ANH NGO - BS20DSY035

LINEAR REGRESSION ANALYSIS

Fish Data

Linear regression analysis

Anh Ngo – BS20DSY035

Bachelor of Data Science 2020

S P Jain School of Global Management

Prof. Suchismita Das

May 02, 2021

Contents

1. Introduction

1.1 Report Objectives

The dataset contains 159 records of common fish species in Finland's market sales. It includes Weight, Length, Height, and Width. In this paper, I will focus on the summary of each statistics and then estimate the weight of fish using linear regression model.

1.2 Problem Statement

There are numerous factors that affect the growth of fish. Hence, without solid understanding about significant factors in the relationship with fish's weight, fishermen may fail to find out the biological changes in the fish. Indeed, there is no doubt that analyzing this dataset is of great importance in fishery assessments.

My observations will be put forward with the help of two software: Microsoft Excel 365 and JMP Pro 15 as statistical tools.

2. Overview

The dataset is built with the purpose of helping fishermen in assessing the variations from the expected weight for the known features of fish.

In this statistical report, various analysis methods from univariate to bivariate method would be used to summarize the data statistics and its distribution. Then, using three criteria as mentioned above to predict weight and evaluate what is the most crucial factor. To that end, numerous hypotheses would be tested to evaluating utility of the model.

The dataset is owned by Aung Pyae.

3. Dataset Exploration

3.1 Dataset Dictionary

Weight: Weight of the fish. It is measured in grams.

Length: Length of the fish from the nose to the end of the tail. It is measured in centimeters.

Height: The maximum height of the fish. It is measured in centimeters.

Width: The maximum width of the fish. It is measured in centimeters.

3.2 Descriptive Analysis

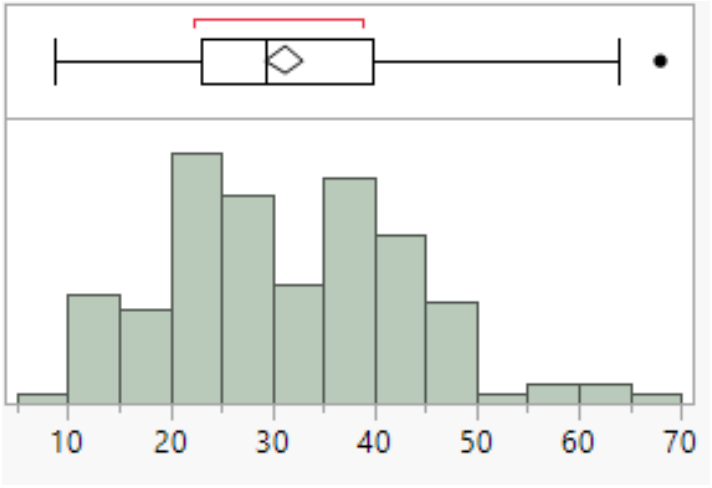
Below those measures have been taken for all four variables independently:

	Length	**Height**	**Width**	**Weight**
Mean	31.227044	8.9709937	4.4174855	398.32642
Std Dev	11.610246	4.2862076	1.6858039	357.97832
Sum	4965.1	1426.388	702.3802	63333.9
Variance	134.79781	18.371576	2.8419347	128148.48
Skewness	0.3915408	0.3971864	0.0049722	1.1044504
Kurtosis	0.0754234	-0.614174	-0.534645	0.8838414
Minimum	8.8	1.7284	1.0476	0
Maximum	68	18.957	8.142	1650
Median	29.4	7.786	4.2485	273
Mode	23.5	2.2139	3.525	300
Range	59.2	17.2286	7.0944	1650
IQR	16.6	6.4414	2.2134	530
Count	159	159	159	159
Upper 95% Mean	33.045615	9.6423638	4.6815414	454.39835
Lower 95% Mean	29.408473	8.2996236	4.1534297	342.25449

