Weight Prediction of Alligators in Central Florida

**G. M. Nazmus Salehin**

**University of North Florida**

**Statistical Methods Project 2**

Table of Contents

**Introduction1**

**Methodology8**

**Results18**

**Discussion19**

**Bibliography20**

# Abstract

The purpose of this paper is to develop a model to predict the weight of an alligator from its length. The datasets are taken from aerial photographs or from visual sightings from air or boat. From that, the length can be estimated quite accurately. However, the alligator’s weight is much more difficult to determine. So, a suitable model is required to predict weight of and alligator. Analysis of the dataset has been conducted and from that a simple regression model has been developed and its validity was determined. After that, a transformed model was compared with the simple linear model and the best model has been chosen. Finally, with the selected model, weight for some new alligator has been predicted from their length.

Chapter 1

# 1.1: Motivation:

Many wildlife populations are monitored by taking aerial photographs. Information about the number of animals and their whereabouts is important to protect certain species and to ensure the safety of surrounding human populations.

It is very difficult to determine the weight an alligator from its length. The goal of this research is to develop a model that would enable us to predict the weight of alligator from its length. The dataset has been collected from the alligator’s captured in Central Florida. Length of the alligators are mostly estimated from aerial photographs or from visual sightings from air or boat.

# 1.2: Background:

Before jumping onto any model, the dataset has been analyzed at first. Though our weight vs length data show non-linearity, a simple linear regression has been fitted to the model to determine any possible prototype pattern in the fitted line, residual plots etc that would help us finding a compatible model. As our analysis with simple linear model proved that this model was not compatible, we proceed by using a transformation approach called Box-Cox approach that automatically selects an optimal transformation on the response variable. We again fitted our data using the new transformed response variable and conducted some test to check the validity of the model. Finally, we predicted weight for the new observations.

The statistical tools that has been employed by this research are as follows:

# Simple Linear Regression:

Simple Linear Regression is a linear regression model with a single explanatory variable. The simple linear regression model can be expressed as:

Where the intercepts and slope are unknown constants and is the random error.

# R-Squared:

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%. Though R-squared is a widely used measure to assess performance of the model, there are some disadvantages. It should not be used to compare models which are in different scales as it depends on transformation that has been used. one must consider R-squared values in combination with residual plots, other statistics, and in-depth knowledge of the subject area.

# ANOVA:

Analysis of variance (ANOVA) is a collection of [statistical models](https://en.wikipedia.org/wiki/Statistical_model) and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a [sample](https://en.wikipedia.org/wiki/Sample_(statistics)). The following equation is the fundamental analysis of variance identity for a regression model:

Or SSTO=SSR+ SSE.

# LACK OF FIT TEST:

In [statistics](https://en.wikipedia.org/wiki/Statistics), a sum of squares due to lack of fit, or more tersely a lack-of-fit sum of squares, is one of the components of a partition of the [sum of squares](https://en.wikipedia.org/wiki/Sum_of_squares_(statistics)) of residuals in an [analysis of variance](https://en.wikipedia.org/wiki/Analysis_of_variance), used in the [numerator](https://en.wikipedia.org/wiki/Numerator) in an [F-test](https://en.wikipedia.org/wiki/F-test) of the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) that says that a proposed model fits well. The other component is the pure-error sum of squares.

The pure-error sum of squares is the sum of squared deviations of each value of the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) from the average value over all observations sharing its [independent variable](https://en.wikipedia.org/wiki/Independent_variable) value(s). These are errors that could never be avoided by any predictive equation that assigned a predicted value for the dependent variable as a function of the value(s) of the independent variable(s). The remainder of the residual sum of squares is attributed to lack of fit of the model since it would be mathematically possible to eliminate these errors entirely.

For lack of fit test, the alternatives are:

= Assumed model fits the data well.

= Assumed model is not good enough.

We then use the general linear test statistic:

Where MSLF denotes the lack of fit mean square and MSPE denotes the pure error mean square.

If , conclude ,

If , conclude , where c is the number of levels and n is the total number of observations in our dataset.

# Residual Diagnostics:

A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate. The residual plots below show three typical patterns.

