A Research on Fish's Weights

Xiheng Tong

Descriptive Abstract:

159 fishes of 7 species are caught and measured. Altogether there are caught from the same lake (Laengelmavesi) near Tampere in Finland.
This data includes 159 randomly selected fishes and their properties in the lake. For each selected property, 7 variables are recorded. These variables include species, weight, three difference lengths, height and width percentage of length from the nose to the end of the tail, sex, etc.

Sources:

http://www.amstat.org/publications/jse/datasets/fishcatch.dat

About the Data:

OBJECT: Fishes caught from a lake in Finland
TYPE: Sample
SIZE: N = 159, 7 variables

Variable Descriptions:

- **Observation #** - Observation number ranges from 1 to 159
- **Species** (x_1) - Numeric

Code	Fish name
1	Bream
2	Whitefish
3	Roach
4	Parkki
5	Smelt
6	Pike
7	Perch

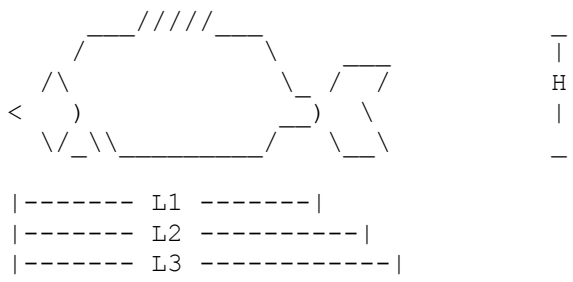
- **Weight** (y) - Weight of the fish (in grams)
- **Length1** (x_2) - Length from the nose to the beginning of the tail (in cm)
- **Length2** (x_3) - Length from the nose to the notch of the tail (in cm)
- **Length3** (x_4) - Length from the nose to the end of the tail (in cm)
- **Height%** (x_5) - Maximal height as % of Length3
- **Width%** (x_6) - Maximal width as % of Length3
- **sex** (x_7) - 1=male, 0=female

Notes:

Height = Height%*Length3/100
Width = Width%*Length3/100
$\because$ Weight may have a relationship with volume, which is approximately proportional to $x_4^3 \cdot x_5 \cdot x_6$;we will consider different transformations and multicollinearities.

SAS output

```
The SAS System
20:16 Tuesday, October 17, 2009   1

The REG Procedure
Model: MODEL1
Dependent Variable: wt

Number of Observations Read         159
Number of Observations Used         159


                          Analysis of Variance

                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F

Model                       6      17531007        2921834     167.92    <.0001
Error                     152       2644793          17400
Corrected Total           158      20175800


Root MSE           131.90889    R-Square     0.8689
Dependent Mean     399.93648    Adj R-Sq     0.8637
Coeff Var           32.98246


                      Parameter Estimates

                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|

Intercept     1     -650.13865      104.16120      -6.24      <.0001
species       1      -13.05973       12.42169      -1.05      0.2948
len1          1       25.49849       45.36999       0.56      0.5749
len2          1       23.01724       57.01733       0.40      0.6870
len3          1      -15.33960       27.59597      -0.56      0.5791
height        1        4.84476        2.74908       1.76      0.0800
width         1        9.01988        6.92323       1.30      0.1946

Note: Many fish don't have their genders recorded, so I did not use that data in the
project.
```



Model 1: Use all the data straightforward.

$$\text{Wt} = -650.14 - 13.06 \times \text{species} + 25.50 \times L1 + 23.02 \times L2 - 15.34 \times L3 + 4.845 \times H\% + 9.020 \times W\%$$

We figured out that F-value=167.92. The corresponding P-value<0.0001.

But look at the t-values. For all the five variables, their respective P-values are 0.2948, 0.5749, 0.6870, 0.5791, 0.0800, and 0.1946. All of them are significantly larger than 0.05. Therefore, the respective null hypotheses $\beta_k = 0$ are accepted, for k=1, 2, 3, 4, 5.

