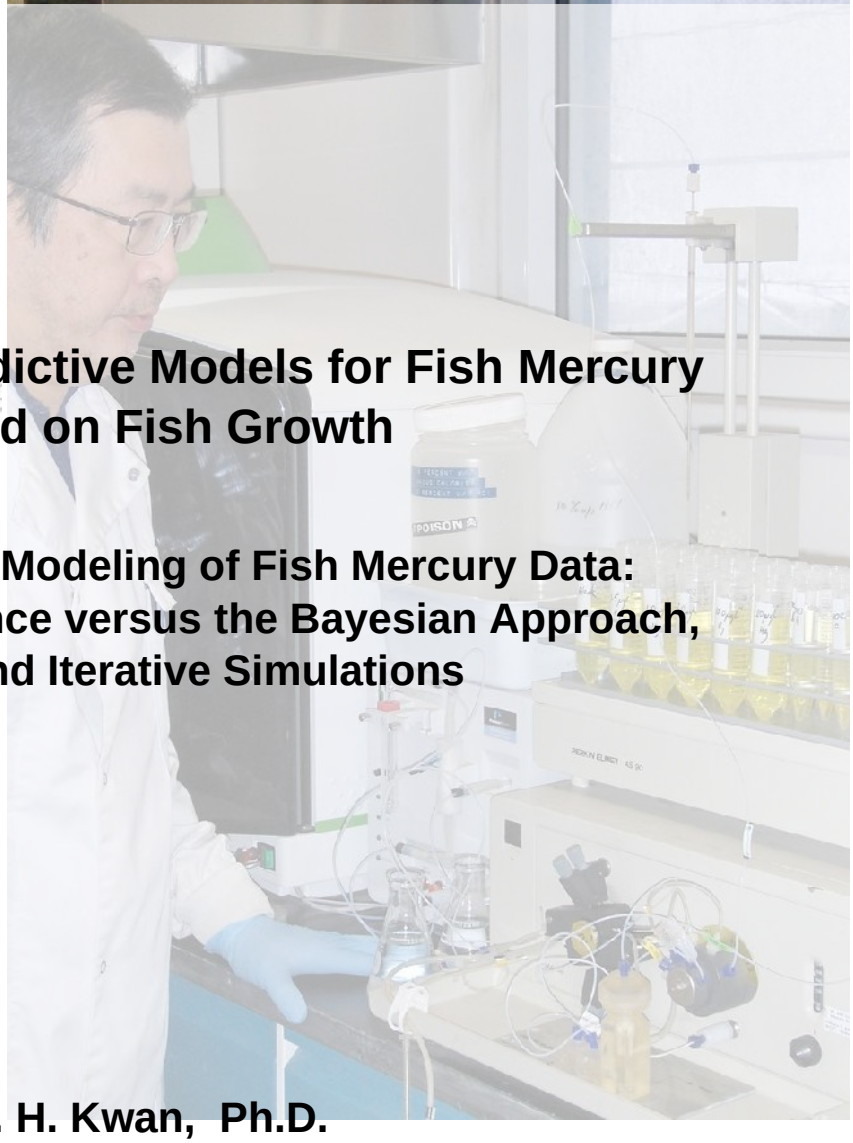


In Pursue of Reliable Predictive Models for Fish Mercury Levels based on Fish Growth

Comparative Regression Modeling of Fish Mercury Data: Classical Frequentist Inference versus the Bayesian Approach, Bootstrapping and Iterative Simulations

Michael K. H. Kwan, Ph.D.

Nunavik Research Centre,

Kuujjuaq, Quebec

Société Makivik
Makivik Corporation

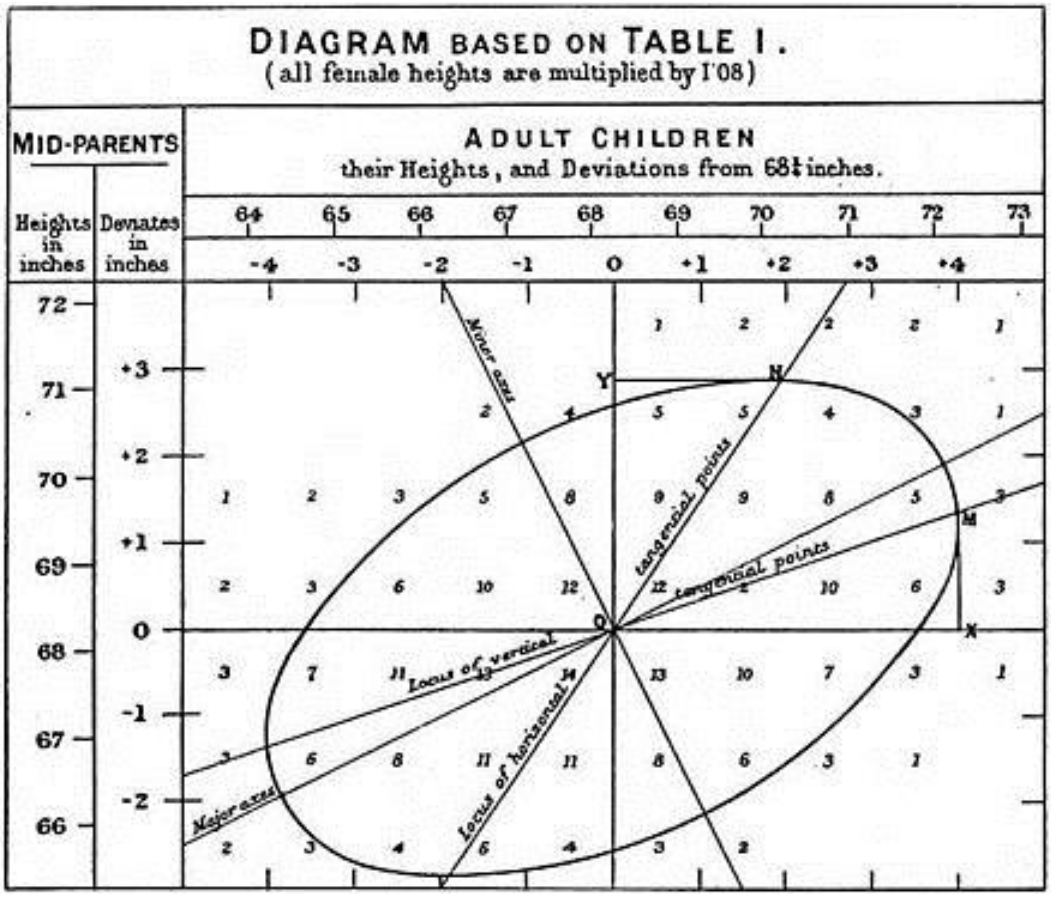
February 2020

“Statistics is the grammar of science.”

Karl Pearson (1857 – 1936)
founder of the discipline of mathematical statistics

Pearson, K. (1892) *“The Grammar of Science.”* Adam & Charles Black, London. 552 pp.

A little bit of history: When regression began in statistics.



The Victorian era statistician Sir Francis Galton first coined the word "**regression**" in the discipline of statistics and invented the use of regression lines in describing correlation relationships in the 1880s at King's College University of London, my own alma mater. Galton observed that adult children's heights tended to deviate less from the mean height than their parents and suggested the concept of "regression towards the mean", giving regression its name (Galton 1886). This diagram from his famous paper illustrated the "locus of horizontal tangential points" passing through the leftmost and rightmost points on the ellipse (which is a level curve of the bivariate normal distribution estimated from the data) is ordinary least squares estimate of the regression of parents' heights on children's heights, while the "locus of vertical tangential points" is the ordinary least squares estimate of the regression of children's heights on parent's heights. The major axis of the ellipse is the total least squares estimate.

Galton, F. (1886). "Regression towards mediocrity in hereditary stature." *The Journal of the Anthropological Institute of Great Britain and Ireland.* **15**: 246–263. 1886

Table of Contents

Summary

- **The structure and properties of the data sets were thoroughly examined in Chapter 1. This has important bearings on the choice of modeling approaches since almost all regression modeling methods have some sorts of pre-requisite assumptions concerning the data structure that needed to be met in order to validate their application. First and foremost, influential outliers were identified and removed using a batch of standard diagnostics such as Cook's distance, Hi – leverage, Studentized residuals and DFFITS. Then a thorough residual analysis was carried out to reveal the nature of the data set: (1) Homoscedasticity (Goldfeld – Quandt Test and various residual plots); (2) Randomness of residuals (Wald – Wolfwitz Runs Test); (3) Independence of residuals (Durbin – Watson test statistic on first – order autocorrelation, auto - correlograms) and (4) Frequency distribution (Normality plots and Anderson – Darling test statistic). The issues of known and unknown confounding (lurking) variables as well as measurement errors in predictor variables (weak exogeneity) were discussed.**

- **The problem of multicollinearity and its restriction on model choice as well as possible remedies were discussed in details in Chapter 2 since the predictor variables in the present case were by nature, collinear. The use of eigenvalues of centered correlations, variance inflation factors, variance proportion and conditional indices in gauging the extent of collinearity followed by the use of auxiliary regression to identify the culprits amongst predictors were demonstrated. The weakness and inability of the ordinary least squares (OLS) model that relies on an unbiased estimator in dealing with multicollinearity was discussed.**

- **The use of shrinkage regularization methods to tackle multicollinearity was demonstrated: Chapter 3 discussed the use of Tikhonov regularization (or ridge regression) by using a biased estimator to carefully control the variance of the model; Chapter 4 utilized the partial least square (PLS) approach relying on decomposition of variables and projection of these variables into latent structures in new hyperplanes so as to nullify the collinear effect.**

- **The difference between the classical frequentist inference statistics and the Bayesian belief revision subjective probability approach using randomized resampling and Markov - chain Monte - Carlo simulation was discussed in Chapter 5. The advantages of the Bayesian approach were highlighted. Bayesian linear regression was applied to model the fish data.**

- **In Chapter 6, hierarchical multiple linear regression (HMLR) was used to investigate the quadratic term of the curvilinear data of the lake whitefish to ascertain the impact of quadratic transformation on the overall model of this species. Subsequently, HMLR was used to fit the lake white fish data into a quadratic polynomial function.**

- **Robust regression modeling using maximum likelihood estimation based on the approach of iteratively reweighting least squares was used in Chapter 7 to generate predictive models of fish mercury level based on the three predictor variables. Unlike conventional approaches that all outliers must be removed before modeling, full data sets were used in robust regression modeling since the method is designed to minimize or even nullify the influence of outliers. The very robust Tukey's bi-square M -estimator was used as the influence function to counter the effect of the presence of outliers.**

- **In the final Chapter, the classical hierarchical multiple linear regression in conjunction with the modern iterative bootstrapping resampling - with - replacement method was used to investigate fish age as the confounding variable on the regression model.**

- **Ultimately all these models generated needed to be put to test and validated using newly collected fish samples from the same watersheds from which the fish data used in constructing the models came from. The three growth parameters (fork length, weight and age) of these new fish samples will be measured and the data fed into these models to calculate the "predicted" fish mercury levels. These predicted fish mercury levels will then be compared with the "actual" mercury levels determined by laboratory analysis of the fish samples to assess the performance of each model (see Epilogue).**

Summary of Regression Models:

Ridge Regression:

Brook trout: $\log_e$ mercury = (-5.4914) + (0.4053) * $\log_e$ fork length + (0.0276) * $\log_e$ weight + (0.1857) * age

Lake trout: mercury = (-0.0268) + ($5.168 * 10^{-4}$) * fork length + ($1.43 * 10^{-4}$) * weight + ($3.915 * 10^{-3}$) * age

Partial Least Squares Regression:

Brook trout: mercury = (0.01982) + ($1.2 * 10^{-4}$) * fork length + ($4.4 * 10^{-5}$) * weight + ($8.13 * 10^{-3}$) * age

Lake trout: mercury = (- 0.05469) + ($4.47 * 10^{-4}$) * fork length + (10^{-4}) * weight + (0.0148) * age

Lake whitefish: mercury = (0.03064) + ($8.6 * 10^{-5}$) * fork length + ($2.2 * 10^{-5}$) * weight + ($3.06 * 10^{-3}$) * age

Bayesian Linear Regression:

Brook trout: $\log_e$ mercury = (-2.169) + (1.716) * $\log_e$ fork length + (-1.484) * $\log_e$ weight + (0.248) * age

Lake trout: mercury = (0.285) + ($4.54 * 10^{-4}$) * fork length + ($1.47 * 10^{-4}$) * weight + (0.002) * age

Quadratic Polynomial Models for Lake Whitefish by Hierarchical Multiple Linear Regression:

$\log_e$ mercury = (-1.11801) + (-0.01221) * fork length + ($2.46 * 10^{-4}$) * fork length^2

$\log_e$ mercury = (-2.64852) + ($-6.19 * 10^{-6}$)) * weight + ($8.04 * 10^{-7}$) * weight^2

$\log_e$ mercury = (-2.50823) + (-0.08413) * age + (0.01408) * age^2

Robust Regression by Maximum Likelihood Estimation:

Brook trout: $\log_e$ mercury = (-2.6380) + (0.00829) * fork length + (-0.5151) * $\log_e$ weight + (0.1916) * age

Lake trout: mercury = (-0.01944) + ($5.16 * 10^{-4}$) * fork length + ($1.42 * 10^{-4}$) * weight + ($4.27 * 10^{-3}$) * age

Lake whitefish: $\log_e$ mercury = (-3.1193) + ($2.16 * 10^{-6}$) * fork length^2 + ($5.16 * 10^{-3}$) * weight + (0.0409) * age

Hierarchical Multiple Linear Regression with Age as Confounding Variable:

Brook trout: $\log_e$ mercury = (-22.340) + (4.866) * $\log_e$ fork length + (-1.531) * $\log_e$ weight + (0.256) * age

Lake trout: mercury = (-0.19) + (0.037) * age

Chapter 1. Data Structures - Homoscedasticity, Normality, Independence, Randomness, Weak Exogeneity and Confounding Variables.

1.1 The Data (see Appendix)

The data set used in this monograph originated from fish samples collected from the Inukjuak River watershed (brook trout and lake whitefish) and Lake Tasialuk (lake trout) in northern Quebec in the summer of 2019 as part of the environmental baseline study of the Innavik Hydroelectric Project. The fish collected and growth parameters measured by technicians from Groupe WSP Global Inc. The mercury concentrations in the fish samples were determined using a standard operating protocol based on cold vapour atomic absorption spectrometry and fish age was determined using otolithes (lake trout and brook trout) and scale (lake whitefish) at Nunavik Research Centre (Kwan 2019).

1.2 Outliers

There is no rigid mathematical definition of what constitute an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise (Zimek and Filzmoser 2018). Outliers (or influential points) can loosely be defined as observations that do not follow the pattern of other observations in the data set. They often exhibit numerically large residuals. Influential outliers cause dependent variables as well as the residuals of the predictor variables deviate from normal frequency distribution resulting the data violating the pre-requisite assumption of the ordinary least squares (OLS) regression model. See Chapter 7 for more on the issues of outliers.

Before proceeding to analysis of residuals and assessing multicollinearity problems, outliers of both the dependent variable and the predictor variables were identified and removed. In the Appendix of data sets, outliers were highlighted in red and they were identified by examining their Studentized deleted residuals, Hi - leverages, Cook's distance and DFFITS. Each of these four diagnostic measures illuminate different aspects of the data, so they do not necessary identify the same observations. Potential outliers were carefully checked and the decision of their removal depended on the outcome of these diagnostic tests. Remember, "Garbage in, Garbage out. Outliers needed to be weeded out!".

Outliers of the dependent variable were identified using Studentized deleted residuals. However, in the case when there was more than one predictor variable, Bonferroni – corrected t – statistic was used to identify outliers in order to control the possible inflation of Type I Error. In this, $100\{1 - \alpha/2n\}$ percentile of Student's t with $(n - p' - 1)$ degrees of freedom (df) was calculated. Where α

is the significant level (usually $\alpha = 0.05$), n is the number of cases, p' is the total number of predictors in the multiple regression (k) plus 1 (p' = k + 1). If the Studentized deleted t residual is greater than the Bonferroni – corrected t – statistic, the corresponding observation is probably an outlier.

Outliers of the predictor variables were identified using Hi - leverage (h_{ii}), Cook's distance and DFFITS. Each individual value of a predictor has a leverage. Hi - leverages measure the joint influence of the predictors by a standardized distance of the i^{th} observation to the other n – 1. A rough indication of an unusual observation is a leverage exceeding 3p'/n. One disadvantage of leverages is that they do not distinguish between high leverage points that are influential in the calculation of partial regression coefficients and those that are not. A measure that is more sensitive to such influential points is the Cook's distance (D). It compares the estimates of the regression coefficients from the full dataset with the estimates of the regression coefficients when the i^{th} observation is omitted; hence, a large value indicates an influential point. The beauty of Cook's distance is that it combines both the Studentized deleted residuals and the leverages: $D = h_{ii} (1 - h_{ii})^{-1} t_i^2 p'$ (Belsley et. al. 2013). The observation is probably an outlier when its Cook's distance exceeds the F statistic with p' and (n – p') degrees of freedom at 0.5 level of significance (i.e. the F – value at the 50th percentile). DFFITS is very similar to Cook's distance. In fact, DFFITS is the square root of p' times Cook's D with the sign of the Studentized deleted t residual attached. Observations with DFFITS values greater than $2\sqrt{(p'/n)}$ are considered large and are probably outliers.

Outliers have a tendency to "nest". This means removing one set of outliers identified alters the regression line; and the new line often uncovers other outliers. Often it needed to go through several cycles of finding and removing and re-running the regression until all outliers are removed from the dataset.

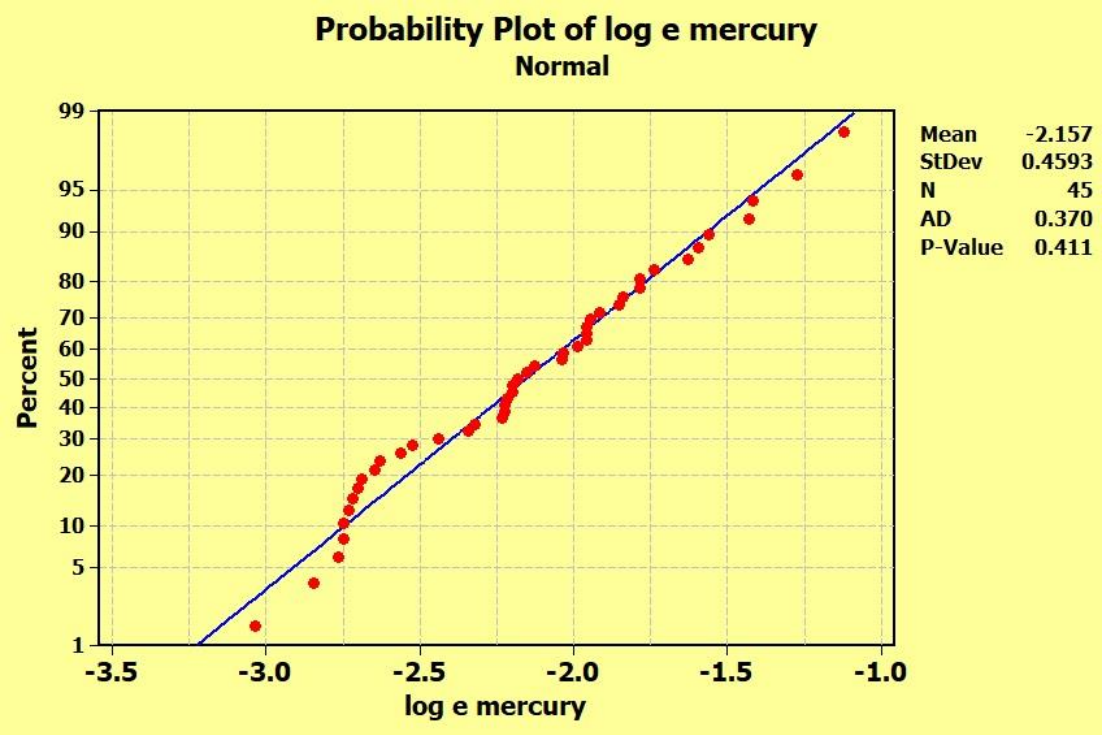
1.3 Studentized deleted residuals for multiple predictor variables

Since there are multiple predictor variables in the datasets, Studentized deleted residuals of each predictor variable were used instead of raw residuals in diagnoses because Studentized deleted residuals are most sensitive in checking for outliers in recognizing the fact that an extreme outlier may influence the estimates of the partial regression coefficients as well as the variance of the residuals. The Studentized deleted residual of an observation is calculated by dividing an observation's deleted

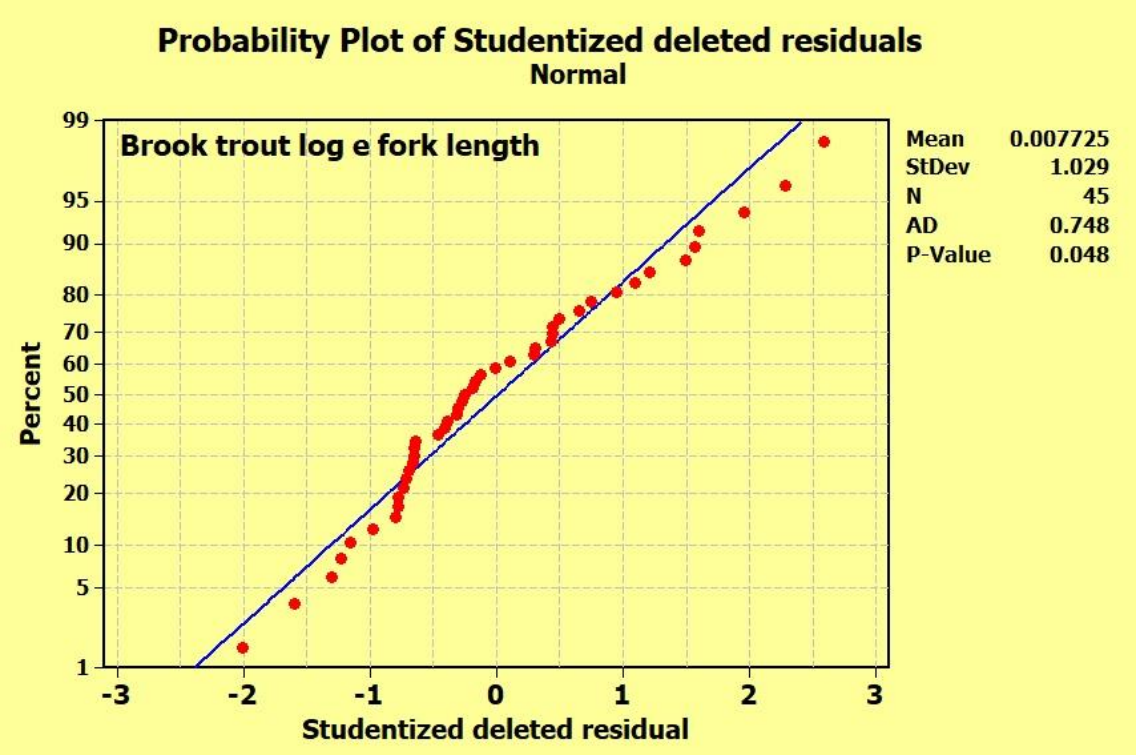
residual by an estimate of its standard deviation. A deleted residual d_i is the difference between y_i and its fitted value in a model that omits the i^{th} observation from its calculations. The observation is omitted to see how the model behaves without this potential outlier. Studentizing residuals is useful because raw residuals can be poor indicators of outliers due to their non-constant variance: residuals with corresponding x-values that are far from mean of x have greater variance than residuals with corresponding x-values closer to mean of x. Studentizing controls for this non-constant variance, and all studentized t deleted residuals have the same standard deviation. Each Studentized deleted residual follows the t distribution with (n – 1 – p) degrees of freedom, where p equals the number of predictor variables in the regression model.

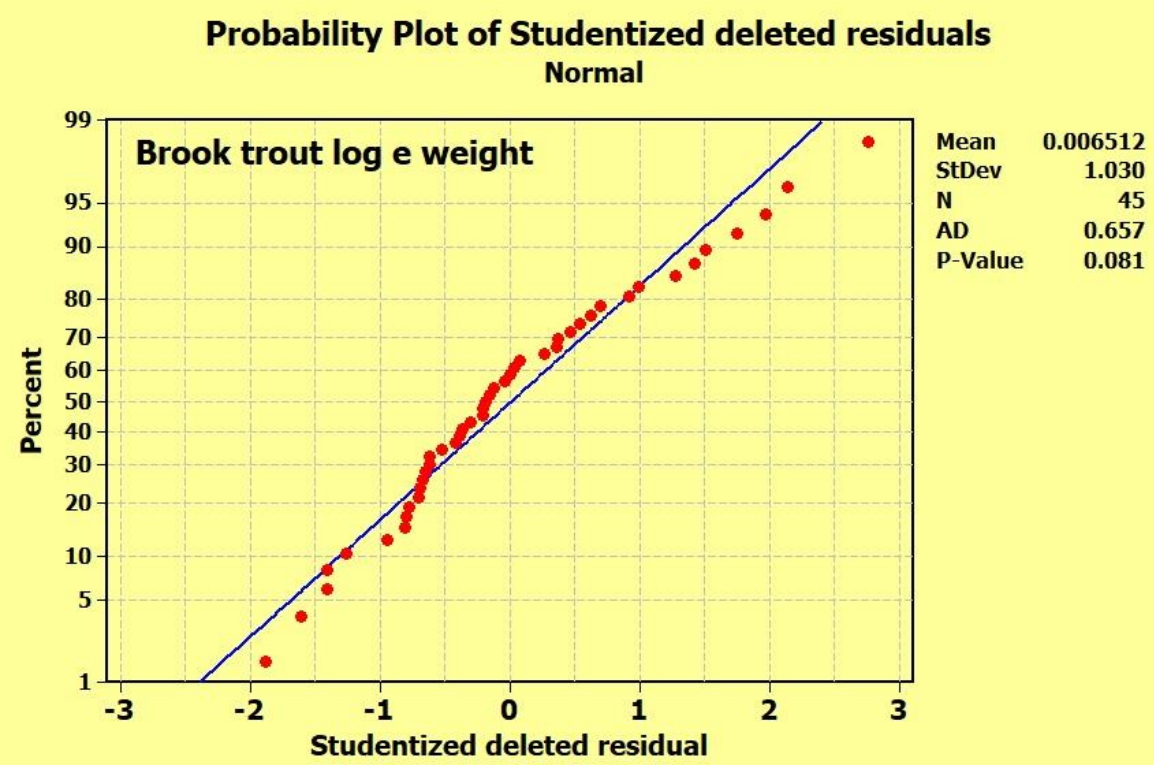
1.4 Normality

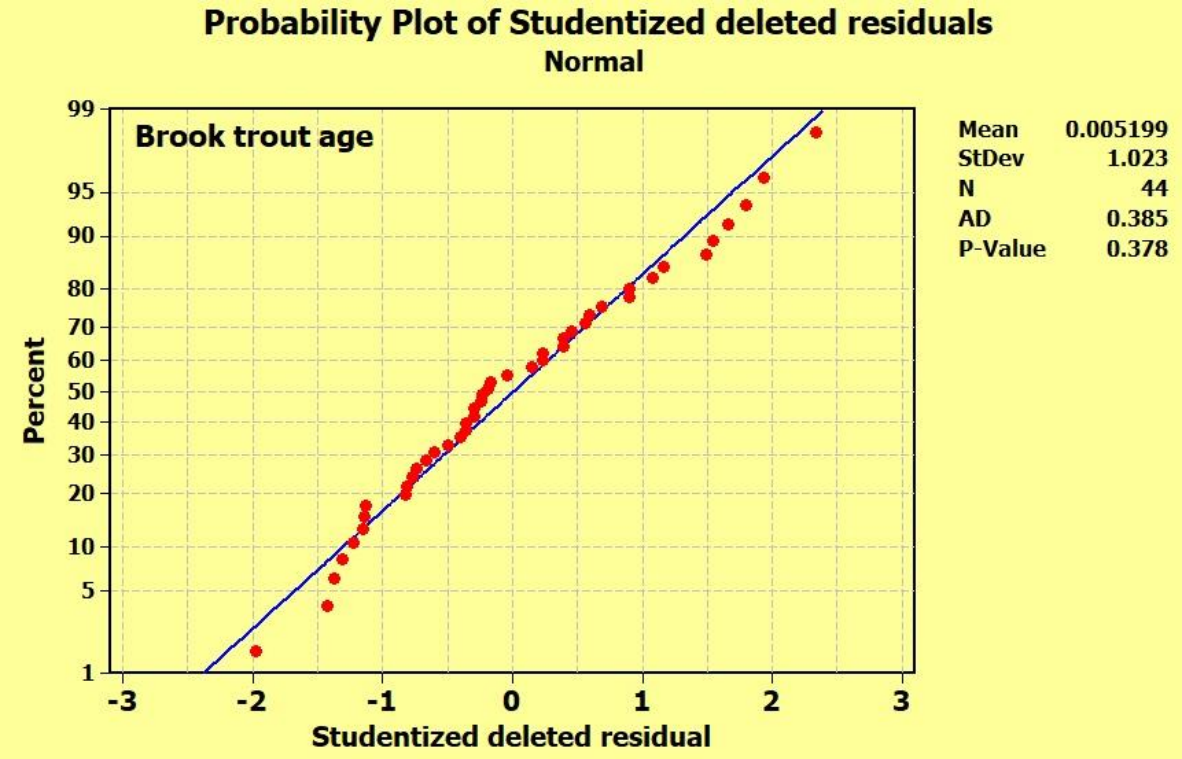
Normal probability plots and Anderson – Darling test statistic (AD) were used to assess the probability distributions of the dependent variable and the Studentized deleted residuals of each predictor variable. Anderson – Darling test compares the empirical cumulative distribution function (ECDF) of the sample data with the distribution expected if the data were normal. If this observed difference is sufficiently large ($p < 0.05$), the test will reject the null hypothesis of population normality.



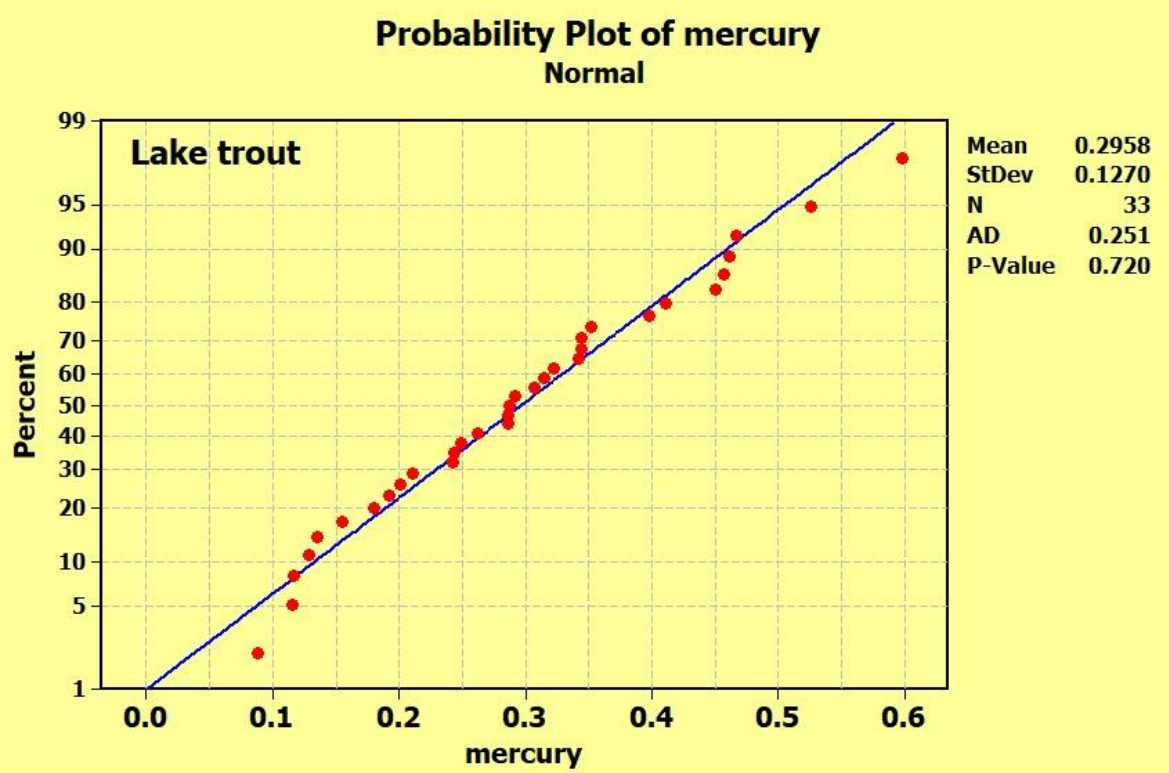
Log $_e$ transformation was needed to render the mercury data of brook trout normally distributed. The Studentized deleted residuals of both fork length and weight of brook trout also needed log $_e$ transformation to become normal. No transformation for brook trout age was needed:

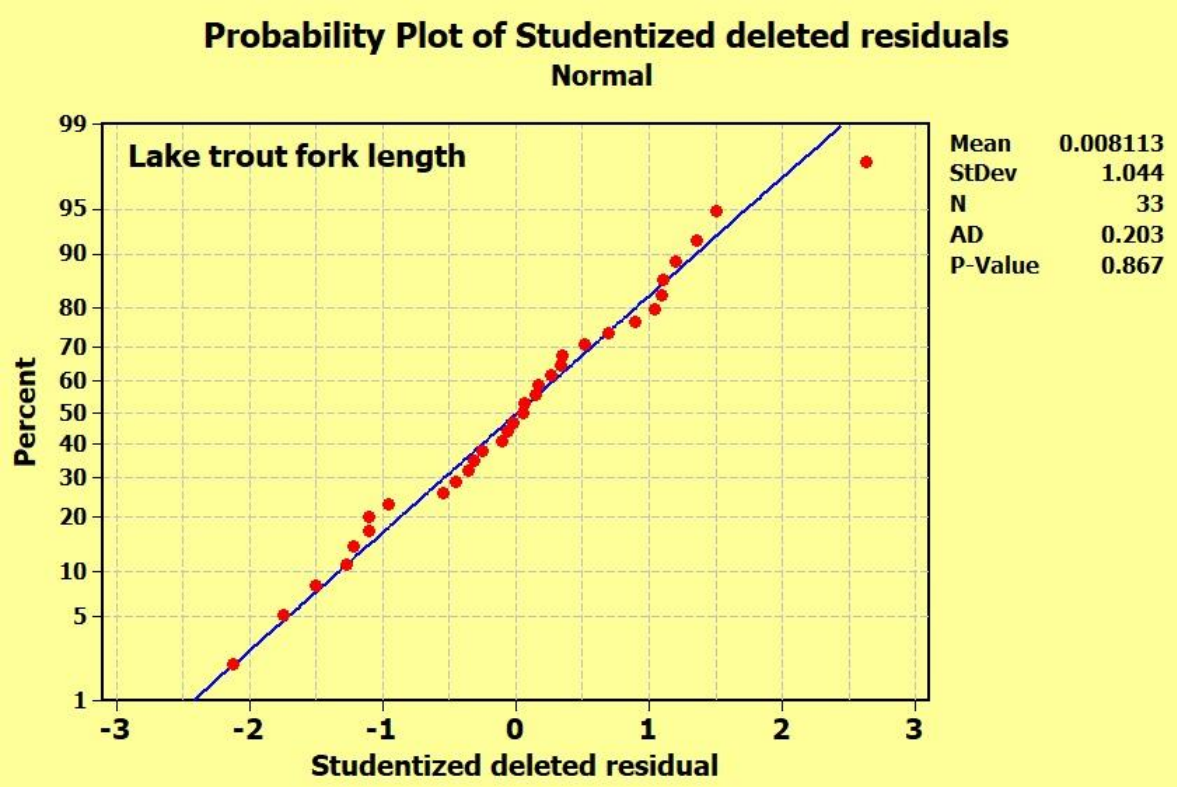


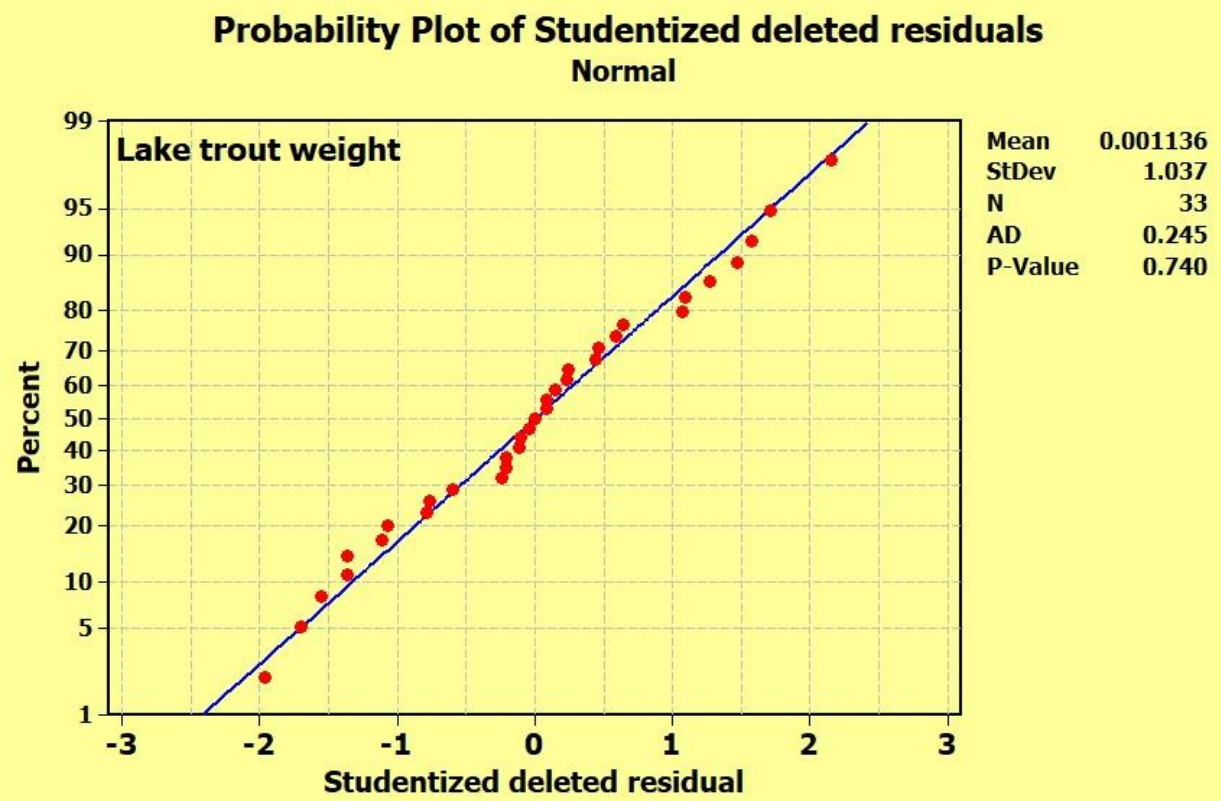


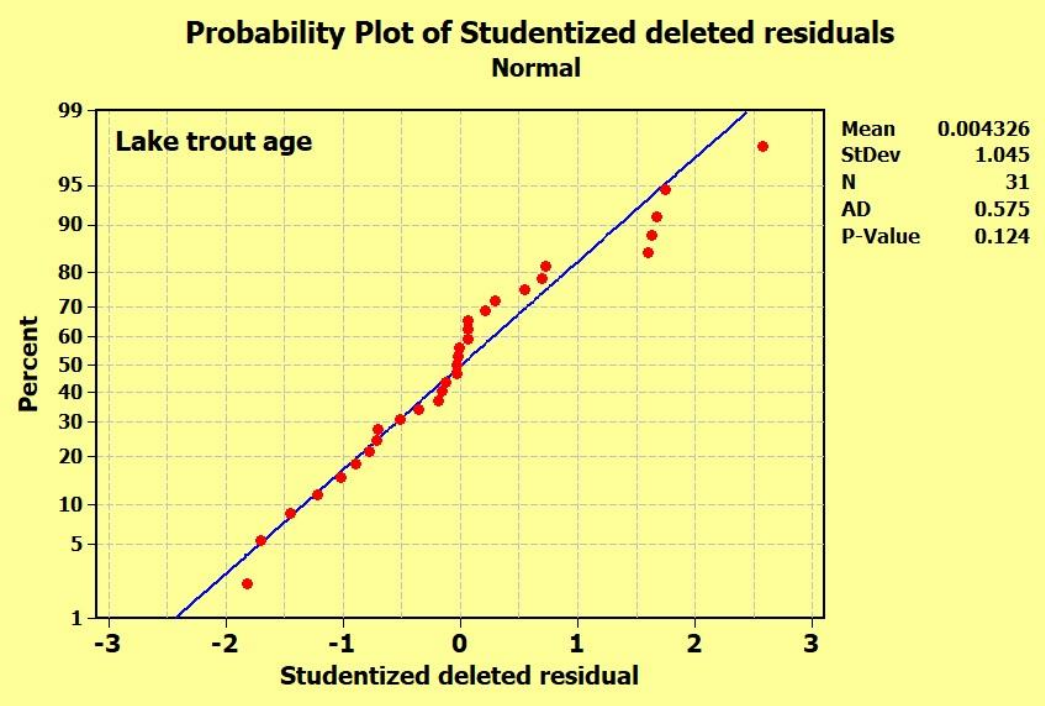


No transformation was needed for the lake trout data:

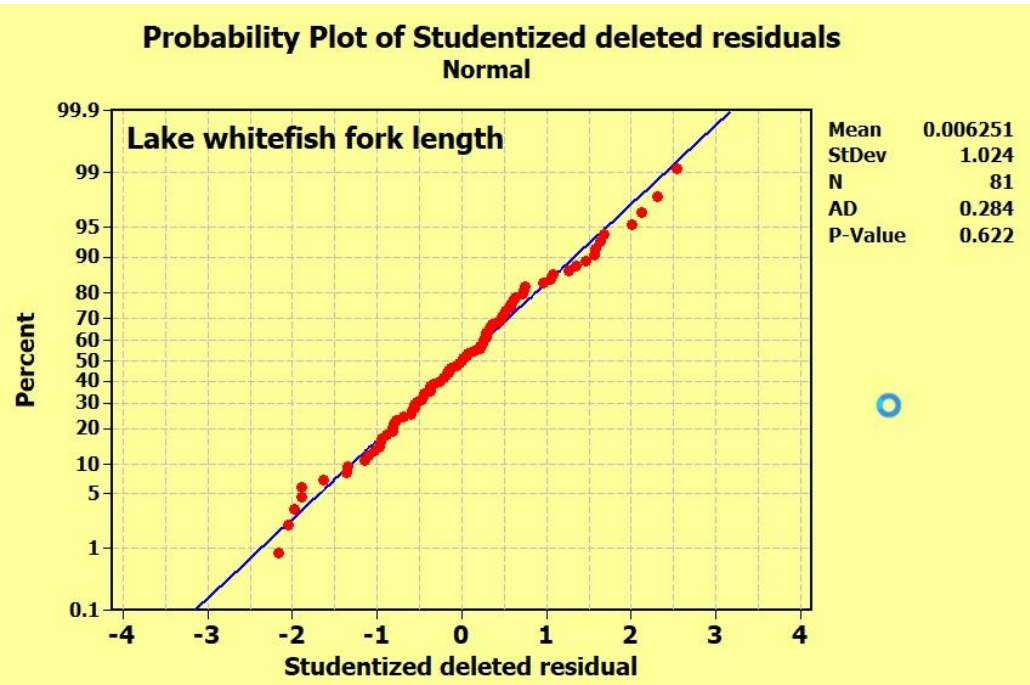


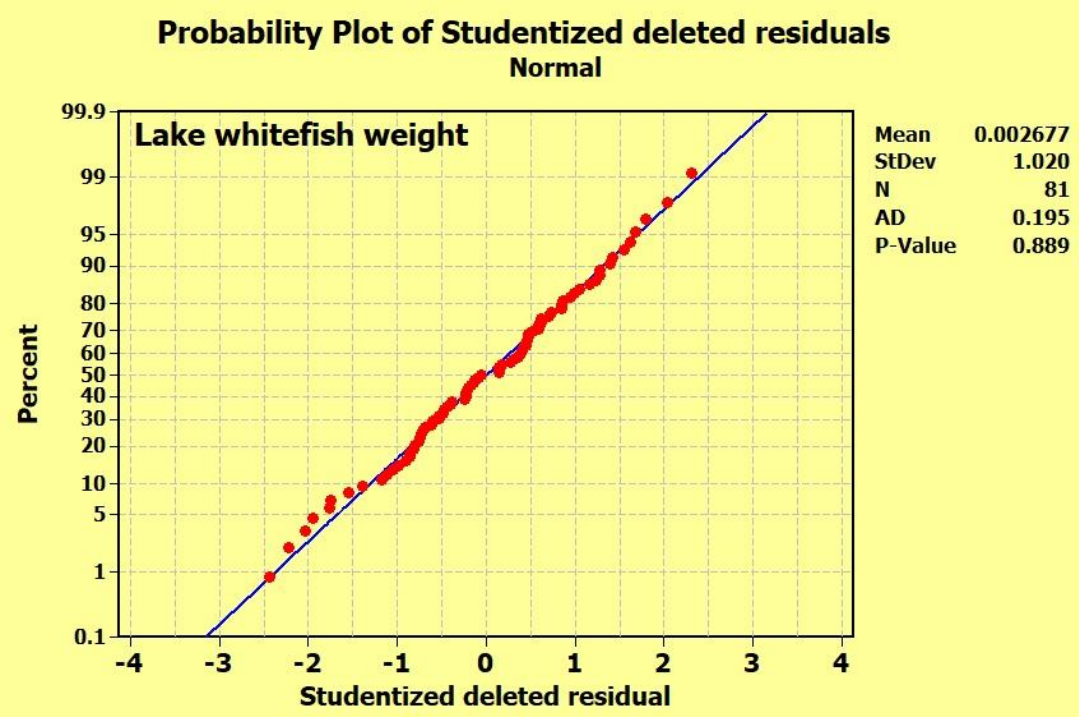

Probability Plot of Studentized deleted residuals
Normal
Lake trout fork length
Mean 0.008113
StDev 1.044
N 33
AD 0.203
P-Value 0.867
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-3
-2
-1
0
1
2
3
Studentized deleted residual