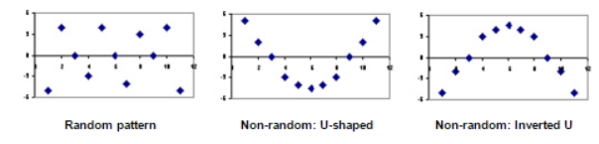


Figure 1.1 Residual Plots for random and no-random pattern

The first plot shows a random pattern (or no pattern), indicating a good fit for a linear model. The other plot patterns are non-random.

# Checking Constancy of Error Variance:

# Brown-Forsythe Test:

The test is based on the variability of the residuals. The larger the error variance, the larger the variability of the residuals will tend to be. To conduct the Brown-Forsythe test, we divide the data set into two groups, according to the level of X, so that one group consists of cases where the X level is comparatively low and the other group consists of cases where X level is comparatively high. If the error variance is either increasing or decreasing with X, the residuals in one group will tend to be more variable than those in the other group. Equivalently, the absolute deviations of the residuals around their group mean will tend to be larger for one group than for the other group. In order to make the test more robust, we utilize the absolute deviations of the residuals around the median for the group. The Brown-Forsythe test then consists simply of the two-sample t test based on test statistic to determine whether the mean of the absolute deviations for one group differs significantly from the mean absolute deviation for the second group.

# Transformations:

# Box-Cox Procedure:

At the core of the Box Cox transformation is an exponent, lambda (λ), which varies from -5 to 5. All values of λ are considered and the optimal value for your data is selected; The “optimal value” is the one which results in the best approximation of a normal distribution curve. The [transformation](https://www.statisticshowto.datasciencecentral.com/transformation-statistics/)of Y has the form

Chapter 2: Methodology

# Data Analysis:

Before applying any model, we must visualize the data first. All the analysis is done in SAS. A good way to look for possible relationships between predictor and response variable is to investigate the scatter plot.

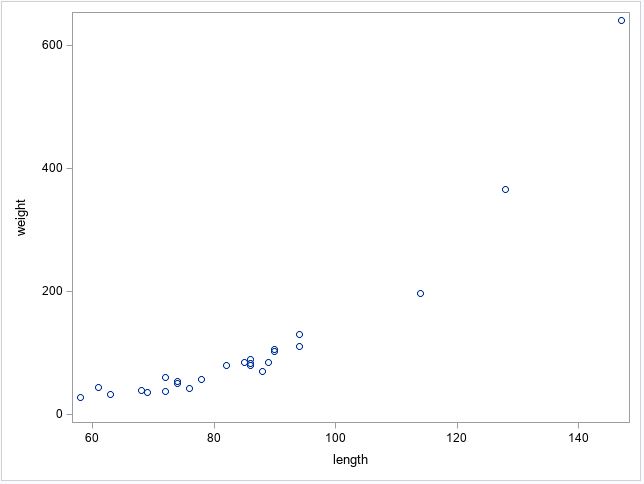


Figure 2.1: Scatter plot between weight and length variable of alligator

From the scatterplot, we can see that the relationship between weight and length is not exactly a straight line. Also we can see some extreme observations which may be outliers.

Next, we draw boxplot and which can help us identifying possible outliers. Boxplot shows us 2 extreme observations. However we don’t have enough evidence to discard these data so we are going to generate our model including these extreme observations.

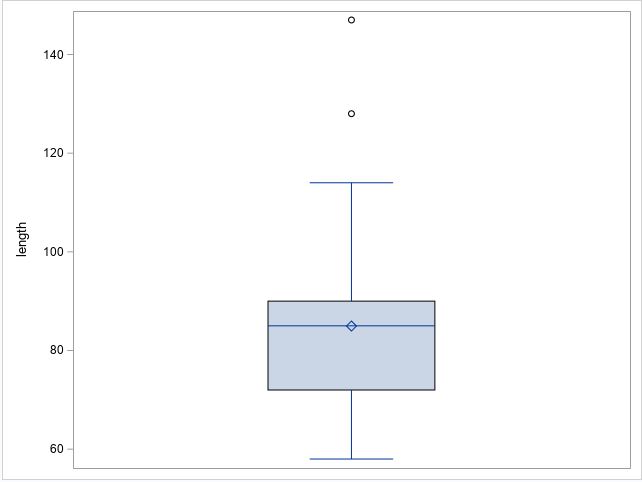


Figure 2.2: Box plot of length variable

Although the relationship between response and predictor variables does not look like a straight line, we first model our data using simple linear regression so that we can compare the output variables with an updated model.

