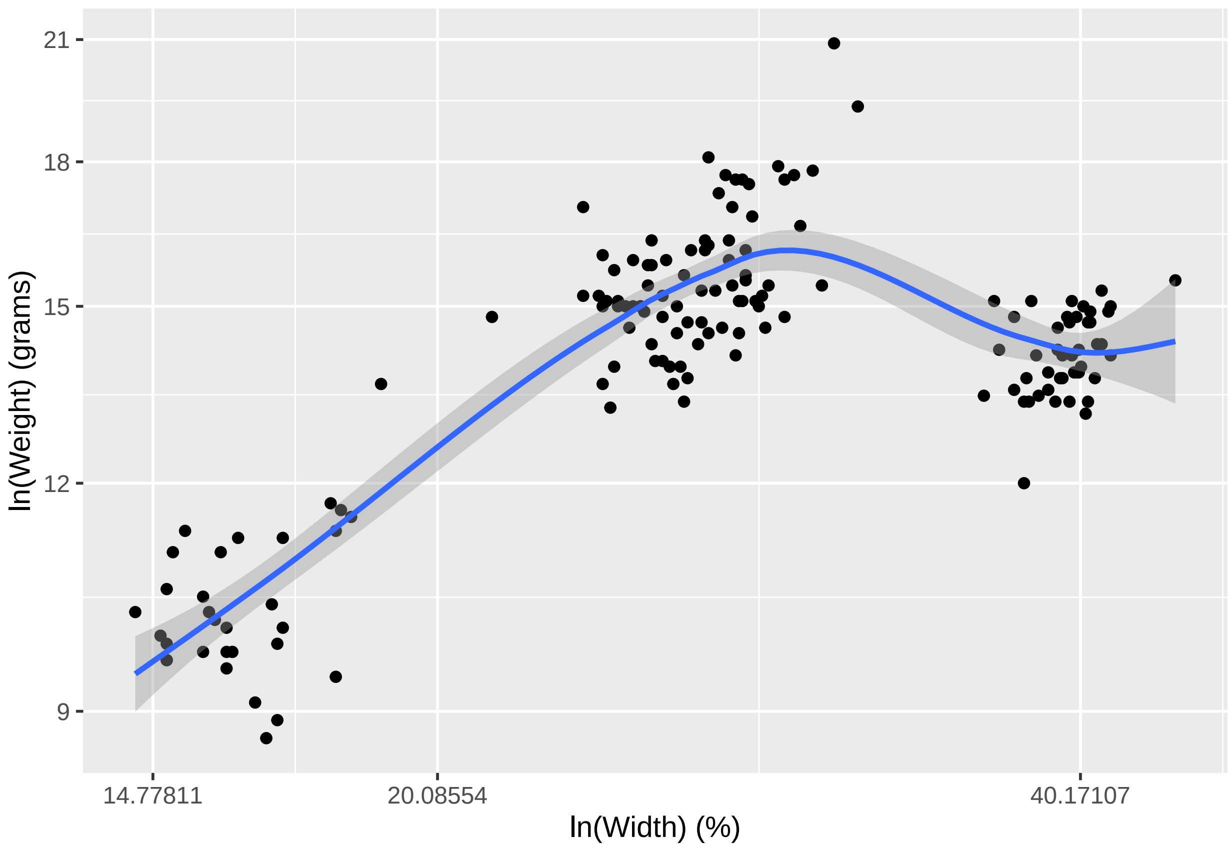
```
##
##  studentized Breusch-Pagan test
##
## data:  lvl_lvl
## BP = 28.102, df = 1, p-value = 1.151e-07

##
##  studentized Breusch-Pagan test
##
## data:  lvl_log
## BP = 6.9663, df = 1, p-value = 0.008306

##
##  studentized Breusch-Pagan test
##
## data:  log_lvl
## BP = 0.0068874, df = 1, p-value = 0.9339

##
##  studentized Breusch-Pagan test
##
## data:  log_log
## BP = 0.18215, df = 1, p-value = 0.6695

## The table was written to the file ’D:/CEU/Data_Analysis_2/DA2_Final_Project/out/logtrans_Weight_Leng
```