4. Exploratory Data Analysis

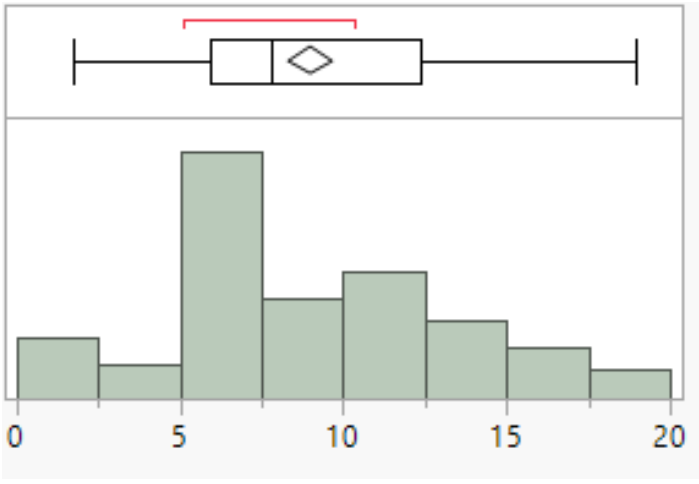
4.1 Univariate Analysis

4.1.1 Length



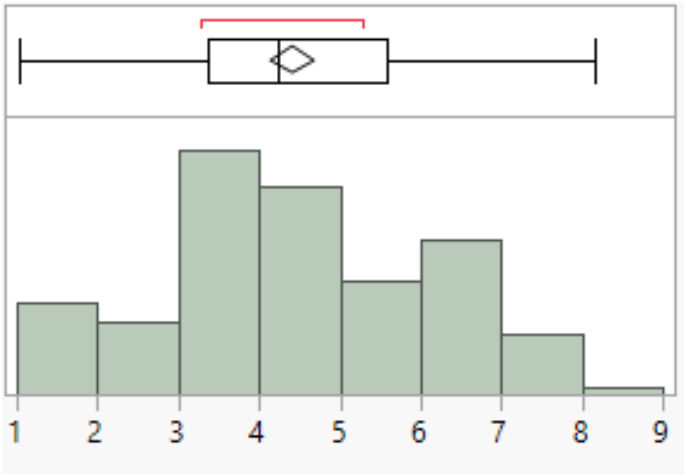
- The average length is about 31 cm with a standard deviation of around 11.6 cm. IQR is 16.6 cm.
- There is one outlier whose value is 68 cm. Consider eliminating this value, the distribution of Length can be approximately normal.

4.1.2 Height



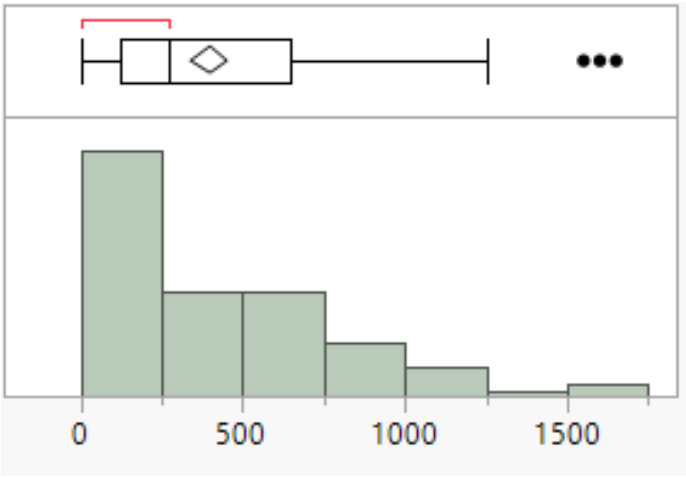
- Length of the right whisker is more than that of the left one, then the distribution is quite right skewed. However, there are no outliers.
- The average height is 8.97 cm, which is relatively higher than the median (7.79 cm). Its standard deviation is around 4.3 cm.

4.1.3 Width



- Width is normally distributed.
- The average width is 4.42 cm with a standard deviation of 1.7 cm. IQR is 2.2 cm.
- There are no outliers. The mean and median values are not much different.