In the meantime, we found out from the SAS output that there were 10 outliers; and in the covariance matrix,
$Cov(species, len3) = 0.8127,\ \ Cov(len1, len2) = -0.8629,\ Cov(len2, len3) = -0.6910;$

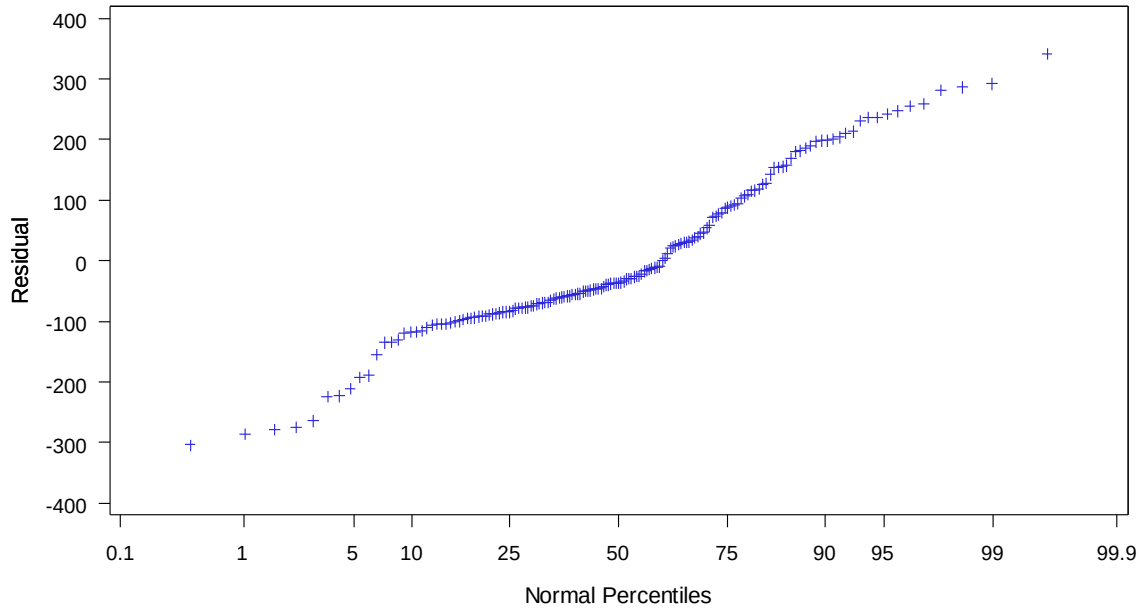
Based on the information provided above, I conclude this model to have the following characteristics:

1. The relatively high value of $R^2(0.8689)$ with few significant t statistics is the one indicator of multicollinearty. Several high values in the correlation matrix suggest the same.

2. The residual plots are pretty scattered. Therefore, we do not need to use the weighted regression method.

3. The normality plot is more like a curve. Errors do not follow the normal distribution.

4. Because Len1, Len2, Len3 have high correlations to one another, I use only Len3 in the following models.



Model 2: The use of volumes.

It's nature to guess that the weight has a linear relationship with the volume, since fish is three-dimensional.

We know that $Height = Length3 \times Height\%, Width = Length3 \times Width\%.$

$$\therefore approximately, Volume \propto \frac{Length3^3 \times Height\% \times Width\%}{100^2}$$

In the project, let's just define that $Volume = \frac{Length3^3 \times Height\% \times Width\%}{100^2}$, since the difference would only be the coefficients of volumes.

EQUATION:

$$Weight = 27.3246 + 0.2186 \times Volume$$

This model has only one variable. We can use either F value or t statistics to judge the fitness of the model.

1. $R^2 = 0.8724, F - Value = 1073.65$. The t-value for *Volume* is 32.77, p-value<0.0001. It's far more acceptable than model 1.