Weight and Width

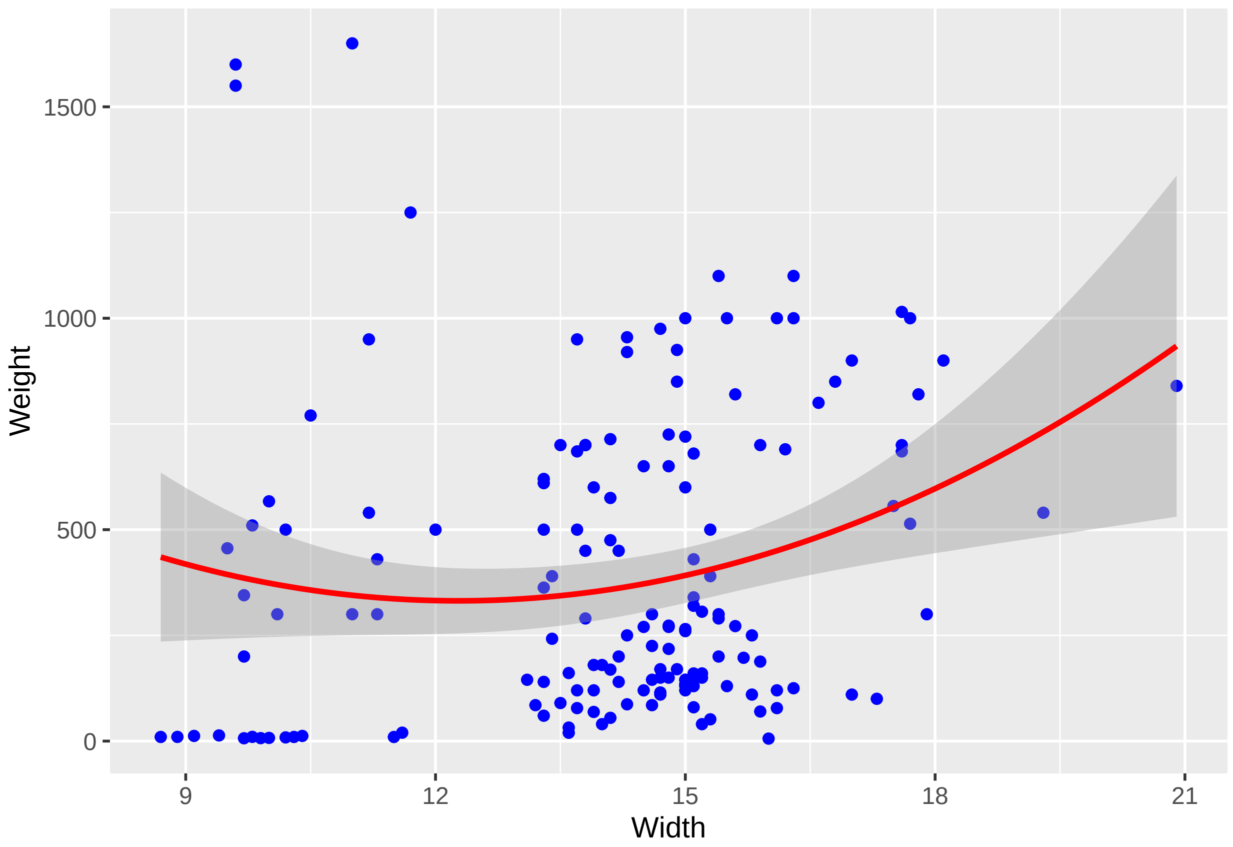
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