# Simple Linear Regression on Data:

Fitting our data to simple linear regression gives us the following results:



Figure 2.3: Simple Linear Regression Results on the Alligator Dataset

We got a low value of 0.8361. Also by looking at the fit plot and normal probability plot it is evident that a simple linear regression model is not applicable here.

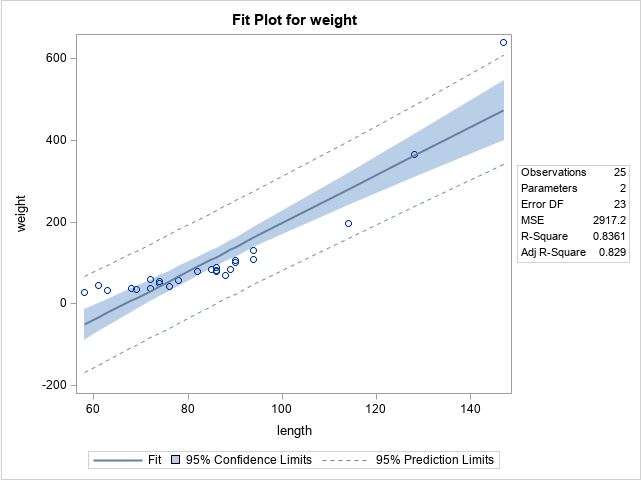


Figure 2.4: Fit plot for the alligator dataset using simple linear regression

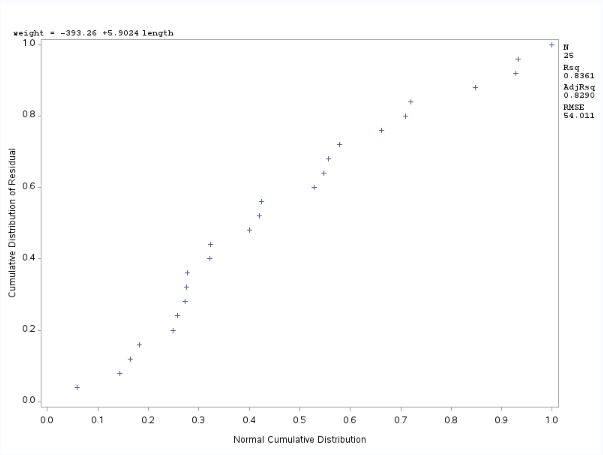


Figure 2.5: Normal Probability plot for the simple linear regression model

Moreover, our simple linear regression model gives us a very high F value of 44.36 in lack of fit test with a very low probability that confirms that the simple linear regression model cannot describe our dataset.

Lastly, we check the residual vs prediction plots which may give us an idea about which model to choose

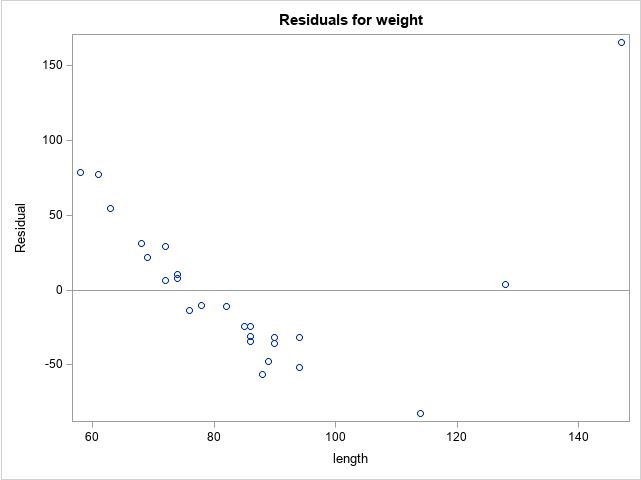


Figure 2.6: Residual vs Prediction plots for the Simple Linear Regression Model

Fig: Residual plot for the simple linear regression model

By looking at the ‘U-shaped’ residual plot, we are confirmed that there must be a nonlinear relationship between the weight and length variable.

A box cox transformation on the weight variable will hep us finding the possible transformation.