2. The residual plot suggests that residuals tend to increase their absolute values when volumes rise up. Therefore, we need to use the weighted regression method to modify the model.

Modified Model 2: using volumes with weighted regression

The equation is:

$$\frac{Weight}{Volume} = 0.2430 + 1.4794 \times \frac{1}{Volume}$$

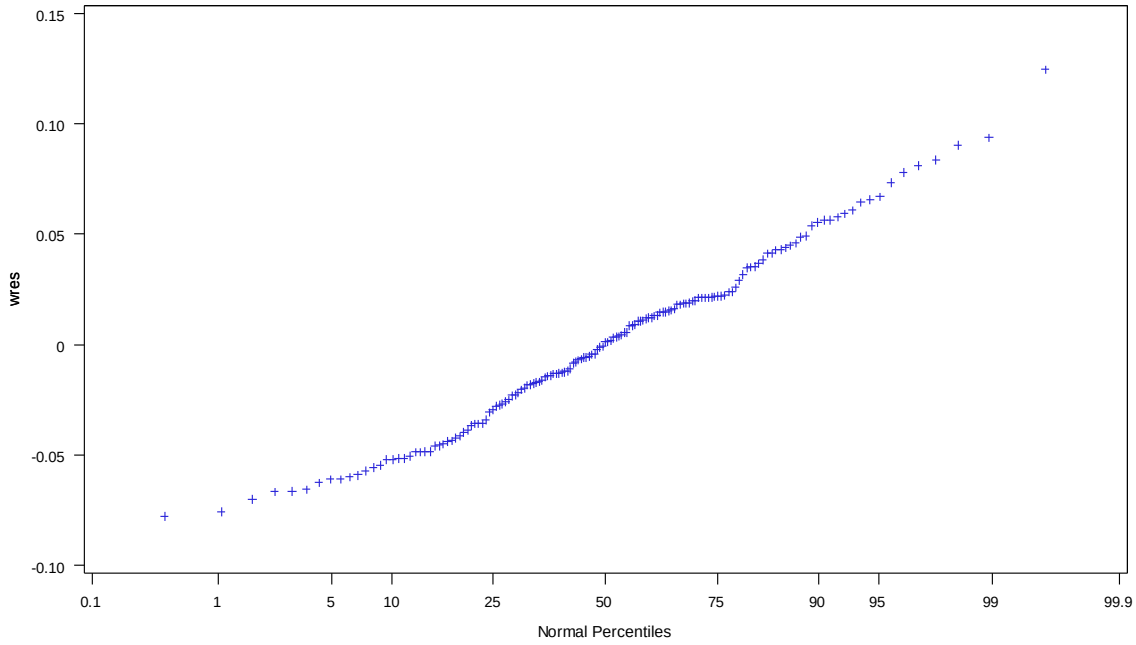
1. $R^2 = 0.9597, F - Value = 3743.62$. $S_\alpha = 0.4075, S_\beta = 0.00397$, both are significantly smaller than those in the previous model.

2. The residual plot is pretty scattered. That indicates the weighted regression is a good fit.

3. $\frac{Weight}{Volume} = density$. So we establish a relationship between *Volume* and *Density*.

4. For the INTERCEPT α, we know the t-value=3.63, the corresponding p-value=0.0004. Therefore, the Null Hypothesis $H_0 : \alpha = 0$ should be rejected. This implies that *density* is not constant, which means for different species, the fish have different densities

5. The normality plot, taking out the outliers, is shown below. We can see is more like a line. The errors follow the normal distribution:



Therefore, I conclude that this model is a good fit for the data.

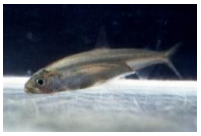
Suppose for different species, there are different linear relationships, since all the seven species have diversified shapes.


Bream


Whitefish


Roach


Parkki


Smelt


Pike


Perch

Model 3: using different models for different species.