4.1.4 Weight

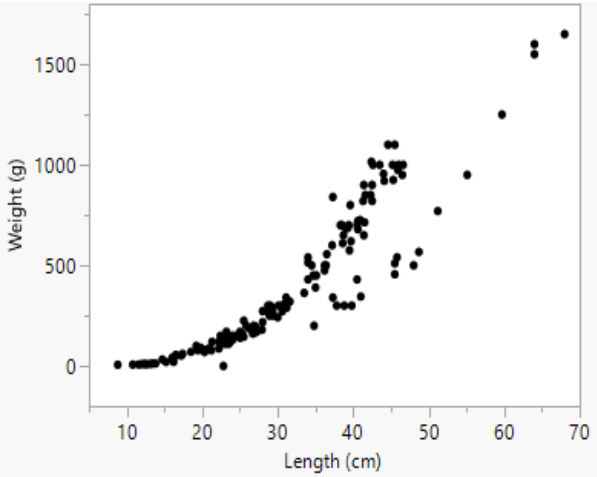


- Most fish have their weight between 0 and 250 grams.
- The distribution is moderately skewed to the right. Indeed, the length of the right whisker is more than that of the left one. Hence, mean value is larger than median value.
- There are three outliers whose values are 1550g, 1600g and 1650g, respectively.

4.2 Bivariate Analysis

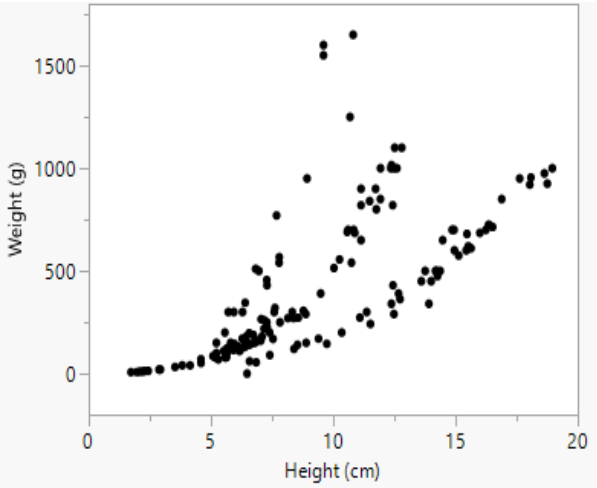
4.2.1 Analyzing relationship between Length/Height/Width and Weight

Length and Weight



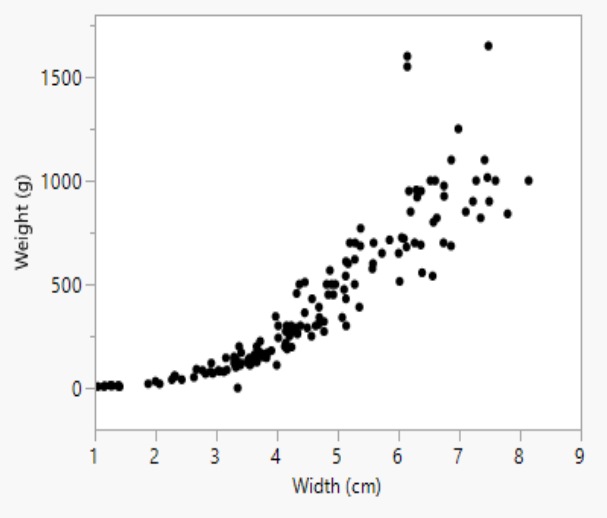
	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.923044	0.896145	0.943184	<.0001*
Covariance	3836.369			
Count	159			

Height and Weight



	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.724345	0.640962	0.790832	<.0001*
Covariance	1111.413			
Count	159			

Width and Weight



	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.886507	0.847847	0.915791	<.0001*
Covariance	534.9901			
Count	159			