# Box Cox Transformation

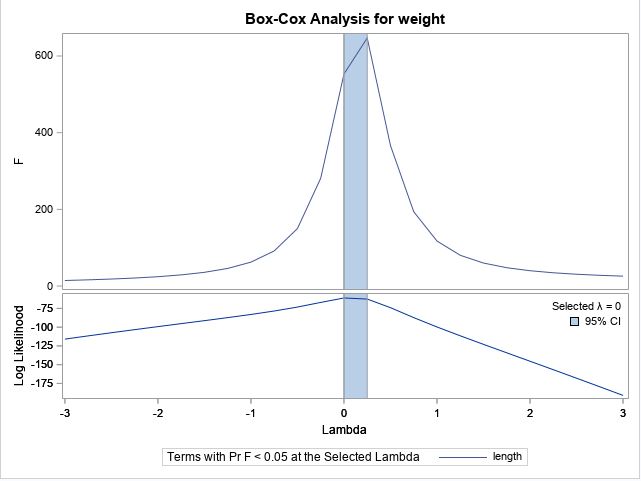


Figure 2.7: Box- Cox Transformation Result on the Alligator Dataset

The box cox transformation gives us a λ value of 0 which suggests a log transformation on weight variable.

# Log Transformation on the Model:

Now we log transform our model so that our model looks like this: where .

Let’s look at the scatter plot of our new weight vs the length variable.

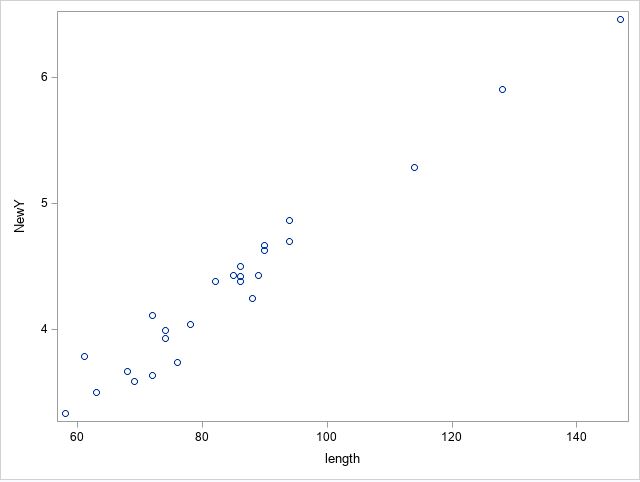


Figure 2.8: Scatter plot of log(weight) vs length

As our data looks reasonably linear now, let’s use simple linear regression model on the transformed y variable. The log transformed model gives us a high R squared value of 0.9601.

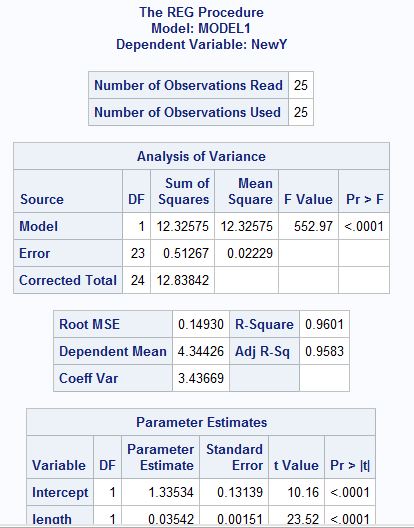


Figure 2.9: Linear Regression results with Log transformation on weight variable

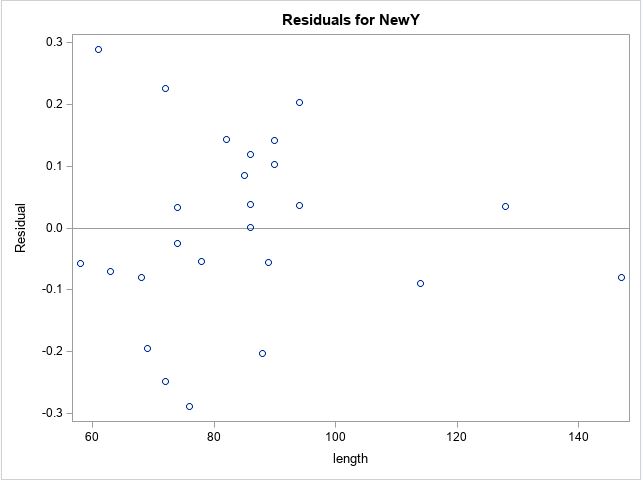


Figure 2.10: Residual plot for the log transformed model

Our new log transformed model seems to have normally distributed residuals. To confirm that, let’s do a Brown and Forsyth test to check for constancy of error variance. For that we divided the total observations to 2 groups of 12 and 13 observations.