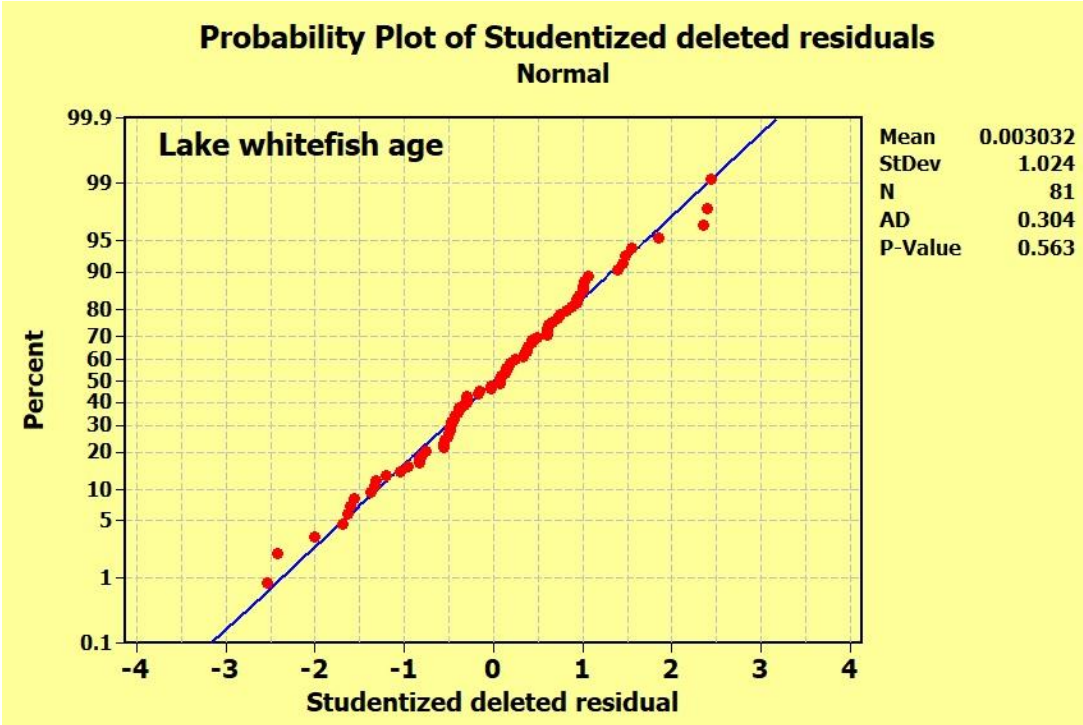

Probability Plot of Studentized deleted residuals
Normal
Lake trout weight
Mean 0.001136
StDev 1.037
N 33
AD 0.245
P-Value 0.740
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-3
-2
-1
0
1
2
3
Studentized deleted residual



For the mercury data of lake whitefish, transformations ($\log_e$, square, square root) failed to render the probability distribution normal. $\log_e$ transformation somewhat improved the distribution (AD = 3.367) in comparison with untransformed mercury data (AD = 8.580) even though it was still not significant ($p < 0.005$). For the $\log_e$ transformed mercury data, Studentized deleted residuals of all three predictor variables exhibited normal probability distributions:



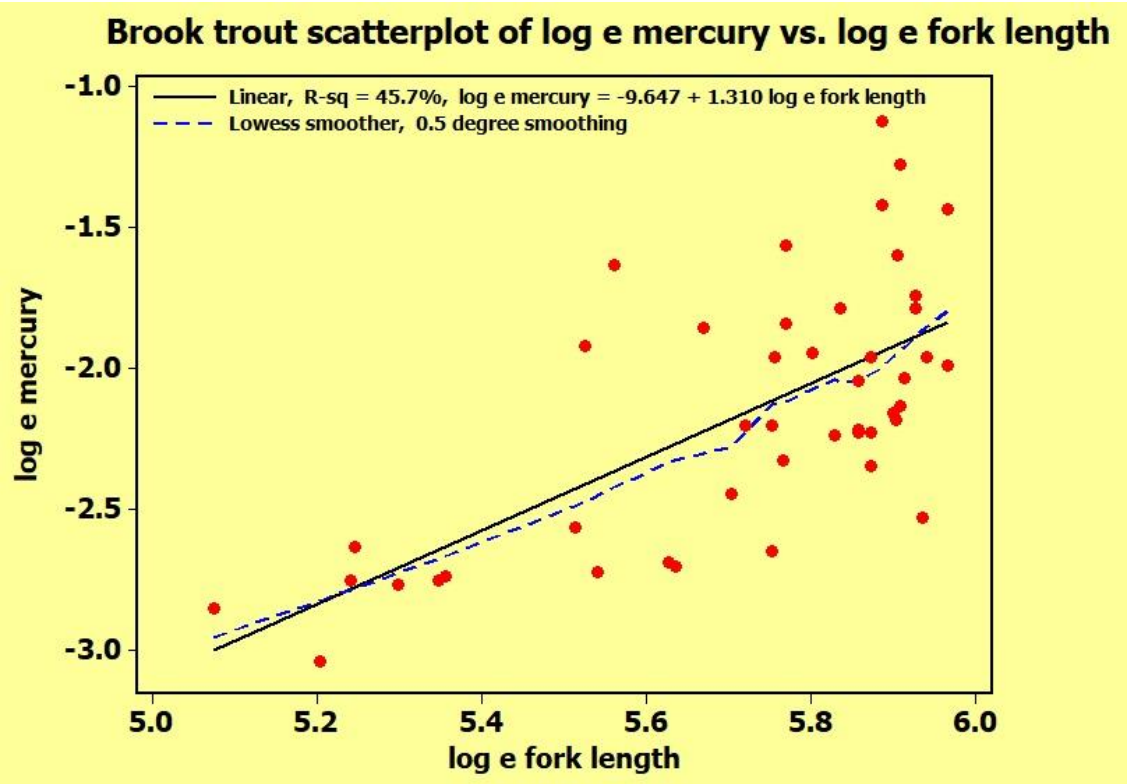

Probability Plot of Studentized deleted residuals
Normal
Lake whitefish weight
Mean 0.002677
StDev 1.020
N 81
AD 0.195
P-Value 0.889
Percent
99.9
99
95
90
80
70
60
50
40
30
20
10
5
1
0.1
-4
-3
-2
-1
0
1
2
3
4
Studentized deleted residual

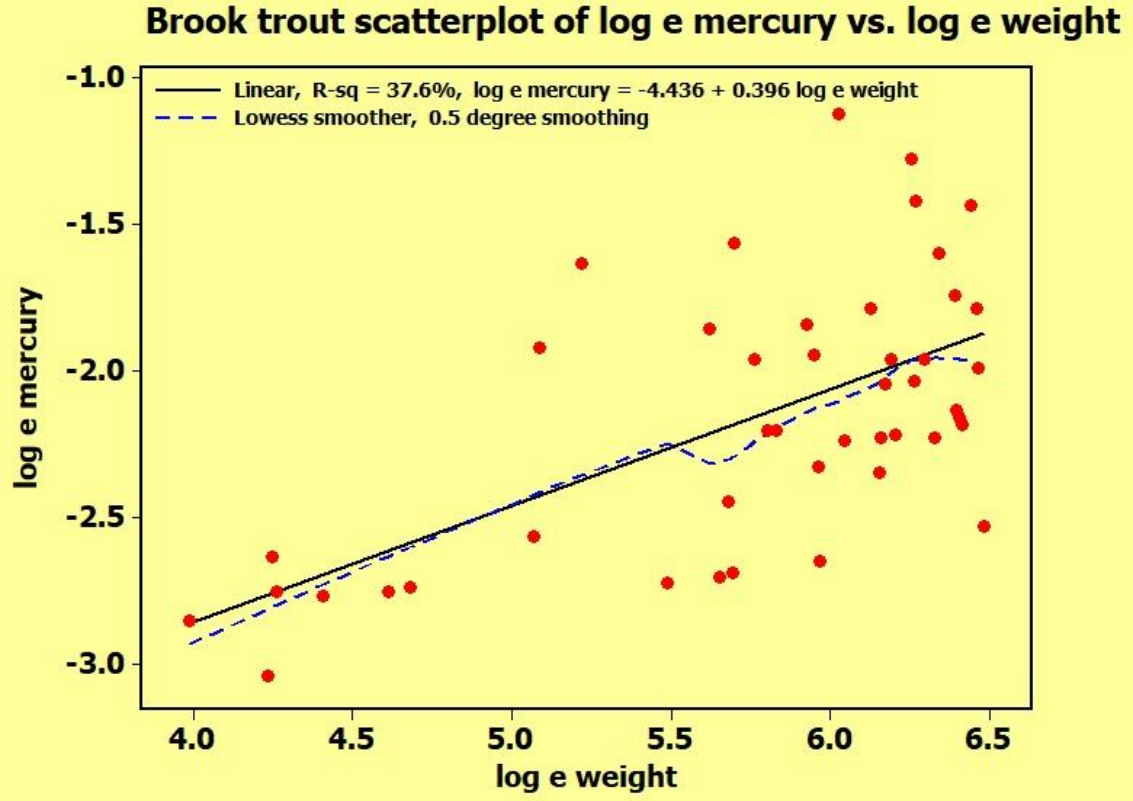

Probability Plot of Studentized deleted residuals
Normal
Lake whitefish age
Mean 0.003032
StDev 1.024
N 81
AD 0.304
P-Value 0.563
Percent
99.9
99
95
90
80
70
60
50
40
30
20
10
5
1
0.1
-4
-3
-2
-1
0
1
2
3
4
Studentized deleted residual

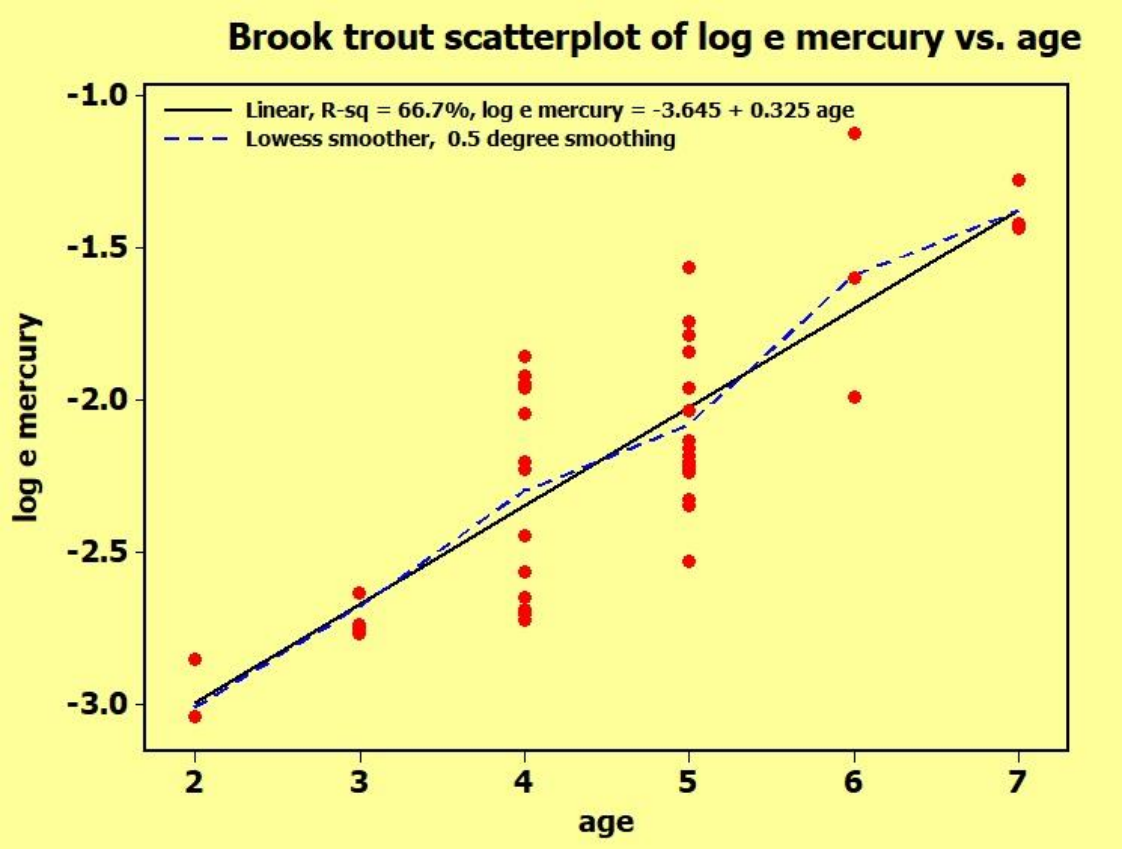
1.5 Relationships between the dependent variable and predictors

Locally-weighted scatterplot smoother (LOWESS) is a common technique for determining a smoothing line that is fitted to the data in order to explore the potential relationships between two variables, without fitting a specified model, such as a regression line or a theoretical distribution. LOWESS is a nonparametric method that combines multiple regression models in a k – nearest – neighbour – based meta model. For each data point, LOWESS performs a "weighted" linear regression, giving points closest to each x-value the greatest weight in the smoothing and limiting the influence of outliers. It is a relatively computationally intensive process by fitting simple models to localized subsets of the data to build up a function that describes the deterministic part of the variation in the data, point by point. The beauty of this method is that it is not required to specify a global function of any form to fit a model to the data, only to fit segments of the data. LOWESS was used here as a "guidance" to suggest which type of classical models (linear, quadratic, etc.) the data are most likely to be described by.

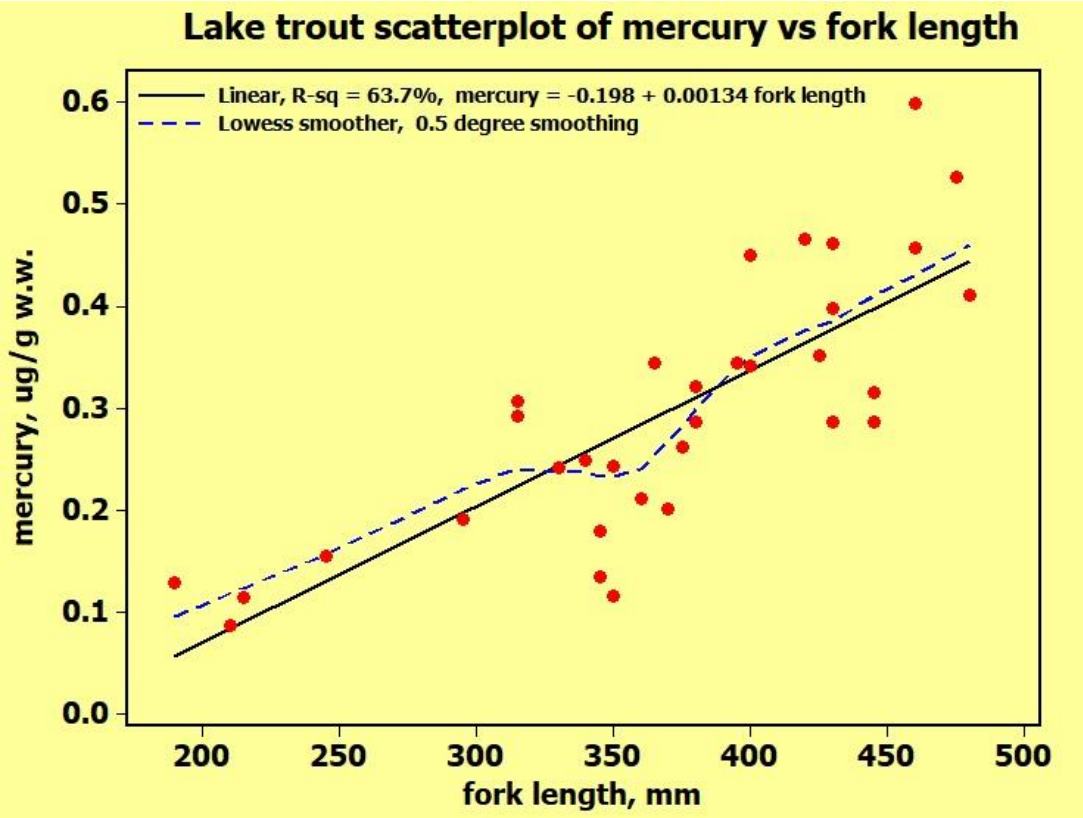
For the brook trout, the relationships between $\log_e$ mercury and $\log_e$ fork length, $\log_e$ weight and age were linear:

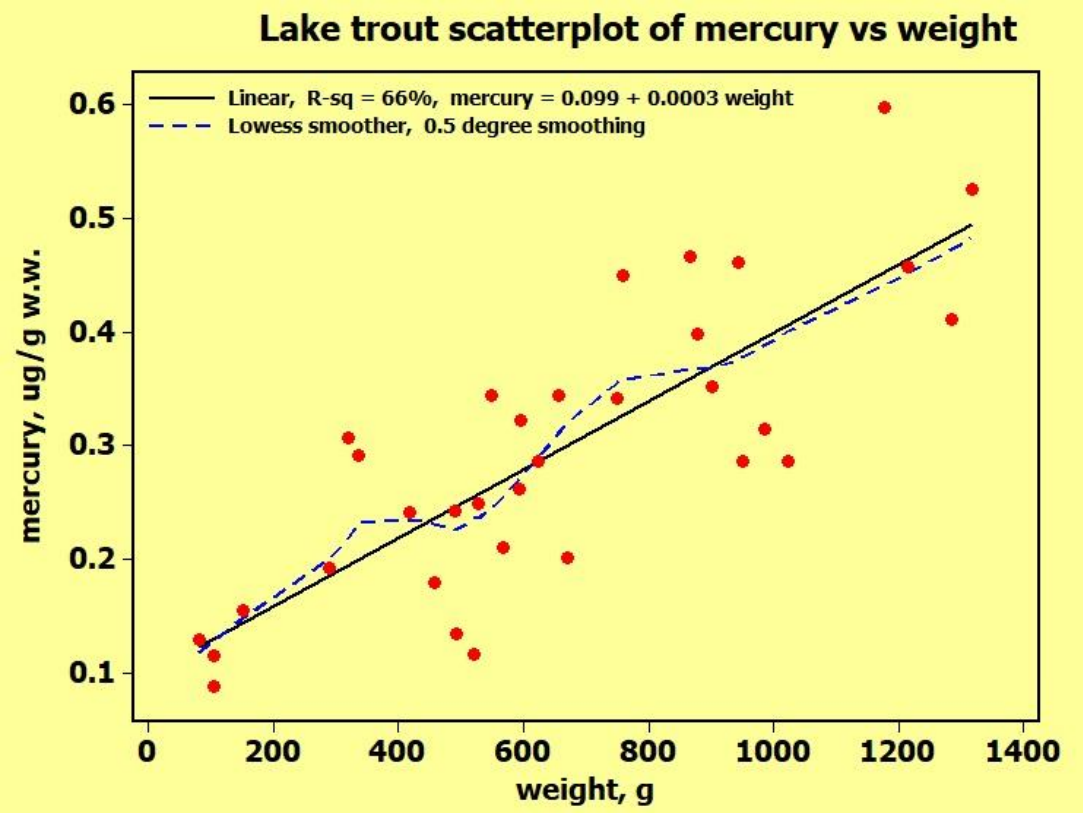


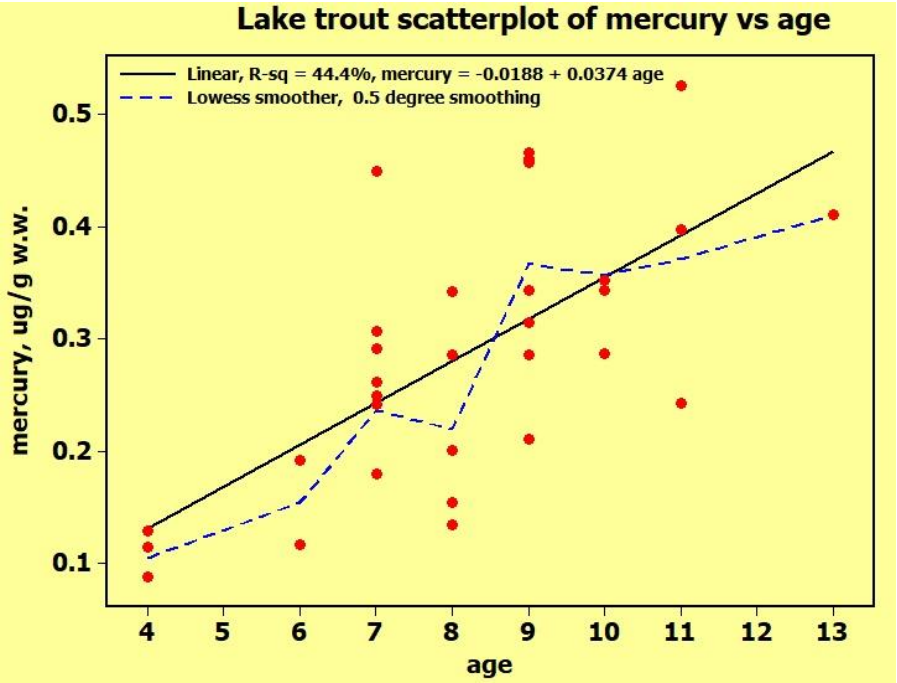

Brook trout scatterplot of log e mercury vs. log e weight
Linear, R-sq = 37.6%, log e mercury = -4.436 + 0.396 log e weight
Lowess smoother, 0.5 degree smoothing
log e mercury
-1.0
-1.5
-2.0
-2.5
-3.0
4.0
4.5
5.0
5.5
6.0
6.5
log e weight


Brook trout scatterplot of log e mercury vs. age
Linear, R-sq = 66.7%, log e mercury = -3.645 + 0.325 age
Lowess smoother, 0.5 degree smoothing
log e mercury
-1.0
-1.5
-2.0
-2.5
-3.0
2
3
4
5
6
7
age

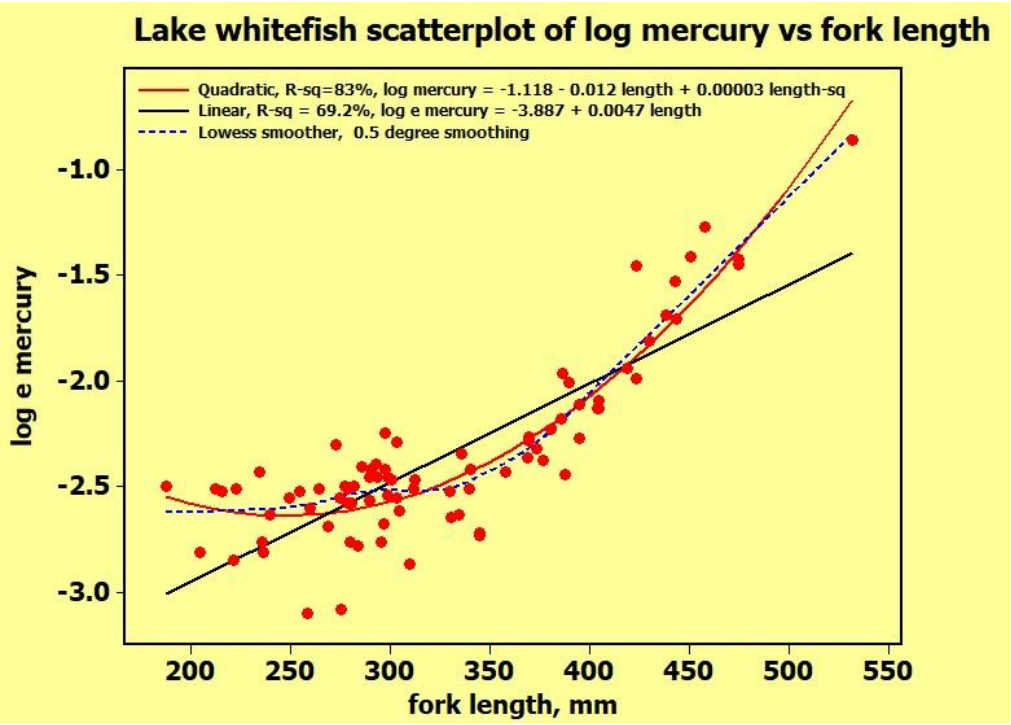
For lake trout, the relationships between mercury and fork length, weight and age were linear:

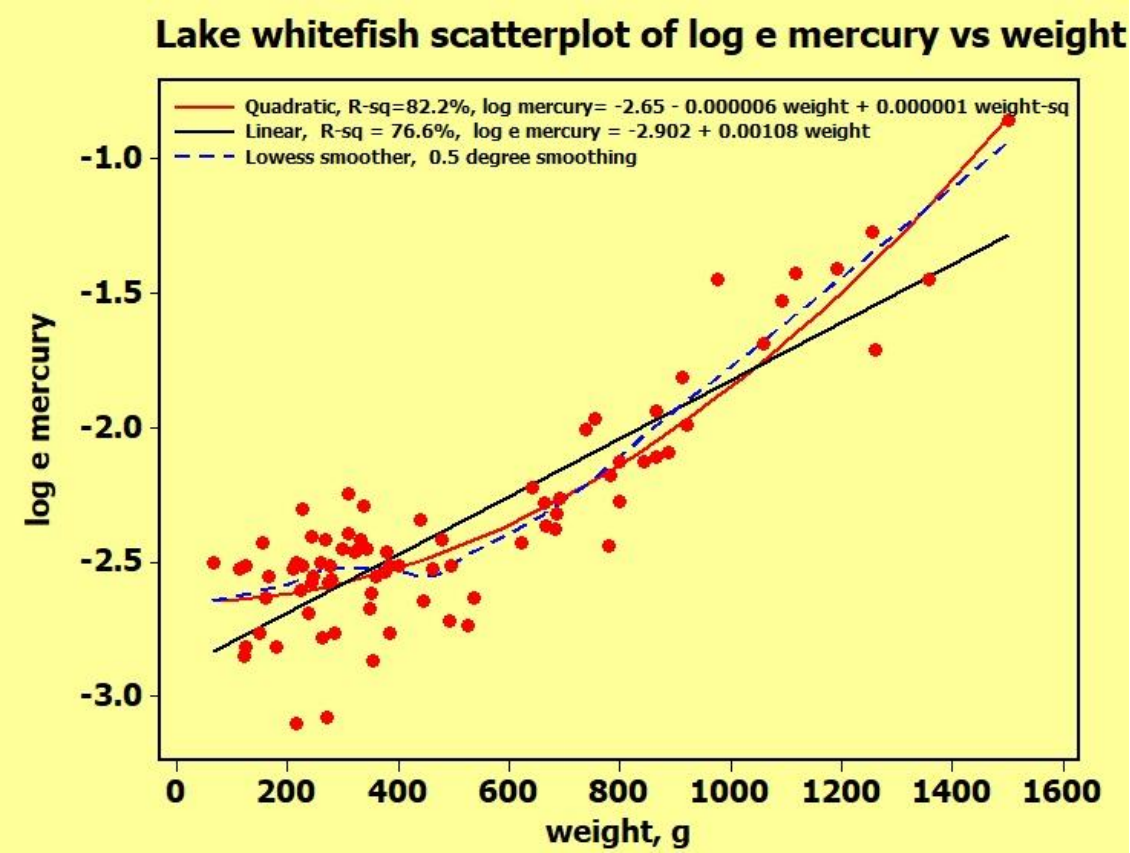


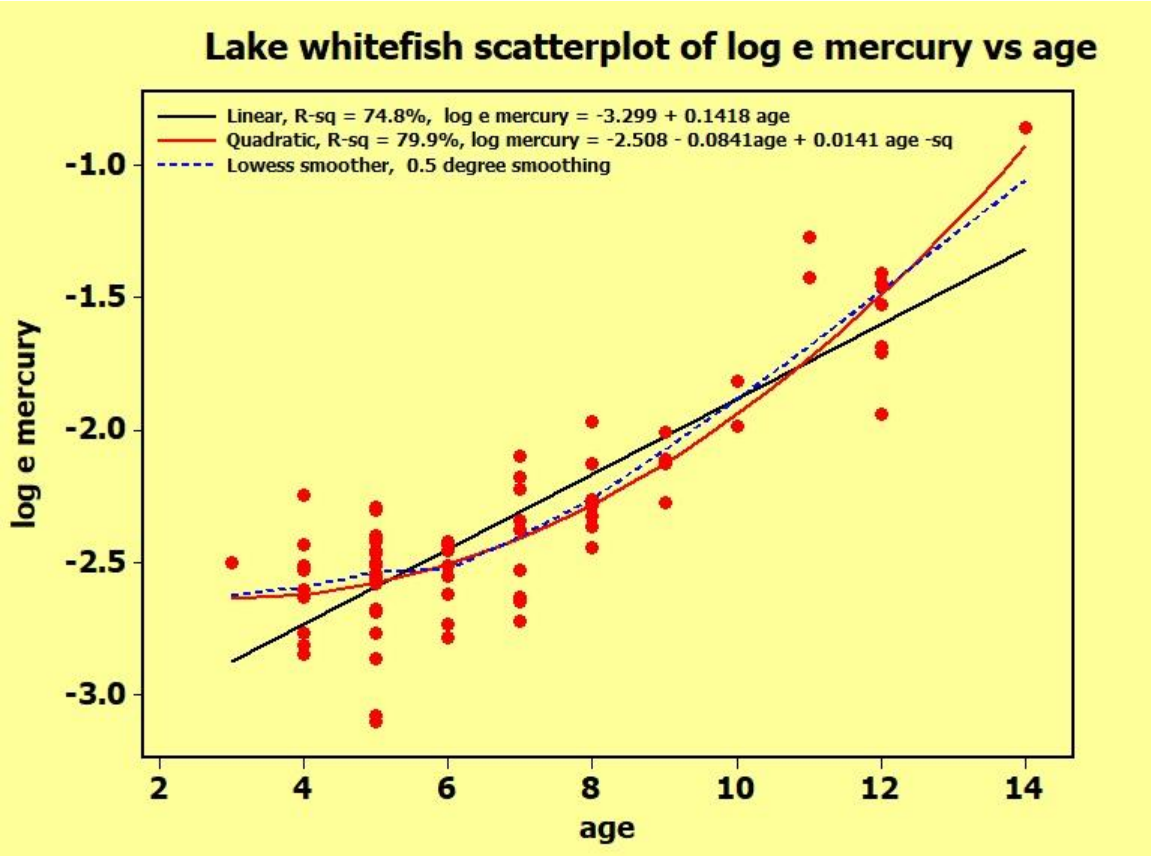


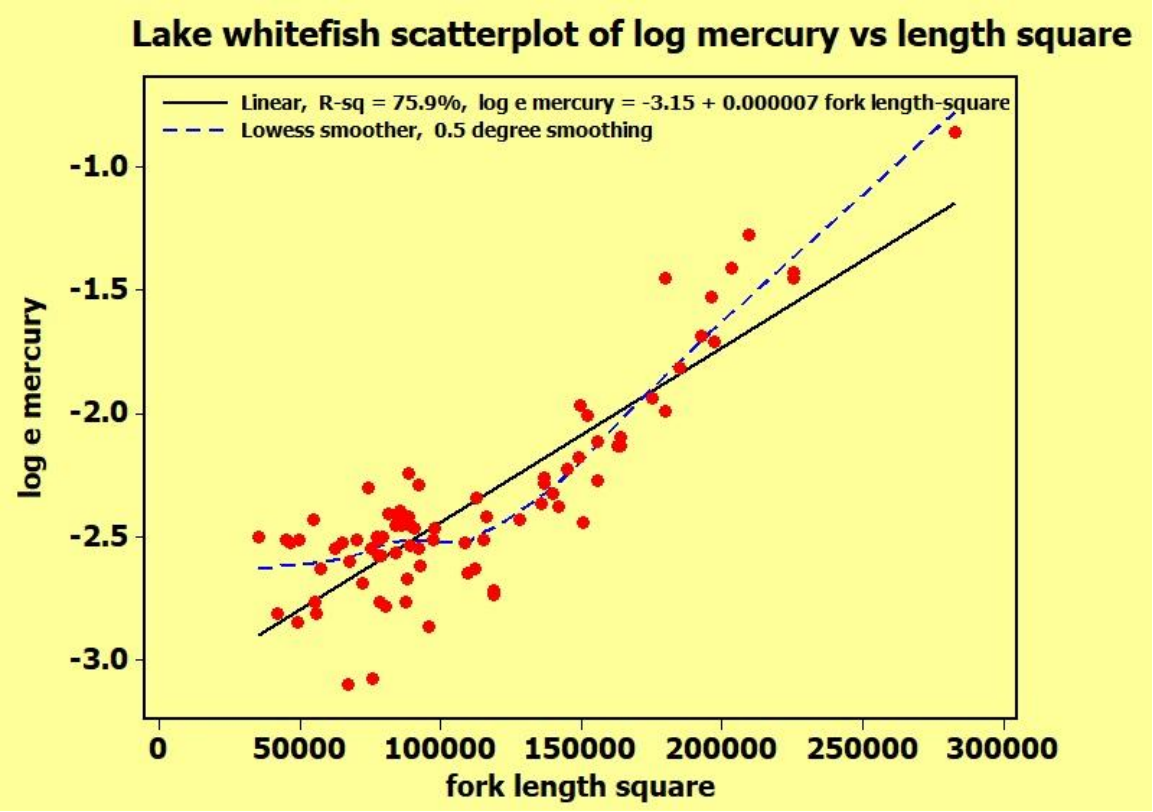


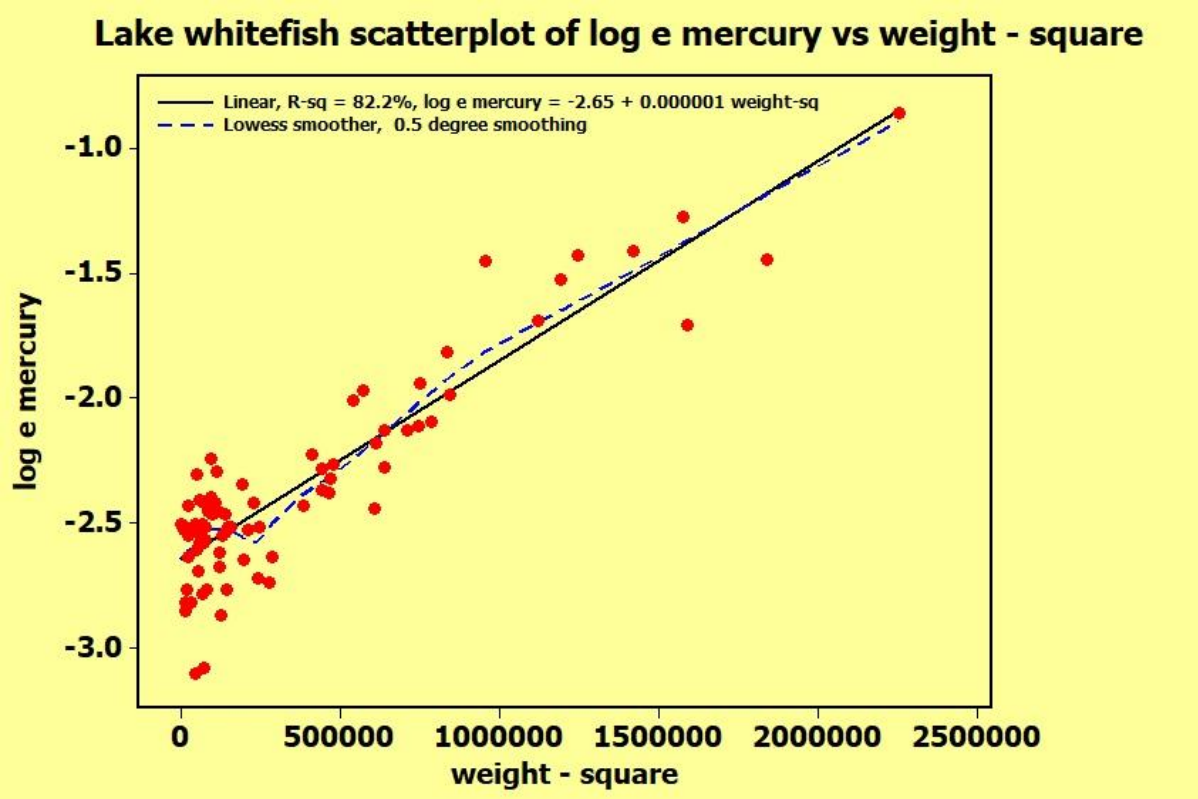
In the case of lake whitefish, the LOWESS smoothing suggested that a quadratic model fitted better than a linear model to describe the relationships between the $\log_e$ mercury and the three (untransformed) predictor variables. Square root transformation of predictor variables still yielded a curvilinear relationship. However, a square transformation greatly rendered both weight and age to approach linearity in their relationship with $\log_e$ mercury, although the relationship between $\log_e$ mercury and fork length square remained curvilinear despite the square transformation:

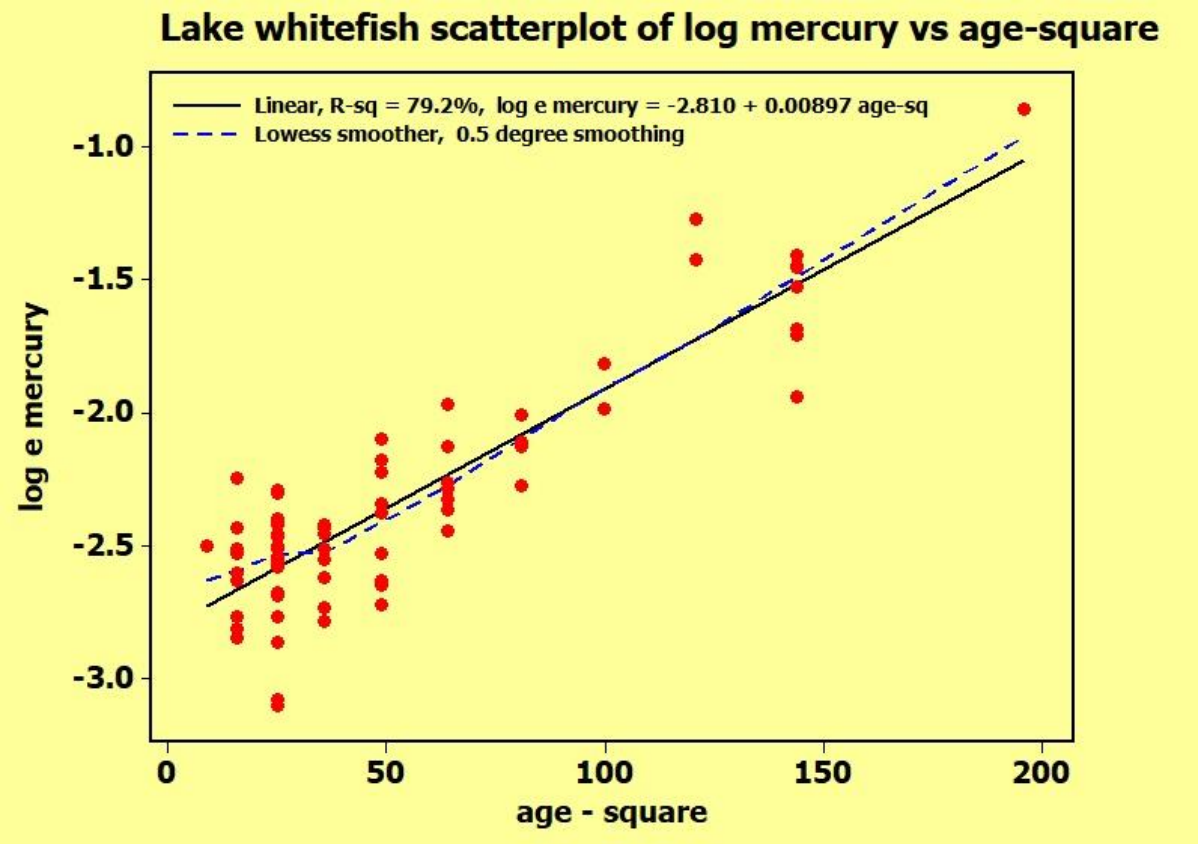



Lake whitefish scatterplot of log e mercury vs weight
Quadratic, R-sq=82.2%, log mercury= -2.65 - 0.000006 weight + 0.000001 weight-sq
Linear, R-sq = 76.6%, log e mercury = -2.902 + 0.00108 weight
Lowess smoother, 0.5 degree smoothing
log e mercury
-1.0
-1.5
-2.0
-2.5
-3.0
0
200
400
600
800
1000
1200
1400
1600
weight, g


Lake whitefish scatterplot of log e mercury vs age
Linear, R-sq = 74.8%, log e mercury = -3.299 + 0.1418 age
Quadratic, R-sq = 79.9%, log mercury = -2.508 - 0.0841age + 0.0141 age -sq
Lowess smoother, 0.5 degree smoothing
log e mercury
-1.0
-1.5
-2.0
-2.5
-3.0
2
4
6
8
10
12
14
age


Lake whitefish scatterplot of log mercury vs length square
Linear, R-sq = 75.9%, log e mercury = -3.15 + 0.000007 fork length-square
Lowess smoother, 0.5 degree smoothing
log e mercury
-1.0
-1.5
-2.0
-2.5
-3.0
0
50000
100000
150000
200000
250000
300000
fork length square


Lake whitefish scatterplot of log e mercury vs weight - square
Linear, R-sq = 82.2%, log e mercury = -2.65 + 0.000001 weight-sq
Lowess smoother, 0.5 degree smoothing
log e mercury
-1.0
-1.5
-2.0
-2.5
-3.0
0
500000
1000000
1500000
2000000
2500000
weight - square



1.6 Homoscedasticity

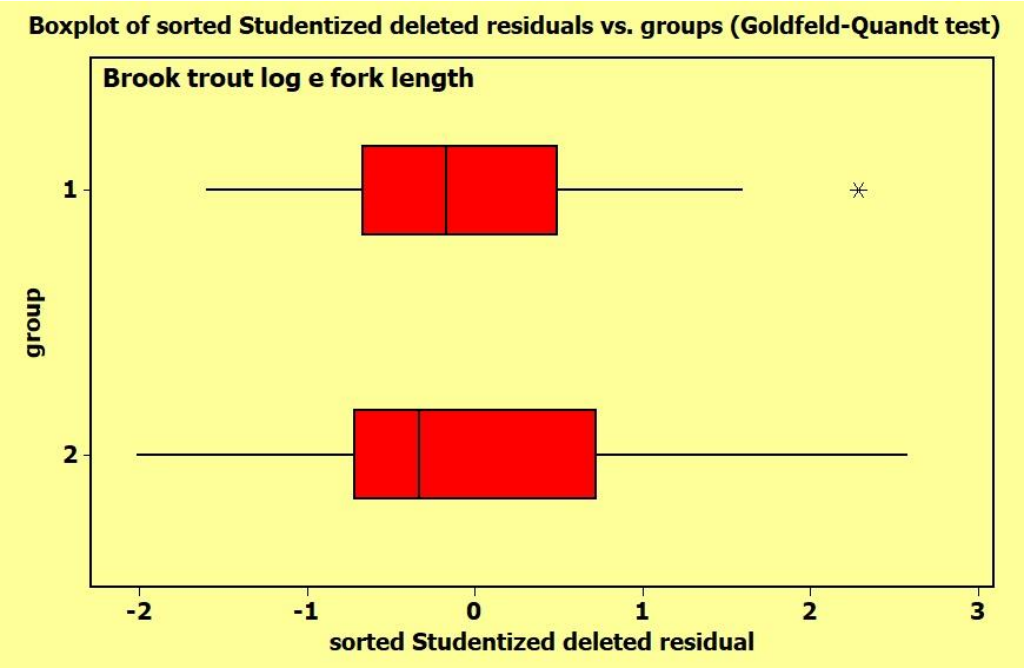
Goldfeld – Quandt test was used to check for homoscedasticity of the Studentized deleted residuals of each predictor variable. The hypothesis to be tested was that the variances of the residuals of the regression model were not constant, but instead were monotonically related to a pre-identified independent variable (Goldfeld and Quandt 1965). The test involved sorting the Studentized deleted residuals from the lowest to the highest values and then divided them into two equal subsets. The two subsets were then subjected to the test for equal variances. The test statistic used was the ratio of the mean square residual errors for the regressions on the two subsets. This test statistic corresponded to an F-test of equality of variances. The test offered a simple and intuitive diagnostic for heteroscedastic errors in a univariate or multivariate regression model. Table 1.1 shows the Levene test statistics and the associated p – values of the Goldfeld – Quandt test. Studentized deleted t residuals versus fitted y values plots as well as Studentized deleted residuals versus predictor plots were showed to visually determine suitability of a linear model, the presence of lurking (confounding) variables and heteroskedastic errors.