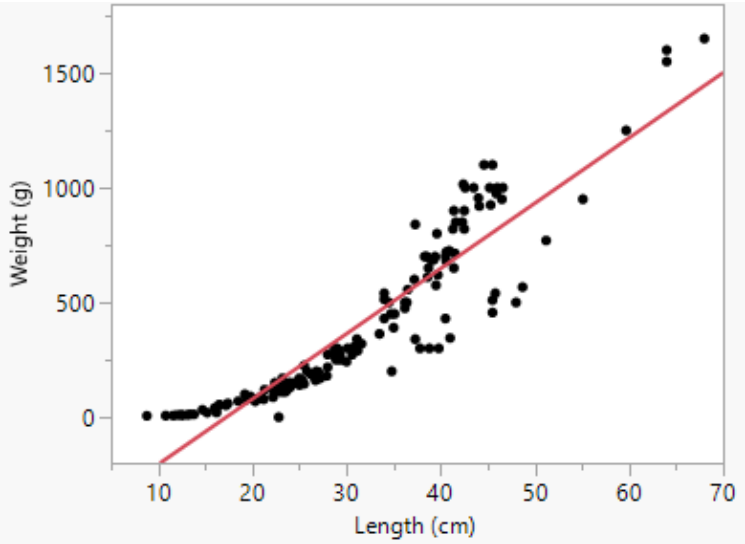
From the previous three scatterplots, we can say that Length, Height, and Width are strongly positively correlated to Weight ($r > 0.7$). This means fish that have high measurements in length, height and width tend to weigh heavier.

4.2.2 Fit Simple Linear Regression

If I assume that for any particular combination of predictor variable values, the distribution of errors is normal with mean 0 and constant variance σ^2, then the simple linear regression model seems appropriate.

4.2.2.1 Length – Weight Linear Regression:

Fit a simple linear regression on x = Length (cm) and y = Weight (g):



Weight = -490.4006 + 28.460171*Length

Summary of Fit of Length – Weight Linear Regression

RSquare	0.852009
Root Mean Square Error	138.1505
Mean of Response	398.3264
Observations	159

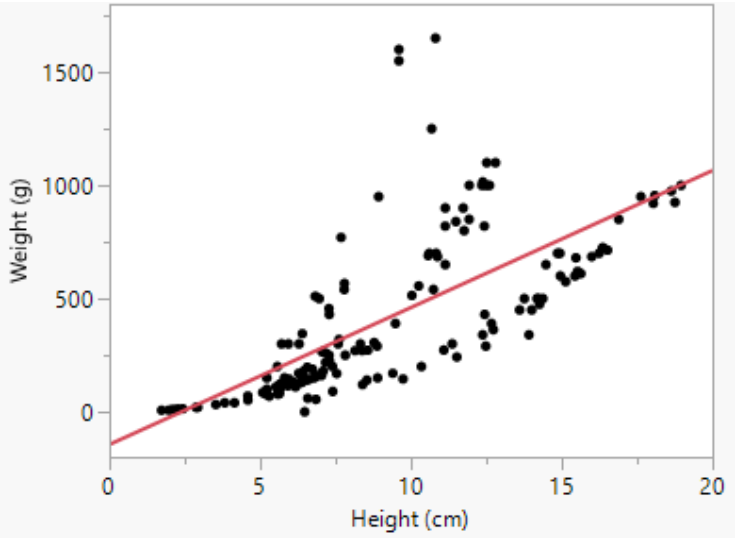
The simple linear regression model is: $\hat{y} = -490.4006 + 28.460171x$

$r^2 \approx 0.85$: Approximately 85% of observed variation in weight can be explained by the simple linear regression model relationship between weight and length.

The magnitude of a typical sample deviation from the least-square line is about 138.

4.2.2.2 Height – Weight Linear Regression:

Fit a simple linear regression on x = Height (cm) and y = Weight (g):



Weight = -144.386 + 60.496351*Height

Summary of Fit of Height – Weight Linear Regression

RSquare	0.524676
Root Mean Square Error	247.5884
Mean of Response	398.3264
Observations	159