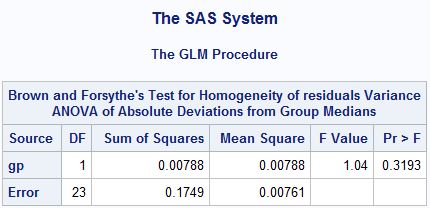


Figure 2.11: Brown and Forsythe Test result on the Log Transformed Model

A p value of 0.3193 concludes that it is reasonable to assume error variance is constant. Still, we investigate the fit diagnostic plot next.

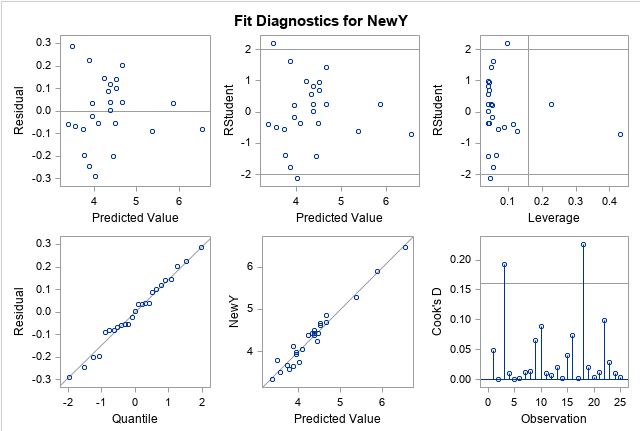


Figure 2.12: Fit Diagnostic plot for log transformed model

The quantile plot is very close to straight line and we are confirmed that a log transformed model is a better choice in this dataset.

A lack of fit test on the model gives a F value of 0.98 with a p value of 0.5539 which states that our model can fit the data well enough.

# Comparison of the Simple Linear Model vs The Log Transformed Model:

So, log transformation on our data looks good. We can compare our log model with the simple linear model side by side to show this.

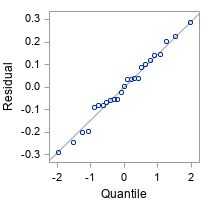
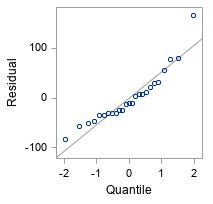


Figure 2.13: Quantile plot of the simple linear model(left) vs Quantile plot of the log transformed model (right)

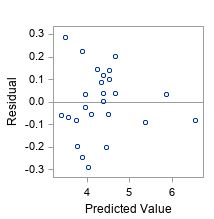
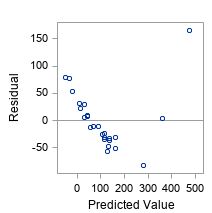
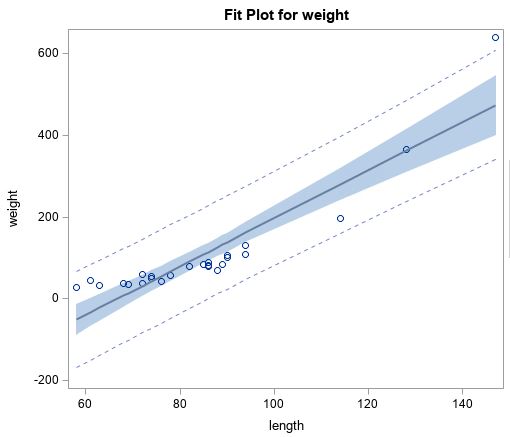


Figure 2.14: Residual vs Prediction plot for simple linear model(left) vs the log transformed model (right)

Lack of Fit test on the simple Linear Model gives us a high F value of 44.36 with a p value less than the significance (<0.001) in contrast to log transformed model, which gives a low F value of 0.98 with a p value of 0.5539, suggesting the log transformed model performs well than the simple linear model.

Lastly, let’s compare the fit plot of 2 model which can give us a visual impression of how the log transformed model outperforms the simple linear model.



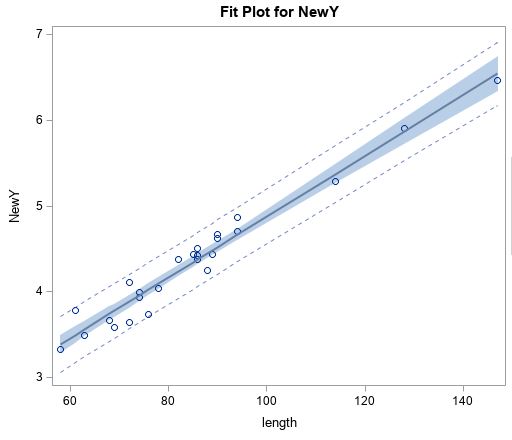


Figure 2.15: Fit plot of the simple linear model (top) vs fit plot of the log transformed model (bottom)

We can therefore use this log transformed model for our prediction of new observations.