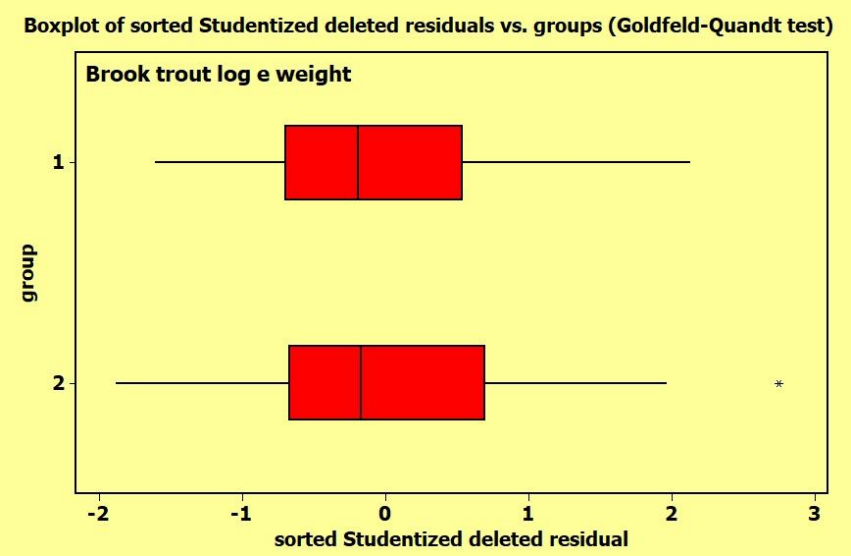
Table 1.1. Goldfeld – Quandt test results.

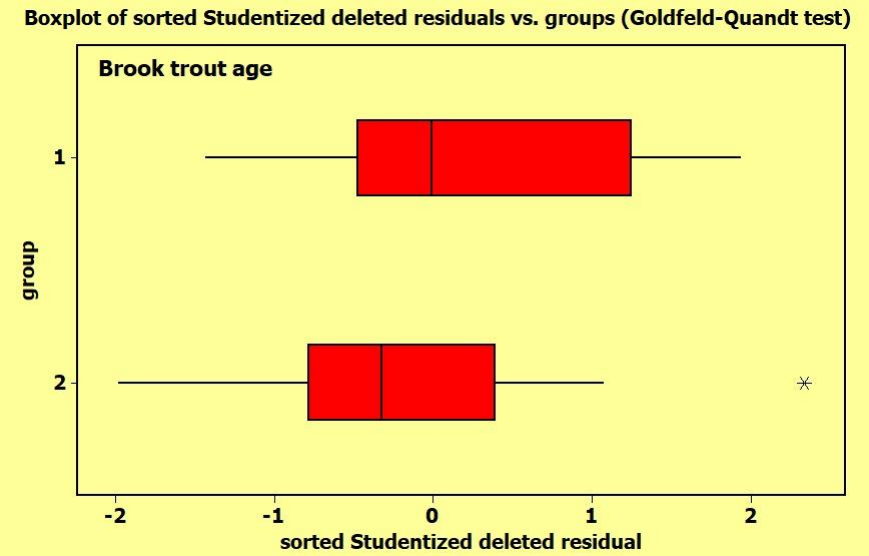
Fish species	Predictor (n)	Levene test statistic	p - value
Brook trout[1]	$\log_e$ fork length (44)	0.12	0.728
	$\log_e$ weight (48)	0.06	0.809
	age (44)	1.16	0.287
Lake trout[2]	fork length (33)	0.09	0.77
	weight (33)	1.19	0.285
	age (31)	0.47	0.499
Lake whitefish[1]	fork length (81)	1.24	0.268
	weight (81)	0.18	0.673
	age (81)	0.11	0.737

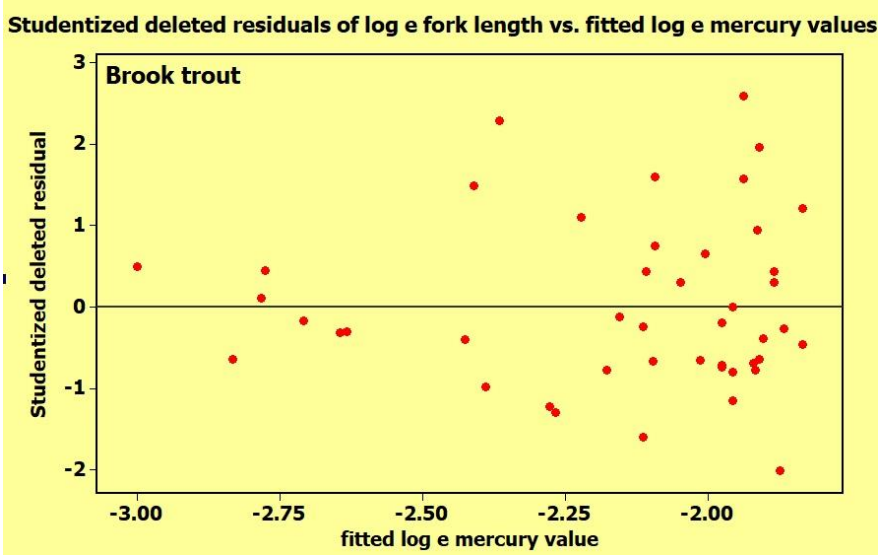
[1] $\log_e$ mercury vs. predictor

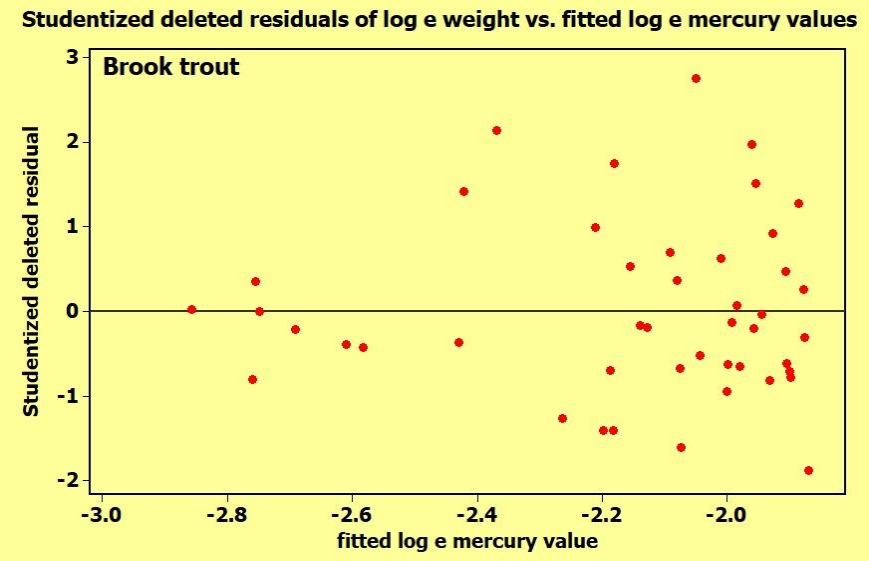
[2] mercury vs. predictor

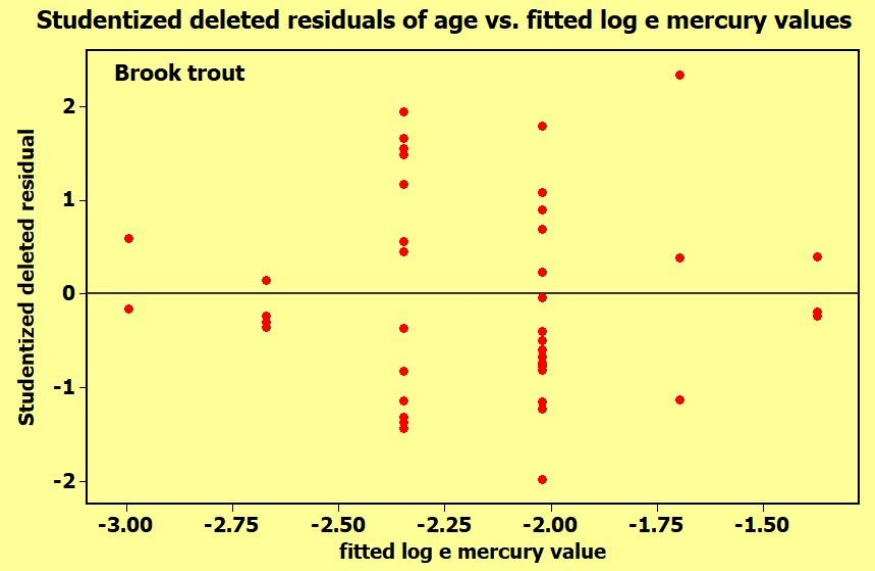


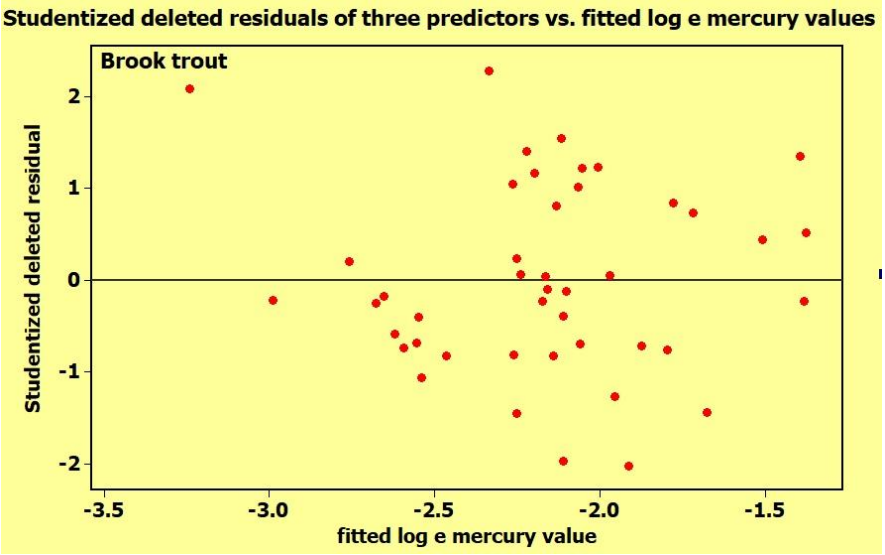

Boxplot of sorted Studentized deleted residuals vs. groups (Goldfeld-Quandt test)
Brook trout log e weight
group
1
2
-2
-1
0
1
2
3
sorted Studentized deleted residual

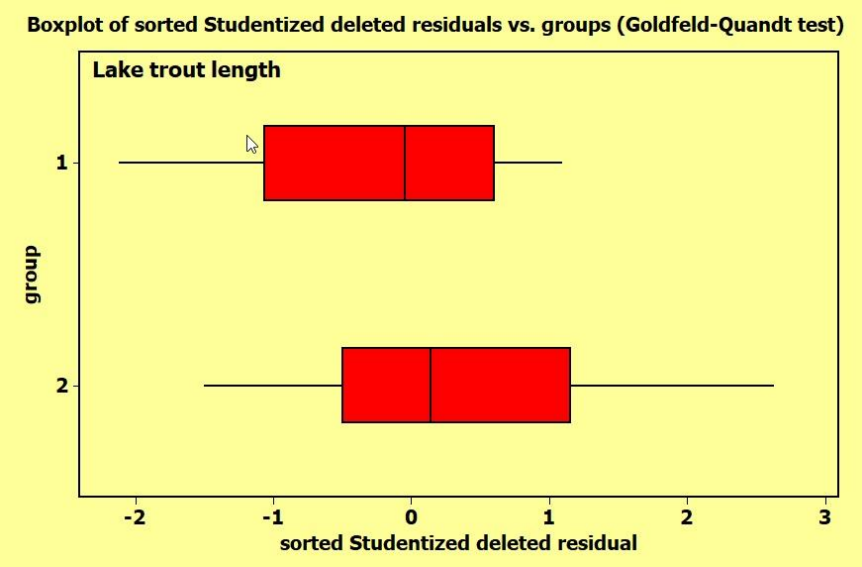

Boxplot of sorted Studentized deleted residuals vs. groups (Goldfeld-Quandt test)
Brook trout age
group
1
2
-2
-1
0
1
2
sorted Studentized deleted residual

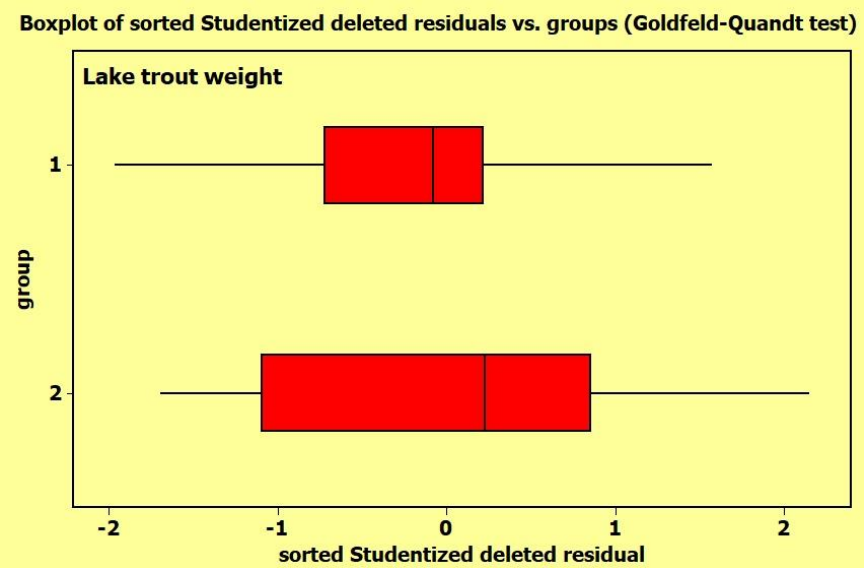

Studentized deleted residuals of log e fork length vs. fitted log e mercury values
Brook trout
Studentized deleted residual
3
2
1
0
-1
-2
-3.00
-2.75
-2.50
-2.25
-2.00
fitted log e mercury value

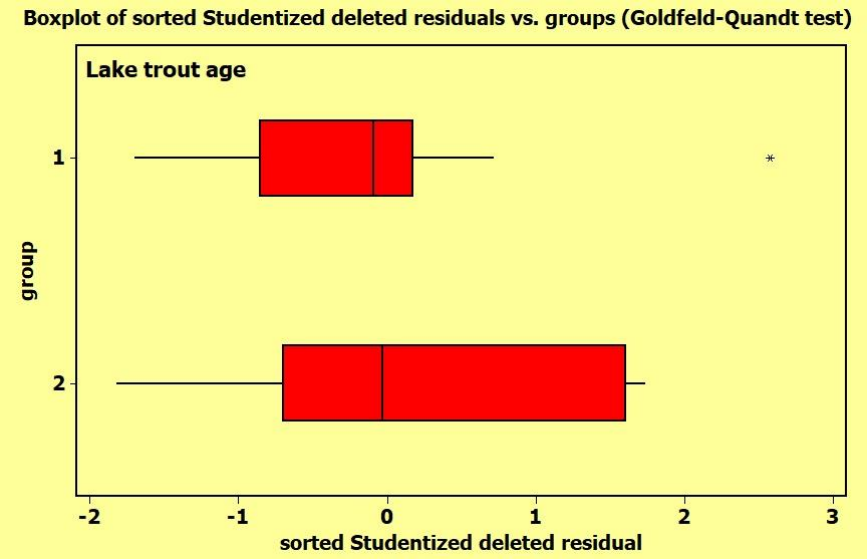

Studentized deleted residuals of log e weight vs. fitted log e mercury values
Brook trout
Studentized deleted residual
fitted log e mercury value
3
2
1
0
-1
-2
-3.0
-2.8
-2.6
-2.4
-2.2
-2.0

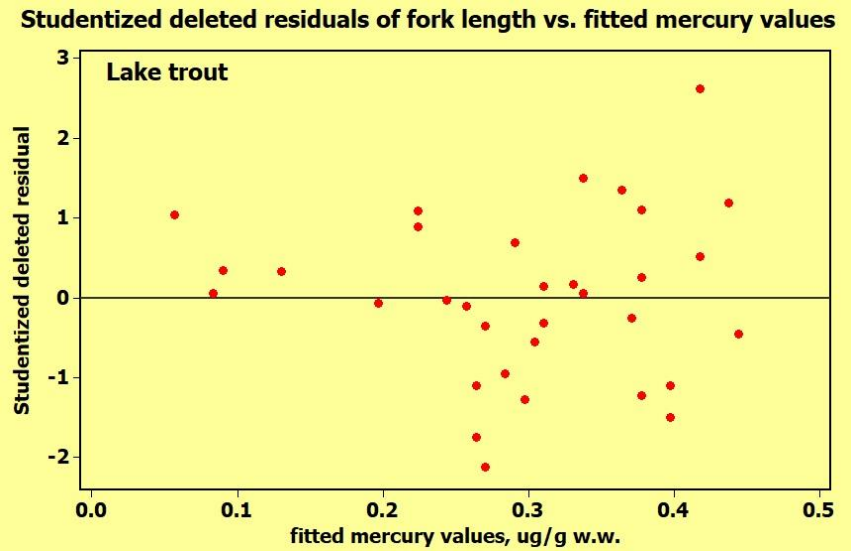

Studentized deleted residuals of age vs. fitted log e mercury values
Brook trout
Studentized deleted residual
fitted log e mercury value
2
1
0
-1
-2
-3.00
-2.75
-2.50
-2.25
-2.00
-1.75
-1.50

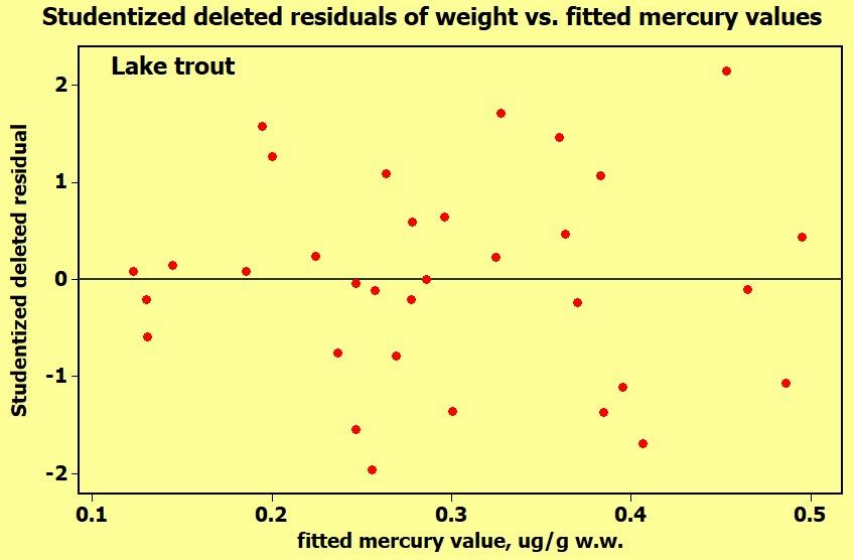

Studentized deleted residuals of three predictors vs. fitted log e mercury values
Brook trout
Studentized deleted residual
fitted log e mercury value
2
1
0
-1
-2
-3.5
-3.0
-2.5
-2.0
-1.5

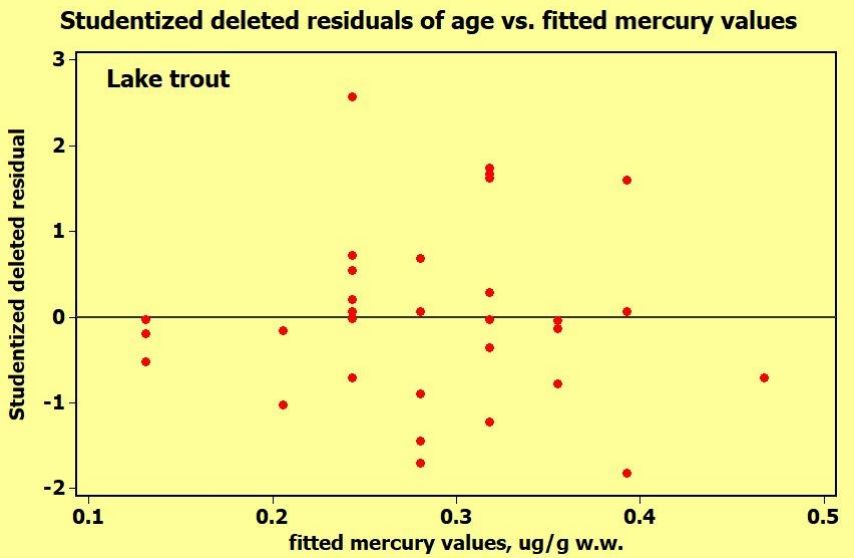

Boxplot of sorted Studentized deleted residuals vs. groups (Goldfeld-Quandt test)
Lake trout length
group
1
2
-2
-1
0
1
2
3
sorted Studentized deleted residual

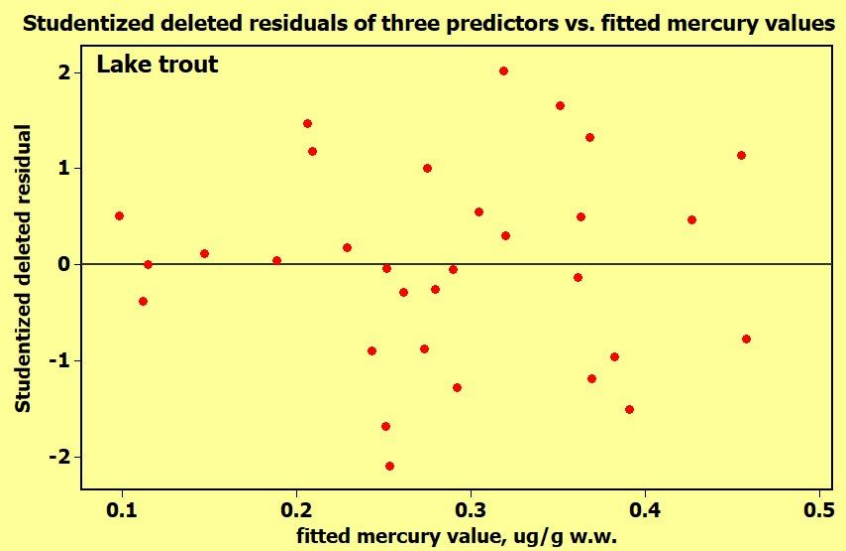

Boxplot of sorted Studentized deleted residuals vs. groups (Goldfeld-Quandt test)
Lake trout weight
group
1
2
-2
-1
0
1
2
sorted Studentized deleted residual


Boxplot of sorted Studentized deleted residuals vs. groups (Goldfeld-Quandt test)
Lake trout age
group
1
2
-2
-1
0
1
2
3
sorted Studentized deleted residual

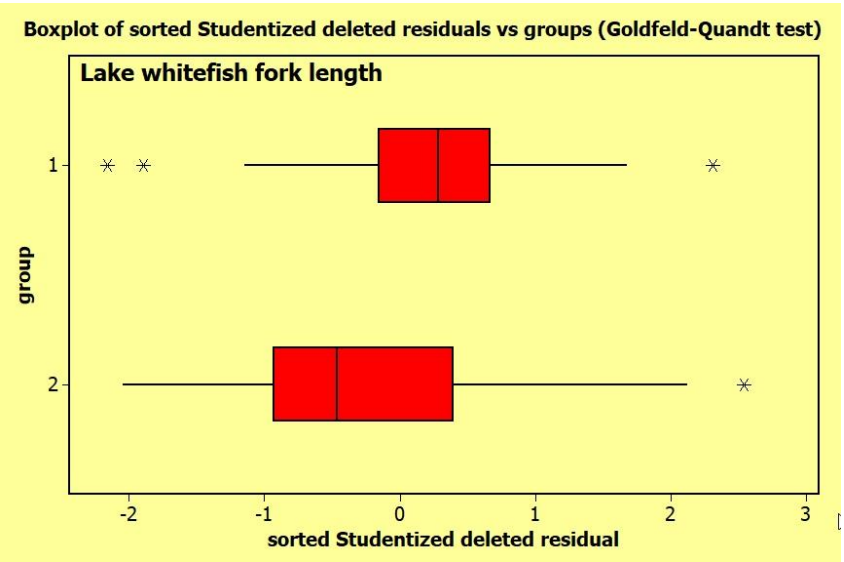

Studentized deleted residuals of fork length vs. fitted mercury values
Lake trout
Studentized deleted residual
fitted mercury values, ug/g w.w.
3
2
1
0
-1
-2
0.0
0.1
0.2
0.3
0.4
0.5

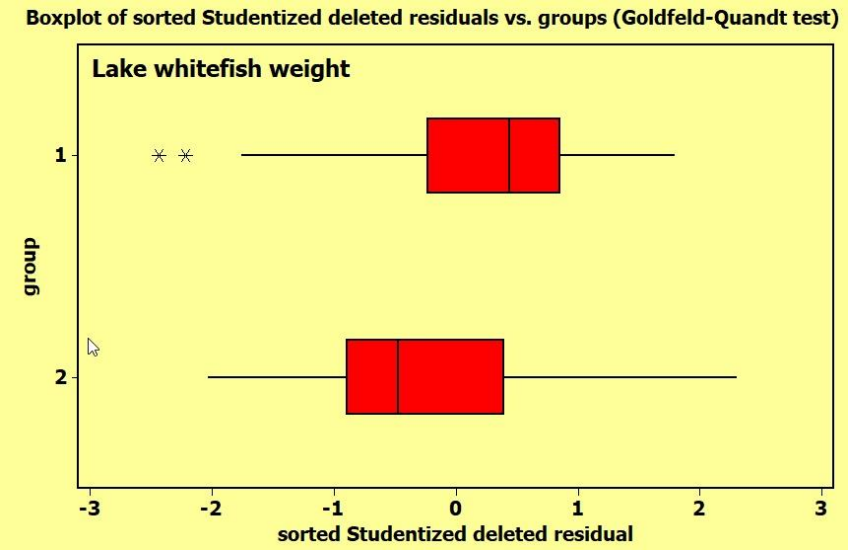

Studentized deleted residuals of weight vs. fitted mercury values
Lake trout
Studentized deleted residual
fitted mercury value, ug/g w.w.
2
1
0
-1
-2
0.1
0.2
0.3
0.4
0.5

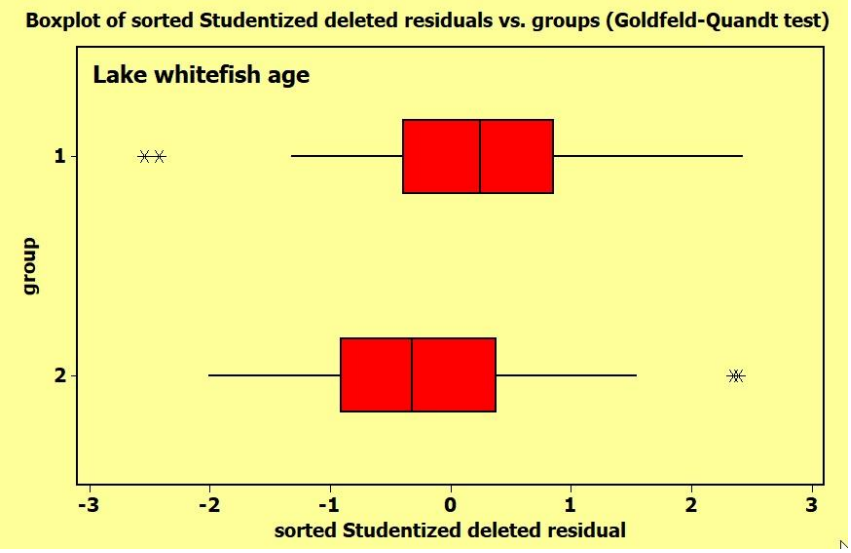

Studentized deleted residuals of age vs. fitted mercury values
Lake trout
Studentized deleted residual
fitted mercury values, ug/g w.w.
3
2
1
0
-1
-2
0.1
0.2
0.3
0.4
0.5

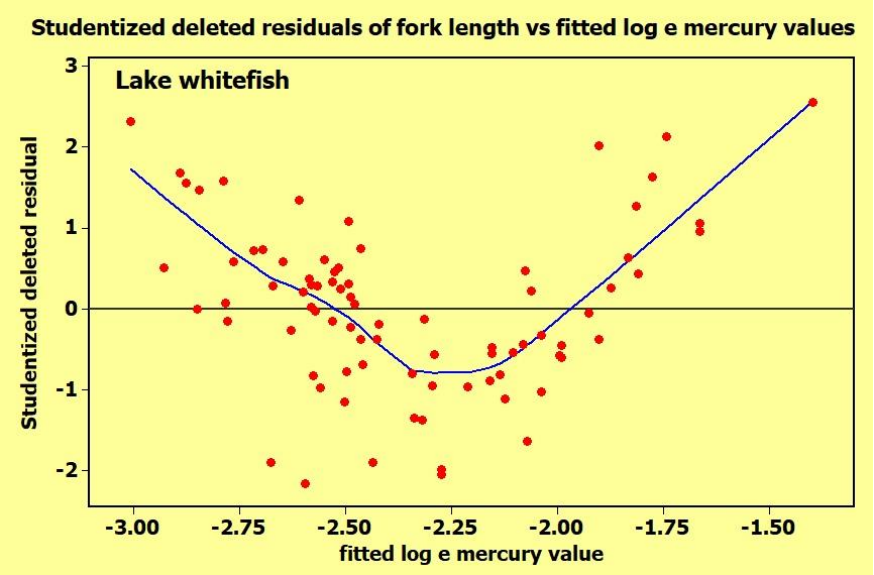


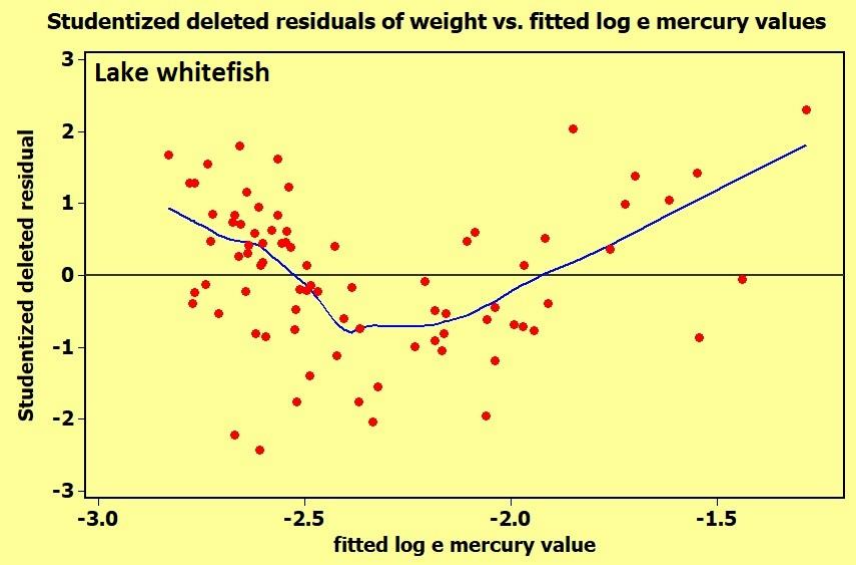
One weakness of the Goldfeld – Quandt test is that it is not very robust to model specification errors (Thursby 1982). In the present case, no model specification error was evident for brook trout and lake trout from examining their scatterplots and their Studentized deleted residuals vs. fits plots. For both species, the relationships between the dependent variable and all three predictors can be described by a linear model; hence, the results of the Goldfeld – Quandt test were valid. However, this is not the case for lake whitefish, both scatterplots of dependent variable versus predictors and the residuals versus fits plots indicated that a simple linear model was inadequate to satisfactorily describe the relationship between the dependent variable and the predictors. Under such circumstance, Goldfeld – Quandt test cannot distinguish between heteroskedastic error structure and an underlying model specification problem such as an incorrect functional form or an omitted (lurking) independent variable. The Goldfeld – Quandt test results of the lake whitefish have to be treated with cautions and might not be interpreted correctly.

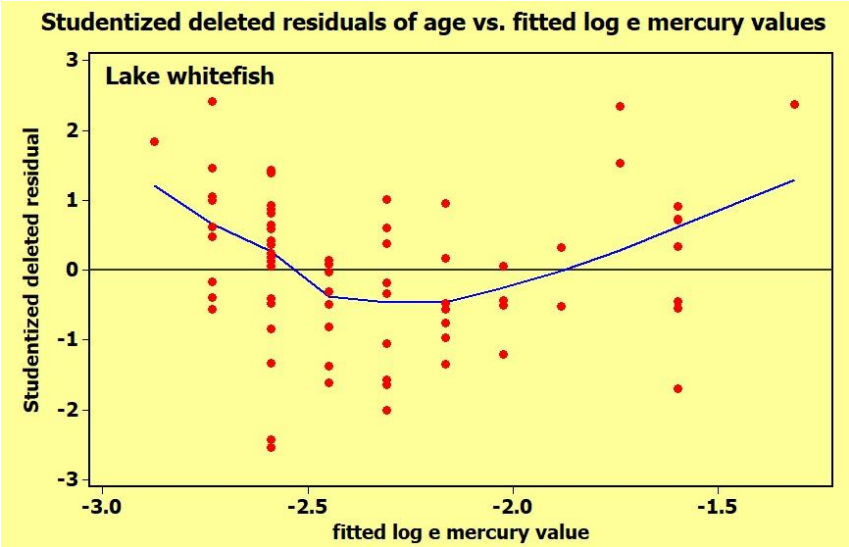


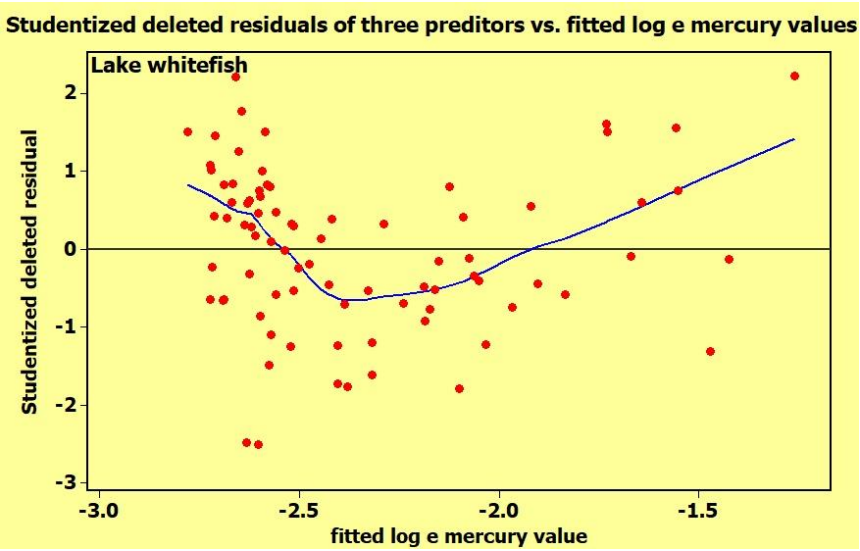

Boxplot of sorted Studentized deleted residuals vs. groups (Goldfeld-Quandt test)
Lake whitefish weight
group
1
2
-3
-2
-1
0
1
2
3
sorted Studentized deleted residual


Boxplot of sorted Studentized deleted residuals vs. groups (Goldfeld-Quandt test)
Lake whitefish age
group
1
2
-3
-2
-1
0
1
2
3
sorted Studentized deleted residual


Studentized deleted residuals of fork length vs fitted log e mercury values
Lake whitefish
Studentized deleted residual
3
2
1
0
-1
-2
-3.00
-2.75
-2.50
-2.25
-2.00
-1.75
-1.50
fitted log e mercury value


Studentized deleted residuals of weight vs. fitted log e mercury values
Lake whitefish
Studentized deleted residual
fitted log e mercury value
3
2
1
0
-1
-2
-3
-3.0
-2.5
-2.0
-1.5


Studentized deleted residuals of age vs. fitted log e mercury values
Lake whitefish
Studentized deleted residual
fitted log e mercury value
3
2
1
0
-1
-2
-3
-3.0
-2.5
-2.0
-1.5


Studentized deleted residuals of three preditors vs. fitted log e mercury values
Lake whitefish
Studentized deleted residual
fitted log e mercury value
2
1
0
-1
-2
-3
-3.0
-2.5
-2.0
-1.5

1.7 Independence and randomness of residuals

Wald – Wolfowitz runs test was used to test for the randomness of the Studentized deleted t residuals. It tested the hypothesis that the elements of a two-valued data sequence were mutually independent (Magel and Wibowo 1997). For both brook trout and lake trout, the runs test yielded a high p-values ($p > 0.05$) for the residuals of all three growth parameters which means that the null hypothesis that the residuals increase or decrease in value randomly at the 0.05 level of significance was accepted (Table 1.2). However, the very low p – values ($p < 0.05$) for the residuals of the three predictors of lake whitefish indicated that the data sequence of the sorted residuals was not mutually independent, hence, not random.

Table 1.2. Results of Wald – Wolfowitz runs test and Durbin – Watson test.

Fish species	Predictor (n)	Wald - Wolfowitz Runs test for randomness, p-value	Durbin – Watson test statistic (DW) for independence
Brook trout[1]	$\log_e$ fork length (44)	0.285	1.995
	$\log_e$ weight (48)	0.285	1.853
	age (44)	0.661	2.068
Lake trout[2]	fork length (33)	0.217	2.323
	weight (33)	0.864	2.369
	age (31)	0.680	2.751
Lake whitefish[1]	fork length (81)	0.001	1.504
	weight (81)	0.019	1.401
	age (81)	0.001	1.472

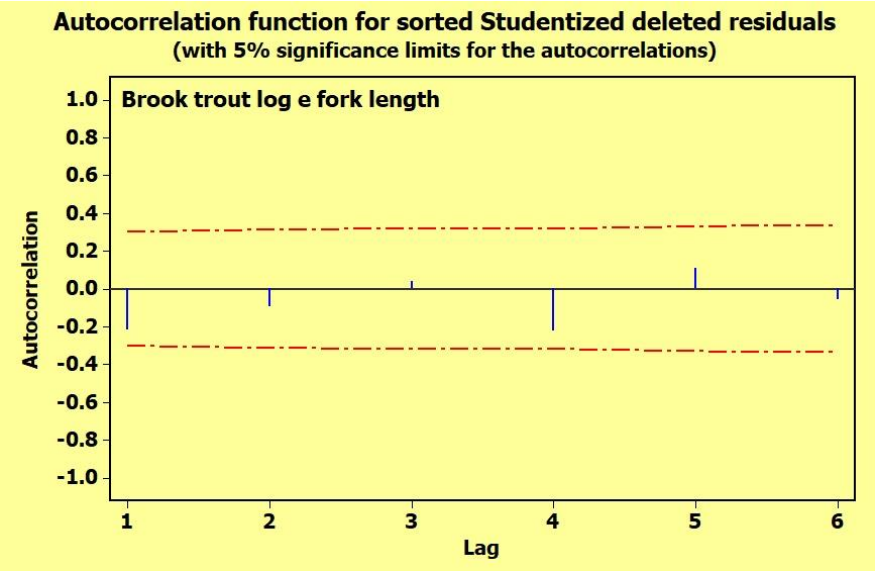
[1] $\log_e$ mercury vs. predictor

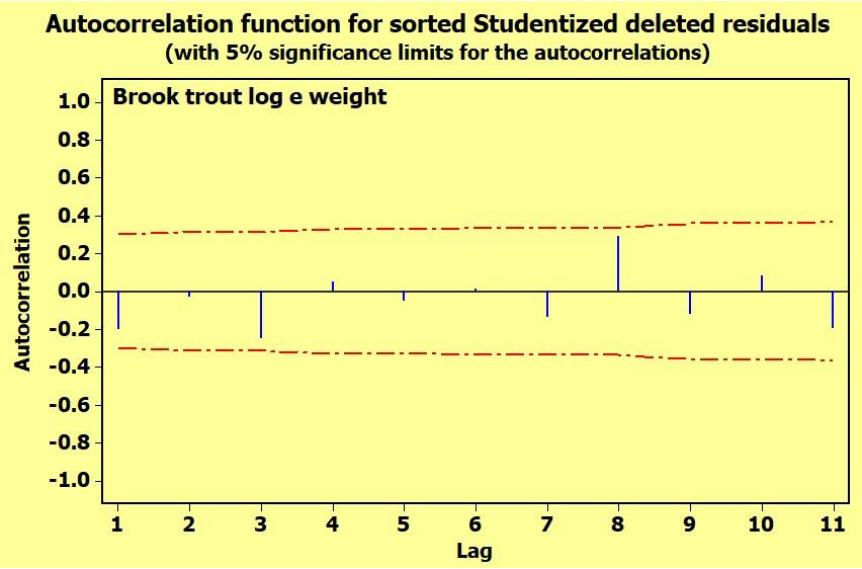
2 mercury vs. predictor

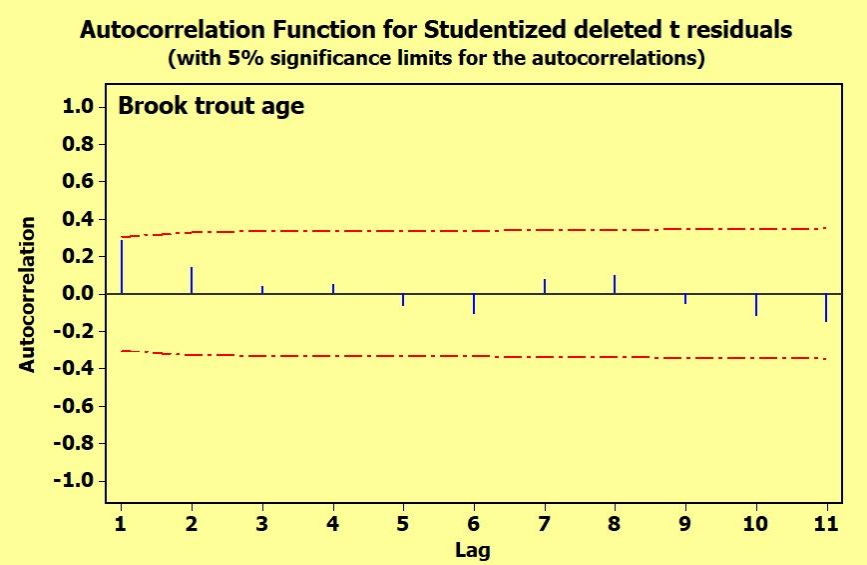
Independence of residuals is an important assumption to validate the linear regression based on the ordinary least squares (OLS) model which assumes that residuals are not correlated with one another. If for example, there are positive correlations between residuals, this tends to inflated the t – values for regression coefficient, rendering the predictor variables appear significant when they may not be (i.e. inflating Type I Error), a phenomenon called “autocorrelation”, which means that adjacent observations are dependent on one another, hence they are not random. To test for the independence

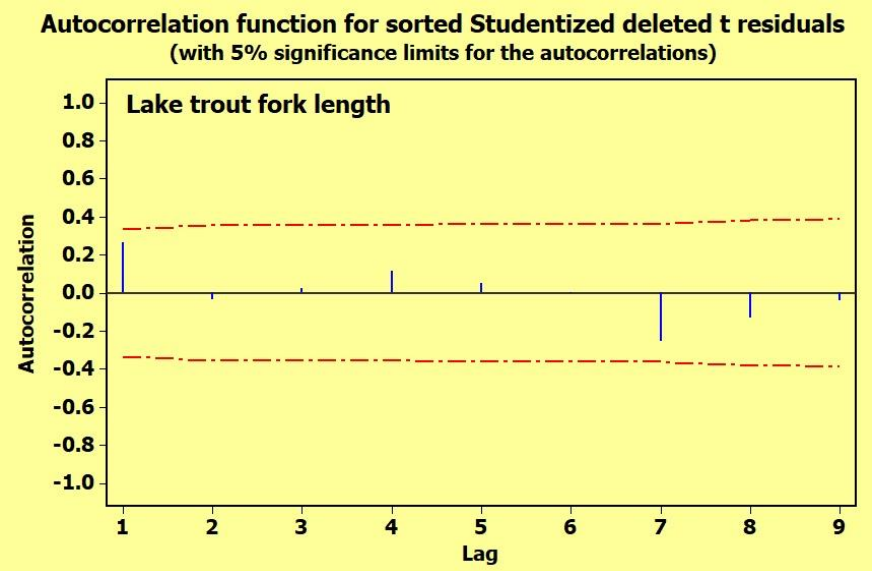
of residuals we tested the sorted Studentized deleted residuals with an autocorrelation plot as well as with the Durbin – Watson test statistic on first- order autocorrelation.

