Munju Kam, Deanna Kwon

Project 3

BAIS:9100 Data and Decisions

Professor David Nembhard

2020/10/19

Executive Summary

Purpose

The purpose of this project was to predict weight estimate of fish using data regression model on a given data.

Explanation of Data and Consideration

Given data contains a spreadsheet with collection of observation of fish identified by its observation number, type of species, weight in grams, lengths measured in cm of 3 different alignments, maximum height in percent, and width measured in percent, both based of measurement "Length 3", and sex of fish observed. The models are created within the limit of this dataset (appending #I-1).

Method

The method of this project is divided into 3 major stages: Data Preparation and Preliminary stage, Experiment and Result stage, and Extraction and Conclusion stage. Methods of each stages are referred to appendix #P through #C.

Results & Conclusion

As a result of numerous trials, we have concluded that Model 3 is the best regression model to predict the weight of fish with given data (appending #E-S1 to #E-S5):

The following is the full model with coefficients included which represents interaction of species with x variables ln(length3), ln(height), and ln(width):

$\mu_{\ln Y}$ = -3.0187 + 0.25132 species2 + 0.12198 species3 + 0.14526 species4 + -0.018313 species5 + 0.068802 species6 + 0.22824 species7 + 1.8498 $\ln(\text{length}_3)$ + 0.65479 ln(height) + 0.52604 ln(width) + 147

Here, we assume Y variable will be converted back to original value with $e^{\ln(\text{weight})}$ function after utilizing the model to predict ln(weight) variable. As you can see in the model, x variables of length1 and length2 were removed from the model due to its insignificance (appending #E-S3, #E-S4 respectively).

We concluded that Model 3 was the best fit model compared to the other regression model trials and the closest fit to our initial assumption of ideal way in predicting weight using only given data. Viewer should consider the outer environment and potential sample bias that may possibly incorporate error (appending #C-1 through #C-4).

Appendix

#I Ideal model

#I-1 The information derives from underlying assumption of weight and volume relationship. As the dataset includes length, width, and height which are all components to define volume, it has been a big factor of consideration along the experiment in predicting weight variable. For farther elaboration, refer to #I-2.

#I-2 Our assumption of ideal fish weight prediction model will look similar to:

> Weight = volume * density * x (certain variable) + constant,
>
> where volume (in cubic cm) = height * length * width and density = mass / volume
>
> However, this dataset does not include any information regarding mass, therefore, we will be transforming and modifying data to build model that will fit closest to the ideal model by using only given data.

#P Data Preparation and Preliminary stage

Data Preparation and Preliminary stage consist of small steps prior to actual experiment to prepare and filter the right data to use for regression model.

#P-1 During the first preparation procedure, fish #14 and #47 were disregarded and were deleted from the dataset due to having null value or 0 for dependent variable.

#P-2 Pivot table analysis was proceeded to distinguish directional framing to whether to perform the regression for each species or as a whole group. There were apparent differences in average weight of differing species. (For example, Species 5, and Species 6 has large gap between average weight) As a result, we decided to perform the separate regression for different species.

Row Labels	Average of Weight
1	626.00
2	531.00
3	160.05
4	154.82
5	11.18
6	718.71
7	382.24
Grand Total	**401.24**

The count of sample species provided in the data set varies, some species does not have enough data to perform the regression analysis. Since the weight on different species had no “NULL” or “N/A” values, we decided to use this data as categorial variable. On the other hand, sex variable

was disregarded as there were partial bias discovered in sample, such that 54.7% of sample had "N/A" or "NULL" values.

Row Labels	Count of Sex
0	55
1	16
NA	86
Grand Total	**157**

#P-2 Results: Shown in two screenshots above

Total Sample Size: n = 157 Fish

Sex Distribution: 1 (Male) = 16, 0 (Female) = 55, NA = 86

Species Distribution: 1 = 34, 2 = 6, 3 = 19, 4 = 11, 5 =14, 6 = 17, 7 = 56

#P3 We have added the derived value column called Height and Width next to the percentage of Height and Width which represent the converted value from percentage of Height and Width value to cm, integrating the calculation specified in the first page of dataset and was added to the dataset in spreadsheet named "Data_Copied."