Chapter 4: Results

# 4.1 Prediction for New Observation

Now that we have selected our model, let’s predict the alligator weight for the observations 70, 80 and 85:

|  |  |  |
| --- | --- | --- |
| Observed length | Predicted weight (rounded to nearest decimals) | 95% CL Predict (rounded to nearest decimals) |
| 70 | 45 | 33 to 63 |
| 80 | 65 | 47 to 89 |
| 95 | 110 | 80 to 151 |

These results are transformed to the original scale because SAS gives us predictions based on log scale as we used log transformations. So, our prediction equation is:

Where intercept and .

Chapter 4: Discussions

We started our model selection process by analyzing the data first. The data set has some possible outliers/ extreme observations which was confirmed by analyzing the box-plot. However, there are no concrete proof of these observations being wrong so we could not discard these observations. We fitted a simple linear regression line on the data set that gave us a R-squared value of only 0.83. Also, looking at the residual plot and normal probability plot, it was evident that the dataset does not follow a linear pattern. Also, our F- test on simple linear regression produces a very high value of 44.36 which indicates a simple linear regression is not compatible with this dataset. After that we performed box-cox procedure on the dataset that pointed us a log transformation is the best transformation on the response variable. As box- cox procedure does not guarantee normality, we again checked our model with quantile plot. The quantile plot seems to be linear which confirms normality. We also performed a Brown- Forsythe test on the new model that produced a p- value of 0.31 which is more than the significance level and confirmed our new model’s constancy of error variance. With our log transformed model we predicted the weight for new alligator observations and found out 95% confidence interval for the new predictions.

Our dataset was small, so we could not split our dataset into training/ testing sets which would help us finding how our model performs with new dataset, however by adding new observations to the dataset would enable us to do so.

Also, there is a good possibility that weight of an alligator not only depend on length, but also on some other variables like age/ gender/ region etc. So, adding more variables may help us finding a better model so that we can experiment with how different predictors interact with the response variable.

We could also fit a polynomial regression to our data, but the dataset is very small and there is a chance that it may overfit our dataset.

# Bibliography

Kutner, M. H., & Nachtsheim, C. J. (n.d.). *Applied Linear Statistical Models.*

*StatisticsHowTo*.(n.d.).Retrieved from https://www.statisticshowto.datasciencecentral.com/

Sunthornjittanon, S. (2015). *Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand.* Portland State University.

*Wikipedia*. (n.d.). Retrieved from https://www.wikipedia.org/

# Appendix

Here is the code that is used in SAS:

**data** alligator;

input length weight;

cards;

94 130

74 51

147 640

58 28

86 80

94 110

63 33

86 90

69 36

72 38

128 366

85 84

82 80

86 83

88 70

72 61

74 54

61 44

90 106

89 84

68 39

76 42

114 197

90 102

78 57

**proc** **print** data=alligator;

**proc** **sgscatter** data=alligator;

plot NewY\*length;

**run**;

/\*BoxPlot\*/

**proc** **sgplot** data=alligator;

vbox length;

**run**;

/\*slr\*/

**proc** **reg**;

model weight=length/lackfit;

output out=two r=residuals p=predicted;

plot r.\*npp.;

plot r.\*p./vref=**0**;

/\*BoxCox\*/

**proc** **transreg** ss2 details;

model boxcox(weight)=identity(length);

**run**;

/\*Log Transformed Model\*/

**proc** **reg**;

model NewY=length/lackfit;

output out=two r=residuals p=predicted;

plot r.\*npp.;

plot r.\*p./vref=**0**;

/\*Perform Brown-Forsythe Test\*/

**proc** **sort** data=alligator;

by length;

**proc** **print** data=alligator;

**DATA** three;

set two;

if(\_n\_)<**13** then gp=**1**;

else gp=**2**;

**proc** **print** data=three;

**proc** **GLM** data=three;

class gp;

model residuals=gp;

means gp/hovtest=bf;

**run**;