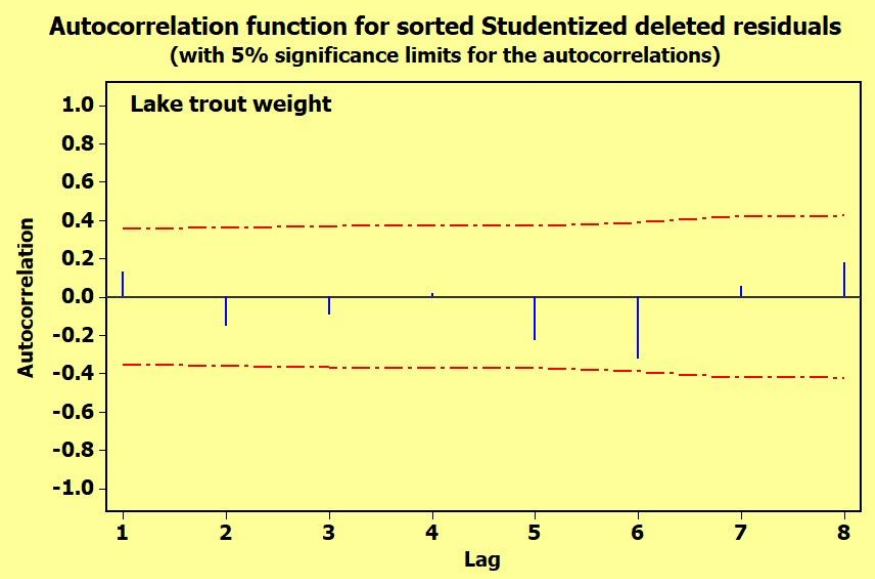
For brook trout and lake trout, autocorrelation plots clearly showed that all the autocorrelations were within the two standard error confidence bound. Hence the null hypothesis that there was no autocorrelation at and beyond a given lag was accepted at the 0.05 level of significance. In other words, the autocorrelation plot confirmed that the mercury data were random and independent of one another. This was supported by the high value of Durbin – Watson test statistic for which the null hypothesis that there was no first-order autocorrelation (autocorrelation at lag 1) in the residuals was accepted at the 0.05 level of significance. The residuals for lake whitefish failed both the Durbin – Watson test and autocorrelation plot which indicated that residuals were correlated with one another.

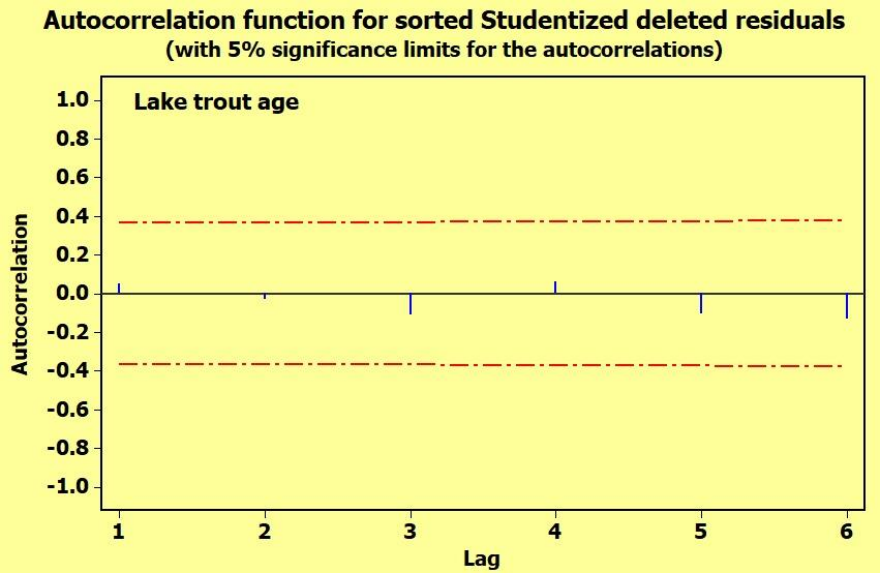


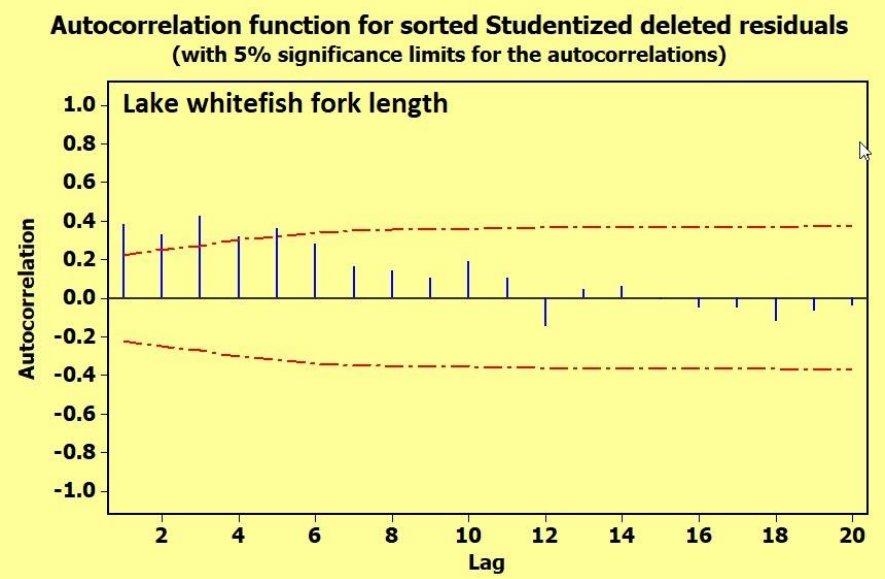


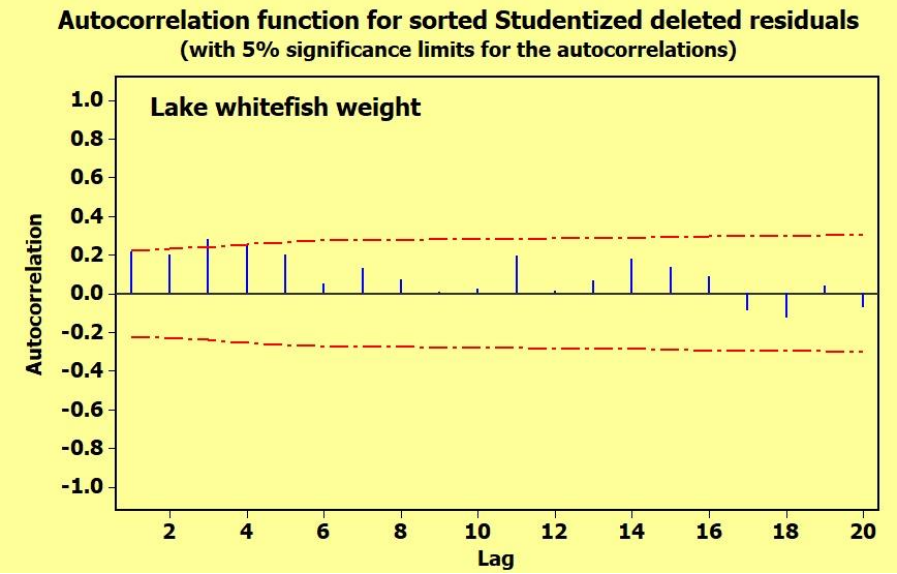

Autocorrelation Function for Studentized deleted t residuals
(with 5% significance limits for the autocorrelations)
Brook trout age
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
2
3
4
5
6
7
8
9
10
11
Lag

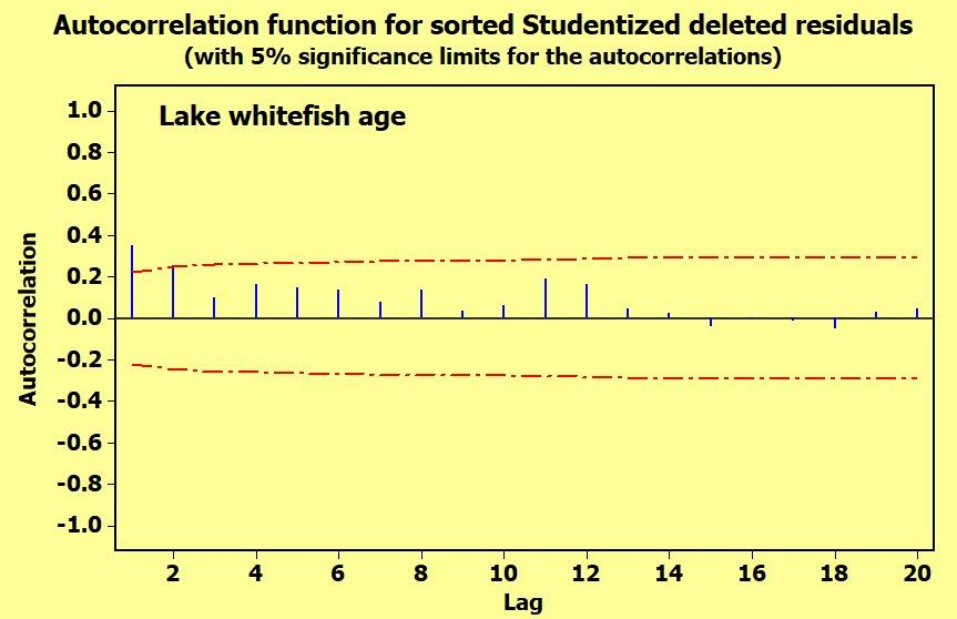

Autocorrelation function for sorted Studentized deleted t residuals
(with 5% significance limits for the autocorrelations)
Lake trout fork length
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
2
3
4
5
6
7
8
9
Lag


Autocorrelation function for sorted Studentized deleted residuals
(with 5% significance limits for the autocorrelations)
Lake trout weight
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
2
3
4
5
6
7
8
Lag


Autocorrelation function for sorted Studentized deleted residuals
(with 5% significance limits for the autocorrelations)
Lake trout age
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
2
3
4
5
6
Lag


Autocorrelation function for sorted Studentized deleted residuals
(with 5% significance limits for the autocorrelations)
Lake whitefish fork length
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag


Autocorrelation function for sorted Studentized deleted residuals
(with 5% significance limits for the autocorrelations)
Lake whitefish weight
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
2
4
6
8
10
12
14
16
18
20
Lag



Considering the curvilinear relationships between the dependent variable and the predictors as well as the V – shaped patterns of the residual plots, the failure of randomness and independence of the residuals for lake whitefish might have been related to one or more possibilities: (1) model specification inadequacy; (2) the presence of lurking (confounding) variable(s) that have significant effects on the dependent variable that were absent in the model; (3) data problems such as that the samples (hence data) were not collected in a random manner or in the same time period. Since the fish were supposed to have been collected in one summer from a number of stations in the Inukjuak River watershed, the data should not have a temporal (time series) nature, autoregressive modeling techniques which often used in analyzing time series data cannot be applied in the present situation with the lake whitefish data. However, instead of utilizing the conventional OLS – based regression analysis which bounded by many pre-requisite assumptions on the data; it is possible to model the lake whitefish data using the partial least squares (PLS) approach which focuses on the data structures of predictors by decomposing the original variables and projecting them into latent structures in new planes before further analyzing them with linear regression. PLS does not restricted by those assumptions that are pre-requisite for OLS - based models as long as dependent variables and predictors are positively correlated in an approximately linear manner. PLS can be used to shed some light on the relationship between mercury and the fish growth parameters for the lake whitefish data set and provide a plausible predictive model. An entirely different approach to deal with the lake whitefish data set is to establish a quadratic polynomial model to describe the relationship between mercury and the three predictors. The final

polynomial regression equation is useful in prediction of mercury in the fish; however, any attempt to interpret the regression coefficients amongst predictors is no longer possible.

1.8 Weak Exogeneity

This essentially means that the predictor variables can be treated as fixed values, rather than random variables. This means, for example that the predictor variables are assumed to be error – free (i.e. not contaminated with measurement errors.). Although this assumption is not realistic in many settings, dropping it leads to the significantly more difficult errors – in – variables models such as Deming regression model (Model II). For linear regressions, if the predictor variables are not error – free, this will lead to underestimation of the regression coefficients known as attenuation bias. In the case of nonlinear regressions, the direction of the bias can be very complicated. In the present context, we have to assume that fork length, weight and age were measured without error in order to proceed with the regression modeling approaches we used in this monograph. Indeed, the determination of the fork length and weight involved simple straightforward measurement which yielded definitive answers. Whereas the determination of fish age using scales or otoliths carried out by a highly experienced technician yielded unambiguous results as well. By and large we can safely assume that the determination of these three predictors were error – free.

1.9 Confounding or lurking variables, the “unknown” in observational situations.

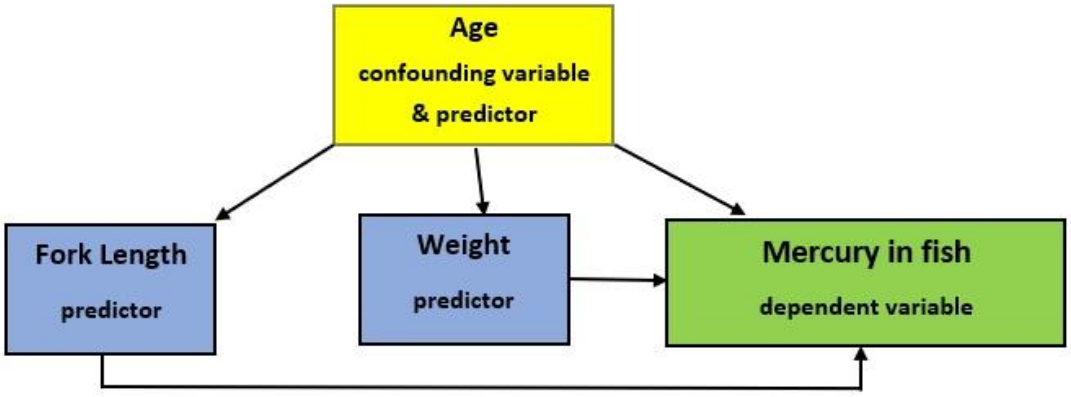
Reality is complex, there are myriad of variables interacting and influencing each other directly and indirectly, some of these variables are relatively obvious and are known to us, but many might be lurking in the background unbeknown to us. In “experimental” situations, we can meticulously plan and design our experiments hoping to control “all” variables known to influence our experimental outcome. The influence of lurking or confounding variables might well be kept in check to good extents under these well – controlled experimental situations. In the “observational” situations such as collecting samples from the real world for analyses of parameters of some sort, the researcher can never be sure that there are not other predictor variables relating and influencing the outcome of the dependent variable in the population sampled. These variables may be unknown to the researcher, difficult to measure, or thought to be irrelevant.

Known and suspected confounding variables can often be controlled and accounted for using various statistical techniques such as analysis of covariance, partial correlation, hierarchical multiple regression, stratified sampling and analysis using methods such as Mantel – Haenszel estimation, etc. Observational situations are much more susceptible to unknown or lurking variables that are confounding. The presence of these confounding variables from the model can often wreak havoc to models rendering estimated coefficients and standard errors unreliable and uninterpretable or worse, lead to model specification problems in cases of regression modeling.

Statistician and creator of SYSTAT™, Leland Wilkinson commented on the problem of lurking variables and residuals very nicely:

"Residuals are not really errors in prediction, rather, they are the location of other structure which has not yet been accounted for the model…….. Fitting a model and looking at the statistical output can only show us what we more or less anticipated was there. We do not learn enough until we remove the effects of everything that we know influences the data and then plot what remains, the residuals. Here we may find the utterly unanticipated." (Wilkinson, et. al. 1996).

In the present case, our previous knowledge regarding bioaccumulation of mercury in fish has given us sound theoretical grounds to the notion that mercury level in fish increases as the fish grows, (i.e. increase in age and size). Length, weight and age are the obvious candidates as predictors in our model for fish mercury level. Whether there are other unknown lurking variables out in the environment that also affecting mercury level in fish, we cannot say; hence they are not in our model. If such a lurking variable exists and exerts significant influence on mercury level in fish, this might well jeopardize the predictive capability of our model and even leads to model specification problems. One interesting fact in our model is that fish age is our "known" confounding variable because the other two predictors (fork length and weight) as well as the dependent variable (fish mercury level) are all affected by fish age:



Chapter 2. The Multicollinearity Problem: Variance Inflation.

In regression modeling an underfitted model can lead to severely biased estimation and prediction. In contrast, an overfitted model can seriously degrade the efficiency of the resulting parameter estimates and predictions. It is often easily fall into a trap when we want to add as many predictors into a regression model in order to improve the predictability of the model for the dependent variable when (1) lacking good and sound theoretical reasons to do so and/or (2) overlooking interactions between predictors. One of the most common problems encountered in multiple regression modeling involving a large number of predictors is multicollinearity amongst predictors due to their linear correlative relationships. Since the ordinary least squares (OLS) model on which linear regression is conventionally based assumes an unbiased estimation of regression coefficients (i.e. OLS doesn't consider which predictor is more important than others); highly significant correlative relationships amongst predictors can create inaccurate estimation of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t – tests for the regression coefficients and can give false p – values. All in all, multicollinearity degrade the predictability of the model as a whole. The problem with OLS model is its "inflexibility" in the sense that it tries to find the **only** set of regression coefficients that "best" fit the data in order to achieve the lowest residual sum of squares. Though OLS gives unbiased estimates and enjoy the minimum variance of all linear unbiased estimators, there is no upper bound on the variance of the estimator and the presence of multicollinearity often produces large variances; hence, a huge price is paid for the unbiasedness property that one achieves by using OLS. In doing so, highly linear correlated predictors will lead to inflation of the standard errors of the regression coefficients. Multicollinearity is a special characteristic of the data matrix, not the underlying statistic model; i.e. it is a data problem and not a statistical problem. Mathematically, the OLS model requires (and assumes) a perfect inversion of the correlation matrix of predictors X^TX (a moment matrix); the presence of highly linearly correlated predictors causes the matrix to have less than full rank and renders a perfect inversion not possible. Instead, the inversion is approximate or "ill – conditioned", and often result in an inaccurately computed inverse matrix.

where $X = \begin{pmatrix} X_{11} & \cdots & X_{k1} \\ \vdots & & \vdots \\ X_{1N} & \cdots & X_{kN} \end{pmatrix}$, a N x (k +1) matrix, where N is the number of observations and k is the number of predictors with N required to be greater than or equal to k + 1. In a multicollinear situation,

the rank of X (and hence the X^TX) is less than k + 1, thus the matrix X^TX is not perfectly invertible. Because of this, the OLS estimator $\beta_{OLS} = (X^TX)^{-1}X^Ty$ is not possible.

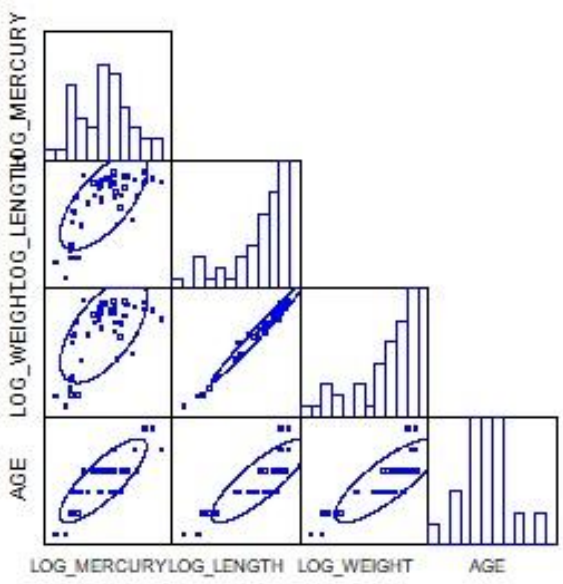
A simple visual way to gain a preliminary idea if multicollinearity is likely a problem is by examining the Spearman correlation matrix of the predictors. Predictors that show significant correlations are likely to have collinear problem.

Spearman correlation matrices clearly show that fork length, weight and age have highly significant positive correlations with one another for brook trout, lake trout and lake whitefish (Figures 2.1, 2.2 and 2.3).

Figure 2.1 Spearman correlation matrix of fork length, weight, age and mercury in brook trout.

Brook trout Spearman correlation matrix

	LOG_MERCURY	LOG_LENGTH	LOG_WEIGHT	AGE
LOG_MERCURY	1.000			
LOG_LENGTH	0.673	1.000		
LOG_WEIGHT	0.575	0.967	1.000	
AGE	0.749	0.855	0.815	1.000

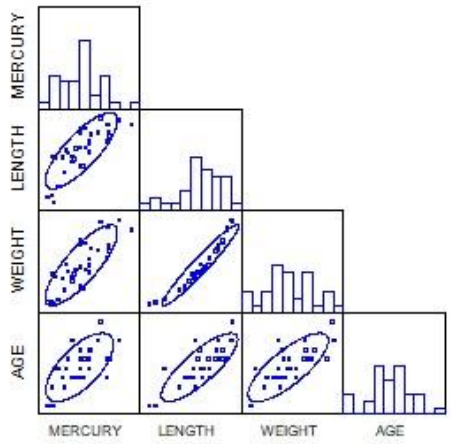


Number of observations: 44

Figure 2.2 Spearman correlation matrix of fork length, weight, age and mercury in lake trout.

Lake trout Spearman correlation matrix

	MERCURY	LENGTH	WEIGHT	AGE
MERCURY	1.000			
LENGTH	0.793	1.000		
WEIGHT	0.780	0.990	1.000	
AGE	0.671	0.780	0.754	1.000

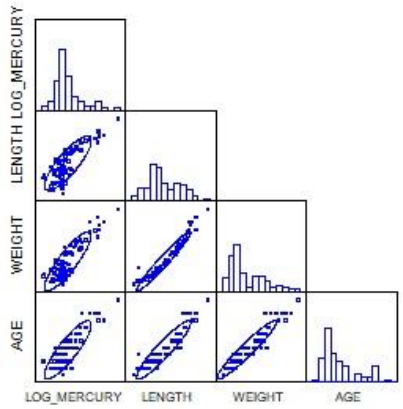


Number of observations: 31

Figure 2.3 Spearman correlation matrix of fork length, weight, age and mercury in lake whitefish.

Lake whitefish Spearman correlation matrix

	LOG_MERCURY	LENGTH	WEIGHT	AGE
LOG_MERCURY	1.000			
LENGTH	0.751	1.000		
WEIGHT	0.725	0.972	1.000	
AGE	0.708	0.916	0.933	1.000



Number of observations: 81

Another commonly used method of detecting the presence of multicollinearity is to look at the variance inflation factor (VIF) associated with each predictor.

The variance of an estimated partial regression coefficient $\mathrm{Var}(\beta_j) = \mathrm{MSE}\, c_{jj}$, where MSE is the mean squared error and c_{jj} is the j^{th} diagonal element of the inverted matrix $(X^TX)^{-1}$. In the multicollinearity situation, the large variances of the coefficients are associated with the large values of the c_{jj} since the mean squared error is not affected. It can be shown that $c_{jj} = \frac{1}{(1 - R_j^2)\sum (x_j - \bar{x}_j)^2}$, where R_j^2 is the coefficient of determination of the regression of x_j on all other predictors in the model. In fact, the $\sum (x_j - \bar{x}_j)^2$ is the denominator of the formula for the variance of the regression coefficient in a simple linear regression. If there is no multicollinearity, $R_j^2 = 0$; then the variance as well as the estimated coefficient is the same for the total and partial regression coefficients. However, correlations amongst predictors cause R_j^2 to increase, effectively increasing the magnitude of c_{jj} and consequently increasing the variance of the estimated coefficient. In other words, the variance of the estimated partial regression coefficient $\mathrm{Var}(\beta_j)$ is inflated by $\frac{1}{(1 - R_j^2)}$. This is the variance inflation factor (VIF) of the j^{th} coefficient. VIF measures how much the variance of an estimated regression coefficient increases if our predictors are correlated. VIF indicates the degree to which the standard errors are inflated due to the level of multicollinearity. As the R – square in the denominator gets closer and closer to one, the variance (and thus VIF) will get larger and larger. VIF = 1 indicates no relationships amongst predictors; VIF > 1 indicates that the predictors have some degrees of correlation. If VIF is close to 10 or even above 10 (hence, an R – square value between one predictor and the rest is 0.90 or greater), the multicollinearity is becoming problematic, i.e. the regression coefficients are poorly estimated and no long accurate (O'Brien 2007). However, it must be pointed out that using VIF = 10 as a "cut-off" point is more to do with a matter of convenience and there is little theoretical basis for that. In fact, the "cut-off" point of VIF that indicates a multicollinearity problem varies from case to case, for example, many authors actually suggested VIF of higher than 4 or 5 as the limit. Montgomery and Peck (1992) suggested that if the VIF is between 5 and 10, the partial regression coefficients are poorly estimated. The choice of the cut-off VIF values should also be evaluated relative to the overall fit of the model under study (i.e. the R-square of the model). If the R – square is very high (e.g. above 0.9), then

the VIF of 10 will not be large enough to seriously affect the estimate of the regression coefficients. On the other hand, if the R – square is low, say below 0.3, then even a relatively low VIF value (e.g. 6) may cause poor estimations of regression coefficients.

A preliminary run of linear regressions based on the OLS model reveal that the VIF values associated with fork length, weight and age for brook trout, lake trout and lake whitefish were above 10 (Tables 2.1, 2.2 and 2.3) with the exception of age for brook trout (VIF = 3.518)

Table 2.1 Linear regression of mercury on fork length, weight and age of brook trout.

```
The regression equation is
log mercury = - 22.3 + 4.87 log length - 1.53 log weight + 0.256 age


44 cases used, 1 cases contain missing values


Predictor      Coef  SE Coef      T      P     VIF
Constant    -22.340    4.370  -5.11  0.000
log length    4.866    1.126   4.32  0.000  60.968
log weight  -1.5310   0.3547  -4.32  0.000  54.347
age         0.25570  0.05608   4.56  0.000   3.518


S = 0.225545   R-Sq = 77.4%   R-Sq(adj) = 75.7%
```

Table 2.2 Linear regression of mercury on fork length, weight and age of lake trout.

```
The regression equation is
log mercury = - 2.70 - 0.00173 fork length + 0.00103 weight + 0.0576 age


Predictor         Coef    SE Coef      T      P     VIF
Constant       -2.6959     0.2788  -9.67  0.000
fork length  -0.001731   0.001223  -1.42  0.161  16.619
weight       0.0010329  0.0003267   3.16  0.002  24.749
age            0.05758    0.02922   1.97  0.052  11.188

S = 0.197654   R-Sq = 78.1%   R-Sq(adj) = 77.3%
```

Table 2.3 Linear regression of mercury on fork length, weight and age of lake whitefish.

```
The regression equation is
log mercury = - 2.70 - 0.00173 Fork length (mm) + 0.00103 Weight (g)
              + 0.0576 age


Predictor             Coef    SE Coef      T      P     VIF
Constant           -2.6959     0.2788  -9.67  0.000
Fork length (mm)  -0.001731   0.001223  -1.42  0.161  16.619
Weight (g)        0.0010329  0.0003267   3.16  0.002  24.749
age                 0.05758    0.02922   1.97  0.052  11.188


S = 0.197654   R-Sq = 78.1%   R-Sq(adj) = 77.3%
```

Another commonly observed effect when multicollinearity occurs is that the signs (directions) associated with the estimated regression coefficients do not make sense when comparing with the direction of the correlation between the predictor and the dependent variable: for examples, the negative regression coefficients associated with weight for brook trout as well as the negative regression coefficient associated with fork length for lake trout and lake whitefish.

Two shortcomings of VIF are (1) there is no meaningful boundary which can be used to distinguish VIF values which are too high from those which are acceptable; (2) VIF is unable to distinguish between several different sets of collinearities amongst predictors, i.e. it can't tell us what is collinear with what. Belsley, Kuh and Welsch (2013) proposed a better procedure using eigenvalues, condition indices together with variance proportions to confirm the presence and the extent of collinearity, then followed by using auxiliary regressions to identify the culprits amongst predictors.

The eigenvalues of the centered correlation tell us how independent each variable is. If all eigenvalues were 1.000, all predictors would be completely independent of each other. When one or more eigenvalues are greater than 1.000, some predictors are correlated and collinearity is a potential problem. For the mercury in brook trout data, fork length (column 1) has the largest eigenvalue (3.962). Condition indices are the most useful indicators of collinearity. It is the square root of the ratio of the largest eigenvalue divided by the eigenvalues of each of the four columns. Belsley et. al. (2013) proposed explicit diagnostic boundaries above which collinearity is harming the regression and below which collinearity is not harmful. Belsley et.al. suggested that condition indices over 15 suggest a **potential** problem and if over 30 indicate a serious problem. MERCURY (column 4, with a condition index of 470.475) confirms that we have a serious problem (Table 2.4):

Table 2.4. Inukjuak watershed brook trout. Regression output of the model $\log_e$ mercury = constant + fork length + $\log_e$ weight + age. (Column 1 = fork length; column 2 = weight; column 3 = age and column 4 = mercury.)

Brook trout
MODEL LOG_MERCURY = CONSTANT + LOG_LENGTH + LOG_WEIGHT + AGE

Eigenvalues of unit scaled X'X

1	2	3	4
3.962	0.035	0.003	0.000

Condition indices

1	2	3	4
1.000	10.628	35.133	470.475

Variance proportions

	1	2	3	4
CONSTANT	0.000	0.000	0.004	0.996
LOG_LENGTH	0.000	0.000	0.000	1.000
LOG_WEIGHT	0.000	0.000	0.058	0.942
AGE	0.001	0.309	0.553	0.137

Dep Var: LOG_MERCURY N: 44 Multiple R: 0.880 Squared multiple R: 0.774

Adjusted squared multiple R: 0.757 Standard error of estimate: 0.226

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	-22.340	4.370	0.000	.	-5.112	0.000
LOG_LENGTH	4.866	1.126	2.537	0.016	4.322	0.000
LOG_WEIGHT	-1.531	0.355	-2.392	0.018	-4.316	0.000
AGE	0.256	0.056	0.643	0.284	4.560	0.000

Effect	Coefficient	Lower 95%	Upper 95%
CONSTANT	-22.340	-31.172	-13.507
LOG_LENGTH	4.866	2.590	7.141
LOG_WEIGHT	-1.531	-2.248	-0.814
AGE	0.256	0.142	0.369

The variance proportions tell us which regression coefficients have been damaged due to collinearity. The variance proportions in the columns measure the independence of each predictor. If each predictor were **completely** independent, then each column would contain a single 1.000 and a set of 0.000s. If several predictors are collinear, most of the variance will show up in a single column. Thus, one column may contain several large variance proportions. To identify collinear predictors, we looked at the column corresponded to the condition index showing collinearity damage (i.e. column 4, with a condition index of 470.475 in Table 2.4 for brook trout). Here $\log_e$ fork length and $\log_e$ weight respectively have 1.000 and 0.945 of their variance which mean that the regression coefficients associated with both fork length and weight were seriously degraded by collinearity. Then we carried out an auxiliary regression of the model: AGE = CONSTANT + LENGTH + WEIGHT to look at the rest of the predictors, since age has the largest condition index (35.133) amongst the three predictors. Auxiliary regression was used here to confirm with predictors were involved. In this auxiliary regression, our

primary concern was to identify the predictors that were significant in predicting AGE. The p – value confirmed that fork length and age were predictors involved in the collinearity Table 2.5).

Table 2.5. Inukjuak watershed brook trout. Auxiliary regression output of the model age = constant + fork length + weight.

Brook trout auxiliary regression
MODEL AGE = CONSTANT + LOG_LENGTH + LOG_WEIGHT

Dep Var: AGE N: 44 Multiple R: 0.846 Squared multiple R: 0.716

Adjusted squared multiple R: 0.702 Standard error of estimate: 0.628

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	-31.495	11.133	0.000	.	-2.829	0.007
LOG_LENGTH	7.454	2.911	1.546	0.019	2.561	0.014
LOG_WEIGHT	-1.146	0.972	-0.712	0.019	-1.180	0.245

Effect	Coefficient	Lower 95%	Upper 95%
CONSTANT	-31.495	-53.978	-9.011
LOG_LENGTH	7.454	1.575	13.333
LOG_WEIGHT	-1.146	-3.108	0.816

Correlation matrix of regression coefficients

	CONSTANT	LOG_LENGTH	LOG_WEIGHT
CONSTANT	1.000		
LOG_LENGTH	-0.998	1.000	
LOG_WEIGHT	0.978	-0.990	1.000

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	40.732	2	20.366	51.615	0.000
Residual	16.177	41	0.395		

Table 2.6. Lake Tiasialuk lake trout. Regression output of the model mercury = constant + fork length + weight + age. (Column 1 = fork length; column 2 = weight; column 3 = age and column 4 = mercury.)

Lake trout
MODEL MERCURY = CONSTANT + LENGTH + WEIGHT + AGE

Eigenvalues of unit scaled X'X

1	2	3	4
3.865	0.116	0.016	0.002

Condition indices

1	2	3	4
1.000	5.764	15.371	46.600

Variance proportions

	1	2	3	4
CONSTANT	0.000	0.026	0.072	0.902
LENGTH	0.000	0.000	0.008	0.992
WEIGHT	0.001	0.081	0.064	0.853
AGE	0.001	0.001	0.912	0.086

Dep Var: MERCURY N: 31 Multiple R: 0.809 Squared multiple R: 0.655

Adjusted squared multiple R: 0.616 Standard error of estimate: 0.073

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	-0.019	0.150	0.000	.	-0.126	0.900
LENGTH	0.000	0.001	0.319	0.063	0.711	0.483
WEIGHT	0.000	0.000	0.459	0.074	1.102	0.280
AGE	0.003	0.011	0.046	0.333	0.234	0.816

Effect	Coefficient	Lower 95%	Upper 95%
CONSTANT	-0.019	-0.327	0.289
LENGTH	0.000	-0.001	0.002
WEIGHT	0.000	-0.000	0.000
AGE	0.003	-0.020	0.025

For the mercury in lake trout data (Table 2.6), fork length (column 1) has the largest eigenvalue (3.865). MERCURY (column 4, with a condition index of 46.6) confirmed that we have a problem. Column 4 of the variance proportions showed that fork length and weight respectively have 0.992 and 0.853 of their variance which mean that the regression coefficients associated with both fork length and weight were seriously degraded by collinearity. Again, age has the largest condition index (15.371) amongst the three predictors, the auxiliary regression model: AGE = CONSTANT + FORK LENGTH + WEIGHT (Table 2.7) showed that fork length and age were possibly involved in collinearity (p = 0.042).

Table 2.7. Lake Tasialuk lake trout. Auxiliary regression output of the model age = constant + fork length + weight.

Lake trout auxiliary regression
MODEL AGE = CONSTANT + LENGTH + WEIGHT

Dep Var: AGE N: 31 Multiple R: 0.817 Squared multiple R: 0.667

Adjusted squared multiple R: 0.643 Standard error of estimate: 1.260

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	-0.371	2.576	0.000	.	-0.144	0.886
LENGTH	0.024	0.011	0.859	0.074	2.136	0.042
WEIGHT	-0.000	0.002	-0.044	0.074	-0.109	0.914

Effect	Coefficient	Lower 95%	Upper 95%
CONSTANT	-0.371	-5.648	4.905
LENGTH	0.024	0.001	0.046
WEIGHT	-0.000	-0.005	0.005

Correlation matrix of regression coefficients

	CONSTANT	LENGTH	WEIGHT
CONSTANT	1.000		
LENGTH	-0.982	1.000	
WEIGHT	0.900	-0.962	1.000

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	89.007	2	44.503	28.016	0.000
Residual	44.477	28	1.588		

For the mercury in lake whitefish data (Table 2.8), fork length (column 1) has the largest eigenvalue (3.816). Column 3 (AGE) has a condition index of 21.738 which suggested a potential collinearity problem. Column 4 (MERCURY) has a condition index of 45.891 which confirmed that we have a problem. In column 4 (MERCURY) of the variance proportions fork length has 0.964 of its variance mean that the regression coefficient associated with fork length was seriously degraded by collinearity. Again, age has the largest condition index amongst the three predictors, the auxiliary regression model: AGE = CONSTANT + FORK LENGTH + WEIGHT (Table 2.9) shows that weight and age were involved in collinearity (p = 0.000).

Table 2.8. Inukjuak watershed lake whitefish. Regression output of the model mercury = constant + fork length + weight + age. (Column 1 = fork length; column 2 = weight; column 3 = age and column 4 = mercury.)

Lake whitefish

MODEL LOG_MERCURY = CONSTANT + LENGTH + WEIGHT + AGE

Eigenvalues of unit scaled X'X

	1	2	3	4
	3.816	0.174	0.008	0.002

Condition indices

	1	2	3	4
	1.000	4.683	21.738	45.891

Variance proportions

	1	2	3	4
CONSTANT	0.000	0.016	0.017	0.967
LENGTH	0.000	0.001	0.035	0.964
WEIGHT	0.001	0.035	0.247	0.718
AGE	0.001	0.001	0.986	0.012

Dep Var: LOG_MERCURY N: 81 Multiple R: 0.884 Squared multiple R: 0.781

Adjusted squared multiple R: 0.773 Standard error of estimate: 0.198

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	-2.696	0.279	0.000	.	-9.671	0.000
LENGTH	-0.002	0.001	-0.307	0.060	-1.415	0.161
WEIGHT	0.001	0.000	0.838	0.040	3.162	0.002
AGE	0.058	0.029	0.351	0.089	1.971	0.052

Effect	Coefficient	Lower 95%	Upper 95%
CONSTANT	-2.696	-3.251	-2.14
LENGTH	-0.002	-0.004	0.001
WEIGHT	0.001	0.000	0.002
AGE	0.058	-0.001	0.116

Table 2.9. Inukjuak watershed lake whitefish. Auxiliary regression output of the model age = constant + fork length + weight.

Lake whitefish auxiliary regression

MODEL AGE = CONSTANT + LENGTH + WEIGHT

Dep Var: AGE N: 81 Multiple R: 0.954 Squared multiple R: 0.911

Adjusted squared multiple R: 0.908 Standard error of estimate: 0.766

Effect	Coefficient	Std Error	Std Coef	Tolerance	t	P(2 Tail)
CONSTANT	2.280	1.049	0.000	.	2.174	0.033
LENGTH	0.003	0.005	0.099	0.061	0.719	0.474
WEIGHT	0.006	0.001	0.858	0.061	6.239	0.000

Effect	Coefficient	Lower 95%	Upper 95%
CONSTANT	2.280	0.192	4.369
LENGTH	0.003	-0.006	0.013
WEIGHT	0.006	0.004	0.009

Correlation matrix of regression coefficients

	CONSTANT	LENGTH	WEIGHT
CONSTANT	1.000		
LENGTH	-0.989	1.000	
WEIGHT	0.928	-0.969	1.000

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	466.238	2	233.119	397.349	0.000
Residual	45.762	78	0.587		

An important issue of multicollinearity is that the estimated regression coefficients may change erratically and significantly in response to small changes in the model or the input data (Belsley 1991). Although multicollinearity does not reduce the predictive power of the model "as a whole" (at least within the range of the data set), it impacts calculations regarding "individual" predictors hence gives erratic results about the relative importance of individual predictors or about which predictors are statistically redundant with respect to others. The estimate of one predictor's impact on the dependent variable while controlling for the other predictors is no longer accurate. One feature of multicollinearity is that the standard errors of the affected coefficients tend to be large. In that case, the test of the hypothesis that the coefficient is equal to zero may lead to a failure to reject a false null hypothesis of no effect of the predictor, i.e. a Type II Error.

In the present context, there are two approaches to remedy the problem of multicollinearity, each has their pros and cons as well as limitations:

(1) By using models other than the classical ordinary least squares (OLS) model such as partial least squares (PLS) regression and robust regression based of maximum likelihood estimators or by a modification of the OLS model using some forms of penalized estimation of the regression coefficients such as regularization techniques (for example ridge regression, Lasso regression and Elastic net) to allow for multicollinearity. It is also possible to reject the classical frequentist inference entirely and adopt the Bayesian approach to carry out the regression modeling, such an approach has rapidly grown in popularity in recent years which based on repeated simulation to determine the posterior probability distribution of the model's parameters based on an assumed form of a prior probability distribution and the estimation of a likelihood function.

(2) By removal of one or more redundant predictors. Computer algorithms such as stepwise regression, best – subset regression and the more recent LASSO regression and Elastic Net can be used to identify redundant predictors for possible exclusion from the regression model. However, it is most important that the final decision on the exclusion of a predictor from the model is based on sounded theoretical reasons instead of solely (and blindly) rely on the outcome of the computer algorithms. The remaining predictors can then be modeled using multiple linear regression.

It is also needed to mention here that another approach that can theoretically remedy the problem of multicollinearity especially when there are large number of predictors is to "redefining" the predictor variables by somehow combining them into principal components using techniques

such as principal component analysis, factor analysis and principal component regression. Hopefully this can reduce the dimensions of the model by combining predictors that are highly correlated. In reality, often the principal components resulted are difficult to interpret to be practically useful in the final model. Such an approach is not particularly useful in the present context of building a model on fish mercury levels based on only three predictors.

Chapter 3. Ridge Regression - Shrinkage Regularization with a Biased Estimator.

Ridge regression, originally known as Tikhonov regularization is often useful in mitigating the problem of multicollinearity in linear regression by carefully conducting a so - called "bias – variance trade – off". The technique is especially useful when there are a large number of predictors and in particular, when the number of observations is small, hence model over – fitting becoming a potential problem. The problem of the unbiased estimates of the OLS model in the face of highly linearly correlated predictors has been discussed in Chapter 2. In ridge regression, a small amount of bias (shrinkage parameter, k) is deliberately, gradually and carefully introduced to penalize the estimation of the regression coefficient in order to reduce the variability of the regression coefficient estimates and mitigate the multicollinearity problem.

The ridge regression equation written in the vector norm format where $\|\beta\|_2 = \sqrt{\beta_0^2 + \beta_1^2 + \cdots + \beta_p^2}$ made up of two components: $\|y - X\beta\|_2^2$ is actually the OLS term and the second component $k\|\beta\|_2^2$ is the penalty term that is what makes the ridge regression works:

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}}{argmin} \|y - X\beta\|_2^2 + k\|\beta\|_2^2$$

In the matrix format, the ridge regression equation is

$$\hat{\beta}^{ridge} = (X^T X + kI)^{-1} X^T Y$$

which is exactly the matrix equation of the OLS model with the penalty term k and its identity matrix $\boldsymbol{I}$ added to it.

The introduction of the shrinkage parameter k increases the residual sum of squares. In ridge regression, we "tune" the value of k to change the model coefficients until the mean squared error is minimized. This "tuning" the value of k has led to some controversy amongst statisticians regarding the subjectivity in the selection of the proper amount of penalty (i.e. bias) being added into the model. With the advancement of computer algorithms, more precise and objective selection of the k value becomes possible in recent years.

Ridge regression can be regarded as a special case of Bayesian linear regression in that the regression coefficients are assumed to be random variables with a specified prior distribution which can bias the solutions for the regression coefficients.

The assumptions of ridge regression are the same as least square regressions (i.e. linear relationships between dependent variable and predictors, homoscedasticity, independence and randomness of residuals) except normality is not to be assumed. In chapter 1, we saw that the brook trout and lake trout data met the assumptions of least square regressions and the three predictors exhibited various degrees of collinearity. The brook trout and lake trout data were analyzed using the ridge regression algorithm of NCSS™ to assess multicollinearity amongst predictors and to correct the problem if necessary.

3.1 Ridge regression analysis of brook trout data:

Correlation Matrix Section

	log_fork_length	log_weight	age	log_mercury
log_fork_length	1.000000	0.990442	0.840286	0.708356
log_weight	0.990442	1.000000	0.818702	0.647358
age	0.840286	0.818702	1.000000	0.816640
log_mercury	0.708356	0.647358	0.816640	1.000000

Pearson correlation coefficients showed that the predictors were significantly correlated with the dependent variable and with each other.

Least Squares Multicollinearity Section

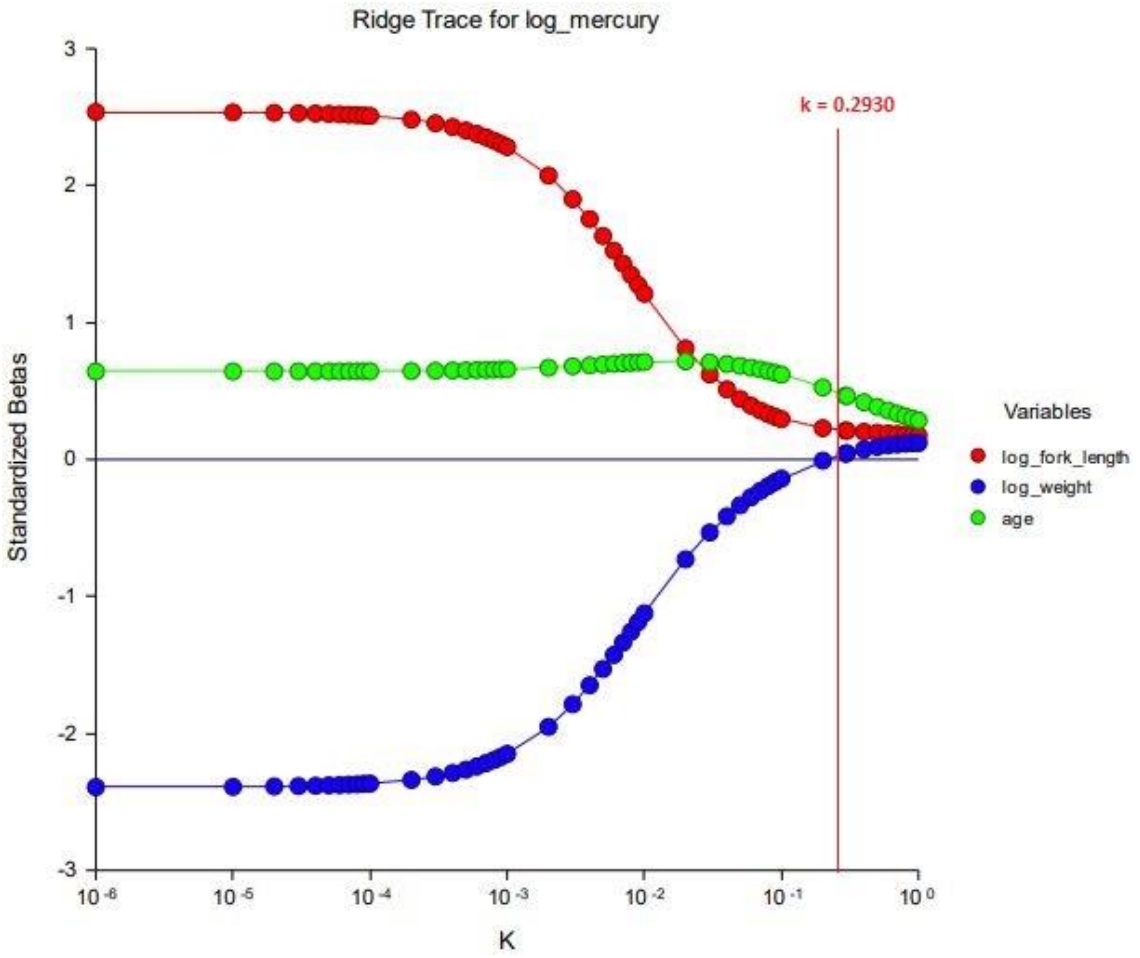
Independent Variable	Variance Inflation	R-Squared Vs Other X's	Tolerance
log_fork_length	60.9665	0.9836	0.0164
log_weight	54.3459	0.9816	0.0184
age	3.5178	0.7157	0.2843

VIF is the reciprocal of $1 - R_x^2$, where R_x^2 is the R^2 obtained when this variable is regressed on the remaining predictors. A VIF of 10 or more for large data sets indicates a multicollinearity problem since the R_x^2 with the remaining X's is 90%. For small data sets, even VIF's of 5 can signify multicollinearity. Log_e fork length and log_e weight have VIF greater than 10 and this indicated a multicollinearity problem. The high R-Squared vs Other X's indicated a lot of overlap in explaining the variation among the remaining predictors.