1. For breams (observation number=35):

$Weight = 72.0889 + 0.1670 \times Volume$

$R^2 = 0.9414, F - Value = 529.67 >> F_{table}$

2. For whitefish (observation number=6):

$Weight = -38.2109 + 0.2784 \times Volume$

$R^2 = 0.9723, F - Value = 140.22 >> F_{table}$

3. For roaches (observation number=20):

$$Weight = 11.7075 + 0.2188 \times Volume$$

$$R^2 = 0.9809, F - Value = 926.83 >> F_{table}$$

4. For parkkis (observation number=11):

$$Weight = 5.331 + 0.2083 \times Volume$$

$$R^2 = 0.9912, F - Value = 1014.73 >> F_{table}$$

5. For smelts (observation number=14):

$$Weight = 3.5209 + 0.1865 \times Volume$$

$$R^2 = 0.9470, F - Value = 214.53 >> F_{table}$$

6. For pikes (observation number=17):

$$Weight = -0.5931 + 0.3328 \times Volume$$

$$R^2 = 0.9196, F - Value = 171.47 >> F_{table}$$

7. For perches(observation number=56):

$$Weight = 15.5523 + 0.2416 \times Volume$$

$$R^2 = 0.9860, F - Value = 3801.23 >> F_{table}$$

We can get the plots of dependent variables against independent variables for all the seven models, and find out that most of them don't fit the line very well. This may be due to the lack of observations. But for perches, the observation number is large enough. It's worthwhile to study this species' relationship.

From the residual plot, we know that this model encountered the similar situation as model 2. So we run a weighted regression for perches.

7. For perches (observation number=56):

$$\frac{Weight}{Volume} = 0.2610 - 0.3471 \times \frac{1}{Volume}$$

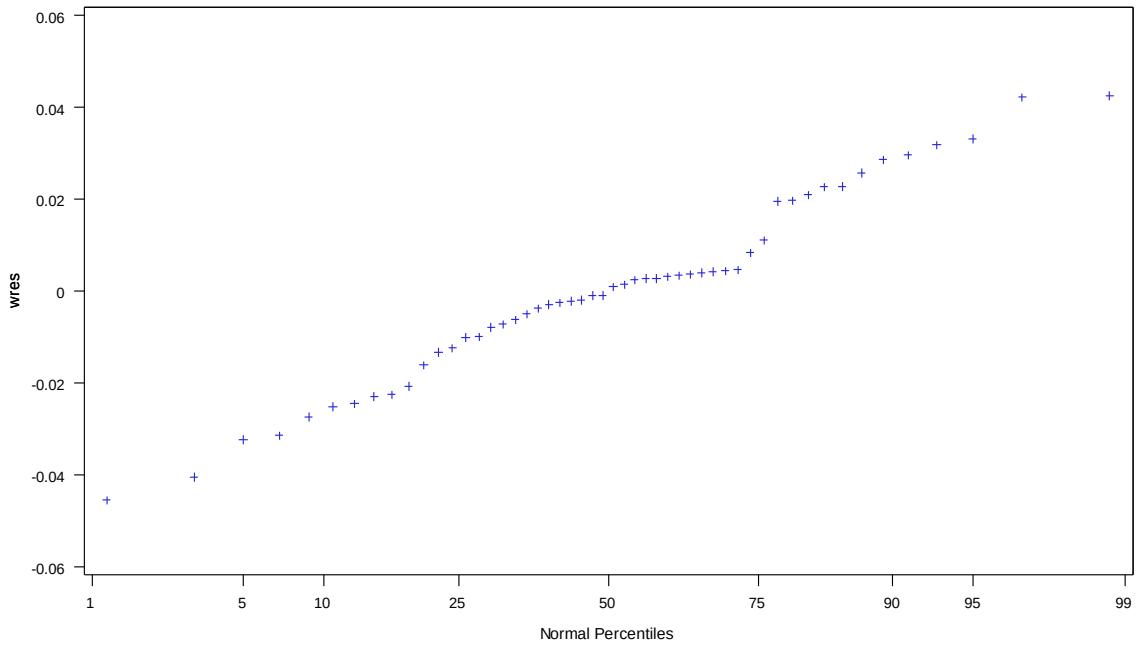
$$R^2 = 0.9895, F - Value = 5065.94 >> F_{table}$$