#E Experiment and Results stage (E)

During the experiment stage, we prominently used backward stepwise regression among the 3 different version of step wise regression to extract insignificant variables. Backward stepwise regression means we ran all potential independent variable in model and remove any variable one-at-a-time based on the p value of t-statistics for each value higher than 0.05. Additionally, null hypothesis test was conducted using t-statistics and p-value of each independent variables during the filtration process, assuming H_0: $\beta_1 = 0$ where Weight and x variable does not have relationship, and H_1: $\beta_1 \neq 0$ where Weight and x variable have relationship. If p-value of a certain variable was less than 0.05, the null hypothesis is rejected which indicated variable's significance in explaining or predicting weight variable. If p-value of a variable was greater than 0.05, we fail to reject null hypothesis, which indicate the independent variable's insignificance in explaining or predicting weight variable. Several trials of regression were conducted consecutively to filtrate variables that had large p-value to best fit the model.

Note:

- For Regression analysis specifically, we have utilized the "Regression" function in the "Data Analysis" tool pack of Microsoft Excel.
- For Correlation analysis specifically, we have utilized the "Correlation" function in the "Data Analysis" tool pack of Microsoft Excel.

Assuming linearity: Model 1

#E-L1 Scatter Plots were created between y variable (Weight) and each x variable data to determine visual representation of relationship. The scatter plots are saved in spreadsheet named "Data_Res." Clearly, Length1, Length2, and Length3 had positive correlation with weight of fish only by looking at the scatter plot.

#E-L2 Correlation Matrix was created to determine correlation for each independent variable to dependent variable. The result is saved in spreadsheet named "Data_Correlation". Due to high number in correlation, we suspect that length1, length2, length3 are not completely independent from each other.

Considering linearity as is, we are interested in estimating the model,

$$Weight_i = \beta_0 + \beta_1\, length1_i + \beta_2\, length2_i + \beta_3\, length3_i + \beta_4\, height_i + \beta_5\, width_i + \varepsilon_i,$$

where weight is the dependent variable, length1, length2, length3, height, and width being independent variables, and ε being error term or residuals of the above model.

#E-L3 Beginning from the result of 1st regression in spreadsheet named "Data_Res", variable that had the largest p-value, which in this case length2 with p-value of 0.721 > 0.05, was removed and 2nd trial of regression was proceeded in spreadsheet named "B_DataL1L3HW."

#E-L4 By looking at the result of 2nd regression (appending #E-L1), variable width indicated its insignificance with largest p-value of 0.259 > 0.05. Therefore, width was removed and the final trial of regression was proceeded in spreadsheet named "Model1"

#E-L5 In the final trial, the p-value for all x variables of length1, length3 and height turned out to be under 0.05.

#E-L6 As a result, length1, length3, and height are the major significant variables to predicting weight in this particular linear regression model using backward step-wise regression.

Model 1 Results:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941721572
R Square	0.88683952
Adjusted R Square	0.884620687
Standard Error	121.8786076
Observations	157

Model 1: After last variable remover, there is no variable with p value higher than 0.05 and we can see the adjusted R square has increased compared to the original regression model

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	17811342.5	5937114.168	399.6873767	3.78735E-72
Residual	153	2272722.434	14854.39499		
Total	156	20084064.94			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-490.0473439	28.10620906	-17.4355546	3.4787E-38	-545.5736983	-434.5209895	-545.5736983	-434.5209895
Length1	67.99973148	13.14093859	5.174647991	7.06877E-07	42.03862047	93.96084249	42.03862047	93.96084249
Length3	-38.8664487	12.41690559	-3.13012356	0.002093241	-63.39716712	-14.33573028	-63.39716712	-14.33573028
Height (cm)	35.62302633	5.412830566	6.581219548	7.01522E-10	24.92949069	46.31656197	24.92949069	46.31656197

$\text{Weight} = \beta_0 + \beta_1 \text{ length1} + \beta_2 \text{ length2} + \beta_3 \text{ length3} + \beta_4 \text{ height} + \beta_5 \text{ width} + \varepsilon.$

Weight = -490.05 + 67.999 length1 + (-38.866) length2 + 35.623 height + 153

From spreadsheet named "Model1," we discovered that all of the p-value of x variables comply the rule (under 0.05) and adjusted R^2 is around 0.885 with F Statistics having p-value of 3.7874E-176 < 0.05 which indicates the variables' joint significance of determining weight variable.