Eigenvalues of Correlations

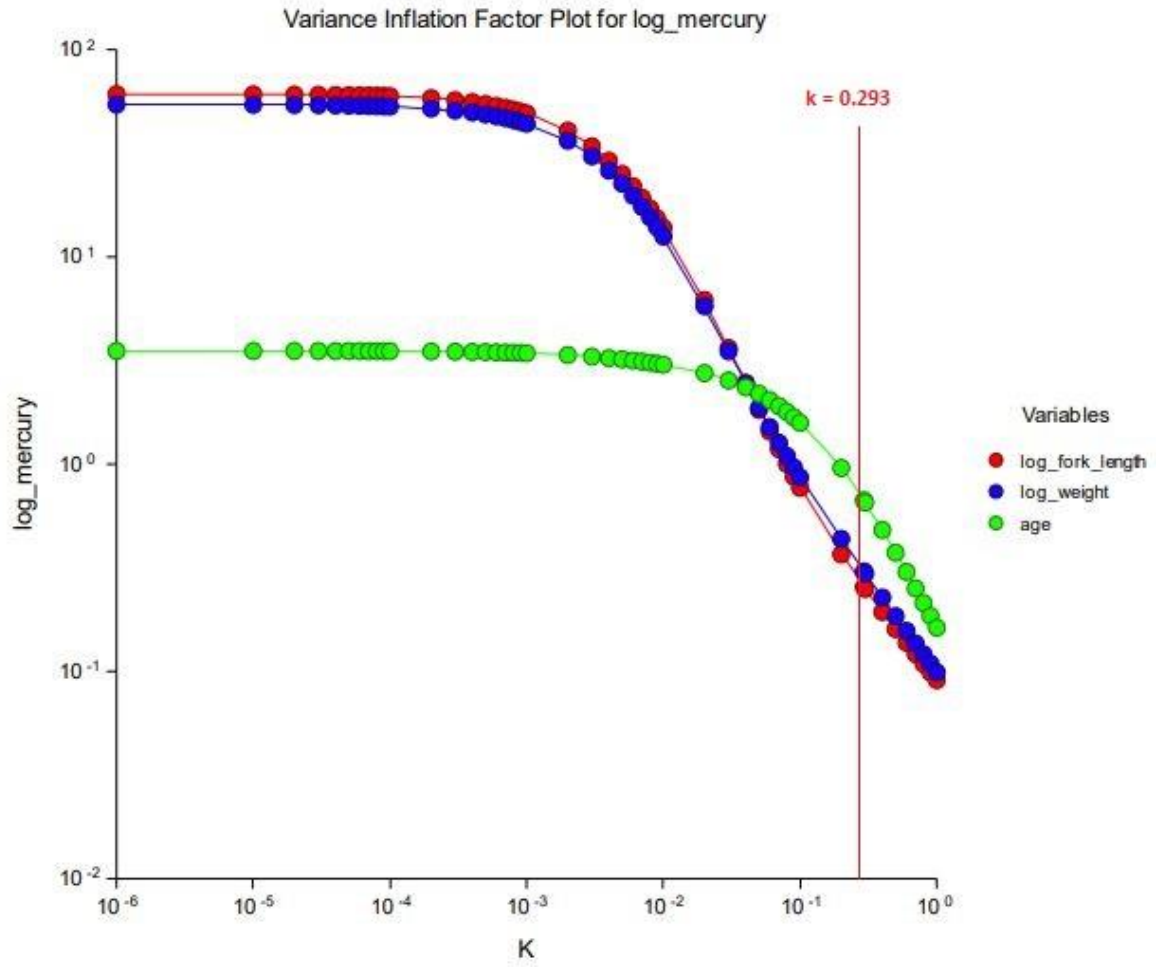
No.	Eigenvalue	Incremental Percent	Cumulative Percent	Condition Number
1	2.768575	92.29	92.29	1.00
2	0.222652	7.42	99.71	12.43
3	0.008774	0.29	100.00	315.56

The eigenvalues of correlations table give an eigenvalue analysis of the predictors after they have been centered and scaled. The sum of eigenvalues is equal to the number of predictors (i.e. 3). Eigenvalues as well as the incremental percent near zero such as the third eigenvalue here indicated a multicollinearity problem in the data. When collinearity is completely absent, all the three eigenvalues would have equal incremental percent. The condition number is the largest eigenvalue divided by each corresponding eigenvalue. Since the eigenvalues are really variances, the condition number is a ratio of variance. Condition numbers greater than 1000 indicate a severe multicollinearity problem while condition number between 100 and 1000 indicate a moderate multicollinearity problem such as the one here associated with the third eigenvalue.



One of the main obstacles and indeed scepticism by many statisticians in the use of ridge regression is in choosing an appropriate value of k. The formulas for calculating the value of k that minimizes the total mean squared error of the set of regression coefficients are functions of the unknown values of the population coefficients. Using the least square coefficients in these formulas

would render their estimation unreliable when multicollinearity is present. Hoerl and Kennard (1970), the inventors of ridge regression, suggested a graphical way to estimate the value of k which they called the ridge trace. The ridge regression coefficients are calculated for a set of values of k. The regression coefficients are standardized and plotted against the values of k. The standardized regression coefficients often vary widely for smaller values of k; and as k continues to increase, the standardized regression coefficients ultimately "settle down" or stabilized and gradually drift towards zero. Without the help of computer algorithm, the value of k that the regression coefficients begin to stabilize has to be picked manually, often quite subjectively. The difficulty stemmed from the fact that the smallest value of k should be selected in order to introduce the smallest amount of bias needed after which the regression coefficients have seem to stabilize as we do not want to introduce an unnecessary amount of bias in the estimation of the regression coefficients. The ridge trace and more importantly the variance inflation factor plot computed by software algorithms such as the NCSS™ greatly helps the selection of the k value. The algorithm has suggested an optimal k value of 0.293 in the present case.



The variance inflation factor plot showed the impact of k on the variance inflation factors. At around k = 0.293, the variance inflation factors have dropped way below 5 (for small data sets) or 10 (for large data sets) and the impact due to multicollinearity was removed. However, 0.293 was definitely not the lowest k value when VIF dropped below 5 or 10. Further examination of the variance inflation factor section and the k – analysis section computed by the NCSS™ algorithm suggested that k values as low as 0.02 and 0.03 were sufficient to lower the VIF and removed multicollinearity. The optimal k value of 0.293 computed by the algorithm was unnecessary high.

Variance Inflation Factor Section

k	log_fork_length	log_weight	age
0.000000	60.9665	54.3459	3.5178
0.000010	60.8292	54.2243	3.5169
0.000020	60.6924	54.1032	3.5159
0.000030	60.5561	53.9825	3.5149
0.000040	60.4202	53.8621	3.5140
0.000050	60.2848	53.7422	3.5130
0.000060	60.1499	53.6227	3.5121
0.000070	60.0154	53.5036	3.5111
0.000080	59.8814	53.3849	3.5102
0.000090	59.7478	53.2666	3.5092
0.000100	59.6147	53.1487	3.5083
0.000200	58.3078	51.9914	3.4990
0.000300	57.0438	50.8721	3.4899
0.000400	55.8210	49.7891	3.4810
0.000500	54.6375	48.7410	3.4724
0.000600	53.4916	47.7262	3.4639
0.000700	52.3818	46.7434	3.4556
0.000800	51.3066	45.7911	3.4475
0.000900	50.2646	44.8683	3.4395
0.001000	49.2543	43.9735	3.4317
0.002000	40.6482	36.3511	3.3614
0.003000	34.1391	30.5853	3.3020
0.004000	29.0971	26.1184	3.2502
0.005000	25.1119	22.5871	3.2040
0.006000	21.9073	19.7470	3.1620
0.007000	19.2919	17.4286	3.1233
0.008000	17.1294	15.5111	3.0872
0.009000	15.3209	13.9071	3.0532
0.010000	13.7930	12.5515	3.0209
0.020000	6.1890	5.7930	2.7528
0.030000	3.6337	3.5073	2.5365
0.040000	2.4651	2.4518	2.3498
0.050000	1.8286	1.8697	2.1852
0.060000	1.4401	1.5091	2.0386
0.070000	1.1833	1.2667	1.9072
0.080000	1.0029	1.0935	1.7888
0.090000	0.8703	0.9638	1.6817
0.100000	0.7692	0.8632	1.5844
0.200000	0.3675	0.4356	0.9588
0.293000	0.2558	0.3035	0.6698
0.300000	0.2503	0.2968	0.6540
0.400000	0.1938	0.2270	0.4817
0.500000	0.1603	0.1849	0.3742
0.600000	0.1377	0.1567	0.3022
0.700000	0.1214	0.1364	0.2513
0.800000	0.1089	0.1210	0.2137
0.900000	0.0990	0.1089	0.1851
1.000000	0.0908	0.0991	0.1626

K Analysis Section

k	R2	Sigma	B'B	Ave VIF	Max VIF
0.000000	0.7739	0.2255	12.5750	39.6101	60.9665
0.000010	0.7737	0.2256	12.5479	39.5235	60.8292
0.000020	0.7736	0.2257	12.5209	39.4372	60.6924
0.000030	0.7735	0.2257	12.4939	39.3512	60.5561
0.000040	0.7734	0.2258	12.4671	39.2655	60.4202
0.000050	0.7732	0.2259	12.4403	39.1800	60.2848
0.000060	0.7731	0.2259	12.4136	39.0949	60.1499
0.000070	0.7730	0.2260	12.3871	39.0101	60.0154
0.000080	0.7729	0.2260	12.3606	38.9255	59.8814
0.000090	0.7728	0.2261	12.3342	38.8412	59.7478
0.000100	0.7726	0.2262	12.3079	38.7572	59.6147
0.000200	0.7714	0.2268	12.0496	37.9327	58.3078
0.000300	0.7702	0.2274	11.7998	37.1353	57.0438
0.000400	0.7691	0.2279	11.5581	36.3637	55.8210
0.000500	0.7679	0.2285	11.3242	35.6169	54.6375
0.000600	0.7668	0.2291	11.0977	34.8939	53.4916
0.000700	0.7657	0.2296	10.8783	34.1936	52.3818
0.000800	0.7646	0.2301	10.6657	33.5151	51.3066
0.000900	0.7636	0.2306	10.4597	32.8574	50.2646
0.001000	0.7625	0.2311	10.2600	32.2199	49.2543
0.002000	0.7532	0.2357	8.5581	26.7869	40.6482
0.003000	0.7453	0.2394	7.2701	22.6755	34.1391
0.004000	0.7385	0.2425	6.2717	19.4885	29.0971
0.005000	0.7327	0.2452	5.4819	16.9677	25.1119
0.006000	0.7275	0.2476	4.8463	14.9388	21.9073
0.007000	0.7229	0.2497	4.3269	13.2813	19.2919
0.008000	0.7188	0.2515	3.8971	11.9093	17.1294
0.009000	0.7151	0.2532	3.5371	10.7604	15.3209
0.010000	0.7117	0.2547	3.2325	9.7885	13.7930
0.020000	0.6886	0.2647	1.7036	4.9116	6.1890
0.030000	0.6746	0.2706	1.1743	3.2258	3.6337
0.040000	0.6643	0.2748	0.9213	2.4222	2.4651
0.050000	0.6558	0.2783	0.7757	1.9612	2.1852
0.060000	0.6486	0.2812	0.6810	1.6626	2.0386
0.070000	0.6421	0.2838	0.6139	1.4524	1.9072
0.080000	0.6362	0.2861	0.5635	1.2951	1.7888
0.090000	0.6308	0.2882	0.5238	1.1719	1.6817
0.100000	0.6257	0.2902	0.4914	1.0722	1.5844
0.200000	0.5862	0.3051	0.3290	0.5873	0.9588
0.293000	0.5589	0.3150	0.2646	0.4097	0.6698
0.300000	0.5571	0.3157	0.2610	0.4004	0.6540
0.400000	0.5331	0.3241	0.2214	0.3009	0.4817
0.500000	0.5124	0.3312	0.1945	0.2398	0.3742
0.600000	0.4940	0.3374	0.1747	0.1989	0.3022
0.700000	0.4773	0.3429	0.1592	0.1697	0.2513
0.800000	0.4621	0.3479	0.1465	0.1479	0.2137
0.900000	0.4480	0.3524	0.1358	0.1310	0.1851
1.000000	0.4349	0.3566	0.1266	0.1175	0.1626

The K – analysis section provides a summary of various statistics that went into the choice of k.

Since the least squares solution maximizes R – squared (**R2**), the largest value of R – squared occurs when k is zero (i.e. an unbiased OLS model). The k value selected should not stray too far from the value of **R2** when k is zero. Same applied to the square root of the mean squared error **(Sigma**) that least squares minimize this value.). The k value selected should not stray too far from the value of **Sigma** when k is zero. Seemingly, the lower k values (0.02 or 0.03) worked better than the optimal k computed (k = 0.293) in these respects.

Ridge regression assumes that the value of the sum of the squared standardized regression coefficients (**B'B**) is too large and so the method tries to reduce **B'B**. The k value selected should correspond to where the value of **B'B** becoming stabilized. It seems that following the initial big drop, B'B became stabilized at k = 0.02 or 0.03.

Ave VIF and **Max VIF** are the average of variance inflation factors and the maximum variance inflation factor respectively. The k value selected should correspond to where **Max VIF** is below 10 for large data sets or below 5 for small data sets. In this case, k = 0.02 and k = 0.03 could have accomplished that as well.

All in all, the k value of 0.293 seems unnecessarily high since we prefer to select the smallest k value (hence smallest bias introduced to the estimation) that can mitigate the multicollinearity problem. The reason for the computer algorithm to optimize the k value as 0.293 was that this is the lowest k value to maintain the regular partial ridge regression coefficient associated with $\log_e$ weight to remain positive (0.02758). As we knew that there was a positive correlation between $\log_e$ mercury and $\log_e$ weight, a negative regular partial ridge regression coefficient associated with $\log_e$ weight indicated multicollinearity was still affecting the model. k = 0.02 and k = 0.03 yielded negative regular partial ridge regression coefficients associated with $\log_e$ weight of -0.4664 and -0.3416 respectively; hence cannot be used.

Ridge vs. Least Squares Comparison Section for k = 0.293000

Independent Variable	Regular Ridge Coeff's	Regular L.S. Coeff's	Stand'zed Ridge Coeff's	Stand'zed L.S. Coeff's	Ridge Standard Error	L.S. Standard Error
Intercept	-5.491363	-22.33982				
log_fork_length	0.4053442	4.865543	0.2114	2.5374	0.1018325	1.12571
log_weight	0.02757564	-1.530959	0.0431	-2.3923	0.0370207	0.354716
age	0.1856737	0.255702	0.4669	0.6430	0.03417391	0.05607698
R-Squared	0.5589	0.7739				
Sigma	0.3150	0.2255				

Ridge Regression Coefficient Section for k = 0.293000

Independent Variable	Regression Coefficient	Standard Error	Stand'zed Regression Coefficient	VIF
Intercept	-5.491363			
log_fork_length	0.4053442	0.1018325	0.2114	0.2558
log_weight	0.02757564	0.0370207	0.0431	0.3035
age	0.1856737	0.03417391	0.4669	0.6698

The regular (unstandardized) ridge partial regression coefficients (**Regular Ridge Coeff's**, when k = 0.293) and the least square partial regression coefficients (**Regular L.S. Coeff's**, when k = 0) associated with the three predictors indicated how much change in the dependent variable ($\log_e$ mercury) occurs for a one-unit change in a particular predictor when the other two predictors were held constant. Note that the **Regular L.S. Coeff's** associated with $\log_e$ weight was negative (-1.530959) as a result of multicollinearity. By applying a k value of 0.293, the **Regular Ridge Coeff's** associated with $\log_e$ weight became positive (0.02757564), which make sense since $\log_e$ mercury and $\log_e$ weight were positively correlated albeit not a very strong one (Pearson coefficient = 0.647).

The standardized ridge coefficients (**Stand'zed Ridge Coeff's**) are calculated as $b_{j,std} = b_j\left(\frac{S_{x_j}}{S_y}\right)$ where Sy and S_{xj} are the standard deviations for the dependent variable and the corresponding j^{th} predictor. **Stand'zed Ridge Coeff's** can be used to gauge the relative importance of each predictor in influencing the dependent variable. In the present case, fish age followed by fork length were important predictors for mercury, whereas weight was a relatively unimportant predictor.

One objective of ridge regression is to reduce the standard error of the regression coefficient, hence making their estimates more precise. In the present case, ridge regression reduced the standard errors of the regression coefficients associated with $\log_e$ length and $\log_e$ weight almost ten times; whereas ridge regression has relatively small effect on reducing the standard error of the regression coefficient associated with fish age. As it was noted that age was not affected by multicollinearity, the VIF for age was very low.

The coefficient of determination (**R – Squared**) for the ridge regression model was quite low (0.5589), which indicated that only about 56% of the variation in the fish mercury was explained by the three predictors in the ridge regression model after suppressing the effect of multicollinearity. If we have decided to include all the three predictors in the regression model, ridge regression analysis yielded the following predictive model:

$\log_e$ mercury = (-5.4914) + (0.4053) * $\log_e$ fork length + (0.0276) * $\log_e$ weight + (0.1857) * age

However, one might incline to drop weight from the model because (1) the standardized ridge regression coefficient for weight suggested that it was a relatively unimportant predictor for mercury;

(2) the positive correlations between $\log_e$ mercury and $\log_e$ fork length as well as between $\log_e$ mercury and age were stronger than that between $\log_e$ mercury and $\log_e$ weight; (3) the inflation of the k value in order to render the ridge regression coefficient of $\log_e$ weight positive and (4) weight account for a substantial amount of multicollinearity problem (VIF = 54.3). Once having weight removed from the model, ridge regression analysis showed that multicollinearity was no longer a problem: VIF for both $\log_e$ fork length and age were below 10 and the condition number turned out to be only 11.5. Ridge regression was no longer needed to model the regression of $\log_e$ mercury on $\log_e$ fork length and age; a multiple linear regression based on the OLS model can be used, perhaps with stepwise or best – subset regressions to screen the predictors beforehand.

3.2 Ridge regression analysis of lake trout data:

Correlation Matrix Section

	fork_length	weight	age	mercury
fork_length	1.000000	0.962477	0.816490	0.799070
weight	0.962477	1.000000	0.782632	0.802770
age	0.816490	0.782632	1.000000	0.666265
mercury	0.799070	0.802770	0.666265	1.000000

Pearson correlation coefficients showed that the predictors were significantly correlated with the dependent variable and with each other.

Least Squares Multicollinearity Section

Independent Variable	Variance Inflation	R-Squared Vs Other X's	Tolerance
fork_length	15.7924	0.9367	0.0633
weight	13.5857	0.9264	0.0736
age	3.0012	0.6668	0.3332

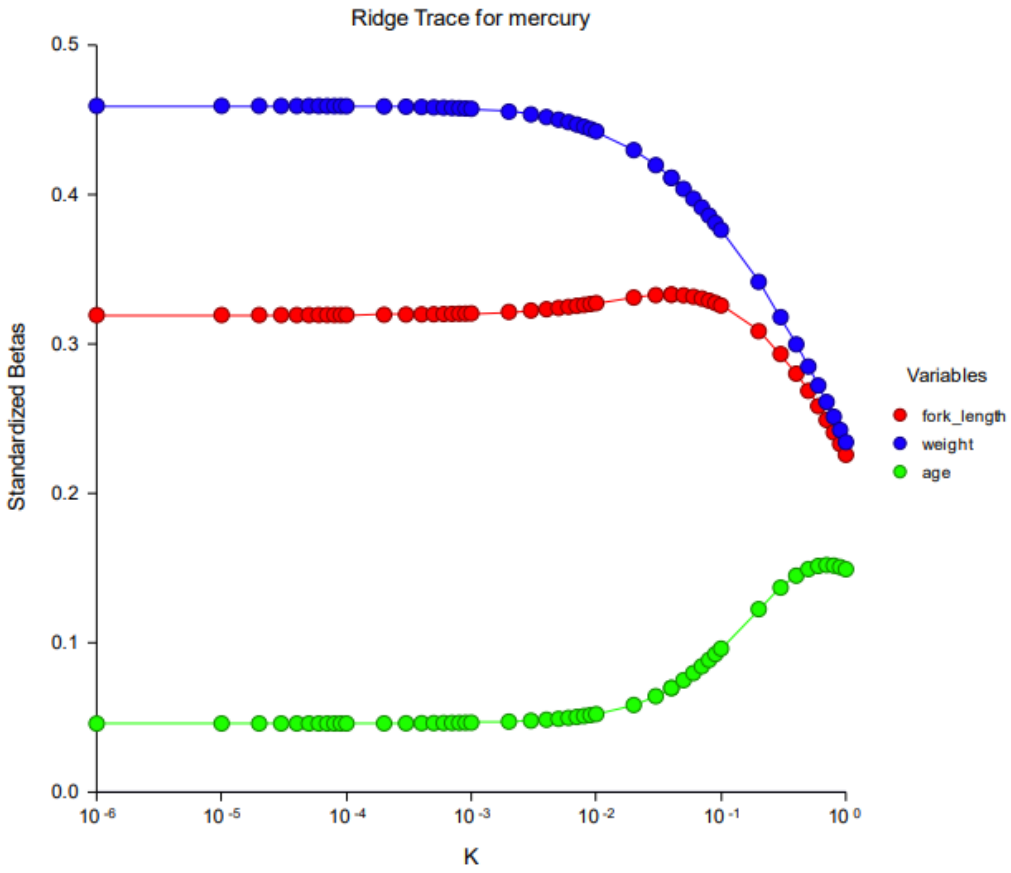
Fork length and weight have VIF just over 10 and this indicated the presence of multicollinearity. The high **R-Squared vs Other X's** indicated a lot of overlap in explaining the variation among the remaining predictors.

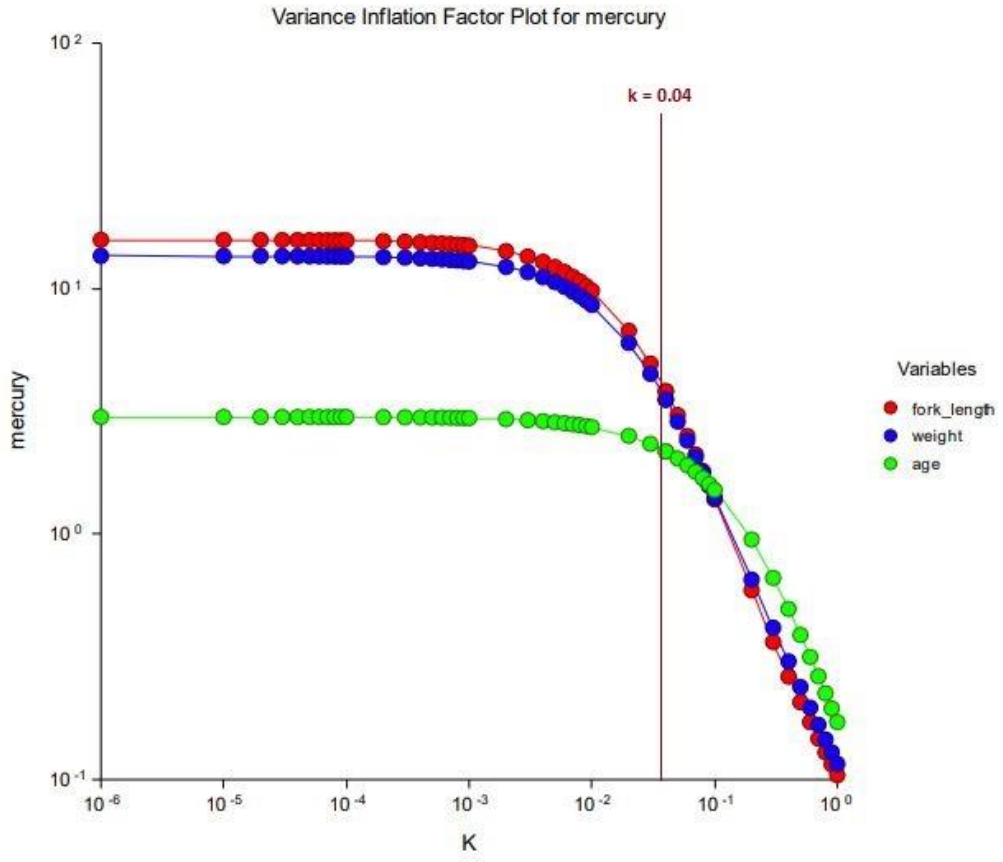
Eigenvalues of Correlations

No.	Eigenvalue	Incremental Percent	Cumulative Percent	Condition Number
1	2.710200	90.34	90.34	1.00
2	0.254183	8.47	98.81	10.66
3	0.035618	1.19	100.00	76.09

The third eigenvalue near zero indicated the presence of multicollinearity in the data. When collinearity is completely absent, all the three eigenvalues would have equal incremental percent which was not the case here. However, the condition numbers associated with all three eigenvalues were below 100 which indicated that the extent of multicollinearity was not a serious problem.

In the present case, the ridge trace showed that the standardized ridge regression coefficients remained fairly stable as the k value increased from zero to about 0.01. Based on the ridge trace alone it seems not possible to selected a k value for the model. However, the variance inflation factor plot, the variance inflation factor section table and the k – analysis section table suggested a k value of 0.04 was sufficient to render the VIF values associated with all three predictors below 5 (VIF cut-off for small data sets as it was in the present case).




Variance Inflation Factor Plot for mercury
k = 0.04
mercury
K
Variables
fork_length
weight
age

Variance Inflation Factor Section

k	fork_length	weight	age
0.000000	15.7924	13.5857	3.0012
0.000010	15.7838	13.5785	3.0009
0.000020	15.7752	13.5714	3.0005
0.000030	15.7666	13.5642	3.0002
0.000040	15.7580	13.5571	2.9999
0.000050	15.7495	13.5499	2.9996
0.000060	15.7409	13.5428	2.9993
0.000070	15.7324	13.5356	2.9990
0.000080	15.7238	13.5285	2.9987
0.000090	15.7153	13.5214	2.9984
0.000100	15.7067	13.5143	2.9980
0.000200	15.6219	13.4435	2.9949
0.000300	15.5377	13.3733	2.9918
0.000400	15.4542	13.3037	2.9887
0.000500	15.3714	13.2346	2.9856
0.000600	15.2893	13.1662	2.9826
0.000700	15.2078	13.0982	2.9795
0.000800	15.1270	13.0309	2.9765
0.000900	15.0469	12.9641	2.9734
0.001000	14.9675	12.8978	2.9704
0.002000	14.2073	12.2636	2.9406
0.003000	13.5051	11.6776	2.9117
0.004000	12.8553	11.1350	2.8836
0.005000	12.2527	10.6316	2.8564
0.006000	11.6927	10.1638	2.8298
0.007000	11.1716	9.7281	2.8040
0.008000	10.6857	9.3216	2.7787
0.009000	10.2320	8.9419	2.7541
0.010000	9.8076	8.5866	2.7300
0.020000	6.7367	6.0082	2.5142
0.030000	4.9502	4.4985	2.3319
0.040000	3.8174	3.5338	2.1732
0.040000	3.8174	3.5338	2.1732
0.050000	3.0524	2.8765	2.0329
0.060000	2.5103	2.4063	1.9074
0.070000	2.1113	2.0566	1.7943
0.080000	1.8084	1.7883	1.6919
0.090000	1.5726	1.5771	1.5988
0.100000	1.3851	1.4073	1.5136
0.200000	0.5903	0.6525	0.9510
0.300000	0.3638	0.4155	0.6636
0.400000	0.2628	0.3030	0.4959
0.500000	0.2069	0.2383	0.3889
0.600000	0.1715	0.1964	0.3160
0.700000	0.1470	0.1672	0.2638
0.800000	0.1291	0.1456	0.2249
0.900000	0.1153	0.1291	0.1951
1.000000	0.1043	0.1159	0.1715

K Analysis Section

k	R2	Sigma	B'B	Ave VIF	Max VIF
0.000000	0.6546	0.0734	0.3152	10.7931	15.7924
0.000010	0.6546	0.0734	0.3152	10.7877	15.7838
0.000020	0.6546	0.0734	0.3152	10.7824	15.7752
0.000030	0.6546	0.0734	0.3152	10.7770	15.7666
0.000040	0.6546	0.0734	0.3151	10.7717	15.7580
0.000050	0.6546	0.0734	0.3151	10.7663	15.7495
0.000060	0.6546	0.0734	0.3151	10.7610	15.7409
0.000070	0.6546	0.0734	0.3151	10.7557	15.7324
0.000080	0.6546	0.0734	0.3151	10.7503	15.7238
0.000090	0.6546	0.0734	0.3151	10.7450	15.7153
0.000100	0.6546	0.0734	0.3151	10.7397	15.7067
0.000200	0.6546	0.0734	0.3150	10.6868	15.6219
0.000300	0.6545	0.0734	0.3149	10.6343	15.5377
0.000400	0.6545	0.0734	0.3148	10.5822	15.4542
0.000500	0.6545	0.0734	0.3146	10.5306	15.3714
0.000600	0.6544	0.0734	0.3145	10.4793	15.2893
0.000700	0.6544	0.0734	0.3144	10.4285	15.2078
0.000800	0.6544	0.0734	0.3143	10.3781	15.1270
0.000900	0.6543	0.0734	0.3142	10.3281	15.0469
0.001000	0.6543	0.0734	0.3141	10.2785	14.9675
0.002000	0.6540	0.0734	0.3131	9.8038	14.2073
0.003000	0.6537	0.0735	0.3121	9.3648	13.5051
0.004000	0.6534	0.0735	0.3111	8.9580	12.8553
0.005000	0.6531	0.0735	0.3101	8.5802	12.2527
0.006000	0.6528	0.0736	0.3092	8.2288	11.6927
0.007000	0.6524	0.0736	0.3082	7.9012	11.1716
0.008000	0.6521	0.0736	0.3074	7.5954	10.6857
0.009000	0.6518	0.0737	0.3065	7.3093	10.2320
0.010000	0.6515	0.0737	0.3056	7.0414	9.8076
0.020000	0.6485	0.0740	0.2978	5.0864	6.7367
0.030000	0.6456	0.0743	0.2910	3.9269	4.9502
0.040000	0.6427	0.0746	0.2849	3.1748	3.8174
0.040000	0.6427	0.0746	0.2849	3.1748	3.8174
0.050000	0.6399	0.0749	0.2793	2.6539	3.0524
0.060000	0.6371	0.0752	0.2742	2.2746	2.5103
0.070000	0.6344	0.0755	0.2695	1.9874	2.1113
0.080000	0.6317	0.0758	0.2651	1.7629	1.8084
0.090000	0.6291	0.0760	0.2609	1.5828	1.5988
0.100000	0.6265	0.0763	0.2570	1.4353	1.5136
0.200000	0.6024	0.0787	0.2269	0.7313	0.9510
0.300000	0.5808	0.0808	0.2059	0.4810	0.6636
0.400000	0.5611	0.0827	0.1894	0.3539	0.4959
0.500000	0.5428	0.0844	0.1756	0.2780	0.3889
0.600000	0.5259	0.0860	0.1638	0.2280	0.3160
0.700000	0.5100	0.0874	0.1534	0.1927	0.2638
0.800000	0.4952	0.0887	0.1441	0.1666	0.2249
0.900000	0.4812	0.0899	0.1358	0.1465	0.1951
1.000000	0.4680	0.0910	0.1282	0.1306	0.1715

Ridge vs. Least Squares Comparison Section for k = 0.040000

Independent Variable	Regular Ridge Coeff's	Regular L.S. Coeff's	Stand'zed Ridge Coeff's	Stand'zed L.S. Coeff's	Ridge Standard Error	L.S. Standard Error
Intercept	-0.02682871	-0.01893847				
fork_length	0.0005167878	0.0004955938	0.3330	0.3194	0.0003487851	0.0006974561
weight	0.0001430434	0.0001598161	0.4112	0.4594	7.522487E-05	0.0001450119
age	0.003914717	0.002578799	0.0697	0.0459	0.009520899	0.01100001
R-Squared	0.6427	0.6546				
Sigma	0.0746	0.0734				

It is rather surprising the standardized ridge coefficients (**Stand'zed Ridge Coeff's)** suggested that fork length and weight were more important predictors for mercury than fish age in the case of

lake trout. Ridge regression did not result in substantial reduction in standard errors of the regression coefficients associated with any of the three predictors. Hence, applying ridge regression in the present case did not improve the estimation of the regression coefficients. This is also confirmed by the similarity between the ridge coefficients and the least square coefficients for all three predictors. The ridge regression model and the least square regression model also have very similar coefficients of determination (**R-Squared**) and the square roots of the mean square error (**Sigma**).

Ridge Regression Coefficient Section for k = 0.040000

Independent Variable	Regression Coefficient	Standard Error	Stand'zed Regression Coefficient	VIF
Intercept	-0.02682871			
fork_length	0.0005167878	0.0003487851	0.3330	3.8174
weight	0.0001430434	7.522487E-05	0.4112	3.5338
age	0.003914717	0.009520899	0.0697	2.1732

Nonetheless, application of the ridge regression model with a k value of 0.04 did reduce the VIF for fork length from 15.7924 to 3.8174; the VIF for weight from 13.5857 to 3.5338 and the VIF for age from 3.0012 to 2.1732.

The ridge regression model with a k value of 0.04 is

mercury = $(-0.0268) + (5.168 * 10^{-4}) *$ fork length $+ (1.43 * 10^{-4}) *$ weight $+ (3.915 * 10^{-3}) *$ age

All in all, it seems that in the case of lake trout, the extent of collinearity amongst the three predictors was relatively small and did not constitute a problem; the application of ridge regression did not contribute much to improve the precision in the estimation of the regression coefficients.

Chapter 4. Partial Least Squares Regression – Projection to Latent Structures in New Planes.

Partial least squares (PLS) fits linear models based on linear combinations called components (or factors) of the predictor variables. These components are obtained in a sequential way that attempts to maximize the covariance between the predictor variables (Xs) and the dependent variables (Ys). In this way, PLS exploits the correlation between predictor variables and dependent variables to reveal underlying latent structures. The components address the combined goals of explaining variation of the dependent variables and the variation of the predictor variables. PLS regression combined the features of multiple linear regression and principal component analysis (PCA). First, PLS regression applies PCA to **both** Xs and Ys and finds the fundamental relations between two matrices (Xs and Ys) by projecting the observable and predicted variables to a new hyperplane as latent structures. The decomposition of Xs and Ys matrices are made so as to maximize the covariance between the projection of the Xs matrix and the projection of the Ys matrix. This is followed by a linear regression step to predict the values of the dependent variables using the decomposition of the predictors' matrices.

The model of PLS is linear – X_i are the k predictors and Y_j are the p dependent variables; for each sample n, the value of y_{nj} is:

$$y_{nj} = \sum_{i=0}^{k} \beta_i x_{ni} + \varepsilon_{nj}$$

The model different from multiple linear regression in the way β_i are found. The results of the PLS regression is the model equation $Y = X\beta + \varepsilon$ showing the β coefficient that gives the relationship between X and Y variables.

The matrix of the regression coefficient of Y on X with h factors (components) generated by the PLS algorithm is given by the equation:

$$\beta = Wh(P'_h W_h)^{-1} C'_h$$

where Y is the matrix of the dependent variables;

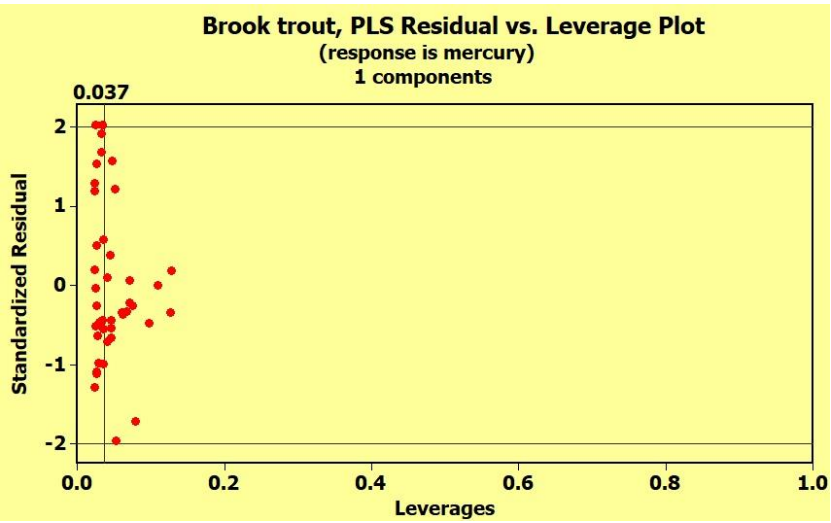
X is the matrix of the predictors;

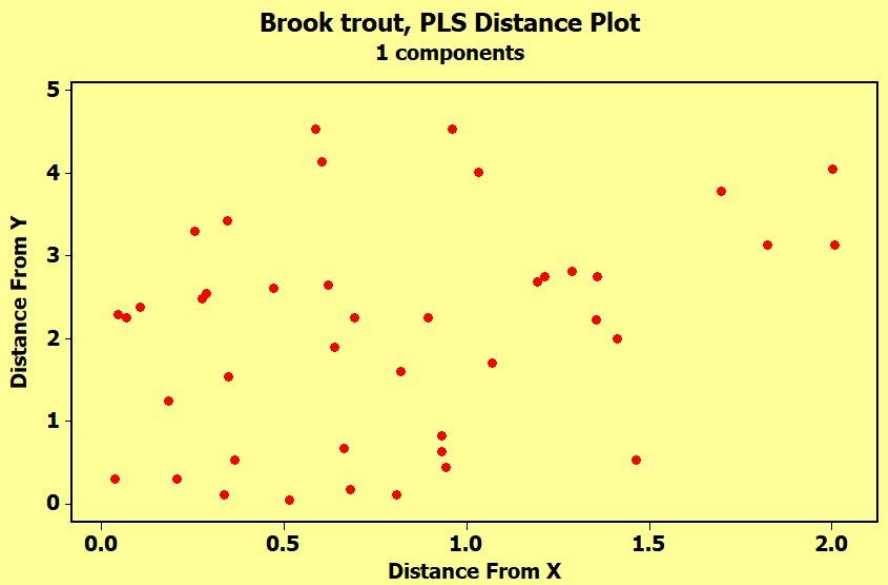
W, W_h, C_h, P_h are orthogonal loading matrices generated by the PLS algorithm.

PLS regression is particularly useful when predictors are highly collinear or in situations when there are more predictors than observations; hence, OLS regression either fails or produces erratic coefficients with high standard errors. The technique is extensively used in chemometrics and spectral analysis. The most commonly used PLS algorithm is the nonlinear iterative partial least square (NIPALS) algorithm developed by Herman Wold (1975) who invented PLS regression. The algorithm reduces the number of predictors using PCA to extract a set of uncorrelated components that describe maximum correlation among the predictors and the dependent variables. Leave – one – out cross validation method was used by the algorithm here to identify the smallest set of components that provide the maximum predictive ability. It then performs least square regression on these components. In the present study, the fish mercury and growth data were analyzed using the NIPALS algorithm from the JMP™ and Minitab™ statistics software.

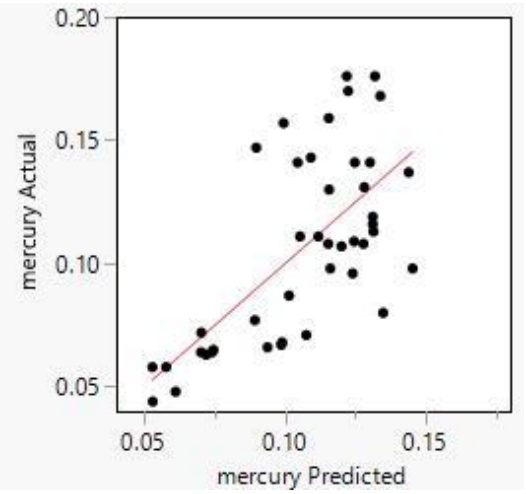
4.1 PLS regression analysis of brook trout data:

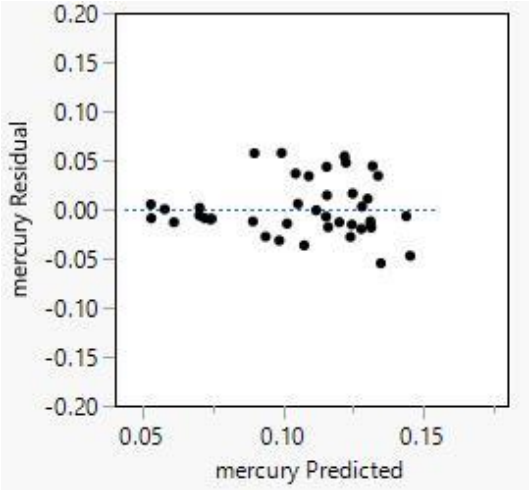
Untransformed **full** data set of the brook trout mercury and fish growth data (see Appendix) was examined for the presence of outliers using residuals versus leverages plots provided by the algorithms. Residuals were used to detect outliers in the dependent variable, while leverages were used to detect outliers in the predictor variables. Since outliers have a tendency to "nest", residuals vs. leverages plots were ran a few times to ensure outliers were identified and subsequently removed from the data set. Before running the PLS regression analysis, 11 outliers were removed from the brook trout data set either because they were outside the ± 2.00 standardized residuals bound or they have extreme leverage values. After weeding out the 11 outliers the residuals vs. leverages plot and the distance plot were as follow. Note that points in the distance plot were randomly scattered and no obvious clustering was evident:

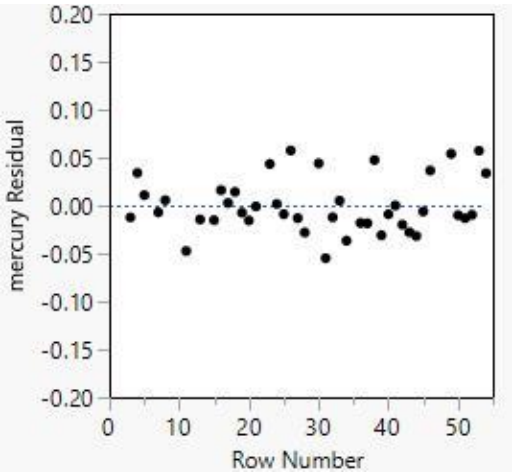


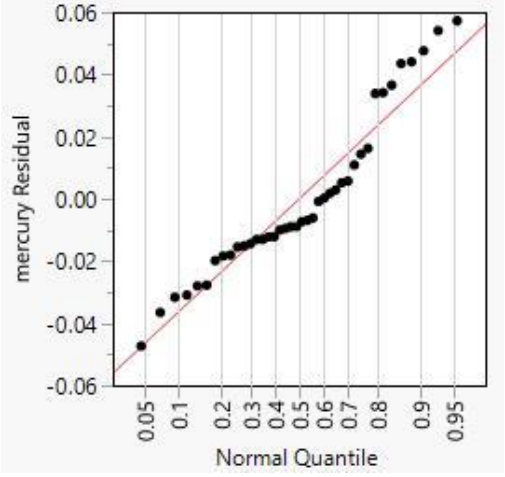


Four Diagnostic plots:









Brook trout (11 outliers removed, n = 42)
PLS Regression: mercury versus fork length, weight, age

Method

Cross-validation	Leave-one-out
Components to evaluate	User specified
Number of components evaluated	3
Number of components selected	1

Analysis of Variance for mercury

Source	DF	SS	MS	F	P
Regression	1	0.0274653	0.0274653	33.41	0.000
Residual Error	40	0.0328808	0.0008220		
Total	41	0.0603461			

Model Selection and Validation for mercury

Components	X Variance	Error	R-Sq	PRESS	R-Sq (pred)
1	0.945004	0.0328808	0.455130	0.0357319	0.407884
2		0.0324743	0.461865	0.0398885	0.339004
3		0.0321843	0.466671	0.0397412	0.341446

Coefficients

	mercury	mercury standardized
Constant	0.0198197	0.000000
fork length	0.0001204	0.240969
weight	0.0000440	0.231967
age	0.0081322	0.220774

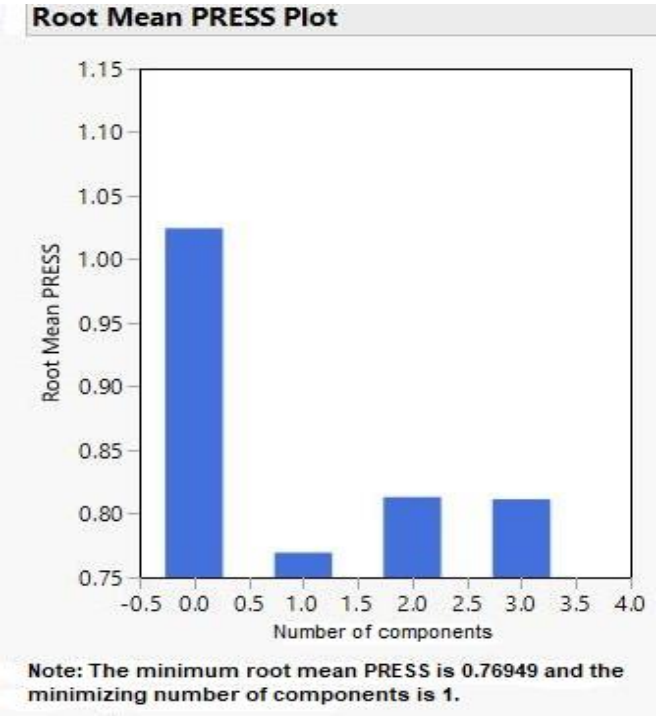
Using leave – one – out cross validation method, NIPALS algorithm selected the one – component model as the optimal model which has the highest R – Sq (pred) of 0.407884. For two – or three – component models, the R – Sq (pred) decrease measurably and might run the risk of overfitting, less robust model and poor predictivity. The ANOVA table showed that the model was significant (p = 0.000). The X Variance indicated the amount of variance in the predictors that was explained by the model. In this case, the one – component model explained 94.5% of the variance in the predictors.