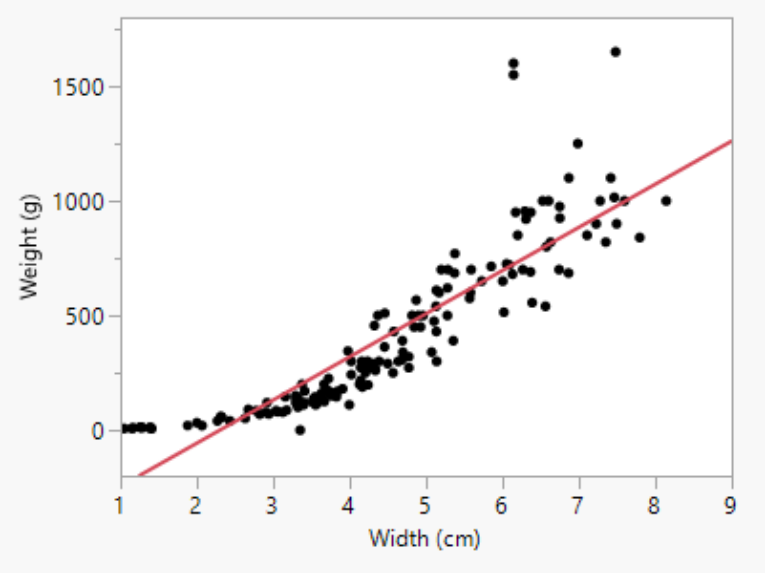
The simple linear regression model: $\hat{y} = -144.386 + 60.496351x$

$r^2 \approx 0.52$: Approximately 52% of variation in observed y values that is explained by the fitted model. And $s_e \approx 248$, which is very large.

Hence, the model does not fit the dataset well.

4.2.2.3 Width – Weight Linear Regression

Fit a simple linear regression on x = Height (cm) and y = Weight (g):



Weight = -433.2589 + 188.24855*Width

Summary of Fit of Width – Weight Linear Regression

RSquare	0.785894
Root Mean Square Error	166.169
Mean of Response	398.3264
Observations	159

The simple linear regression model: $\hat{y} = -433.2589 + 188.24855x$

$r^2 \approx 0.8$: Approximately 80% of the observed variation in y can be attributed to the linear relationship with Width.

The magnitude of a typical sample deviation from the least-square line is about 166.

5. Multiple linear regression

From Bivariate Analysis part, it seems that there is a linear relationship between the independent variables (Length, Height and Width) and the dependent variable (Weight).

Now, I will analyse the multiple linear regression to get better understanding about this relationship.

5.1 Fit multiple linear regression

A multiple linear regression model takes the form:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

where: $\hat{y}$ = Weight, x_1 = Length, x_2 = Height, x_3 = Width

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-521	29.33136	-17.76	<.0001*	-578.9408	-463.0592
Length	19.44453	1.812478	10.73	<.0001*	15.864185	23.024875
Height	3.8534309	3.848711	1.00	0.3183	-3.749262	11.456124
Width	62.832602	14.55995	4.32	<.0001*	34.071066	91.594138

From the above table, we have: $\hat{y} = -521 + 19.44453x_1 + 3.8534309x_2 + 62.832602x_3$

We can see that the three independent variables (Length, Height, and Width) are positively linked to Weight:

- When Height and Width are fixed, the mean increase in weight associated with a 1-cm increase in length is approximately 19g.
- When Length and Width are fixed, the mean increase in weight associated with a 1-cm increase in height is about 4g.
- When Length and Height are fixed, the mean increase in weight associated with a 1-cm increase in width is nearly 63g.

→ Width is the most crucial factor and hence, is particularly useful for predicting weight of fish with the same length and height.

From the table, we can also construct 95% Confidence Interval for β_i:

$$15.864185 \leq \beta_1 \leq 23.024875$$
$$-3.749262 \leq \beta_2 \leq 11.456124$$
$$34.071066 \leq \beta_3 \leq 91.594138$$

5.2 Evaluate the utility of the multiple linear regression model

Summary of Fit of Multiple Linear Regression

RSquare	0.877841
Root Mean Square Error	126.323
Mean of Response	398.3264
Observations	159

The coefficient of multiple determination is $R^2 \approx 87.8\%$, which means about 87.8% of variation in observed y values is explained by the fitted model. And the value of s_e is much lower (about 126) compared to the three simple linear regressions above.

To that end, the multiple linear regression is the most appropriate model and hence successful in relating y to the predictors.