Looking back to "Data_Res" tab of the excel, residual plots for x variable are implying that the linear model might not be a great fit for this data. Therefore, we have decided to perform another regression with transformation in data set to find the better fit, continuing to "Assuming non-linearity: Model 2" section.

Assuming non-linearity: Model 2

#E-N1 As mentioned above, by looking at the residuals plots on the spreadsheet tab "Data_Res," we can easily observe a parabolic shape for residual plots that stands out for variables length1, length2, and length3. Here, we have decided to apply logarithmic scales to rectify its characteristics.

#E-N2 New columns ln(L1), ln(L2), ln(L3), ln(H), and ln(W) were created with applying ln() function to values of all x variables of length1, length2, length3, height in cm, and width in cm respectively. The data is saved in spreadsheet named "Data_Copied_ln."

#E-N3 A regression model with residual plots of transformed x was proceeded and was created in spreadsheet named "Data_Res_lnx."

#E-N4 The results from the first regression and residual plots of transformed x (appending #E-N3), were unsatisfactory as residual plots still expressed parabolic characteristic.

#E-N5 New column with transformation in y variable, ln(Weight) was created with application of ln() function to values of y variable. A regression model with residual plots of transformed y was proceeded and was created in spreadsheet named "Data_Res_lny."

#E-N6 The results from the first regression and residual plots of transformed y (appending #E-N5), were unsatisfactory as residual plots still expressed parabolic characteristic.

#E-N7 A regression model with residual plots of log-log model, of both x and y variables transformed to logarithmic scales, was created in spreadsheet named "Data_Res_lnxlny."

#E-N8 Log-log model of both x and y variables in spreadsheet "Data_Res_lnxlny" (appending #E-N7), made a major difference in residual plots which display random placement.

#E-N9 Correlation Matrix was created to determine correlation for each independent variable to dependent variable. The result is saved in spreadsheet named "Data_Correl_ln".

Considering non-linearity, we are interested in estimating the model,

$$\mu_{\ln Y} = \beta_0 + \beta_1 \ln(\text{length}_1) + \beta_2 \ln(\text{length}_2) + \beta_3 \ln(\text{length}_3) + \beta_4 \ln(\text{height}) + \beta_5 \ln(\text{width}) + \varepsilon,$$

where $\mu_{\ln Y}$ = ln(weight), β_0 = coefficient constant, ln(length1), ln(length2), ln(length3), ln(height), ln(width) = independent or predictor variables, and ε for error terms or residuals of above model. Here, we assume Y variable will be converted back to original value after a valid model is extracted, using $e^{\ln(\text{weight})}$ function.

#E-N10 Beginning from the result of 1st regression in spreadsheet named "Data_Res_lnxlny", variable that had the largest p-value, ln(L1) with p-value of 0.495 > 0.05, was removed and 2nd trial of regression was proceeded in spreadsheet named "B_Data_ln(L2L3HW)."

#E-N11 Followed by result of 2nd regression (appending #E-N10), variable ln(L3) indicated its insignificance with largest p-value of 0.300 > 0.05. Therefore, ln(L3) was removed, and the 3rd trial of regression was proceeded in spreadsheet named "Model2."

#E-N12 All transformed x-variables in 3rd regression are in the spreadsheet "Model2" (appending #E-N11), which are ln(L2), ln(H), and ln(W), indicated its significance in explaining the transformed Y in this model with p-value < 0.05.

Model 2 Results:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.997654357
R Square	0.995314216
Adjusted R Square	0.995222337
Standard Error	0.091929345
Observations	157

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	274.6487736	91.5495912	10832.98332	6.4602E-178
Residual	153	1.293003694	0.008451005		
Total	156	275.9417773			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2.03239192	0.114290862	-17.78262837	4.58871E-39	-2.25818384	-1.806600001	-2.25818384	-1.806600001
ln(L2)	1.503377994	0.050117364	29.9971481	5.76808E-66	1.404366615	1.602389373	1.404366615	1.602389373
ln(H)	0.627788012	0.030666449	20.4714937	1.14493E-45	0.567203673	0.688372351	0.567203673	0.688372351
ln(W)	0.883748437	0.058745231	15.04374771	5.68335E-32	0.767691927	0.999804946	0.767691927	0.999804946