However, the t-value for the intercept is -0.53. The corresponding p-value=0.5974. So we have to accept the null hypothesis $H_0 : \alpha = 0$. And after deleting the outliers, the model becomes:

$$\frac{Weight}{Volume} = 0.2577.$$

This is a lovely equation. We finally find out, that for perches, the density is almost a constant.

The normality plot also implies the errors approximately follow the normal distribution.



So, this model is somewhat optimal.

Model 4: Box-Cox transformation (textbook Page277, we assume $\lambda = \frac{1}{3}$)

This is another model I found out well fit for the data. The model is also established by len3, weight% and width%.

The only difference is to make a transformation $W3 = \sqrt[3]{Weight}$ and assume:

$W3 = \alpha + \beta_1 \times Len3 + \beta_2 \times Height\% + \beta_1 \times Width\% + \varepsilon$

I did this power transformation because this is obviously a non-linear model, and 1/3 seems to be a reasonable power, since the fish is three-dimensional.

After deleting 7 outliers(#41,#96,#103,#104,#122,#153,#155), we get:

$W3 = -3.0934 + 0.1939 \times Length3 + 0.03726 \times Height\% + 0.1829 \times Width\%$

1. The C (p) test suggests that the use of all three variables is better than disposing any of them.

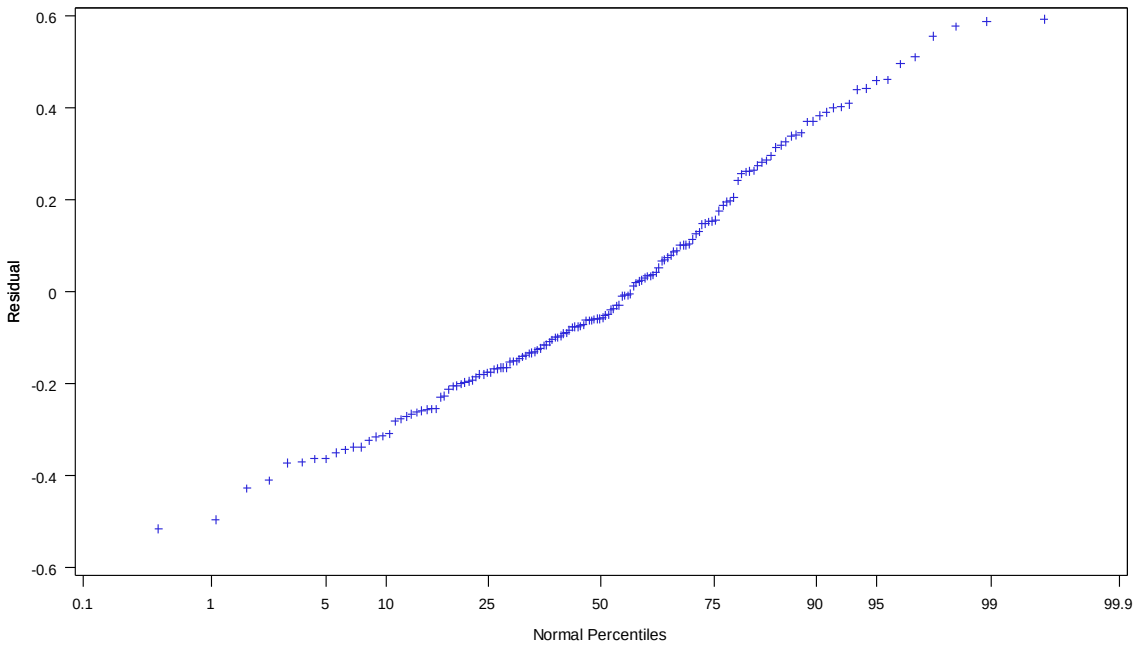
2. $R^2 = 0.9886,$ And all the corresponding p-values for t-statistics are smaller than 0.0001. So all the null hypotheses are rejected, including *intercept*=0.