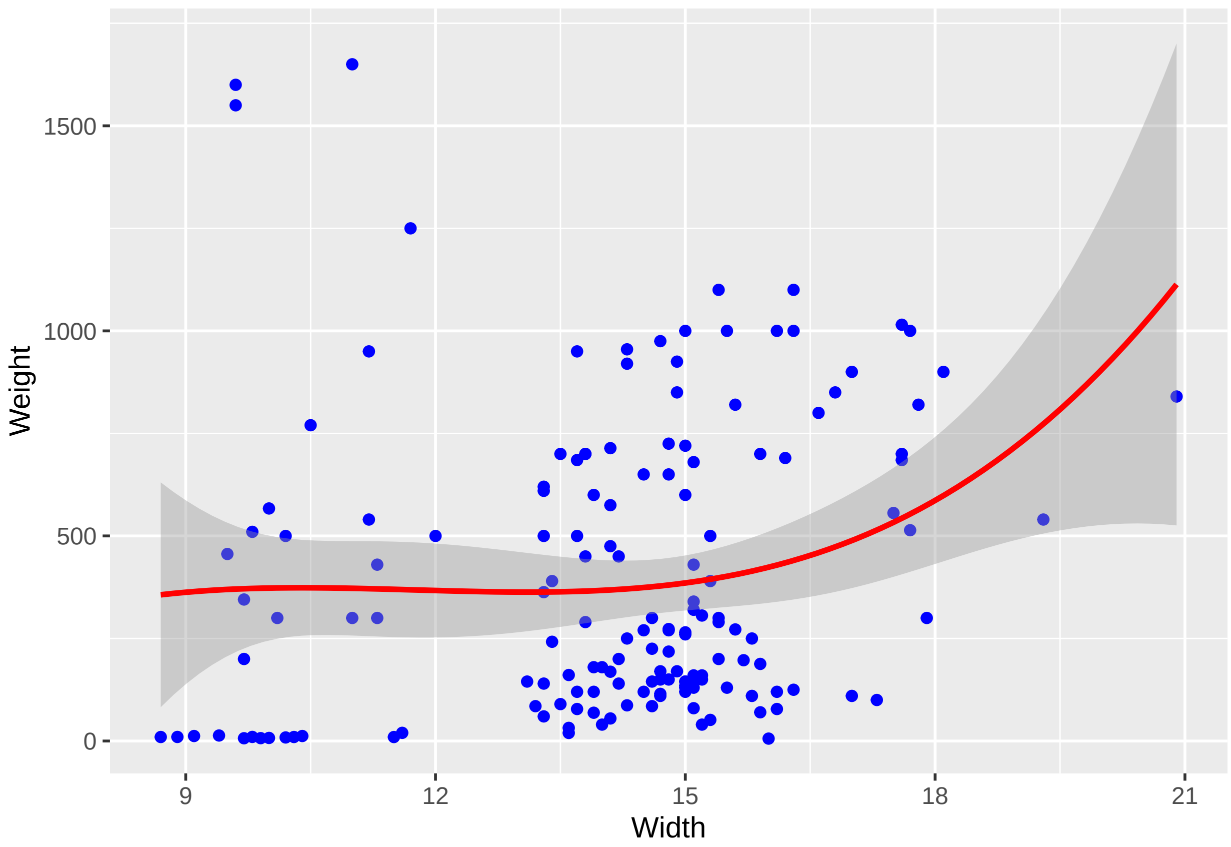


```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
##
##  studentized Breusch-Pagan test
##
## data:  log_lvl
## BP = 40.838, df = 1, p-value = 1.654e-10

##
##  studentized Breusch-Pagan test
##
## data:  log_log
## BP = 42.731, df = 1, p-value = 6.282e-11

## The table was written to the file 'D:/CEU/Data_Analysis_2/DA2_Final_Project/out/logtrans_Weight_Widt
```


Weight
1500
1000
500
0
9
12
15
18
21
Width



```
## The table was written to the file 'D:/CEU/Data_Analysis_2/DA2_Final_Project/out/model_comparison_Leng
```

	reg1	reg2	reg5	reg8
(Intercept)	-5.27 ***	-5.58 ***	-6.46 ***	-5.87 ***
	(0.29)	(0.17)	(0.09)	(0.13)
ln_Final_Length	3.17 ***	3.03 ***	3.00 ***	3.02 ***
	(0.08)	(0.05)	(0.02)	(0.04)
Height		0.03 ***	0.02 ***	0.02 ***
		(0.00)	(0.00)	(0.01)
Width			0.09 ***	0.04 ***
			(0.00)	(0.01)
Species1				-0.28 **
				(0.10)
Species2				0.02
				(0.03)
Species3				-0.10 ***
				(0.02)
Species4				-0.13
				(0.10)
Species5				-0.33 ***
				(0.05)
Species6				-0.24 ***
				(0.07)
nobs	157	157	157	157
r.squared	0.95	0.97	0.99	1.00
adj.r.squared	0.95	0.97	0.99	1.00
statistic	1399.67	2186.99	6642.68	4658.18
p.value	0.00	0.00	0.00	0.00
df.residual	155.00	154.00	153.00	147.00
nobs.1	157.00	157.00	157.00	157.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00

*** p < 0.001; ** p < 0.01; * p < 0.05.