Model Comparison Summary

Method	Number of rows	Number of components	Percent Variation Explained for Cumulative X	Percent Variation Explained for Cumulative Y	Number of VIP > 0.8
NIPALS	42	1	94.500438	45.51301	3

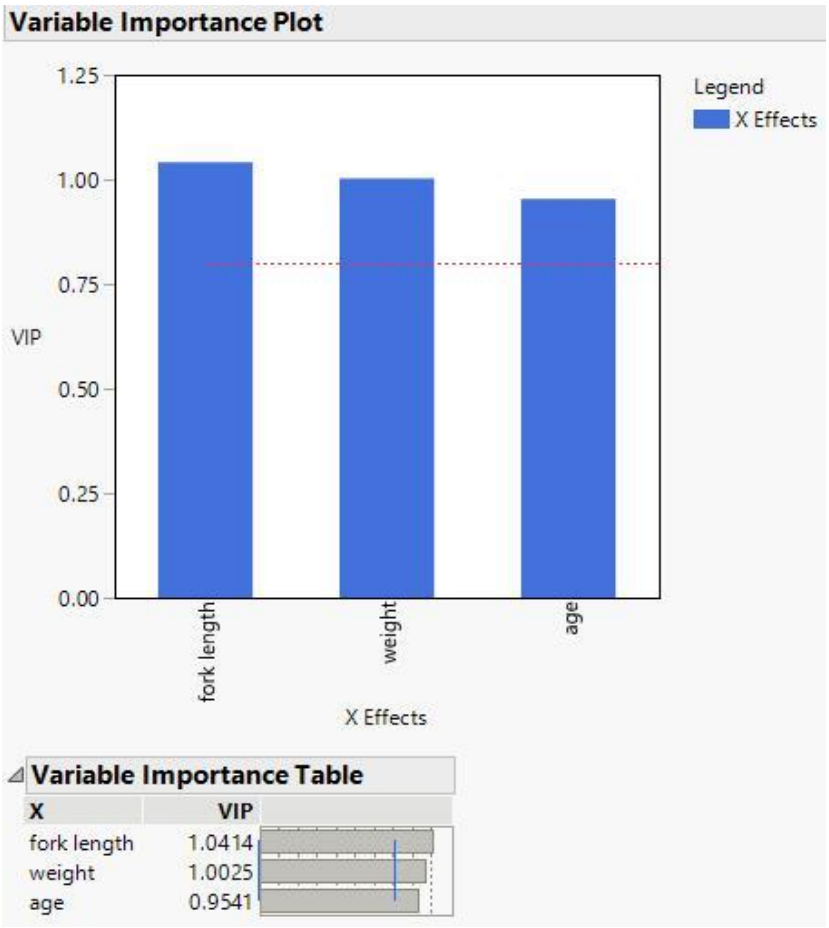
Cross Validation with Method=NIPALS

Number of components	Root Mean PRESS	van der Voet T^2	Prob > van der Voet T^2	Q^2	Cumulative Q^2	R^2X	Cumulative R^2X	R^2Y	Cumulative R^2Y
0	1.024390	7.512963	0.0030*	-0.049375	-0.049375	0.000000	0.000000	0.000000	0.000000
1	0.769491	0.000000	1.0000	0.407884	0.407884	0.944989	0.944989	0.455254	0.455254
2	0.813017	4.179729	0.0360*	0.339004	0.608614	0.039844	0.984833	0.007793	0.463047
3	0.811514	3.439118	0.0640	0.341446	0.742251	0.015167	1.000000	0.004606	0.467652

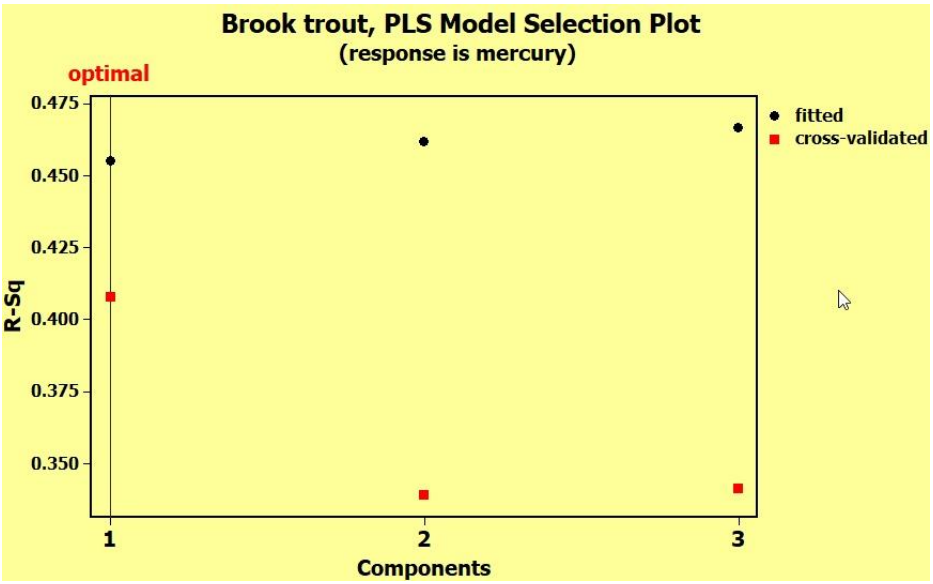


The root mean PRESS (predicted residual sum of squares) was minimum (0.769491) for the one – component model. However, the van der Voet's T^2 statistic suggested that both two – and three – component models (p – values of 0.036 and 0.064 respectively) were not different significantly (at 0.1 level of significance) from the one – component model with the minimum PRESS value. For the one – component optimal model, 94.5% of the variation in X (R^2X) and 45.5% of the variation in Y (R^2Y) were explained by the model. The Number of VIP > 0.8 (the number of model effects with variable importance for projection values greater than the 0.8 threshold) showed that all three predictors (fork length, weight and age) were important in the one – component model. Q2 which is a measure of the predictive ability of a model was higher for the one – component model than both the two – and three – component models.

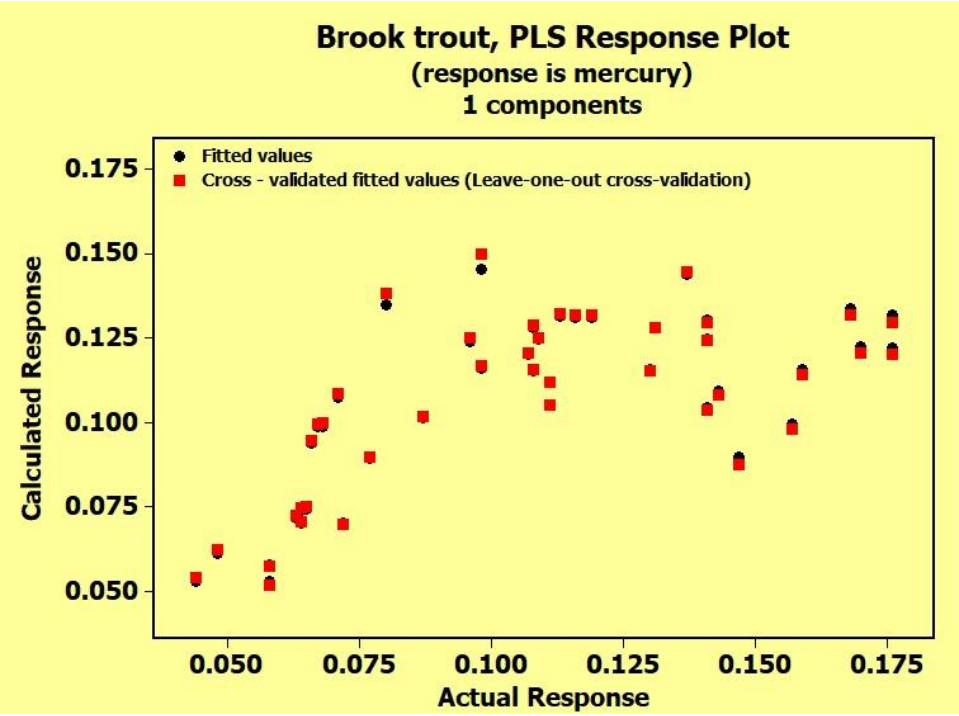
The variable importance plot graphed the VIP (variable importance in the projection) for each of the predictors. The variable importance table showed the VIP scores. Based only on VIP scores, fork length, weight and age were pretty much equally important in modeling the mercury in brook trout. The VIP scores for all three predictors were above the 0.8 threshold. This was also illustrated by their relatively similar standardized regression coefficients: fork length (0.240969), weight (0.231967) and age (0.220774). All three predictors have positive correlations with mercury in brook trout.



Based on the results of the leave – one – out cross validation method and the van der Voet's T^2 statistic, the model selection plot indicated the optimal model has one component. The R – Sq for the cross – validated mercury data dropped significantly for the two – and three – component models:



The response plot indicated that the model fitted the data adequately. Although there were differences between the fitted and cross – validated fitted response, none were severe enough to indicate an extreme leverage point.



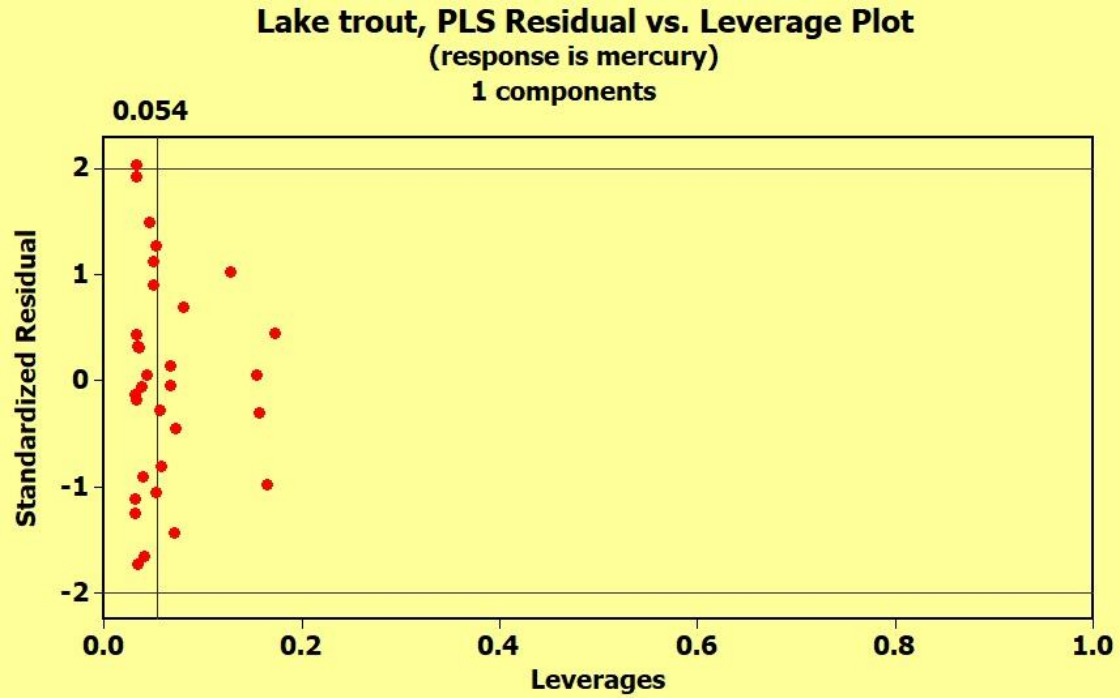
Cross – validation is very important procedure in PLS regression. The objective of cross – validation is to test the model's ability to predict **new** data that were not used in estimating the model, in order to flag problems such as overfitting and to give an insight on how well the model will generalize to an unknown data set (Cawley and Talbot 2010). Cross – validated fitted values indicate how well the model **predicts** data. The process consists of multiple rounds of repeated calculations which can be computationally intensive depending on the size of the data set: the data set is partitioned into two approximately equal – size subsets; analysis is performed on one subset (training set) and validation of the analysis is carried out on the other subset (testing set). Many rounds of analysis and validation are performed to reduce variability and the results are averaged to give an estimate of the model's predictive performance (Kohavi 1995). In PLS, the cross – validated fitted value is the predicted response for each observation in the data set, calculated individually so that the observation can be excluded from the model used to calculate the predicted response for that observation. The cross – validated fitted values are calculated during cross – validation and vary based on how many observations are omitted each time the model is recalculated.

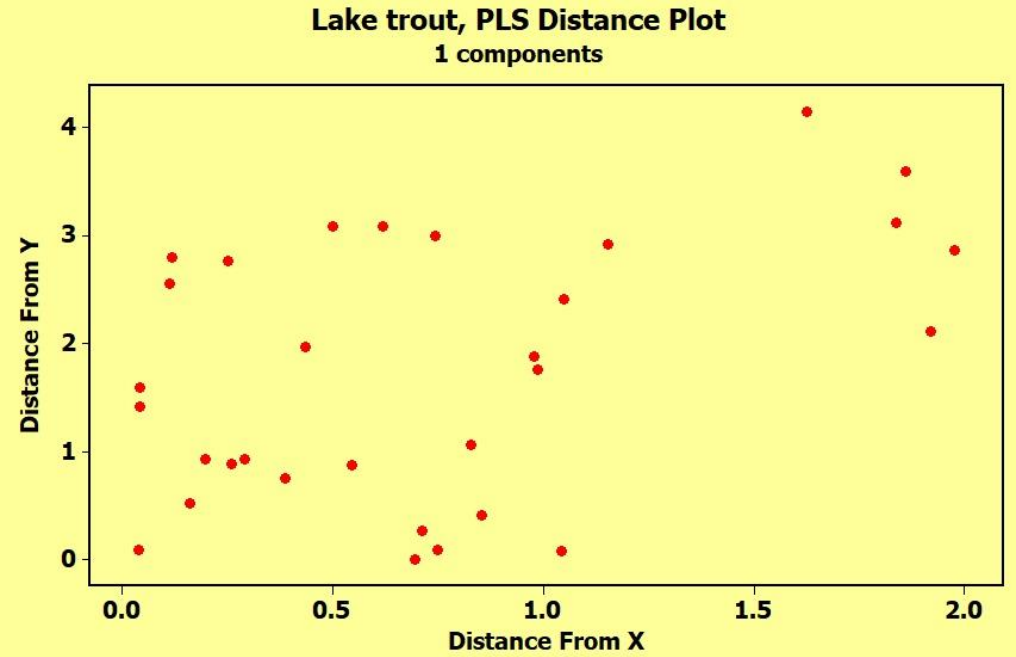
In PLS regression, the emphasis is on developing predictive model. Unlike other regression techniques such as stepwise regression, best-subset regression, and LASSO regression; PLS is not usually used to screen out predictors that are not useful in explaining the depending variables. Fork length, weight and age were all included in the PLS model for mercury in brook trout:

mercury = $(0.01982) + (1.2 * 10^{-4}) *$ fork length $+ (4.4 * 10^{-5}) *$ weight $+ (8.13 * 10^{-3}) *$ age

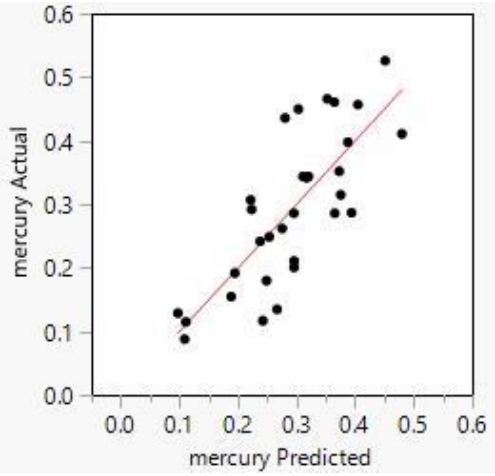
4.2 PLS regression analysis of lake trout data:

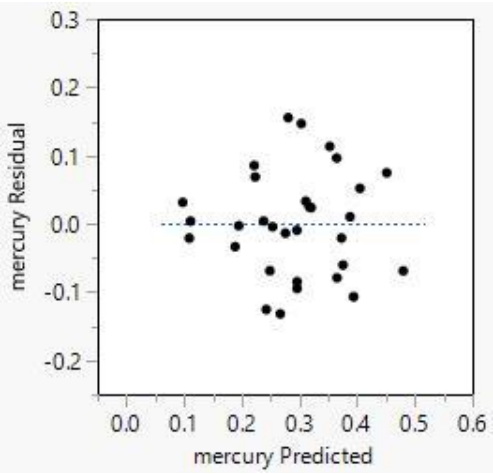
Four outliers were identified and removed from the lake trout data set using the residuals vs. leverages plot provided by the PLS algorithm and after which the final residuals vs. leverages plot and the distance plot are as follow:

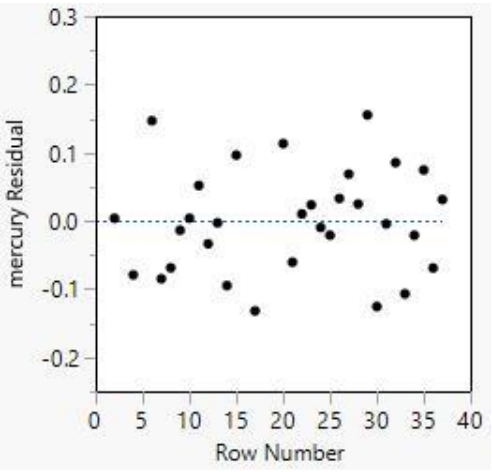


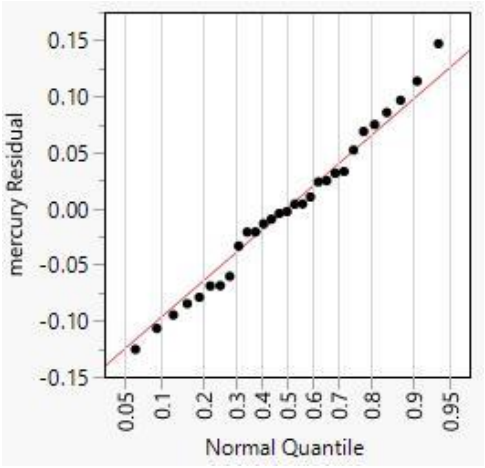


Four diagnostic plots:









Lake trout (4 outliers removed, n = 31)
PLS Regression: mercury versus fork length, weight, age

Method

Cross-validation	Leave-one-out
Components to evaluate	User specified
Number of components evaluated	3
Number of components selected	1

Analysis of Variance for mercury

Source	DF	SS	MS	F	P
Regression	1	0.266366	0.266366	44.37	0.000
Residual Error	29	0.174092	0.006003		
Total	30	0.440458			

Model Selection and Validation for mercury

Components	X Variance	Error	R-Sq	PRESS	R-Sq (pred)
1	0.922102	0.174092	0.604747	0.194433	0.558565
2		0.170805	0.612210	0.203354	0.538313
3		0.170495	0.612914	0.213475	0.515333

Coefficients

mercury

	mercury	standardized
Constant	-0.0546936	0.000000
fork length	0.0004476	0.281683
weight	0.0000991	0.277974
age	0.0148045	0.249368

The three models cross – validated yielded very similar R -Sq (pred) values. Nonetheless, NIPALS algorithm selected the one – component model as the optimal model which has a higher R – Sq (pred) of 0.558565 slightly higher than that of the two – and three component models. The ANOVA table showed that the model was significant (p = 0.000). The X Variance indicated the amount of variance in the predictors that was explained by the model. In this case, the one – component model explained 92.2% of the variance in the predictors. Also, the root mean PRESS of the one – component model was only slightly lower than that of the two – component model. However, the van der Voet's T^2 statistic suggested that the two – component model (p – values of 0.4000) was different significantly (at 0.1 level of significance) from the one – component model with the minimum PRESS value. The three – component model (p = 0.088) might also be different from the one – component model. The van der Voet's T^2 statistic supported the notion that the one – component model was optimal. For the one –

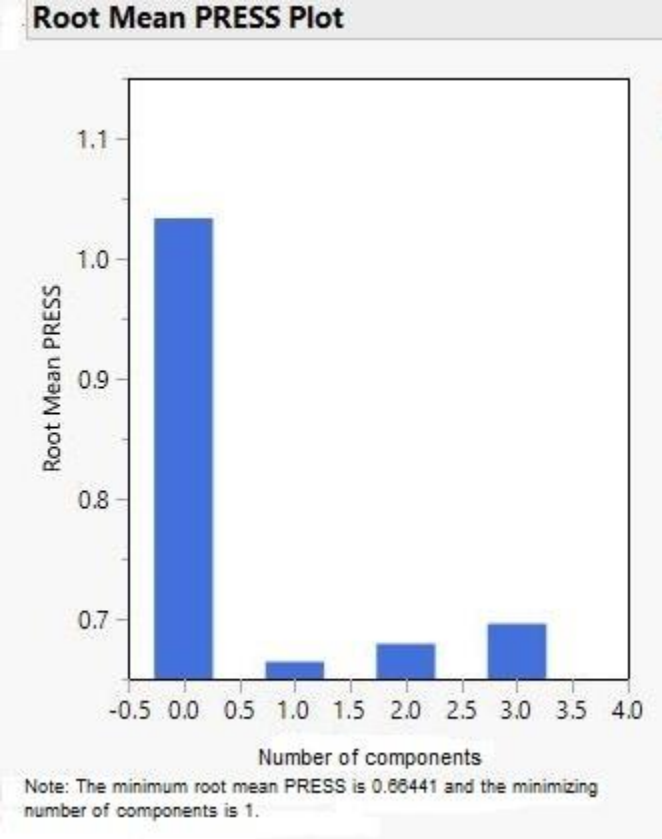
component optimal model, 92.2% of the variation in X (R^2X) and 60.5% of the variation in Y (R^2Y) were explained by the model.

Model Comparison Summary

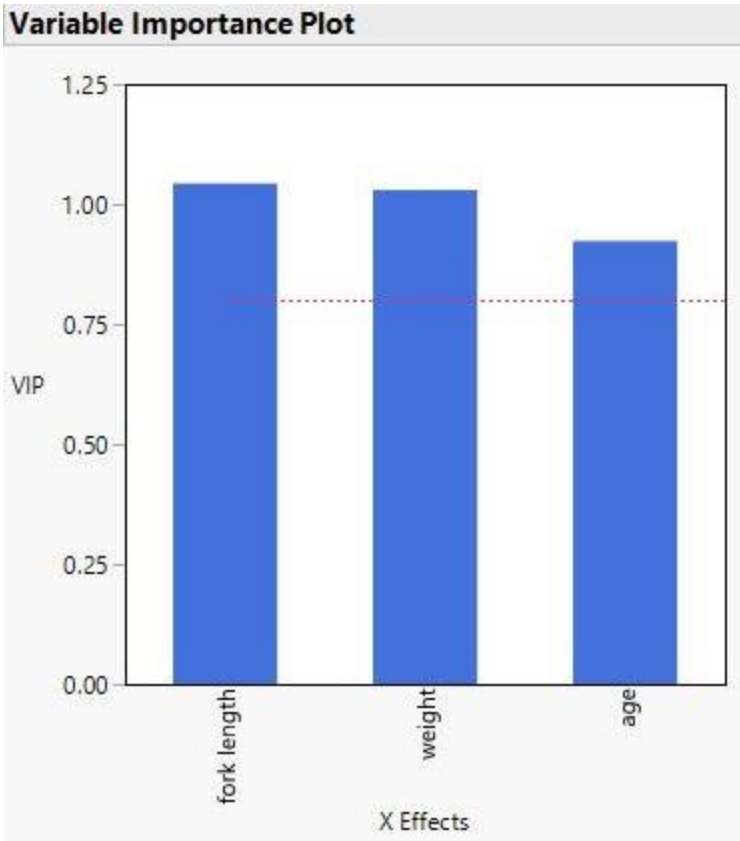
Method	Number of rows	Number of components	Percent Variation Explained for Cumulative X	Percent Variation Explained for Cumulative Y	Number of VIP > 0.8
NIPALS	31	1	92.210232	60.474706	3

Cross Validation with Method=NIPALS

Number of components	Root Mean PRESS	van der Voet T^2	Prob > van der Voet T^2	Q^2	Cumulative Q^2	R^2X	Cumulative R^2X	R^2Y	Cumulative R^2Y
0	1.033333	8.271719	0.0020*	-0.067778	-0.067778	0.000000	0.000000	0.000000	0.000000
1	0.664405	0.000000	1.0000	0.558565	0.558565	0.922053	0.922053	0.604749	0.604749
2	0.679476	0.869948	0.4000	0.538313	0.796195	0.065281	0.987334	0.007928	0.612678
3	0.696181	2.606249	0.0880	0.515333	0.901223	0.012666	1.000000	0.000985	0.613662



The Number of VIP > 0.8 (the number of model effects with variable importance for projection values greater than the 0.8 threshold) showed that all three predictors (fork length, weight and age) were important in the one – component model. Q2 which is a measure of the predictive ability of a model was slightly higher for the one – component model than both the two – and three – component models.

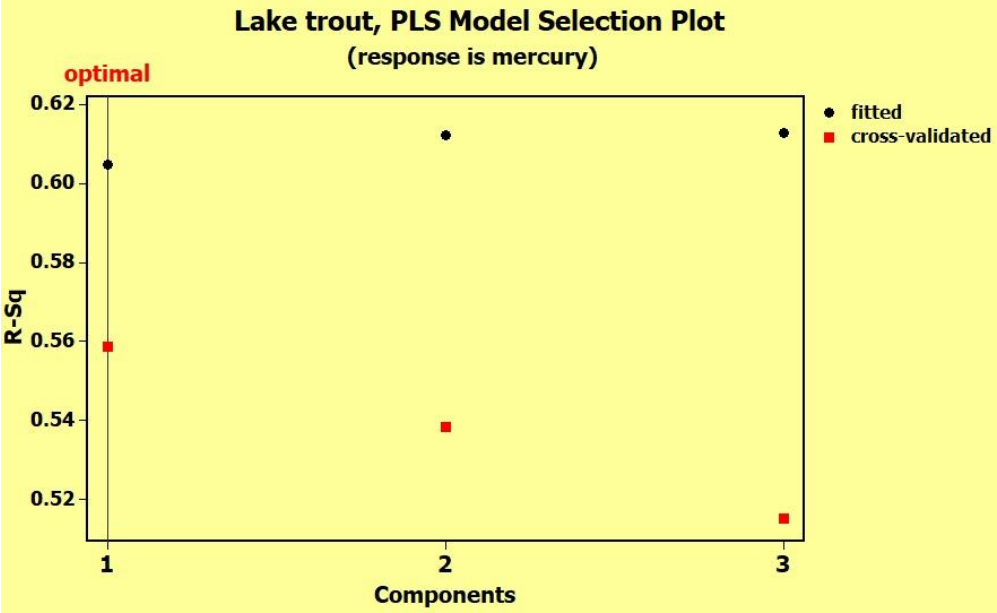


Variable Importance Table

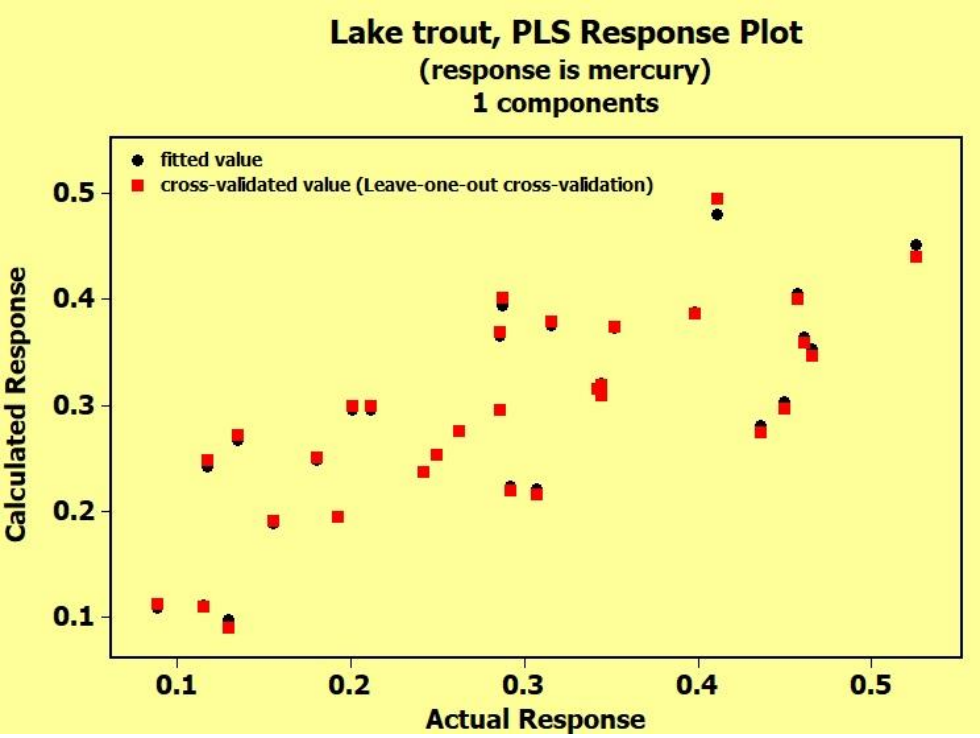
X	VIP
fork length	1.0430
weight	1.0293
age	0.9234

Based on VIP scores, fork length, weight and age were pretty much equally important in modeling the mercury in lake trout. The VIP scores for all three predictors were above the 0.8 threshold. This was also illustrated by their relatively similar standardized regression coefficients: fork length (0.281683), weight (0.277974) and age (0.249368). All three predictors have positive correlations with mercury in lake trout.

Based on the results of the leave – one – out cross validation method and the van der Voet's T^2 statistic, the model selection plot indicated the optimal model has one component. The R – Sq for the cross – validated mercury data dropped for the two – and three – component models:



The response plot indicated that the model fitted the data adequately. Although there were differences between the fitted and cross – validated fitted response, none were severe enough to indicate an extreme leverage point.

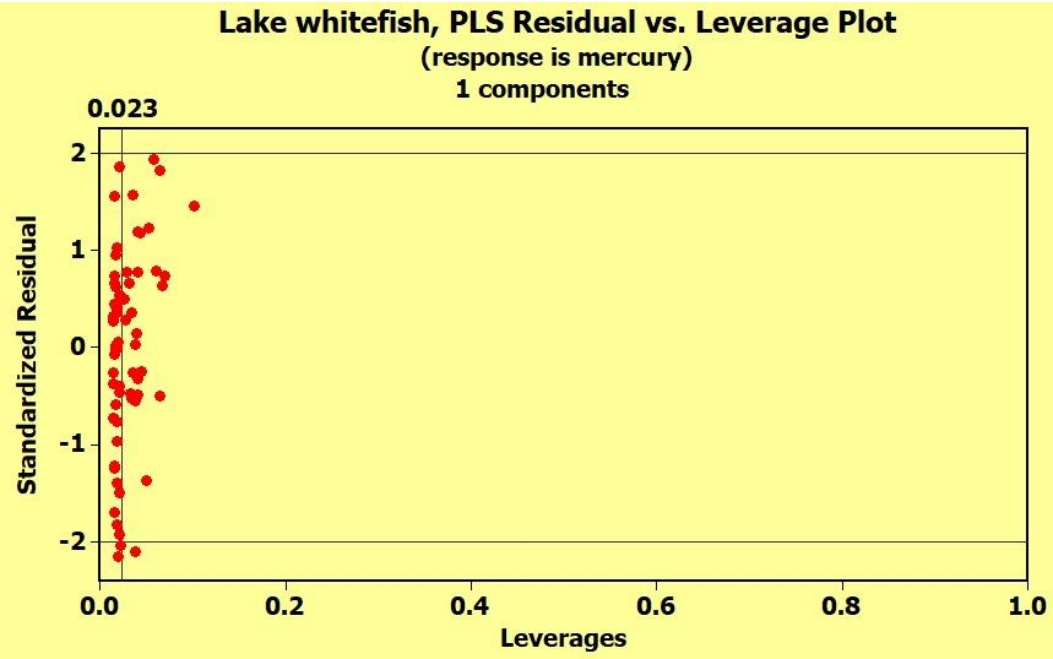


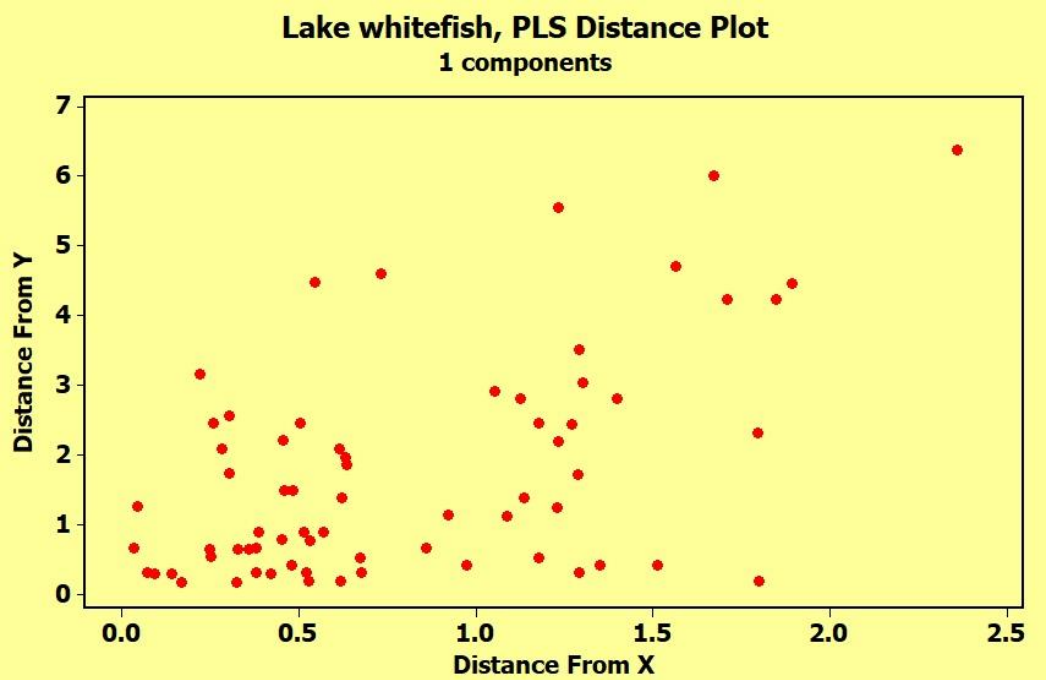
Fork length, weight and age were all included in the PLS model for mercury in lake trout:

mercury = (-0.05469) + ($4.47 * 10^{-4}$) * fork length + (10^{-4}) * weight + (0.0148) * age

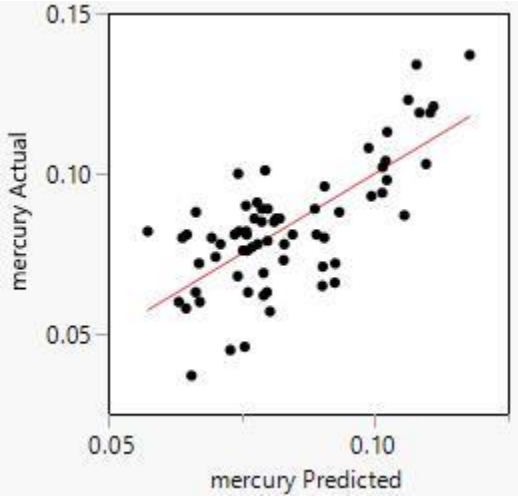
4.3 PLS regression analysis of lake whitefish data:

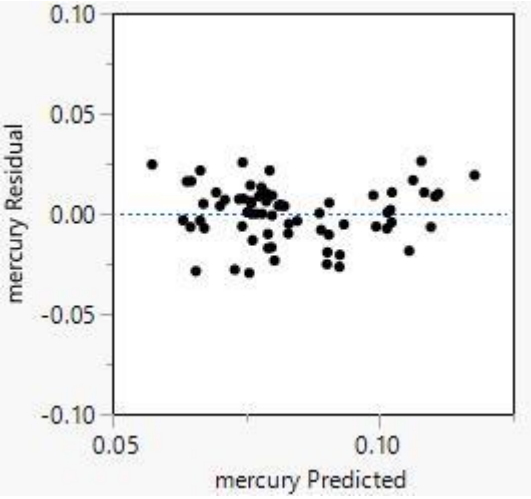
Untransformed full data set of the lake whitefish mercury and fish growth data was examined for the presence of outliers using residuals versus leverages plots. In total, 19 outliers were identified and removed after several rounds of running the residuals vs. leverage plots. After weeding out the outliers the final residuals vs. leverages plot and the distance plot were as follow:

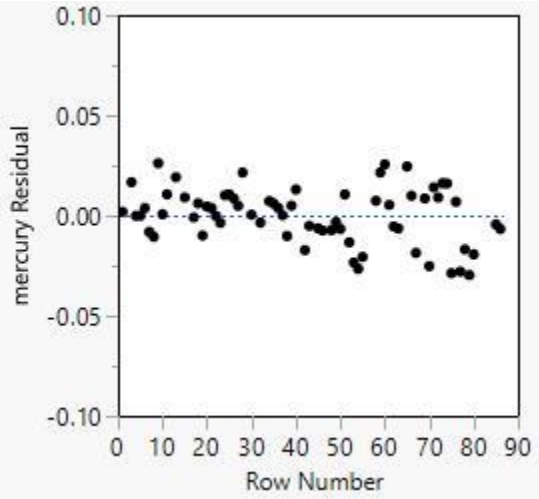


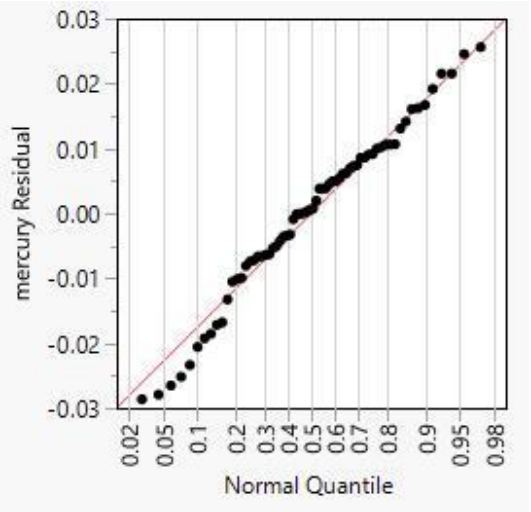


Four diagnostic plots:

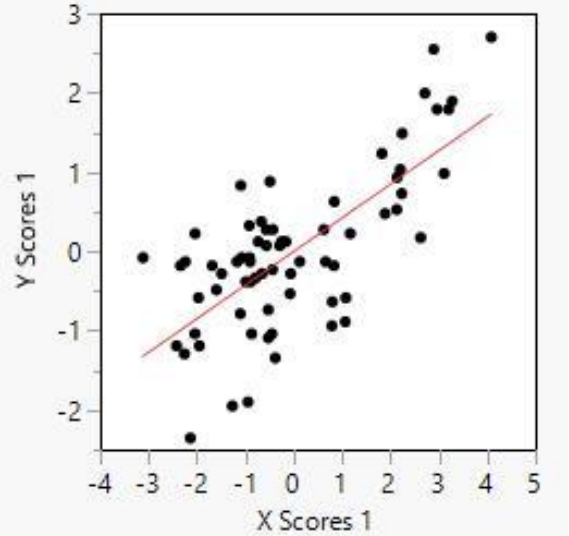








X-Y Scores Plots



Lake whitefish (19 outliers removed, n = 68)
PLS Regression: mercury versus fork length, weight, age

Method

Cross-validation	Leave-one-out
Components to evaluate	Set
Number of components evaluated	3
Number of components selected	1

Analysis of Variance for mercury

Source	DF	SS	MS	F	P
Regression	1	0.0135577	0.0135577	70.70	0.000
Residual Error	66	0.0126572	0.0001918		
Total	67	0.0262149			

Model Selection and Validation for mercury

Components	X Variance	Error	R-Sq	PRESS	R-Sq (pred)
1	0.960384	0.0126572	0.517176	0.0135394	0.483520
2		0.0124735	0.524183	0.0142038	0.458177
3		0.0123958	0.527148	0.0140900	0.462518

Coefficients

	mercury	mercury standardized
Constant	0.0306402	0.000000
fork length	0.0000859	0.243796
weight	0.0000219	0.250950
age	0.0030605	0.239013

Similar to the lake trout data set, the three models cross – validated yielded very similar R -Sq (pred) values. Nonetheless, NIPALS algorithm selected the one – component model as the optimal model which has a higher R – Sq (pred) of 0.483520 slightly higher than that of the two – and three component models. The ANOVA table showed that the model was significant (p = 0.000). The X Variance indicates the amount of variance in the predictors that is explained by the model. In this case, the one – component model explained 96.0% of the variance in the predictors. Also, the root mean PRESS of the one – component model was only slightly lower than that of the two – component model. However, the van der Voet's T^2 statistic suggested that the three – component model (p – values of 0.2080) was different significantly (at 0.1 level of significance) from the one – component model with the minimum PRESS value. The two – component model (p = 0.0690) was not significantly different from the one – component model. The similarity of R – Sq (pred) and the root mean PRESS between the one component

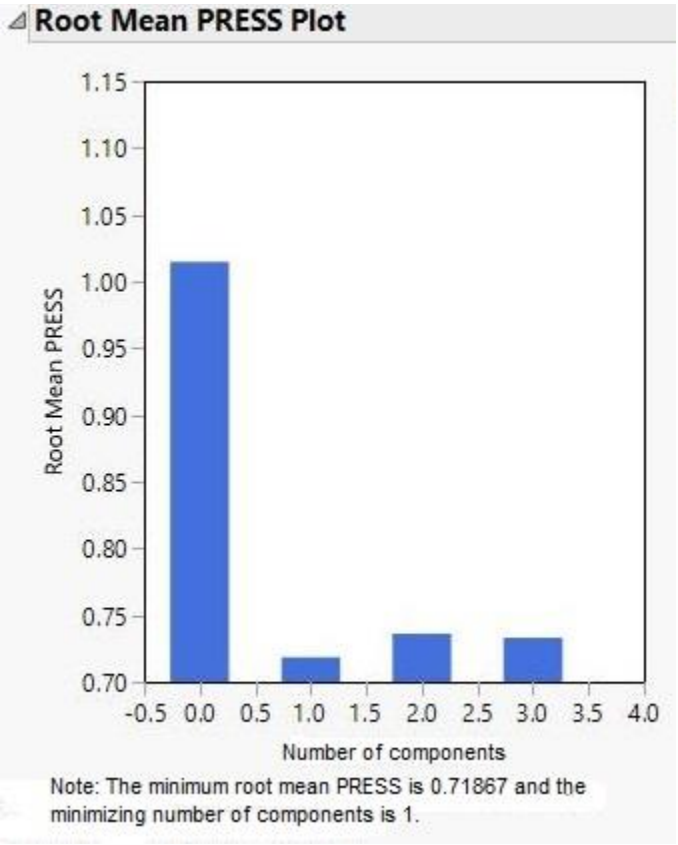
model and the two – component model and the van der Voet's T^2 statistic results suggested that a two – component model was also a possibility and was unlikely leading to overfitting. Nonetheless, a one – component model was selected. For the one – component optimal model, 96.0% of the variation in X (R^2X) and 51.7% of the variation in Y (R^2Y) were explained by the model.