5.3 Confidence Interval of the Error Variance σ^2

The residual sum of squares is estimated as

$$\text{SSResid} = \sum_{i=1}^{159} (y_i - \hat{y}_i)^2 = 2473414$$

Because 4 parameters need to be estimated, an unbiased estimator of σ^2 is:

$$\hat{\sigma}^2 = \frac{\text{SSResid}}{159 - 4} = 15958$$

Choosing significant value $\propto = 0.05$, we get $\chi^2_{155,0.025} = 191.3623$ and $\chi^2_{155,0.975} = 122.4228$

Now, I construct 95% Confidence Interval for σ^2:

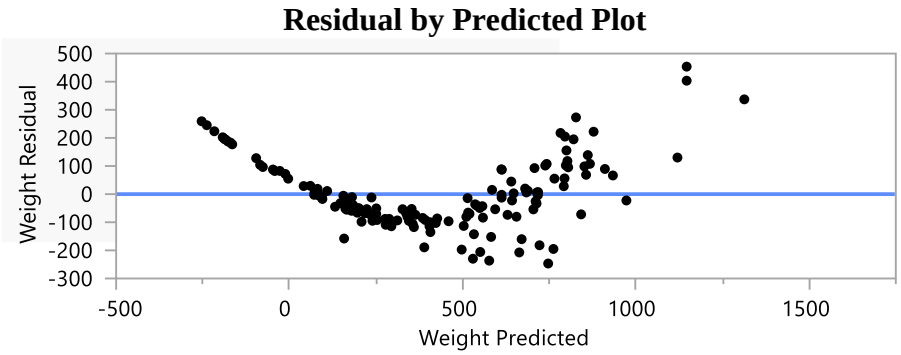
$$\frac{(n-k)\hat{\sigma}^2}{\chi^2_{n-k,\alpha/2}} \leq \sigma^2 \leq \frac{(n-k)\hat{\sigma}^2}{\chi^2_{n-k,(1-\frac{\alpha}{2})}}$$

$$\Leftrightarrow \frac{(159-4) \times 15958}{191.3623} \leq \sigma^2 \leq \frac{(159-4) \times 15958}{122.4228}$$

$$\Leftrightarrow 12925.691 \leq \sigma^2 \leq 20204.488$$

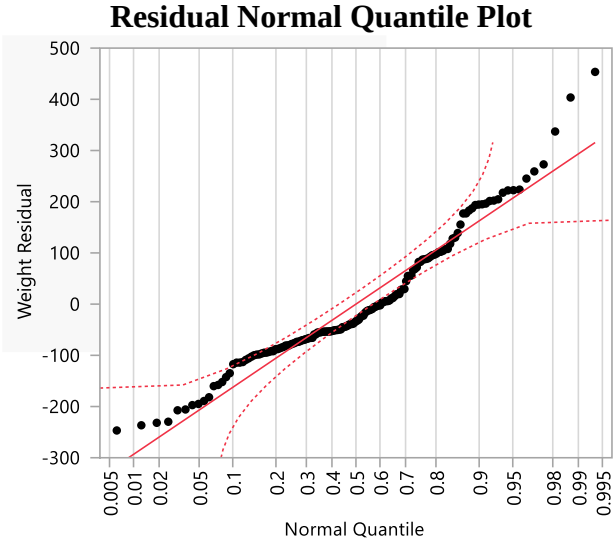
5.4 Residual Analysis

Analyzing the residuals assists us in checking four basic assumptions of linear regression model.



As you can see from the "Residual plot" above, the points follow a curve pattern, suggesting a better fit for a nonlinear model.

However, the mean residual (the blue line) is around the value 0, hence, the assumption that $\mu_\varepsilon = 0$ is reasonable.



Looking at "Residual Normal Quantile Plot", most of the points fall outside the red border. Hence, the assumption that the distribution of residual at any particular x value is normal is not valid.

As a result, the four assumptions for linear regression model are not reasonable.

6. Hypothesis Testing

6.1 The Model Utility F Test for Multiple Regression

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_1: \beta_i \neq 0,$ for atleast one i.

ANOVA Table:

Analysis of Variance

Source of Variation	DF	Sum of Squares	Mean Square	F Ratio
Regression	3	17774045	5924682	371.2787
Residual	155	2473414	15958	**Prob > F**
Total	158	20247459		<.0001*

Following the ANOVA table, we get $F = 371.2787$

Choosing significant value $\propto = 0.05$, $F_{3,\,155,\,0.05} = 2.6629$

➔ $F = 371.2787 > F_{3,\,155,\,0.05} = 2.6629$

Hence, we have strong evidence to reject H_0, which means there is at least one independent variable that affects the response.