3. Based on the plots, there clearly is a linear relationship between W3 and Len3

4. Residual plots are pretty scattered. There is no need to assume heteroscedasticity.

5. Look at the correlation matrix, the highest correlation is between height% and width%, $Cov(Height\%, Width\%) = -0.4437$. There are almost no linear relationships among the independent variables.

6. The normality plot is shown below:



It's safe to say that the errors follow the normal distribution.

Model 5: Box-Cox transformation (textbook Page277, we assume $\lambda = 0$)

The only difference is to make a transformation $\ln wt = \log(Weight)$ and assume:
$\ln wt = \alpha + \beta_1 \times Len3 + \beta_2 \times Height\% + \beta_1 \times Width\% + \varepsilon$
After deleting 3 outliers (#76, #103, and #104), we get:
$\ln wt = -0.47084 + 0.0996 \times Length3 + 0.02539 \times Height\% + 0.1466 \times Width\%$

1. The C (p) test suggests that the use of all three variables is better than disposing any of them.

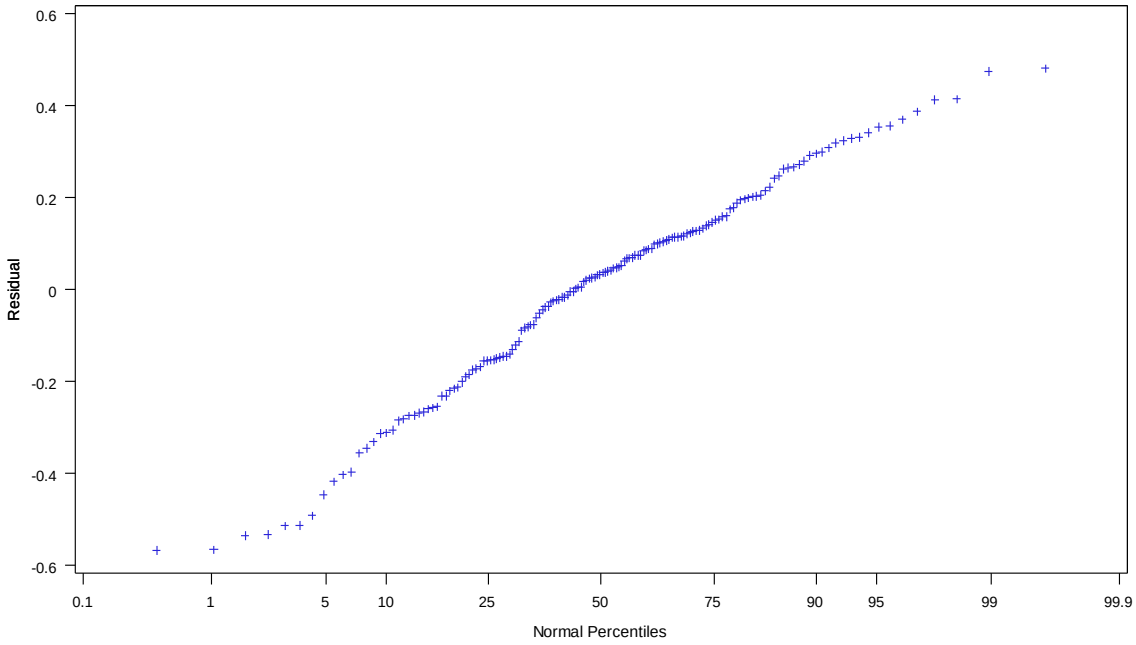
2. $R^2 = 0.9666,$ And all the corresponding p-values for t-statistics are smaller than 0.0003. So all the null hypotheses are rejected, including *intercept*=0.