$\mu_{\ln Y} = \beta_0 + \beta_1 \ln(\text{length}_1) + \beta_2 \ln(\text{length}_2) + \beta_3 \ln(\text{length}_3) + \beta_4 \ln(\text{height}) + \beta_5 \ln(\text{width}) + \varepsilon.$

$\mu_{\ln Y} = -2.0324 + 1.5034 \ln(\text{length}_2) + 0.6278 \ln(\text{height}) + 0.8837 \ln(\text{width}) + 153.$

From spreadsheet named "Model2," we discovered that all of the p-value of x variables comply to the rule (under 0.05) and adjusted R^2 is around 0.995 with F Statistics having p-value of 6.4602E-176 < 0.05 which indicates the variables' joint significance of determining ln(weight) variable.

We assume that adjusted R^2 for this model is accurate but think that adding the categorical variable "Species" will improve the regression model. We will be conducting another model based on the model 2 and add categorical variable species, continuing to "Integrating Species Category: Model 3" section.

Integrating Species Category: Model 3

#E-S1 As integration of categories requires a new sheet of dataset, a new spreadsheet named "Data_Reg_wSpecies" was created. In this data set, we have the log-log model of x and y transformed variable data set from Model 2 with categories divided by species. As the earlier exploration of average weight in species differ, we assume that adding the categorical variable in the model 2 will make the best fit for this data. Species column Species1- Species7 were added to categorize the object (1 to represent the corresponding species and 0 otherwise).

#E-S2 1st regression was conducted with making scatter plots of the transformed x and transformed y variables, interacting with species 2 through 7 in spreadsheet "Data_Res_Spec."

Considering categorical interaction, we are interested in estimating the model,

$\mu_{\ln Y} = \beta_0 + \beta_1 \text{ species2} + \beta_2 \text{ species3} + \beta_3 \text{ species4} + \beta_4 \text{ species5} + \beta_5 \text{ species6} + \beta_6 \text{ species7} + \beta_7 \ln(\text{length}_1) + \beta_8 \ln(\text{length}_2) + \beta_9 \ln(\text{length}_3) + \beta_{10} \ln(\text{height}) + \beta_{11} \ln(\text{width}) + \varepsilon,$

where μ_{lnY} = ln(weight), β_0 = coefficient constant, $\beta_{1\text{-}6}$ = coefficient for species2 through 7 (as species1 was set to dummy variable in this model), ln(length3), ln(height), ln(width), and ε for error terms or residuals. Here, we assume Y variable will be converted back to original value with $e^{ln(weight)}$ function.

#E-S3 Beginning from the result of 1st regression in spreadsheet named "Data_Res_Spec," variable that had the largest p-value, ln(L2) with p-value of 0.128 > 0.05, was removed and 2nd trial of regression was proceeded in spreadsheet named "B_Data_Sepc_ln(L1L3HW)."

#E-S4 Followed by result of 2nd regression (appending #E-S4), variable ln(L1) indicated its insignificance with largest p-value of 0.281 > 0.05. Therefore, ln(L1) was removed, and the 3rd trial of regression was proceeded in spreadsheet named "Model3."

#E-S5 All transformed x-variables left in 3rd regression in spreadsheet "Model3" (appending #E-S4), which are ln(L3), ln(H), and ln(W), indicated its significance in explaining the transformed Y in this model with p-value < 0.05.

Model 3 Results:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.998233988
R Square	0.996471095
Adjusted R Square	0.996255039
Standard Error	0.081389813
Observations	157

Model 3: We set the species 1 as dummy variable, set x and y as log-log regression, and added the categorical variable species to perform the regression. Among all the model, we found model 3 is best working model with high adjusted R square