6.2 The Model Utility t Test

Now, I conduct the model utility t test to evaluate whether there is a useful linear relationship between each individual independent variable and the response.

$H_0: \beta_i = 0, \text{for } i = 1,2,3$

$H_1: \beta_i \neq 0$

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-521	29.33136	-17.76	<.0001*	-578.9408	-463.0592
Length	19.44453	1.812478	10.73	<.0001*	15.864185	23.024875
Height	3.8534309	3.848711	1.00	0.3183	-3.749262	11.456124
Width	62.832602	14.55995	4.32	<.0001*	34.071066	91.594138

The estimated standard deviation of the statistic b (s_b) is shown in the column 'Std Error' above.

We also use statistic $t = \frac{b-\beta}{s_b}$ to check their effect on model. Following the Parameter Estimates table, we get the test statistic for x_1 is $t_1 = 10.73$. Similarly, $t_2 = 1.00$, $t_3 = 4.32$

Choosing significant value $\propto = 0.05$, $t_{157,\,0.025} = 1.9752$

We have:

$|t_1| > t_{157,\,0.025}$ and $|t_3| > t_{157,\,0.025}$. Hence, we have strong evidence to reject H_0, which means Length and Width both affect Weight.

On the other hand, $|t_2| < t_{157,\,0.025}$. Hence, it fails to reject H_0 or say: there is no useful linear relationship between Height and Weight.

7. Conclusion

In short, from descriptive and graphical analysis, we can conclude that the three mentioned factors are strongly positively correlated to the weight of fish. However, there are several factors which I have not considered in this paper also play a crucial role in relation to weight such as species of fish, environmental conditions, etc.

Even though the multiple linear regression in my report can be used to predict the weight of fish at the given three measurements, the idea should not be generalized because the dataset that I take is collected in one certain place: Finland.

Also, as I stated above, four assumptions of linear regression model are not satisfied, hence, it is better to fit a nonlinear model to this dataset in order to predict the target precisely.

This paper gives only part of a greater work, by focusing on the statistical area, that has to be done in order to predict the variations in the fish's weight.

Appendix

Weight	Length	Height	Width
242	30	11.52	4.02
290	31.2	12.48	4.3056
340	31.1	12.3778	4.6961
363	33.5	12.73	4.4555
430	34	12.444	5.134
450	34.7	13.6024	4.9274
500	34.5	14.1795	5.2785
390	35	12.67	4.69
450	35.1	14.0049	4.8438
500	36.2	14.2266	4.9594
475	36.2	14.2628	5.1042
500	36.2	14.3714	4.8146
500	36.4	13.7592	4.368
340	37.3	13.9129	5.0728
600	37.2	14.9544	5.1708
600	37.2	15.438	5.58
700	38.3	14.8604	5.2854
700	38.5	14.938	5.1975
610	38.6	15.633	5.1338
650	38.7	14.4738	5.7276
575	39.5	15.1285	5.5695
685	39.2	15.9936	5.3704
620	39.7	15.5227	5.2801
680	40.6	15.4686	6.1306
700	40.5	16.2405	5.589
725	40.9	16.36	6.0532
720	40.6	16.3618	6.09
714	41.5	16.517	5.8515
850	41.6	16.8896	6.1984
1000	42.6	18.957	6.603
920	44.1	18.0369	6.3063
955	44	18.084	6.292
925	45.3	18.7542	6.7497
975	45.9	18.6354	6.7473
950	46.5	17.6235	6.3705
40	16.2	4.1472	2.268
69	20.3	5.2983	2.8217
78	21.2	5.5756	2.9044
87	22.2	5.6166	3.1746
120	22.2	6.216	3.5742
0	22.8	6.4752	3.3516
110	23.1	6.1677	3.3957
120	23.7	6.1146	3.2943