3. Based on the plots, there roughly is a linear relationship between W3 and Len3

4. Residual plots are pretty scattered. There is no need to assume heteroscedasticity.

5. Look at the correlation matrix, the highest correlation is between height% and width%, $Cov(Height\%, Width\%) = -0.4407$. There are almost no linear relationships among the independent variables.

6. The normality plot is shown below:



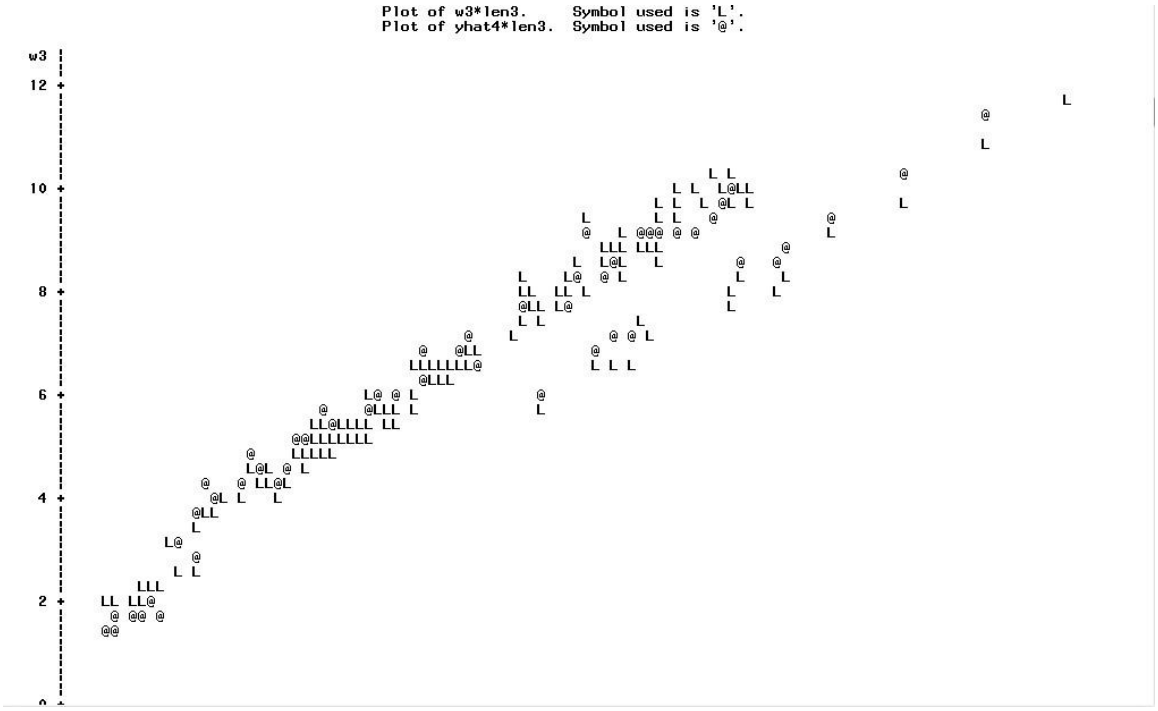
It’s almost a line.