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	274.9680049	30.552	4612.10884	1.8126E-175
Residual	147	0.973772354	0.00662		
Total	156	275.9417773			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-3.018708117	0.248184407	-12.1632	5.45295E-24	-3.509178409	-2.528237825	-3.509178409	-2.528237825
Species2	0.251322809	0.060093747	4.18218	4.94183E-05	0.132563544	0.370082074	0.132563544	0.370082074
Species3	0.121978196	0.059066804	2.06509	0.040671193	0.00524841	0.238707982	0.00524841	0.238707982
Species4	0.145261499	0.032417706	4.48093	1.48433E-05	0.08119655	0.209326449	0.08119655	0.209326449
Species5	-0.018313543	0.101802279	-0.17989	0.857484116	-0.219498598	0.182871512	-0.219498598	0.182871512
Species6	0.068801669	0.119731839	0.57463	0.566419001	-0.167816378	0.305419717	-0.167816378	0.305419717
Species7	0.228239325	0.064202561	3.55499	0.000508821	0.101360086	0.355118564	0.101360086	0.355118564
ln(L3)	1.849838917	0.143634159	12.8788	6.94989E-26	1.565984307	2.133693527	1.565984307	2.133693527
ln(H)	0.654787694	0.137709696	4.75484	4.69142E-06	0.382641205	0.926934182	0.382641205	0.926934182
ln(W)	0.526040254	0.10493706	5.01291	1.52135E-06	0.318660141	0.733420367	0.318660141	0.733420367

$\mu_{lnY} = \beta_0 + \beta_1$ species2 $+ \beta_2$ species3 $+ \beta_3$ species4 $+ \beta_4$ species5 $+ \beta_5$ species6 $+ \beta_6$ species7 $+ \beta_7$ ~~$\ln(length_1)$~~ $+ \beta_8$ ~~$\ln(length_2)$~~ $+ \beta_9 \ln(length_3) + \beta_{10} \ln(height) + \beta_{11} \ln(width) + \varepsilon$.

μ_{lnY} = -3.0187 + 0.25132 species2 + 0.12198 species3 + 0.14526 species4 + -0.018313 species5 + 0.068802 species6 + 0.22824 species7 + 1.8498 $\ln(length_3)$ + 0.65479 ln(height) + 0.52604 ln(width) + 147

From spreadsheet named "Model3," we discovered that all of the p-value of x variables comply the rule (under 0.05) and adjusted R^2 is around 0.996 with F Statistics having p-value of 1.8126E-175 < 0.05 which indicates the variables' joint significance of determining ln(weight) variable.

In order to conduct the regression with categorical variable, we set the species 1 to dummy variable and conducted the regression based on data created previously (appending #E-S1). Among all of the previous model conducted, we are able to see that adjusted R square has improved (0.996) all of the x variables are under 0.05. Based on our experiment, the model 3 is best working model for fish data.

#C Extraction and Conclusion stage (C)

Extraction and conclusion process of extracting the best fitted model with validity and accuracy was proceeded after considering various concepts and errors that may have impacted the current data in the dataset, as well as the meaning to statistical analysis of the dataset.

#C-1 By looking at the coefficients, the type of relationship can be easily identified. The positive or negative number indicates positive relationship and negative relationship respectively. Positive relationship indicates as one unit of independent variable increases, so does the weight. On the other hand, negative relationship indicates as one unit of independent variable increases, the dependent variable decreases and vice versa.

#C-2 As a result of multiple analysis, we concluded that the best working model is model 3 with categorical variable of fish species. However, we are aware of potential bias in model because the collected sample of fish did not have even distribution of samples per species, therefore the model might be biased in such way. When using this model to reference the weight of fish, potential bias in the sample should be noted.

#C-3 Additionally, viewer should consider that there may be other environmental errors, such that growth factor in lengths, width, or height over time. By looking at the differences in fish lengths between the smallest and the largest in varying in different species, the existence of errors may be identified. Application of logarithmic scale seemed the best fit to explain growth factor in this case of the model, due to easier comparison between smaller parts and specificity.

#C-4 After taking error terms and other conditions in consideration, model 3 with transformation in both x and y variable, integrated with interaction of categorical variable, which in this case species, was selected as best performing model. Model 3 has high adjusted R^2 value of 0.996 among all models with significance F and its p - value of $1.8126E\text{-}175 < 0.05$. Therefore, we concluded that Model 3 was the best fit model in all different regression model trials and the closest fit to our initial assumption of ideal formula in predicting weight using only given data (appending #I-1 and #I-2 respectively).