Weight	Length	Height	Width
85	20.8	5.1376	3.0368
85	21	5.082	2.772
110	22.5	5.6925	3.555
115	22.5	5.9175	3.3075
125	22.5	5.6925	3.6675
130	22.8	6.384	3.534
120	23.5	6.11	3.4075
120	23.5	5.64	3.525
130	23.5	6.11	3.525
135	23.5	5.875	3.525
110	23.5	5.5225	3.995
130	24	5.856	3.624
150	24	6.792	3.624
145	24.2	5.9532	3.63
150	24.5	5.2185	3.626
170	25	6.275	3.725
225	25.5	7.293	3.723
145	25.5	6.375	3.825
188	26.2	6.7334	4.1658
180	26.5	6.4395	3.6835
197	27	6.561	4.239
218	28	7.168	4.144
300	28.7	8.323	5.1373
260	28.9	7.1672	4.335
265	28.9	7.0516	4.335
250	28.9	7.2828	4.5662
250	29.4	7.8204	4.2042
300	30.1	7.5852	4.6354
320	31.6	7.6156	4.7716
514	34	10.03	6.018
556	36.5	10.2565	6.3875
840	37.3	11.4884	7.7957
685	39	10.881	6.864
700	38.3	10.6091	6.7408
700	39.4	10.835	6.2646
690	39.3	10.5717	6.3666
900	41.4	11.1366	7.4934
650	41.4	11.1366	6.003
820	41.3	12.4313	7.3514
850	42.3	11.9286	7.1064
900	42.5	11.73	7.225
1015	42.4	12.3808	7.4624
820	42.5	11.135	6.63

150	24.7	5.8045	3.7544
145	24.3	6.6339	3.5478
160	25.3	7.0334	3.8203
140	25	6.55	3.325
160	25	6.4	3.8
169	27.2	7.5344	3.8352
161	26.7	6.9153	3.6312
200	26.8	7.3968	4.1272
180	27.9	7.0866	3.906
290	29.2	8.8768	4.4968
272	30.6	8.568	4.7736
390	35	9.485	5.355
270	28.7	8.3804	4.2476
270	29.3	8.1454	4.2485
306	30.8	8.778	4.6816
540	34	10.744	6.562
800	39.6	11.7612	6.5736
1000	43.5	12.354	6.525
55	16.5	6.8475	2.3265
60	17.4	6.5772	2.3142
90	19.8	7.4052	2.673
120	21.3	8.3922	2.9181
150	22.4	8.8928	3.2928
140	23.2	8.5376	3.2944
170	23.2	9.396	3.4104
145	24.1	9.7364	3.1571
200	25.8	10.3458	3.6636
273	28	11.088	4.144
300	29	11.368	4.234
5.9	8.8	2.112	1.408
32	14.7	3.528	1.9992
40	16	3.824	2.432
51.5	17.2	4.5924	2.6316
70	18.5	4.588	2.9415
100	19.2	5.2224	3.3216
78	19.4	5.1992	3.1234
80	20.2	5.6358	3.0502

1100	44.6	12.8002	6.8684
1000	45.2	11.9328	7.2772
1100	45.5	12.5125	7.4165
1000	46	12.604	8.142
1000	46.6	12.4888	7.5958
200	34.8	5.568	3.3756
300	37.8	5.7078	4.158
300	38.8	5.9364	4.3844
300	39.8	6.2884	4.0198
430	40.5	7.29	4.5765
345	41	6.396	3.977
456	45.5	7.28	4.3225
510	45.5	6.825	4.459
540	45.8	7.786	5.1296
500	48	6.96	4.896
567	48.7	7.792	4.87
770	51.2	7.68	5.376
950	55.1	8.9262	6.1712
1250	59.7	10.6863	6.9849
1600	64	9.6	6.144
1550	64	9.6	6.144
1650	68	10.812	7.48
6.7	10.8	1.7388	1.0476
7.5	11.6	1.972	1.16
7	11.6	1.7284	1.1484
9.7	12	2.196	1.38
9.8	12.4	2.0832	1.2772
8.7	12.6	1.9782	1.2852
10	13.1	2.2139	1.2838
9.9	13.1	2.2139	1.1659
9.8	13.2	2.2044	1.1484
12.2	13.4	2.0904	1.3936
13.4	13.5	2.43	1.269
12.2	13.8	2.277	1.2558
19.7	15.2	2.8728	2.0672
19.9	16.2	2.9322	1.8792