Model 4 and Model 5 are pretty much alike. The only difference for them is the different λ’s. I prefer Model 4 over Model 5 because:

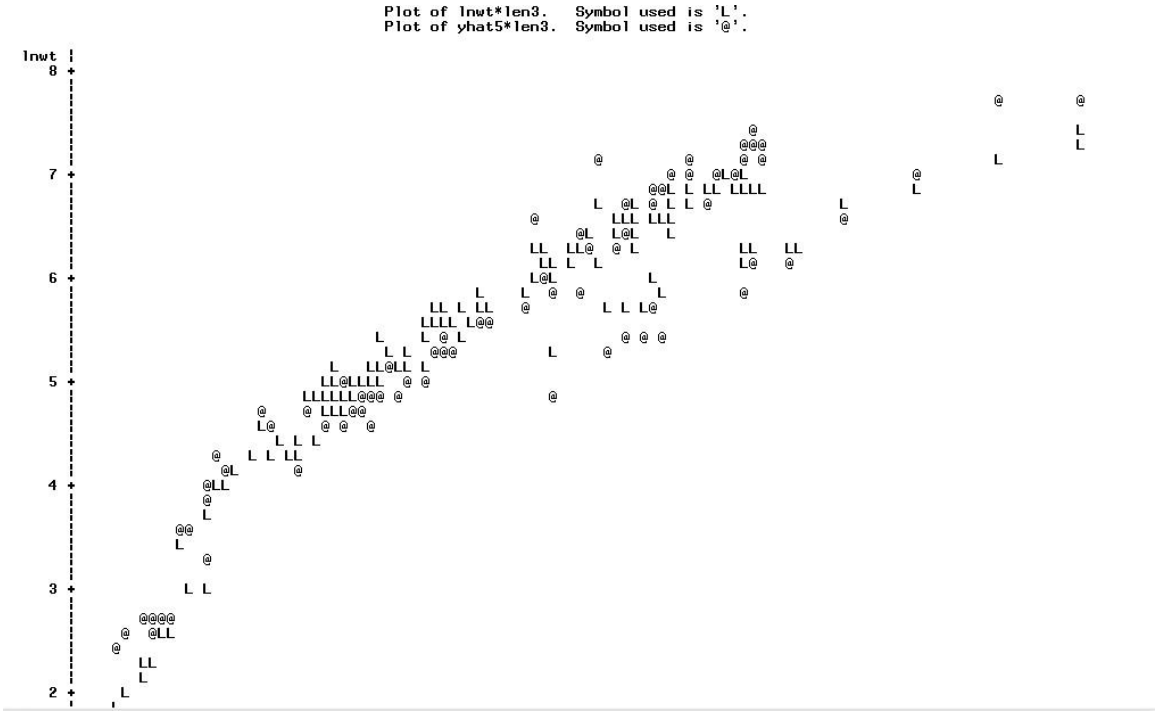
1. The normality plot for Model 4 is a little bit more like a line.

2. Comparing the plots of dependent variables against Length3, Model 4’s has a better linear relationship.

3. Intuitively w3 is a better transformation, since the fish is always three-dimensional.



In Model 4: w3*len3



In Model 5: lnwt*len3

Conclusion:

When I chose to use this subject, I knew specifically that there won't be a linear relationship between the dependent variable and independent variables. This must be a non-linear model project. The evidence substantiated that Model 1 was far from being an optimal model. Because the fish is not plane or unidimensional , I guessed that I could explore some cubic relationships among them. That's how the Model 4 came to my mind. However, I reached the professor in Finland who collected the data for me, and he suggested that I could also use Model 5 to establish the relationship. He also suggested me to read some contents on Box-Cox Transformation, which is the theoretic foundation for Model 4 and 5.

We also learned from high school physics about calculating the density. That's how I discovered Model 2 and 3. In Modified Model 2, the hypothesis $H_0 : \alpha = 0$ was rejected, which concluded that the fish did not have constant density. Those made me wonder if the same species had a close density or not? They have similar appearances; it's natural for me to make that assumption. However, not all the species have enough number of observations, which resulted in the fact that the fitness for Model 3 was only applicable for perches.

Based on all the information provided above, if we don't know the species, we can use either MODIFIED MODEL 2 or MODEL 4 to predict the weight. If we do know the species and the fish happens to be a perch, we can use MODEL 3, since the densities of the fish are relatively constant.