Model Comparison Summary

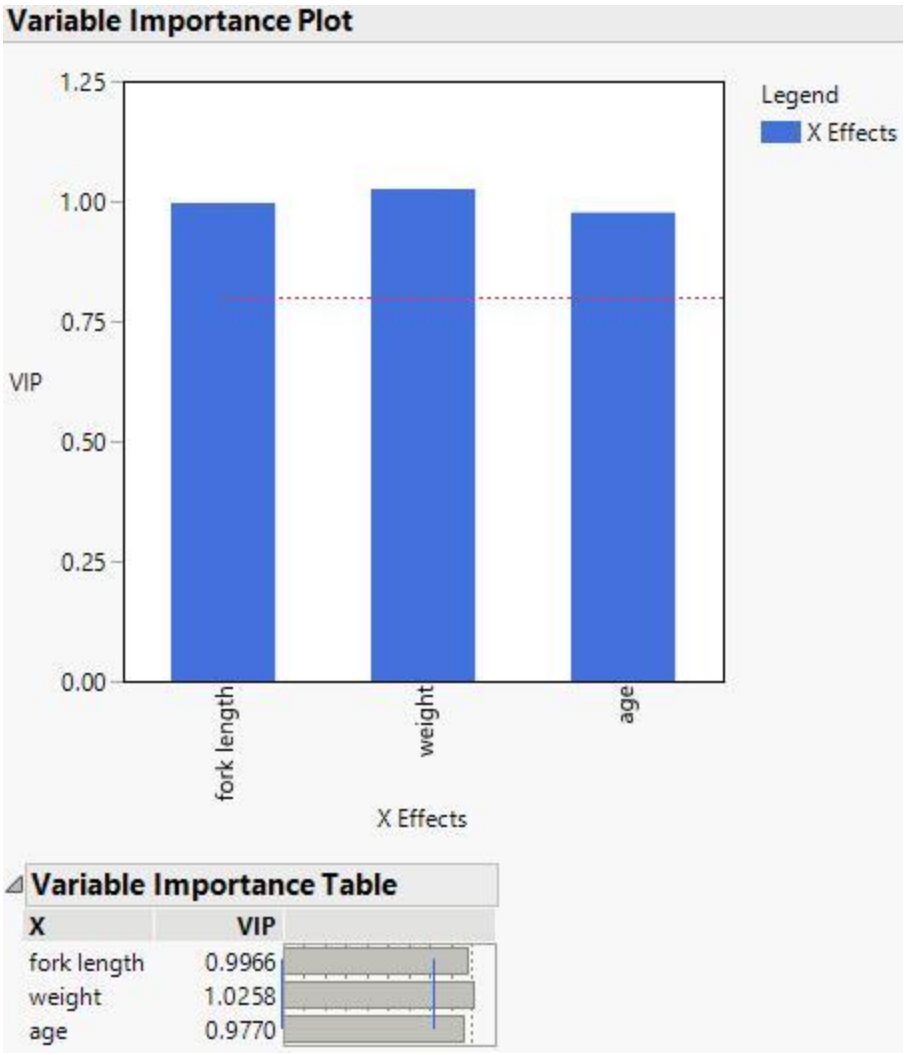
Method	Number of rows	Number of components	Percent Variation Explained for Cumulative X	Percent Variation Explained for Cumulative Y	Number of VIP > 0.8
NIPALS	68	1	96.038377	51.71765	3

Cross Validation with Method=NIPALS

Number of components	Root Mean PRESS	van der Voet T^2	Prob > van der Voet T^2	Q^2	Cumulative Q^2	R^2X	Cumulative R^2X	R^2Y	Cumulative R^2Y
0	1.014925	7.921070	0.0040*	-0.030074	-0.030074	0.000000	0.000000	0.000000	0.000000
1	0.718665	0.000000	1.0000	0.483520	0.483520	0.960384	0.960384	0.517143	0.517143
2	0.736086	3.381464	0.0690	0.458177	0.720160	0.021455	0.981840	0.007353	0.524495
3	0.733132	1.622862	0.2080	0.462518	0.849591	0.018160	1.000000	0.002838	0.527333



The Number of VIP > 0.8 shows that all three predictors (fork length, weight and age) were important in the one – component model. Q2 which is a measure of the predictive ability of a model was only slightly higher for the one – component model than both the two – and three – component models.

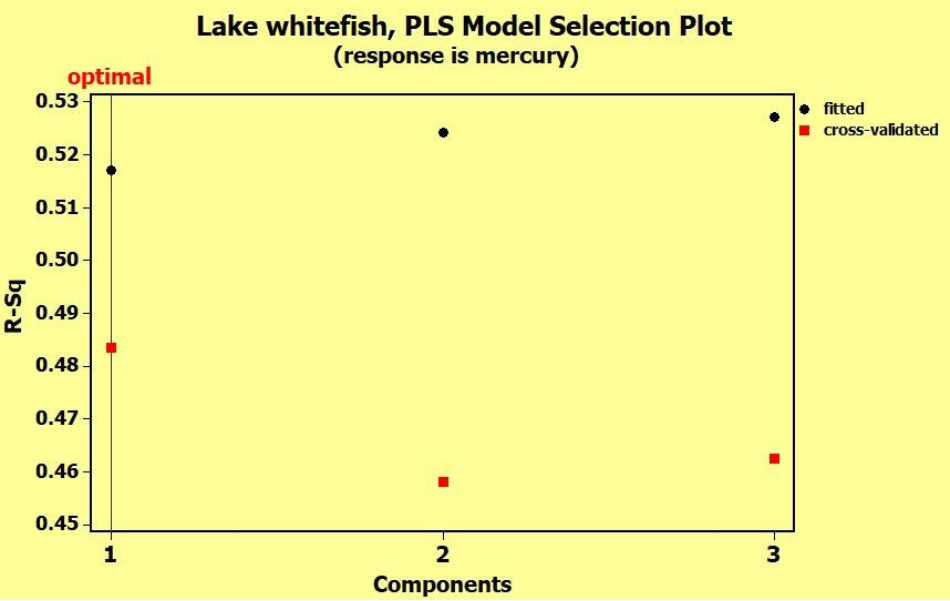


Variable Importance Table

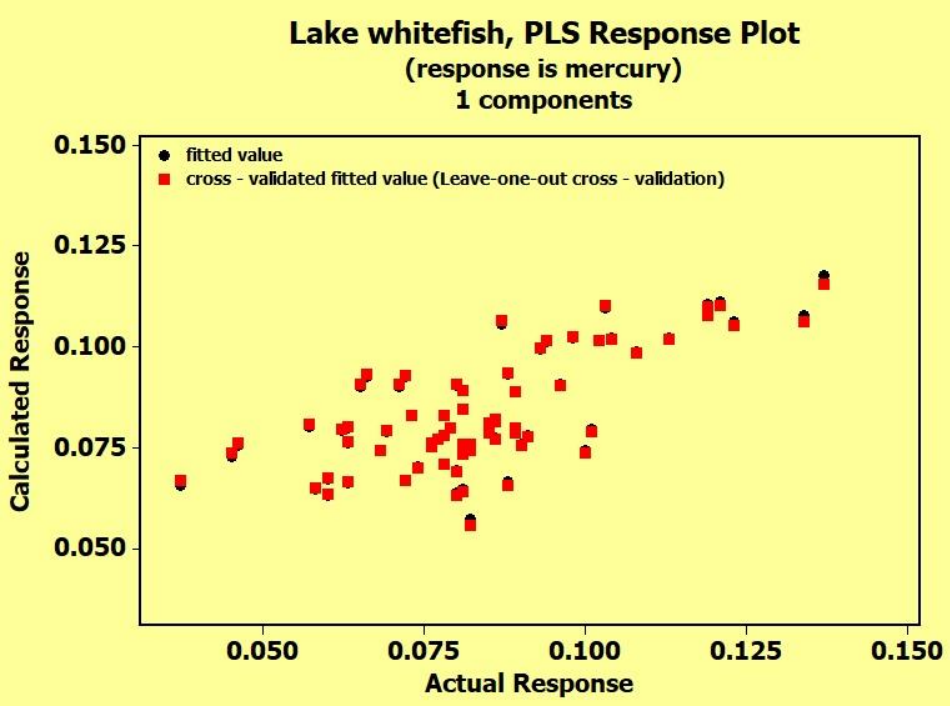
X	VIP
fork length	0.9966
weight	1.0258
age	0.9770

Based on VIP scores, fork length, weight and age were pretty much equally important in modeling the mercury in lake whitefish. The VIP scores for all three predictors were above the 0.8 threshold. This was also illustrated by their relatively similar standardized regression coefficients: fork length (0.243796), weight (0.250950) and age (0.239013). All three predictors have positive correlations with mercury in lake whitefish.

Based on the results of the leave – one – out cross validation method and the van der Voet's T^2 statistic, the model selection plot indicated the optimal model has one component. The R – Sq for the cross – validated mercury data dropped for the two – and three – component models:



The response plot indicated that the model fitted the data very well. Only very few differences between the fitted and cross – validated fitted responses were found:



Fork length, weight and age were all included in the PLS model for mercury in lake whitefish:

mercury = $(0.03064) + (8.6 * 10^{-5}) *$ fork length $+ (2.2 * 10^{-5}) *$ weight $+ (3.06 * 10^{-3}) *$ age

Chapter 5. Bayesian Approach to Regression Modeling: No more levels of significance and p – values, no more testing of the null hypothesis.

In this chapter I rejected the classical frequentist approach to statistical inference which has dominated the field of applied statistics since the beginning and instead adopted the Bayesian rational belief revision approach based on subjective probability in modeling the regression on the fish data. No more setting the "rigid" uncompromising levels of significance and making decisions based on the "tricky" p – value concept which often lands unsuspected researchers and indeed, fellow statisticians into fallacious traps and drawing wrong conclusions due to ballooning of Type I Error without even realizing it. One big practical advantage of the Bayesian approach relative to the frequentist approach is that it actually put a number to quantify the probability or the odds that null hypothesis is true. As for the frequentists, they have to make a clear-cut choice of accepting or rejecting the null hypothesis at a **predetermined** level of significance. Bayesian approach to probability is nothing new. Bayes' theorem as an axiom of the probability theory has been around for over two hundred years (Bayes 1763). It was not until the recent three decades or so Bayesian approach in solving practical statistical problems becoming popularized mainly due to the popularization of high-speed personal computers which greatly benefit the execution of simulation algorithms such as various forms of Markov - chain Monte Carlo (MCMC) simulation which is essential in generating numerical approximations of model's parameters and integrals from the posterior distribution established when solving complex multilevel modeling problems using the Bayesian approach.

5.1 The Bayesian Way:

The key ingredients to a Bayesian analysis are: (1) the prior distribution, which quantifies what is known based on theoretical reasonings of the researcher about the model's parameters and the uncertainty in the values of the unknown model parameters **before** observing the data, and (2) observe the data to formulate the likelihood function, which reflects information about the model's parameters contained in the data. The prior distribution and likelihood function can then be used to estimate (3) the posterior distribution, which represents total knowledge about the model's parameters **after** the data have been observed. The posterior distribution quantifies the uncertainty in the values of the unknown

model's parameters. Summaries of the posterior distribution can then be used to calculate quantities of interest and ultimately to draw inferential conclusions about the model such as point estimates, interval estimates and posterior probabilities of competing hypotheses. In practice, getting anything meaningful out of the posterior distribution is often computationally intensive and time consuming except for very simple models. This indeed was the main obstacle to the popularization of the Bayesian approach until the advancement of high-speed computer – based algorithms in the last three decades or so such as various samplers based on the original MCMC stochastic simulation method. These algorithms construct Markov chains and perform random walks (or draws) simulating samplings from complex posterior distributions to provide numerical approximations to model's parameters and integrals estimated from the probability distribution especially in complex multidimensional hierarchical problems when analytical solutions become intractable or too time consuming to achieve. In the present case, our previous knowledge regarding bioaccumulation of mercury in fish has given us sound theoretical grounds to the notion that mercury level in fish increases as the fish grows, (i.e. increase in age and size) is our ***prior*** in the Bayesian sense. The likelihood function is then established and the posterior distribution estimated. Our belief in our ***prior*** will be revised upon the outcome from the posterior distribution.

5.2 Bayesian Linear Regression:

In Bayesian regression, the regression coefficients are assumed to be random variables with a specified prior distribution. The prior distribution can "bias" the solutions for the regression coefficients. The Bayesian estimation process produces not a single point estimate for the "best" values of the regression coefficients but an entire posterior distribution, completely describing the uncertainty surrounding the predictors. Bayesian regression is used as a kind of regularization technique to bias the model towards assuming uncorrelated residuals.

Predictive inference is a straightforward computation once the posterior distribution has been obtained. Say, we like to make a prediction on a future observation y based on the observed data we already have $\boldsymbol{y}$ = ($y_{1,\ldots\ldots\ldots}y_n$). From the analysis of the data, we have obtained the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. We want to make probabilistic statements about a not – yet observed y , so that we want to compute the *posterior predictive distribution* of y which is $p(y \mid \boldsymbol{y})$. Note that we are not interested

in conditioning on model parameter values, but that we only want to condition on what we have observed: i.e. previous data. The *posterior predictive distribution* can be computed using the equation:

$$p(y|y)=\int p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|y)d\boldsymbol{\theta}$$

which make the appropriate assumption that future data are independent of past data conditional on the model parameters. Therefore, integrating the product of the data model distribution with the posterior distribution with respect to the model parameters produces the *posterior predictive distribution*, which can then be summarized for predictive inference: i.e. we have our **Bayesian linear regression model.** The classical frequentist approach to linear regression makes the assumption that there are enough observations to say something meaningful about regression coefficients. Whereas in the Bayesian approach, the data are supplemented with additional information in the form of a prior probability distribution. The prior belief about the model's parameters is combined with the data's likelihood function according to Bayes theorem to yield the posterior belief about the model's parameters (regression coefficients and the population standard deviation). Like the frequentist's OLS models, there are assumptions that the data need to meet in order for Bayesian linear regression to be valid: (1) the dependent variable has to be continuous; (2) The dependent variable is by and large linearly related to all predictors and the effects of the predictors are additive; (3) the residuals (errors) for each predictor are uncorrelated with each other; (4) the error variance of each predictor is constant across all values of that predictor (i.e. homoscedasticity) and (5) the residuals (errors) of predictors are normally distributed with mean zero.

5.3 Regression Modeling of Fish Data using the Bayesian Approach:

The BAS algorithms package (Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling) written in **R** by Professor Merlise Clyde of Duke University (Clyde et. al. 2010) was used to model the brook trout and lake trout data sets both of which satisfied the assumptions needed to apply Bayesian linear regression.

5.3.1 Bayesian linear regression of brook trout data set:

Model Comparison

Models	P(M)	P(M|data)	BF_M	BF_{01}	R^2
log length + log weight + age	0.250	0.987	235.188	1.000	0.774
age	0.083	0.009	0.101	36.004	0.667
log length + age	0.083	0.001	0.015	236.478	0.669
log weight + age	0.083	0.001	0.015	240.685	0.668
log length + log weight	0.083	6.911e −4	0.008	476.247	0.656
log length	0.083	3.043e −6	3.347e −5	108160.870	0.502
log weight	0.083	1.513e −7	1.664e −6	2.176e +6	0.419
Null model	0.250	6.761e −11	2.028e −10	1.460e +10	0.000

Each row in the table above represent one model. The Null model only contain the intercept. The column P(M) shows the prior probability of the respective models. The P(M|Data) and BF_M provide the posterior model probability and the Bayes factor for model M, respectively. The posterior model probabilities express the probability of a model after seeing the data. The Bayes factor quantifies the strength of evidence produced by the data and hence it is a measure of the data – induced changes from prior model odds to posterior model odds. In a way, the Bayes factor serves a similar role as the p – value in the frequentists' hypothesis testing. The BF_{01} column indicates the Bayes factors of the models in each row compared to the model in the first row. Therefore, the first row always has $BF_{01} = 1$ because the model is compared against itself. Also, the model in the first row is the best model determined by the BAS algorithm that predict the data best. R^2 is the coefficient of determination of the particular model. For the brook trout data, the algorithm suggested that the best model to predict $\log_e$ mercury was one that contains all three predictors: $\log_e$ fork length, $\log_e$ weight and age which clearly out performs all other models: it explained 77.4% of the variance (highest R^2 amongst all models), as the data were over 36 times more likely to occur under this model than under the second – best model that only included age as the predictor ($BF_{01} = 36.004$ associated with the second model).

Posterior Summaries of Coefficients

						95% Credible Interval	
Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	Lower	Upper
Intercept	−2.169	0.034	1.000	1.000	1.000	−2.237	−2.100
log length	4.716	1.108	0.500	0.989	94.155	2.481	6.951
log weight	−1.484	0.349	0.500	0.989	93.909	−2.188	−0.780
age	0.248	0.055	0.500	0.999	1439.309	0.137	0.359

With the coefficient estimates (mean) from the table above, the algorithm suggested the following model:

$\log_e$ mercury = (-2.169) + (1.716) * $\log_e$ fork length + (– 1.484) * $\log_e$ weight + (0.248) * age

It is noteworthy that the estimated coefficient associated with $\log_e$ weight was negative, which seemingly an anomaly since we have shown in the scatterplot of $\log_e$ mercury versus $\log_e$ weight that there was a positive correlation between the two (see Chapter 1) albeit the points were rather scattered (R^2 only 37.6%). It was found that a Bayesian regression model of $\log_e$ weight as the sole predictor yielded a positive estimated coefficient of 0.368 for $\log_e$ weight:

Posterior Summaries of Coefficients

						95% Credible Interval	
Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	Lower	Upper
Intercept	−2.157	0.055	1.000	1.000	1.000	−2.267	−2.046
log weight	0.368	0.075	0.500	1.000	2083.081	0.217	0.520

As soon as $\log_e$ fork length and/or age were added to the model, the estimated coefficient of $\log_e$ weight became negative:

Posterior Summaries of Coefficients

						95% Credible Interval	
Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	Lower	Upper
Intercept	−2.157	0.043	1.000	1.000	1.000	−2.243	−2.070
log weight	−1.841	0.433	0.500	0.998	564.749	−2.713	−0.970
log length	6.726	1.298	0.500	1.000	9063.575	4.109	9.342

Posterior Summaries of Coefficients

						95% Credible Interval	
Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	Lower	Upper
Intercept	−2.169	0.041	1.000	1.000	1.000	−2.249	−2.096
log weight	−0.040	0.098	0.500	0.230	0.299	−0.157	0.065
age	0.333	0.061	0.500	1.000	78490.922	0.236	0.402

When we examined the changes in the Bayes factor (BF_{01}), data were over 6.7 times more likely to occur in a model with age as the only predictor than in a model with age and log_e weight as predictors. Data were over 60,000 times more likely to occur in a model with age as the sole predictor than in a model with log_e weight as the sole predictor:

Model Comparison

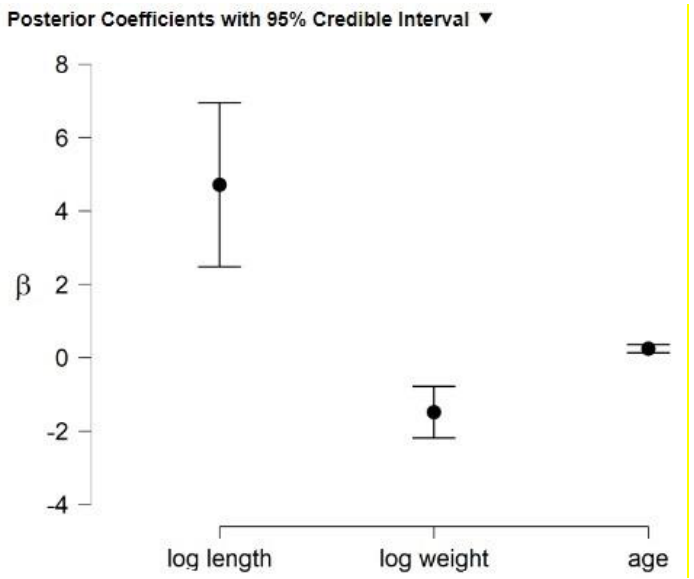
Models	P(M)	P(M|data)	BF_M	BF_{01}	R^2
age	0.167	0.770	16.711	1.000	0.667
log weight + age	0.333	0.230	0.598	6.685	0.668
log weight	0.167	1.274e −5	6.368e −5	60433.582	0.419
Null model	0.333	3.795e −9	7.590e −9	4.056e +8	0.000

For a model with log_e weight and log_e fork length as predictors, data were over 5.4 times more likely to occur in a model with both log_e weight and log_e fork length as predictors than in a model with log_e fork length as the only predictor. Data were over 84 times more likely to occur in a model with log_e weight and log_e fork length as predictors than in a model with log_e weight as the sole predictor:

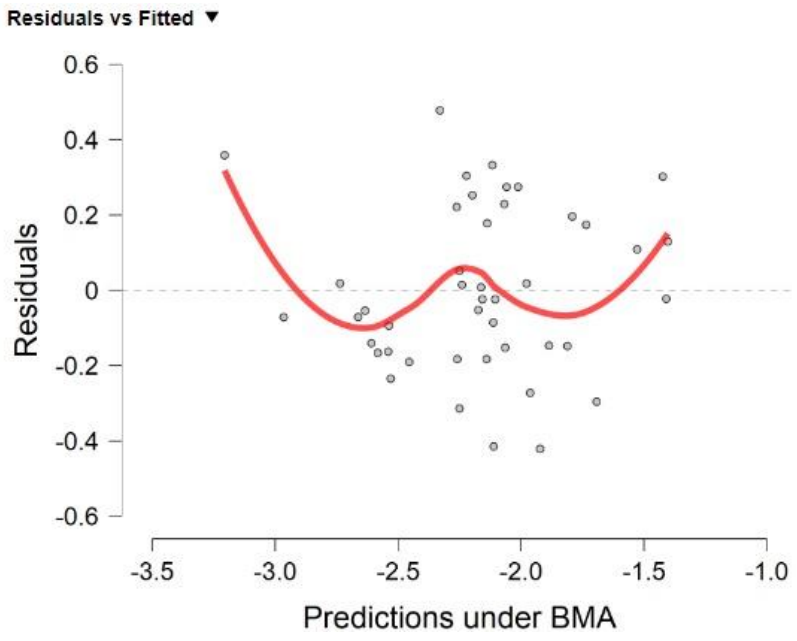
Model Comparison

Models	P(M)	P(M|data)	BF_M	BF_{01}	R^2
log weight + fork length	0.333	0.911	20.409	1.000	0.541
fork length	0.167	0.084	0.458	5.431	0.456
log weight	0.167	0.005	0.027	84.413	0.376
Null model	0.333	5.179e −6	1.036e −5	175838.388	0.000

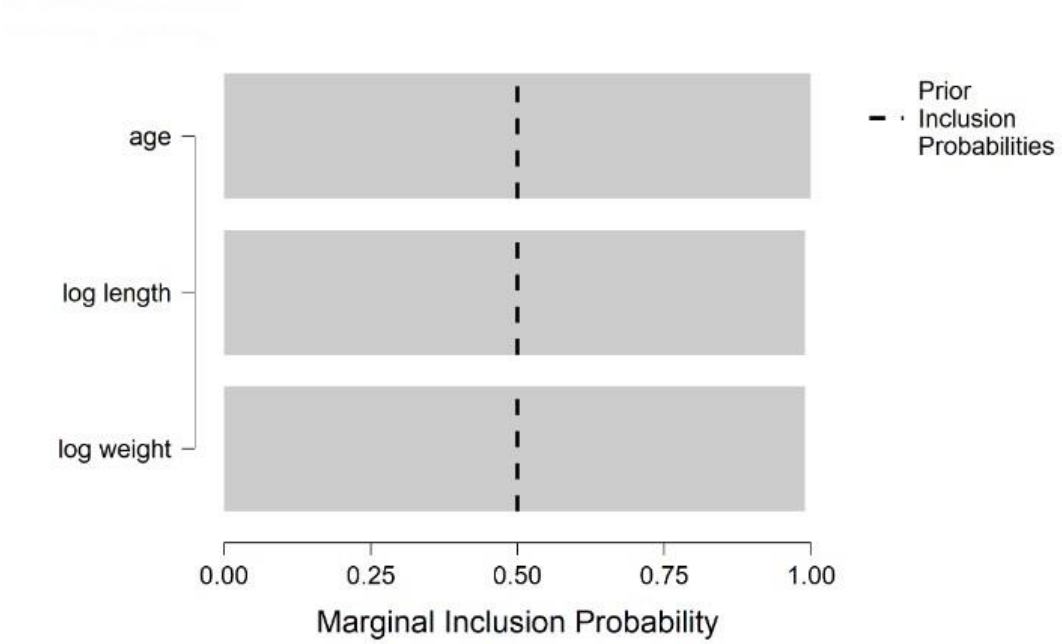
The error bars in the coefficient plot for the three - predictor model below indicated the 95% credible intervals. These were calculated by dropping 5% of the draws from each tail of the marginal distribution of each model's parameter:



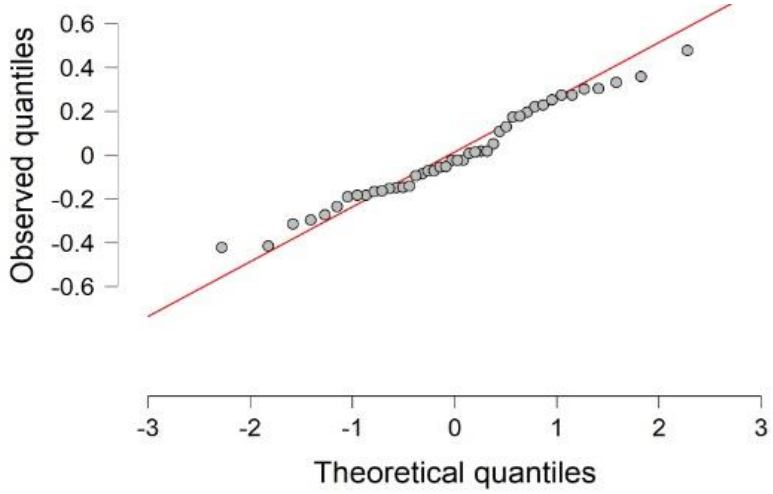
The plot of residuals of the model averaged predictions (BMA) versus the residuals for the three - predictor model looked random without a definitive pattern which suggested the absence of model specification problem or confounding variables:



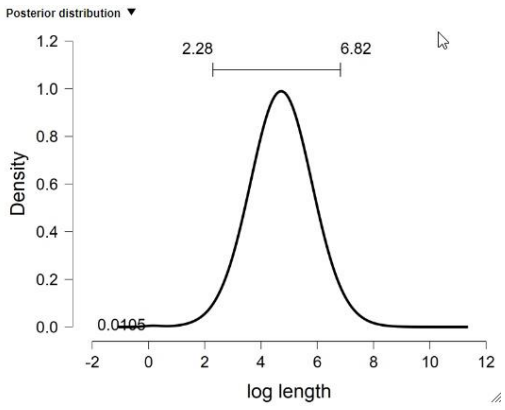
The inclusion probability histogram showed that all three predictors were equally importance in the best model:

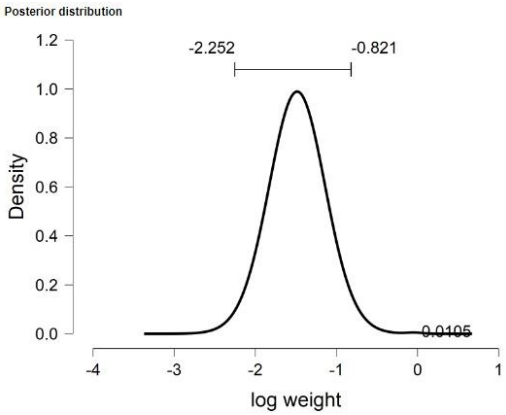


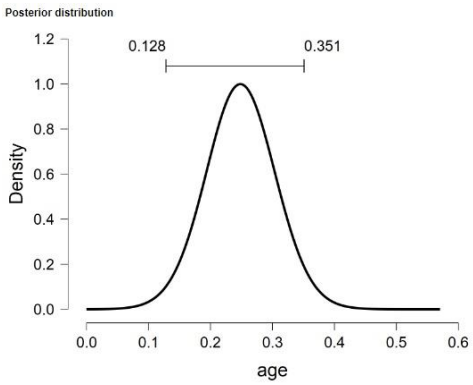
The Q – Q plot of model averaged residuals showed quite nice fits of the best model (the residuals were approximately normally distributed):



The averaged posterior distribution plots below showed the t – approximations of the posterior distributions averaged over all models. Each peak was defined by the upper and lower 95% credible limits. Notice the density for all the three peaks were very similar:







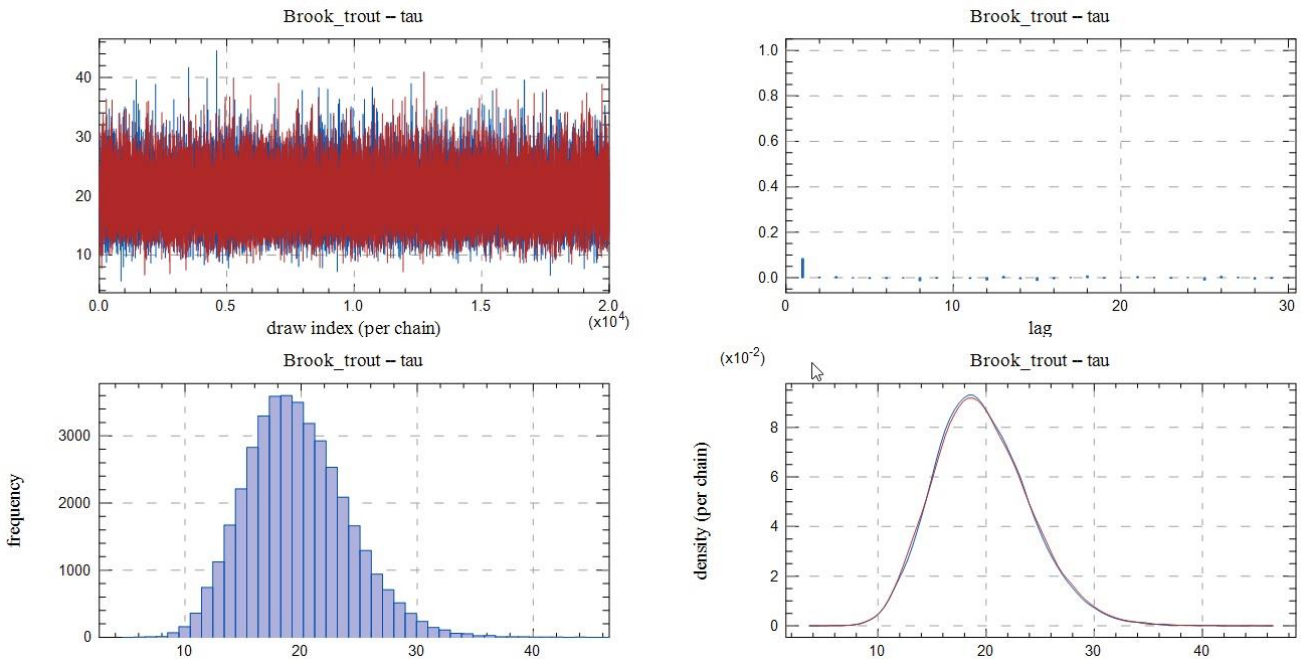
To assess the performance of the Gibb sampler used in doing the MCMC random walks, we used the diagnostics function of the BaySE© econometric software which produced the following table that contained estimates of the MCMC standard errors, $\sqrt{\sigma^2}/G$ for G number of draws (all very low) as well as the inefficiency factors or autocorrelation time (all very close to one) for all parameters:

MCMC Diagnostics: myModel

dataset:	BT_Bayesian+	total retained draws:	20000
lhs variable:	log_mercury	burnin (draws/chain):	10000
# of rhs variables:	4	number of chains:	1
# of observations:	44	thinning parameter:	1
log-ML (Lewis & Raftery):	-26.1403		

	Mean	Spectral density at 0	MCMC sd. error	Relative Numer. Eff.	Inefficiency factor
constant	-21.8784	2.97423	0.0305676	1.06682	0.937366
log_length	4.7474	0.205676	0.00803835	1.02476	0.975837
log_weight	-1.49572	0.0213663	0.00259083	0.983464	1.01681
age	0.258242	0.000435302	0.000369803	1.19084	0.839743
tau	19.6271	3.26604	0.0320321	0.916805	1.09074
sigma_e	0.229964	0.000122996	0.000196571	0.885547	1.12925

Four diagnostic plots were provided for visually assessing the performance of the Gibb sampler. The two subplots at the top presented a history and a correlogram of the draws for the precision parameter tau τ (where $\tau \equiv 1/\sigma^2$) and indicated that these were not autocorrelated. Two Markov chains were used in the draws (red and blue in the first subplot. The two remaining subplots at the bottom presented a histogram and an estimate of the kernel density of the draws. Both of them were smooth, suggesting that the sampler did not get trapped for any considerable amount of draws in specific regions of the sample space.



5.3.2 Bayesian linear regression of lake trout data set:

Model Comparison ▼

Models	P(M)	P(M|data)	BF_M	BF_{01}	R^2
weight	0.083	0.388	6.967	1.000	0.644
fork length	0.083	0.310	4.948	1.250	0.639
fork length + weight	0.083	0.097	1.187	3.980	0.654
weight + age	0.083	0.079	0.942	4.914	0.648
fork length + weight + age	0.250	0.068	0.217	17.220	0.655
fork length + age	0.083	0.057	0.665	6.800	0.639
age	0.083	0.001	0.011	378.450	0.444
Null model	0.250	7.056e −6	2.117e −5	164867.153	0.000

In the case of lake trout, it is interesting to see that the BAS algorithm suggested that the best model only consisted of one predictor: either weight (BF_{01} = 1.000) or fork length (BF_{01} = 1.250). It is surprising to see that the data were 300 times more likely to occur under either these two models than under a model with age as the only predictor (BF_{01} = 378.450). Respectively, the data were 17 and 13 times more likely to occur under the weight alone model and the fork length alone model than the model with all three predictors (BF_{01} = 17.220). However, the coefficients of determination (R^2) of the first six models were very similar hence they all explained well over 60% of variance. The model with age alone as the predictor only explained 44.4% of the variance. The relatively inferiority of age being a predictor in the model in comparison with fork length and weight was largely because of the high degree of variations of age versus mercury as seen in the scatterplot in Chapter 1; which is also evident from the wide error bar of the posterior coefficient associated with age indicating the 95% credible interval depicted below. For the lake trout data, four possible models were presented here:

(1) best model with weight alone as predictor:

Posterior Summaries of Coefficients ▼

Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	95% Credible Interval Lower	95% Credible Interval Upper
Intercept	0.285	0.013	1.000	1.000	1.000	0.259	0.312
weight	2.678e −4	3.772e −5	0.500	0.632	1.715	1.908e −4	3.449e −4

mercury = (0.285) + (2.68 * 10^{-4}) * weight

(2) second best model with for length alone as predictor:

Posterior Summaries of Coefficients

Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	95% Credible Interval	
						Lower	Upper
Intercept	0.296	0.014	1.000	1.000	1.000	0.271	0.327
length	0.001	1.780e −4	0.500	1.000	326302.818	9.599e −4	0.002

mercury = (0.296) + (0.001) * fork length

(3) a model with weight and fork length as predictors:

Posterior Summaries of Coefficients ▼

Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	95% Credible Interval	
						Lower	Upper
Intercept	0.296	0.013	1.000	1.000	1.000	0.266	0.322
length	3.495e −4	6.283e −4	0.500	0.433	0.762	−2.723e −4	0.002
weight	2.106e −4	1.388e −4	0.500	0.786	3.669	0.000	3.644e −4

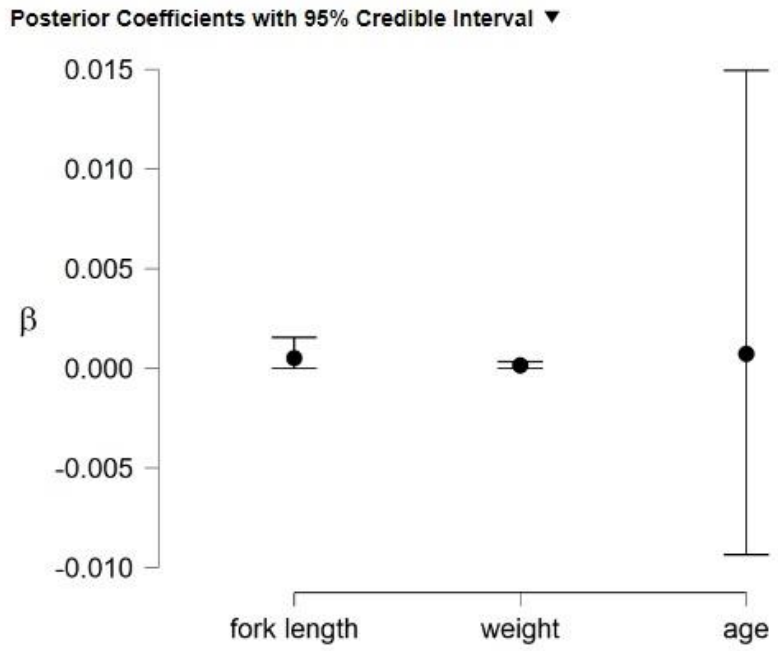
mercury = (0.296) + ($3.5 * 10^{-4}$) * fork length + ($2.11 * 10^{-4}$) * weight

(4) complete model with fork length, weight and age as predictors:

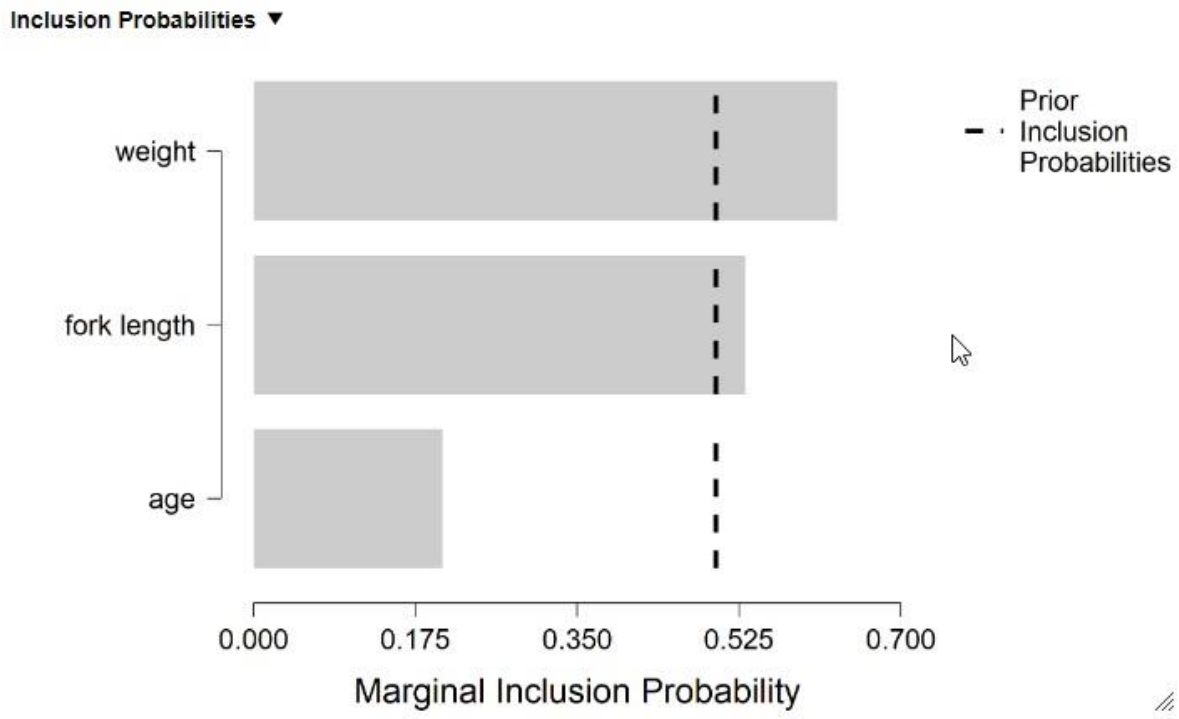
Posterior Summaries of Coefficients

Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	95% Credible Interval	
						Lower	Upper
Intercept	0.285	0.013	1.000	1.000	1.000	0.259	0.311
fork length	4.543e −4	6.677e −4	0.500	0.532	1.138	0.000	0.002
weight	1.465e −4	1.388e −4	0.500	0.632	1.715	−4.594e −6	3.354e −4
age	0.002	0.011	0.500	0.205	0.257	−0.009	0.015

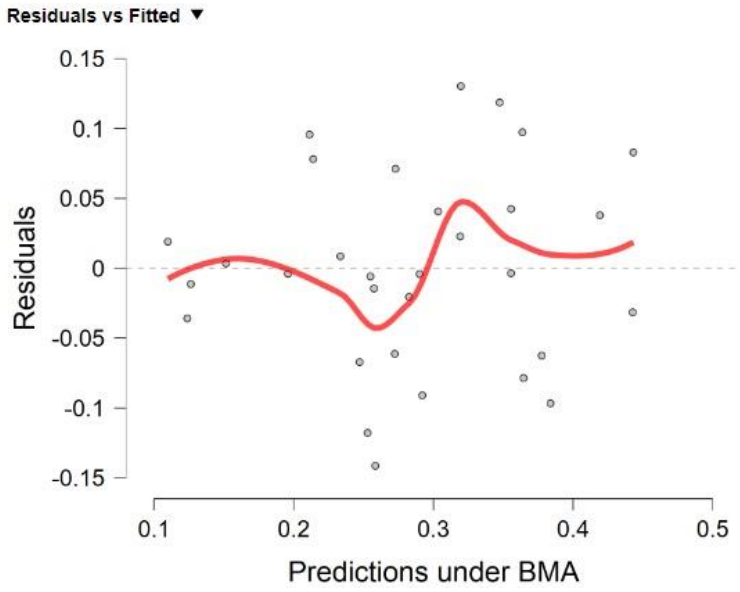
mercury = (0.285) + ($4.54 * 10^{-4}$) * fork length + ($1.47 * 10^{-4}$) * weight + (0.002) * age



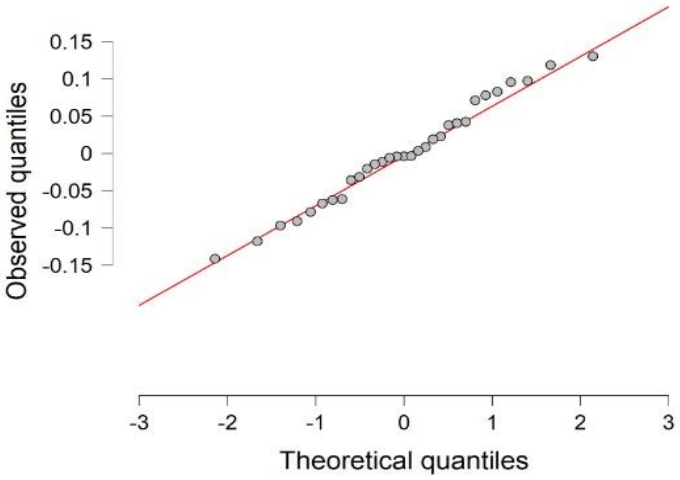
The inclusion probability histogram showed that both weight and fork length were more important than age in the model:



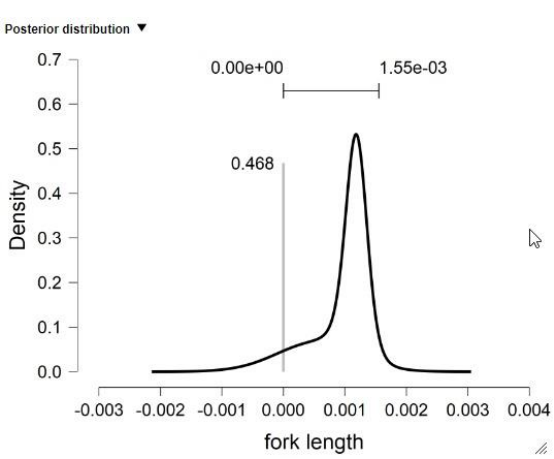
The plot of residuals of the model averaged predictions (BMA) versus the residuals looked random without a definitive pattern which suggested the absence of model specification problem or confounding variables:

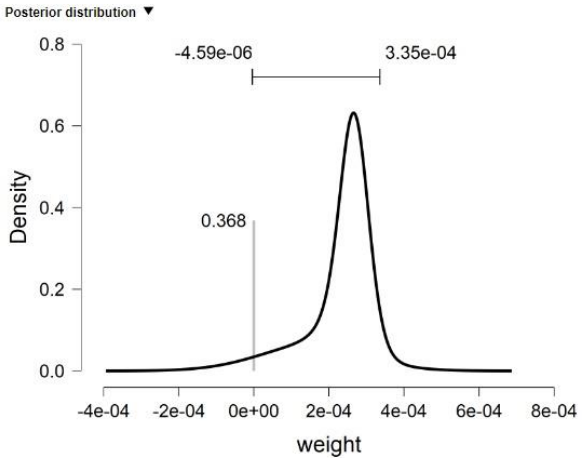


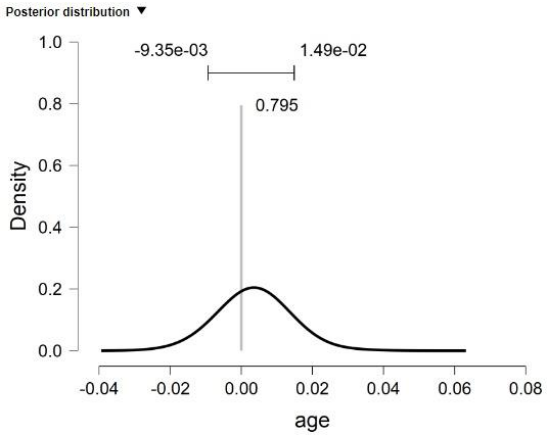
The Q – Q plot of model averaged residuals showed quite nice fits of the model (the residuals were approximately normally distributed):



The averaged posterior distribution plots below showed the t – approximations of the posterior distributions averaged over all models. Each peak was defined by the upper and lower 95% credible limits. Notice the density associated with the averaged posterior distribution of the age as a predictor was much smaller than that of both fork length and weight:







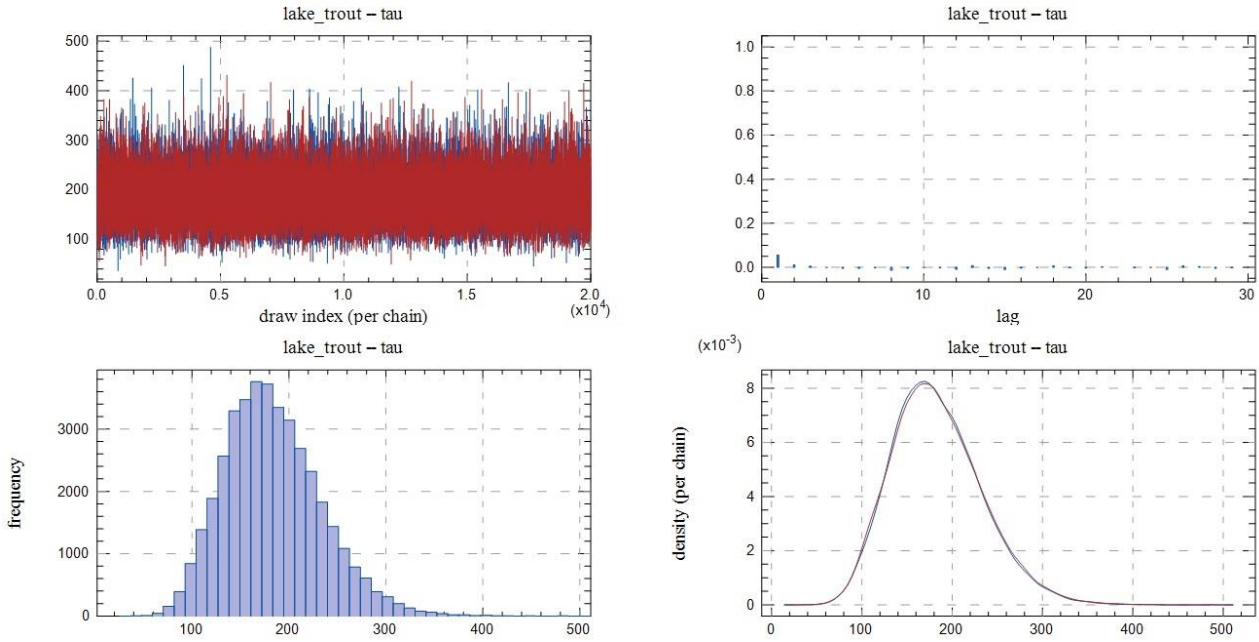
The performance of the Gibb sampler was confirmed by the extremely small MCMC standard errors and the fact that the inefficiency factors were all close to unity:

MCMC Diagnostics: lake_trout

dataset:	LT_Bayesian+	total retained draws:	40000
lhs variable:	mercury	burnin (draws/chain):	10000
# of rhs variables:	4	number of chains:	2
# of observations:	31	thinning parameter:	1
log-ML (Lewis & Raftery):	-8.51574		

	Mean	Spectral density at 0	MCMC sd. error	Relative Numer. Eff.	Inefficiency factor
constant	-0.0189471	0.00434179	0.000825837	0.908238	1.10103
length	0.000495556	8.76678e-008	3.71091e-006	0.975559	1.02505
weight	0.000159696	3.87151e-009	7.7983e-007	0.95312	1.04919
age	0.00259099	2.1156e-005	5.76469e-005	0.995264	1.00476
tau	183.27	471.439	0.272128	0.828356	1.20721
sigma_e	0.0759739	2.21265e-005	5.89544e-005	0.827428	1.20856

The four diagnostic plots also confirmed the performance of the random walk of the MCMC samplers;



Chapter 6. Using Hierarchical Multiple Linear Regression to Investigate the Quadratic Term of the Curvilinear Data of Lake Whitefish – a Quadratic Polynomial Model.

Scatterplots in Chapter 1 have clearly shown that $\log_e$ mercury versus all three fish growth parameters of lake whitefish were better described by a curvilinear function. Square transformation of fish age and weight but not fork length rendered the relationships pretty linear. The curvilinear function observed seemed best described by a quadratic polynomial model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

The curved (V – shaped) patterns of the residuals versus fitted dependent variable plots and the residuals versus predictors plots (see Chapter 1) as well the significant Durbin – Watson statistic also suggested that a linear relationship between the dependent variable and the predictors was not appropriate. However, the bottom line concerning the decision to support (or not to dismiss) that a nonlinear polynomial function was the best model rested ultimately on the theoretical basis of a plausible explanation of such a fit. Polynomial regression can be seen as a special case of multiple linear regression in that the multiple predictor variables are now X, X^2 , X^3, …… X^n . However, it is very unlikely in biological situations that polynomial regression higher than X^3 (cubic function) is plausible. In this chapter I used hierarchical multiple linear regression to investigate the impact of the quadratic term on the model fits of the lake whitefish data.

In hierarchical multiple linear regression (also known as sequential multiple linear regression), independent variables (including the predictors we want to test) are added into the model in the order specified by the researcher based upon certain theoretical or preferential grounds. Independent variables or sets of independent variables are entered in steps known as “blocks” into the statistics algorithm. With each independent variable being assessed in terms of what it adds to the prediction of the outcome of the dependent variable after the previous independent variables entered into model are controlled for. So, this in a way is to control the effects of independent variables that might be potentially confounding or they are covariate – type of variables so that they can be accounted for; hence, to develop a better prediction model for the outcome of the dependent variable.

We enter the potentially confounding variables (i.e. independent variables that we want to control for) into the first block and then those actual predictor variables that we want to use to predict the dependent variable into the second block. Once all the sets of variables are entered, the overall model is assessed in terms of its ability to predict the outcome of the new measures of the dependent variable. The contribution of each block of variables is also assessed so that we can determine very precisely how much influence a particular independent variable may be having.

In the present context, hierarchical multiple linear regression was useful in assessing the influence of the quadratic term (X^2) on the overall model. The linear term of X was entered into the first block and the quadratic term (X^2) was entered into the second block. Hence, two regression models were established: one with the linear term alone and the other with both the linear and the quadratic terms. IBM™ SPSS™ Statistics software is particularly useful for this because the algorithm displaces the "change" in the R^2 of the overall model with and without the quadratic term being added to the model. The NCSS™ statistics software was used here for the curve fitting since it allowed us to use the randomized resampling technique of bootstrapping to compute robust estimates of coefficients, standard errors and confidence intervals. In order to keep things simple, here we established three models, one for each of the three predictors. I have also decided not to include interaction terms in the models to account for effects of particular combinations of predictors. The "complete second order model" can sometime be used to include linear and quadratic terms for two quantitative predictor variables along with interaction terms:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \epsilon$$

More than two predictors can be incorporated into such complex models. But I have decided that at the present stage it is somewhat beyond the scope of this monograph to explore complex models like these.

In the general sense, all the pre – requisite assumptions of OLS regression applied to hierarchical multiple linear regression. However, in the present context, since the quadratic term (X^2) we want to assess was derived from X; hence, they were naturally highly collinear. Also, the linearity assumption did not apply here since we were assessing a curvilinear model. Normality and homoscedasticity of the residuals have been confirmed and outliers have been identified and weeded out (see Chapter 1).

6.1 Fork length:

Model 1 consisted of only the linear term fork length. Model 2 consisted of both the linear term fork length and the quadratic term, (i.e. square fork length).

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	fork_length[b]	.	Enter
2	square_length[b]	.	Enter

a. Dependent Variable: log_mercury

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.832[a]	.692	.688	.231729	.692	177.283	1	79	.000
2	.911[b]	.830	.826	.173062	.139	63.641	1	78	.000

a. Predictors: (Constant), fork_length

b. Predictors: (Constant), fork_length, square_length

For model 1 which only contained the linear term, 69.2% (R Square = 0.692) of the variance in the dependent variable ($\log_e$ mercury) was attributed to the linear term (fork length) of the predictor variable. As much as 83% of the variance in the dependent variable was attributed to Model 2 which was a combination of both the linear term (fork length) and the quadratic term (square fork length). Both models were significant (p = 0.000). More importantly, an addition of 13.9% of the variability of $\log_e$ mercury was accounted for by including the quadratic term into the model (R Square Change = 0.139), which is a lot. The F value of 177.283 under the column F Change was in fact the F value associated with just the linear term (Model 1) **only**. The F Change of 63.641 on the other hand, represented **only** the effect of the quadratic term into the model alone (i.e. not including the F value from the model with the linear term only). The F value of 190.747 associated with Model 2 in the ANOVA table below in fact represented the F value of the model with both the linear term and the quadratic term; and the R Square value associated with this F value was 83%. The ANOVA table also showed that both models were highly significant (p = 0.000). Hence, the slopes of the two models were

significantly different from zero. In the present case, all the p – values were significant and were easy and straightforward to interpret. In fact, the p - value associated with Model 2 in the Model Summary table above was the most important in the present context, which was the "p – value of the increase of R – Square"; i.e. it was the probability of getting the observed increase in R – Square between the equation of Model 2 and the one of a lower power (Model 1) by chance, if the relationship between the predictor and the dependent variable were really of the form described by the equation of a lower power. This p – value actually tested the null hypothesis of the increase of R – Square was only as large as would expect by chance. Since $p < 0.05$, we have to reject this null hypothesis; hence, the increase in R -Square resulted from adding the quadratic term into the model was significant. It is also important to note that before we looked into this p – value associated with the Model 2, the p – value that associated with Model 1 has to be significant first. This Model 1 p – value tested the null hypothesis of there was no relationship between the predictor and the dependent variable. Since this Model 1 p – value was less than the critical level of 0.05, we have to reject this null hypothesis as well and there was a relationship between the predictor and the dependent variable. The R -Square Change and the F Change together with the p – values associated with the two models clearly demonstrated the incremental predictive capacity of the model brought about by including the quadratic term was significant.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.520	1	9.520	177.283	.000[b]
	Residual	4.242	79	.054		
	Total	13.762	80			
2	Regression	11.426	2	5.713	190.747	.000[c]
	Residual	2.336	78	.030		
	Total	13.762	80			

a. Dependent Variable: log_mercury

b. Predictors: (Constant), fork_length

c. Predictors: (Constant), fork_length, square_length

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3.887	.118		-32.950	.000
	fork_length	.005	.000	.832	13.315	.000
2	(Constant)	-1.118	.358		-3.122	.003
	fork_length	-.012	.002	-2.169	-5.722	.000
	square_length	2.455E-005	.000	3.023	7.978	.000

a. Dependent Variable: log_mercury

Unlike linear regression, in polynomial regression when the predictors are raised to higher orders, regression coefficients are largely uninterpretable both in terms of their direction and size. In the present case, it was obvious from the scatterplot that there was a strong positive correlation between $\log_e$ mercury and fork length, yet the coefficient associated with the linear term of fork length was negative. But the "overall" model in describing the relationship between the dependent variable and the predictor was perfectly valid.

Excluded Variables[a]

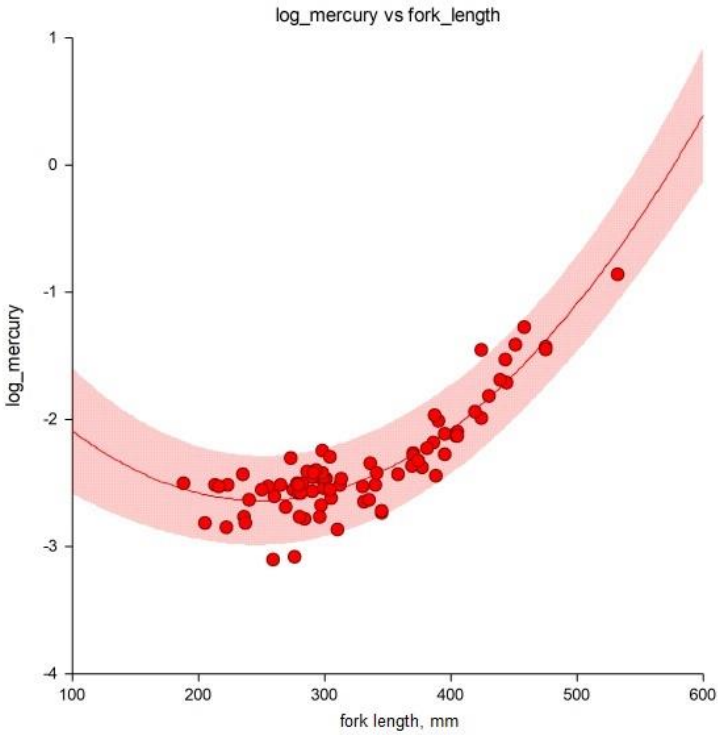
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	square_length	3.023[b]	7.978	.000	.670	.015

a. Dependent Variable: log_mercury

b. Predictors in the Model: (Constant), fork_length

In "normal" multiple linear regression with two or more predictors, a Tolerance lower than 0.1 might indicate collinearity problems amongst predictors. But in the present context, a very low Tolerance was expected (0.015 in this case) since of course, square fork length is fork length multiplies itself! Collinearity Statistics did not apply here in polynomial models.

The final model and the curve fitting depicted below by using NCSS™ statistics software which also carried out bootstrapping resampling processes to compute the confidence interval (grey band around the best – fitted line) of the model:



Dataset Untitled
Y Variable: log_mercury. X Variable: fork_length.
Model Fit: log_mercury=A+B*(fork_length)+C*(fork_length)^2

Model Estimation Section

Parameter Name	Parameter Estimate	Asymptotic Standard Error	Lower 95% C.L.	Upper 95% C.L.
A	-1.11800	0.35816	-1.83104	-0.40497
B	-0.01221	0.00213	-0.01646	-0.00796
C	0.00002	0.00000	0.00002	0.00003

Iterations	8	Rows Read	81
R-Squared	0.830248	Rows Used	81
Random Seed	23380	Total Count	81

Bootstrap Section

Estimation Results		Bootstrap Confidence Limits	
Parameter	**Estimate**	**Conf. Level**	**Lower**	**Upper**
A				
Original Value	-1.11800	0.90000	-1.84083	-0.64636
Bootstrap Mean	-1.05075	0.95000	-1.98643	-0.54330
Bias (BM - OV)	0.06725	0.99000	-2.24927	-0.38082
Bias Corrected	-1.18526			
Standard Error	0.35832			
B				
Original Value	-0.01221	0.90000	-0.01499	-0.00777
Bootstrap Mean	-0.01264	0.95000	-0.01552	-0.00682
Bias (BM - OV)	-0.00043	0.99000	-0.01636	-0.00520
Bias Corrected	-0.01179			
Standard Error	0.00217			
C				
Original Value	0.00002	0.90000	0.00002	0.00003
Bootstrap Mean	0.00003	0.95000	0.00002	0.00003
Bias (BM - OV)	0.00000	0.99000	0.00001	0.00003
Bias Corrected	0.00002			
Standard Error	0.00000			

Bootstrapping is a modern computer – intensive iterative resampling – with – replacement simulation method that has become available in recent years as extensive computer power has become popularized. It computes standard errors and confidence intervals for regression coefficients and predicted values in situations when standard regression assumptions are on shaky grounds or simply not valid. The bootstrapping method can be applied to many of the statistics that are computed in regression analysis. The only assumption made when using bootstrapping is that the sample approximate the population fairly well. Hence, bootstrapping only works well for relatively large sample size. In the present bootstrapping run, 3000 draws of bootstrap samples were used; hence, provided 3000 estimates of the coefficients and the confidence limits! We drawn 3000 bootstrap samples of size n from the original samples with replacement; hence each observation was selected more than once. For each of the 3000 bootstrap samples, the regression results were computed, stored and used in the final computation of various regression statistics. The reflection method was used to calculate the bootstrap confidence intervals in which confidence limits were formed by reflecting the percentiles of the bootstrap values. In the Bootstrap Section table above, the Original Value is the parameter estimate obtained from the complete original data set **without** bootstrapping. The Bootstrap Mean is the average of the parameter estimates of the 3000 bootstrap samples. The Bias (BM – OV) is the estimate of the bias in the original estimate. It is computed by subtracting the Original Value from the Bootstrap Mean. The Bias Corrected is an estimate of the parameter that has been corrected for its bias. The correction is made by subtracting the estimated bias from the original parameter estimate. The Standard Error term is the bootstrap method's estimate of the standard error of the parameter estimate. It is simply the

standard deviation of the parameter estimate computed from the bootstrap estimate. Note that the Parameters A, B and C in the bootstrap table (i.e. the constant and the two regression coefficients associated with the linear and quadratic terms) all have the same signs as their corresponding upper and lower confidence limits calculated by bootstrapping and these three parameters were all within their corresponding upper and lower confidence limits.

The equation of the final quadratic polynomial model for fork length is:

$\log_e$ mercury = (-1.1180) + (-0.01221) * fork length + ($2.46 * 10^{-4}$) * fork length^2

One important thing regarding the use of any regression model, particularly nonlinear models is that the model is only valid for new data that are within the range of the original data that used in estimating the model. Extrapolation beyond such a range (in either direction) can be problematic. Weir things might happen that do not make theoretical sense, especially in the case of nonlinear models of higher orders.

6.2 Weight

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	weight[b]	.	Enter
2	square_weight[b]	.	Enter

a. Dependent Variable: log_mercury

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.875[a]	.766	.763	.201948	.766	258.445	1	79	.000
2	.906[b]	.822	.817	.177463	.056	24.303	1	78	.000

a. Predictors: (Constant), weight

b. Predictors: (Constant), weight, square_weight

For Model 1, as much as 76.6% of the variance in the dependent variable was attributed to the linear term of fish weight. The F value associated with Model 1 was 258.445 and there was a significant

relationship between $\log_e$ mercury and weight (p = 0.000). Model 2 (a combination of weight and square weight) accounted for 82.2% of the variance of $\log_e$ mercury. Although by including the quadratic term of weight seemingly only increased the R – Square by 5.6%, nonetheless such an increase was statistically significant (p = 0.000). The F value associated with the quadratic term only was 24.303. The ANOVA table below shows that both models were significant (p = 0.000). The F value associated with Model 2 containing both weight and square weight was 179.491.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.540	1	10.540	258.445	.000[b]
	Residual	3.222	79	.041		
	Total	13.762	80			
2	Regression	11.306	2	5.653	179.491	.000[c]
	Residual	2.456	78	.031		
	Total	13.762	80			

a. Dependent Variable: log_mercury

b. Predictors: (Constant), weight

c. Predictors: (Constant), weight, square_weight

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.902	.041		-71.147	.000
	weight	.001	.000	.875	16.076	.000
2	(Constant)	-2.649	.063		-42.264	.000
	weight	-6.185E-006	.000	-.005	-.027	.978
	square_weight	8.043E-007	.000	.911	4.930	.000

a. Dependent Variable: log_mercury

Again, as in the case with fork length, the estimated coefficient associated with the linear term of weight was negative despite the positive correlation between $\log_e$ mercury and weight. No interpretation of coefficients associated with predictors for a polynomial model was made.

Excluded Variables[a]

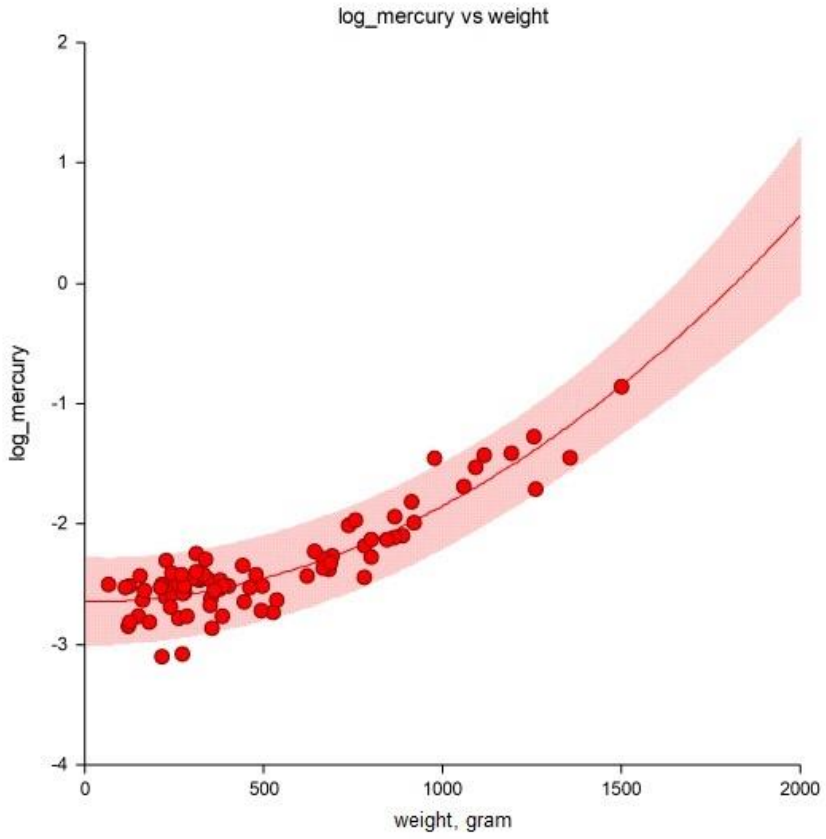
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	square_weight	.911[b]	4.930	.000	.487	.067

a. Dependent Variable: log_mercury

b. Predictors in the Model: (Constant), weight

The low Tolerance here as expected since weight and square weight were highly positively correlated.

The final model and the curve fitting depicted below by using NCSS™ statistics software which also carried out bootstrapping resampling processes to compute the confidence interval (grey band around the best – fitted line) of the model:



Dataset Untitled
Y Variable: log_mercury. X Variable: weight.
Model Fit: log_mercury=A+B*(weight)+C*(weight)^2

Model Estimation Section

Parameter Name	Parameter Estimate	Asymptotic Standard Error	Lower 95% C.L.	Upper 95% C.L.
A	-2.64854	0.06267	-2.77330	-2.52378
B	-0.00001	0.00023	-0.00046	0.00045
C	0.00000	0.00000	0.00000	0.00000

Iterations	4	Rows Read	81
R-Squared	0.821503	Rows Used	81
Random Seed	26512	Total Count	81

In the present bootstrapping run, 3000 draws of bootstrap samples were used; hence, provided 3000 robust estimates of the coefficients and the confidence limits using the reflection method:

Bootstrap Section

Estimation Results		Bootstrap Confidence Limits	
Parameter	Estimate	Conf. Level	Lower	Upper
A				
Original Value	-2.64854	0.90000	-2.75532	-2.54550
Bootstrap Mean	-2.64626	0.95000	-2.77864	-2.52300
Bias (BM - OV)	0.00228	0.99000	-2.83581	-2.48034
Bias Corrected	-2.65082			
Standard Error	0.06433			
B				
Original Value	-0.00001	0.90000	-0.00035	0.00041
Bootstrap Mean	-0.00003	0.95000	-0.00042	0.00053
Bias (BM - OV)	-0.00002	0.99000	-0.00057	0.00078
Bias Corrected	0.00001			
Standard Error	0.00023			
C				
Original Value	0.00000	0.90000	0.00000	0.00000
Bootstrap Mean	0.00000	0.95000	0.00000	0.00000
Bias (BM - OV)	0.00000	0.99000	0.00000	0.00000
Bias Corrected	0.00000			
Standard Error	0.00000			

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.

The equation of the final quadratic polynomial model for weight is:

$\log_e$ mercury = (-2.64852) + (-6.19 x 10^{-6}) * weight + (8.04 x 10^{-7}) * weight^2

6.3 Age

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	age[b]	.	Enter
2	square_age[b]	.	Enter

a. Dependent Variable: log_mercury

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.865[a]	.748	.744	.209709	.748	233.931	1	79	.000
2	.894[b]	.799	.794	.188307	.051	19.978	1	78	.000

a. Predictors: (Constant), age

b. Predictors: (Constant), age, square_age

For Model 1, 74.8% of the variance in the dependent variable was attributed to the linear term of fish age. The F value associated with Model 1 is 233.931 and there was a significant relationship between

$\log_e$ mercury and age (p = 0.000). Model 2 (a combination of age and square age) accounted for 79.9% of the variance of $\log_e$ mercury. Although by including the quadratic term of age seemingly only increased the R – Square by 5.1%; nonetheless such an increase was statistically significant (p = 0.000). The F value associated with the quadratic term alone was 19.978. The ANOVA table below shows that both models were significant (p = 0.000), i.e. the slopes of their regression lines were significantly different from zero. The F value associated with Model 2 containing both weight and square weight was 155.053.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.288	1	10.288	233.931	.000[b]
	Residual	3.474	79	.044		
	Total	13.762	80			
2	Regression	10.996	2	5.498	155.053	.000[c]
	Residual	2.766	78	.035		
	Total	13.762	80			

a. Dependent Variable: log_mercury

b. Predictors: (Constant), age

c. Predictors: (Constant), age, square_age

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3.299	.066		-49.965	.000
	age	.142	.009	.865	15.295	.000
2	(Constant)	-2.508	.187		-13.437	.000
	age	-.084	.051	-.513	-1.643	.104
	square_age	.014	.003	1.396	4.470	.000

a. Dependent Variable: log_mercury

Again, as in the case with fork length and weight, the estimated coefficient associated with the linear term of age was negative despite the positive correlation between $\log_e$ mercury and age. No interpretation of coefficients associated with predictors for a polynomial model was made.

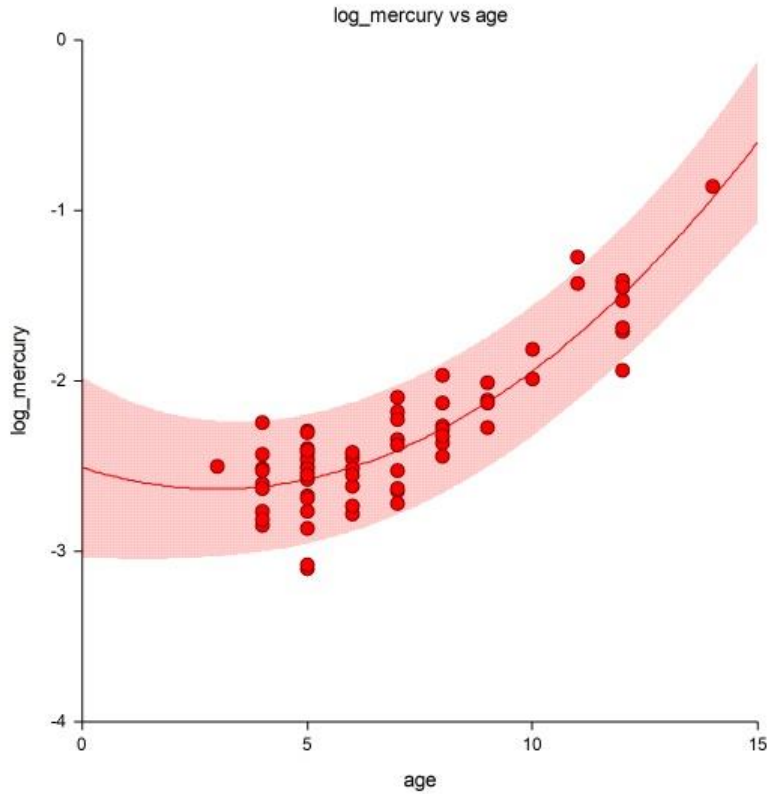
Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	square_age	1.396[b]	4.470	.000	.452	.026

a. Dependent Variable: log_mercury

b. Predictors in the Model: (Constant), age

The very low Tolerance here as expected since age and square age were highly positively correlated.



The curve fitting depicted above by using NCSS™ which also carried out bootstrapping resampling to compute the confidence interval (grey band around the best – fitted line) of the model.

Dataset Untitled
Y Variable: log_mercury. X Variable: age.
Model Fit: log_mercury=A+B*(age)+C*(age)^2

Model Estimation Section

Parameter Name	Parameter Estimate	Asymptotic Standard Error	Lower 95% C.L.	Upper 95% C.L.
A	-2.50823	0.18667	-2.87987	-2.13660
B	-0.08413	0.05122	-0.18609	0.01784
C	0.01408	0.00315	0.00781	0.02035

Iterations	7	Rows Read	81
R-Squared	0.799024	Rows Used	81
Random Seed	27380	Total Count	81

Bootstrap Section

Estimation Results			Bootstrap Confidence Limits	
Parameter	Estimate	Conf. Level	Lower	Upper
A				
Original Value	-2.50823	0.90000	-2.77227	-2.21375
Bootstrap Mean	-2.51604	0.95000	-2.82049	-2.13396
Bias (BM - OV)	-0.00781	0.99000	-2.92005	-1.97429
Bias Corrected	-2.50042			
Standard Error	0.17132			
B				
Original Value	-0.08413	0.90000	-0.16478	-0.01262
Bootstrap Mean	-0.08215	0.95000	-0.18270	0.00085
Bias (BM - OV)	0.00198	0.99000	-0.22233	0.02906
Bias Corrected	-0.08611			
Standard Error	0.04629			
C				
Original Value	0.01408	0.90000	0.00971	0.01915
Bootstrap Mean	0.01396	0.95000	0.00874	0.02036
Bias (BM - OV)	-0.00011	0.99000	0.00641	0.02275
Bias Corrected	0.01419			
Standard Error	0.00290			

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.

3000 draws of bootstrap samples were used in the bootstrapping runs; hence, provided 3000 robust estimates of the coefficients and the confidence limits using the reflection method. The Parameters A, B and C in the bootstrap table (i.e. the constant and the two regression coefficients associated with the linear and quadratic terms) all have the same signs as their corresponding upper and lower confidence limits calculated by bootstrapping and these three parameters were all within their corresponding upper and lower confidence limits.

The equation of the final quadratic polynomial model for age is:

$\log_e$ mercury = (-2.50823) + (-0.08413) * age + (0.01408) * age^2

Applying hierarchical multiple linear regression to investigate the impact of quadratic terms of the predictors on the overall polynomial model clearly showed that inclusion of the quadratic term of fork length resulted in greater improvement of the predictive capability of the regression model than in the case of weight and age.

Chapter 7. Robust Linear Regression using M – Estimators to Suppress the Effects of the Presence of Outliers on Inflating Residuals.

When we are fitting a regression model, while most of the observations fit the model and meet the Gauss – Markov conditions of linear model at least approximately, some observations do not. This situation might occur if (1) there is something wrong with those few anomalous observations (e.g. error in sampling and/or measurement, or mistake in entering data) or (2) the fitted model is not appropriate (i.e. a model specification problem). Because of these two very different reasons for the existence of observations that apparently do not belong to the model, we have two very different purposes trying to identify them: (1) to protect the model from any observation that does not belong to the model (i.e. outliers) and thus adversely affecting the model; and (2) to find the shortcomings of the model itself that we have fitted.

Assuming our model is correct and it follows a linear function. In general there are two very different approaches in dealing with outliers: (1) Run a regression to get the residuals and apply a whole bunch of diagnostic tests to the residuals to identify the outliers and then to throw them out and re-run the regression model as mentioned in Chapter 1; or (2) subject the entire data set to a robust regression procedure using robust estimators that are more resistant to the inflation of residuals due to the presence of outliers; hence minimizing their impact on the coefficient estimates. The first approach can run into problems when a large number of possible outliers are found relative to the size of the data set. Removing them reduces the degrees of freedom, but more importantly it runs the risk of model specification problems, i.e. the function that describe the model changes after the outliers have been removed. This problem is exacerbated by the fact that there is no generally agreed criteria as to how many outliers are too many for this first approach to handle safely.

OLS estimator is not robust because it is highly susceptible to the inflation of residuals due to the presence of outliers since OLS estimator weight all residuals equally when calculating the regression line by minimizing the sum – of - squares of residuals. Sum – of – squares of residuals of outliers are huge and hence influential enough to throw off the regression model. Outliers violate the assumption of normally distributed residuals in OLS regression. They tend to distort the least square coefficients by having more influence than they should. In OLS regression, the weight attached to each observation would be about 1/N in a data set of N observations. Outlying observations may receive substantially higher weight than 1/N and seriously distort the estimated coefficients. When the model has only one or

two predictor variables, outliers can often be spotted visually on a scatterplot. However, in complex models with many predictors, outliers can often be hidden from view.

Robust regression use estimators that "down – weight" the influence of outliers. Outliers are given less weight hence are less important and have less contribution in the estimation process of the model. Some robust estimators actually remove extreme outliers altogether from the modeling process. Robust Regression is an iterative procedure, it conducts its own residual analysis that seeks to identify outliers by the amount of weight assigned to each observation so as to minimize their impact on the estimation of the regression coefficients. Robust regression algorithms use influence functions (e.g. Huber M – estimator and Tukey's Bisquare M – estimator) to determine the amount of weight assigned to each observation. The algorithmic iterative process is often computationally intensive for complex models with large number of observations. It often takes a number of iterations for the coefficients to stabilize and the absolute values of the residuals to converge. This is one main reason why robust regression only starting to become more popular when high – speed computers becoming more available.

7.1 Maximum Likelihood Estimation (M – Estimation)

M – estimation is the most commonly used robust regression technique. It replaces the square of residuals (ε_i^2) used in OLS by another function of residuals $H(\varepsilon_i)$. In the case of OLS, $H(\varepsilon_i) = \varepsilon_i^2$. In M – estimation we need a $H(\varepsilon_i)$ to have the following properties: (1) non – negative, $H(\varepsilon_i) \geq 0$; (2) the function of zero is actually zero, $H(0) = 0$, i.e. if the residual is zero, it will not contribute to the sum; (3) the function should be symmetrical, $H(- \varepsilon_i) = H(\varepsilon_i)$; the function should be monotonic $| \varepsilon_i | > | \varepsilon_j |$ then $H(\varepsilon_i) > H(\varepsilon_j)$, hence if the residual is bigger then the function should be bigger; hence large residuals got panelized more than small residuals and (4) most importantly the function $H(\varepsilon_i)$ needed to be continuous derivative with respect to the coefficients so enable us to be more effective (numerical stability) in finding the minima of the sum of squares.

In M – estimation, we take the derivative of H with respect to the residual $\frac{\partial \mathrm{H}}{\partial \varepsilon_i}$ multiply by the derivative with respect to the coefficient which is simply the value of the predicted variable X_{ki} . Sum all

the data points and set it equal to zero and find the value of the parameter that minimize the summation of the function to obtain the estimate of the regression coefficients:

$$\frac{\partial S}{\partial \beta_k} = 0 \rightarrow \sum_{i=1}^{n} \frac{\partial H}{\partial \varepsilon_i} x_{ki} = 0$$

So, this is done for all the parameters and we can solve the simultaneous equations.

Now we can define the weight w_i as $w_i = \frac{1}{\varepsilon_i} \frac{\partial H}{\partial \varepsilon_i}$, hence, $w_i \varepsilon_i = \frac{\partial H}{\partial \varepsilon_i}$

Now we have $\frac{\partial S}{\partial \beta_k} = 0 \rightarrow \sum_{i=1}^{n} w_i \varepsilon_i x_{ki} = 0$

which is in fact a weighted linear regression. This in fact forms the scheme with which we can use to carry out iterative process of M – estimation: first set the weight w_i as 1 to start with. Carry out a linear regression and obtain a set of residuals. This is in fact an OLS regression (iteration 0). Then use these residuals to calculate another w_i using the formula above. Then we plug the weight in and do another weighted linear regression which give us another set of residuals and we have iteration 1. The process is repeated for a number of iterations by the computer algorithm until the coefficients stabilize and the absolute values of the residuals converge. This approach is known as iteratively reweighted least squares (IRLS). The repeated running of iterations actually overcome the masking nature of outliers when many predictor variables are present in the model.

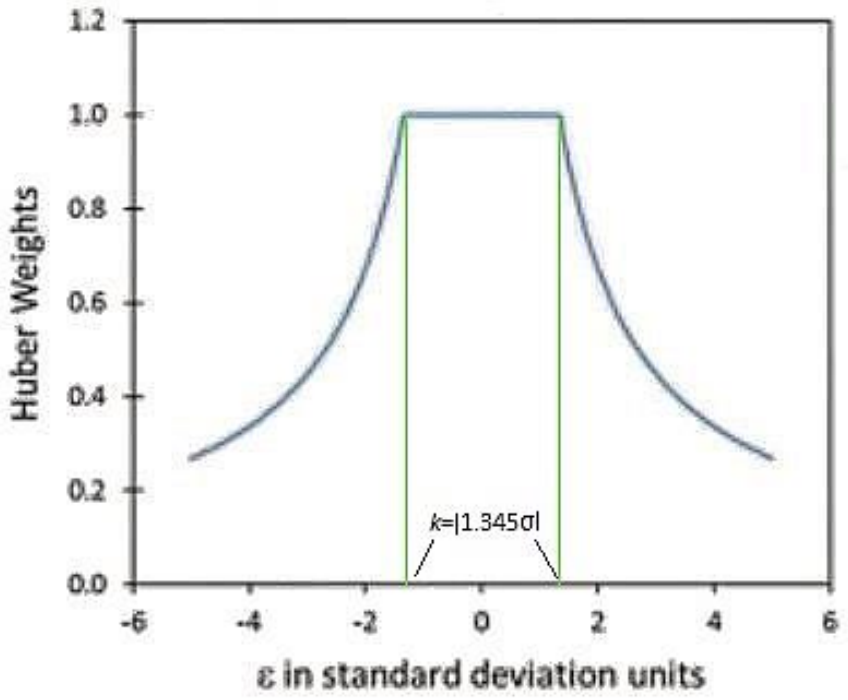
Now we have a scheme to carry out M – estimation. Next, we have to find a good function for H(ε) that we can use. This is called the influence function and there are a few of them developed for use in maximum likelihood estimation different in their efficiency and robustness. Influence functions are special curves relating the weight to be assigned to each observation to the residuals measured in their standard deviation units. The most commonly used functions are the Huber M – estimator and the Tukey's Bisquare M – estimator. Huber M – estimator tends to have better convergence properties than Tukey's Bisquare M – estimator, but Tukey's is more robust than Huber. These are the two influence functions used in the NCSS™ algorithm that we are going to use to model our fish data.

With the Huber M – estimator, what we do is to make our penalty function for having residuals H(ε) to go with the residual squares (just like the OLS estimator) **but** only out to a certain value (*k*) of the residual $|\varepsilon| \leq k$. Once the residual is above that then we will do an absolute value of the residual and

weighted linearly passes the value k, i.e. ($k|\varepsilon| - k^2/2$ for $|\varepsilon| > k$). Hence, the Huber M – estimator is represented by the following:

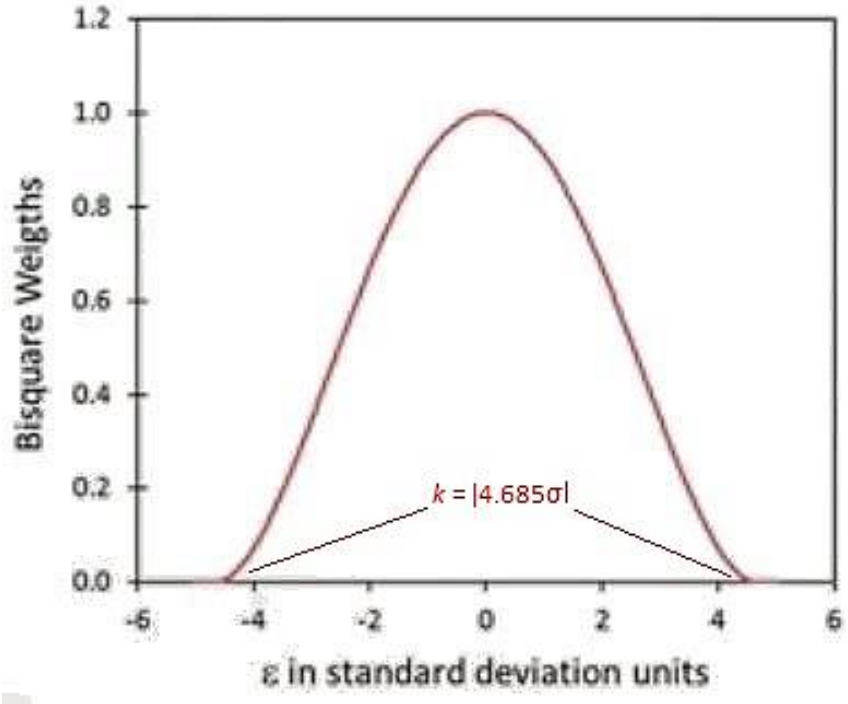
$$H(\varepsilon) = \begin{cases} \varepsilon^2/2 & for\ |\varepsilon| \le k \\ k|\varepsilon| - k^2/2 & for\ |\varepsilon| > k \end{cases}$$

The most commonly used value k that Peter Huber selected when developing the M – estimator (Huber 1981) is k = 1.345 multiply the standard deviation of the residuals (σ). The Huber M – estimator weight all the data point to be 1.0 until their residuals is 1.345σ away from the zero mean, then it starts weighting them lower and lower. k is known as the truncation constant (or tuning constant) which is the cut – off point on the influence function designating when an observation's weight should be reduced:



This k value renders the M – estimation as much as 95% as efficient as the OLS estimator, i.e. only a mere 5% loss in asymptotic efficiency! All in all, the Huber M – estimator weighting outliers much lower than the OLS estimator since once |ε| is above 1.345σ, the absolute value of the deviation instead of the square of the deviation is used in the weighting. Once we got a set of residuals, we Studentized these residuals using the median absolute deviation (MAD) as our estimative scale which allow us to have a robust estimate of Studentized residuals as well.

The Tukey's Bisquare (or Tukey's Biweight) M – estimator is even more robust than the Huber M – estimator because after you get out some distance away from the zero mean of the residuals measured in standard deviation units, the weighting becomes a constant (i.e. levelled off). That distance k from the zero mean was defined by John Tukey as its truncation constant $k = 4.685\sigma$. The Bisquare weighting immediately starts to drop off as the residuals begin to deviate from the zero mean. When the residuals reach 4.685σ, the weighting becomes zero (i.e. the Bisquare M – estimator just ignores those data points with residuals higher than or lower than 0.4685σ). Hence even very extreme outliers would not affect the regression model at all:



Apart from protecting the regression model from the influence of the presence of outliers, M – estimators can be used as a diagnostic tool to identify outliers by looking at the weight assigned to each data point. Data points associated with very low weight are probably outliers. In the case of Bisquare M - estimator, the weighting might even reach zero for those extreme outliers and in such cases these data points are ignored.

There are other even more robust estimators such as the least trimmed square (LTS) estimator and the least median square (LMS) estimator which are used in the area of regression modeling known as bounded influence regression which is often used in fully automated machine learning algorithms

free from any human intervention in the realm of artificial intelligence design. However, this area is beyond the scope of this chapter.

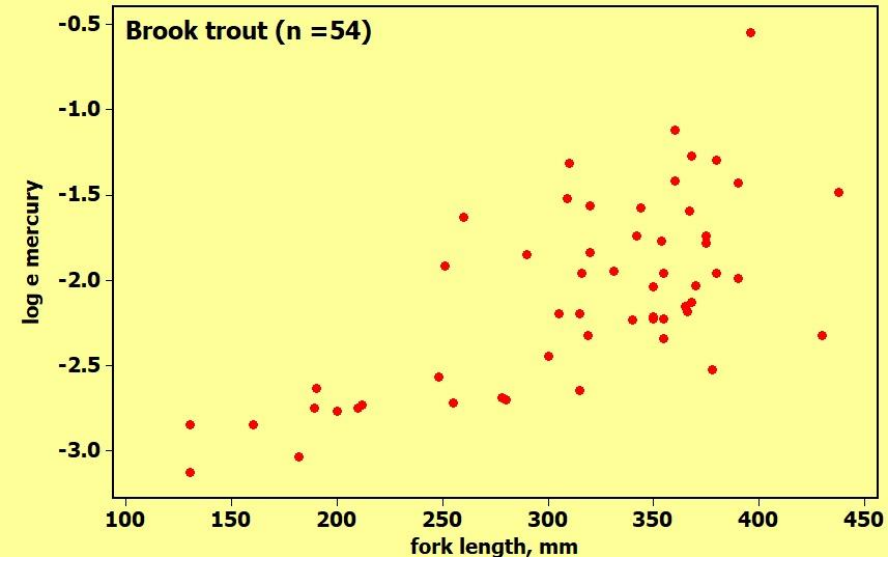
7.2 Robust regression modeling of the fish data.

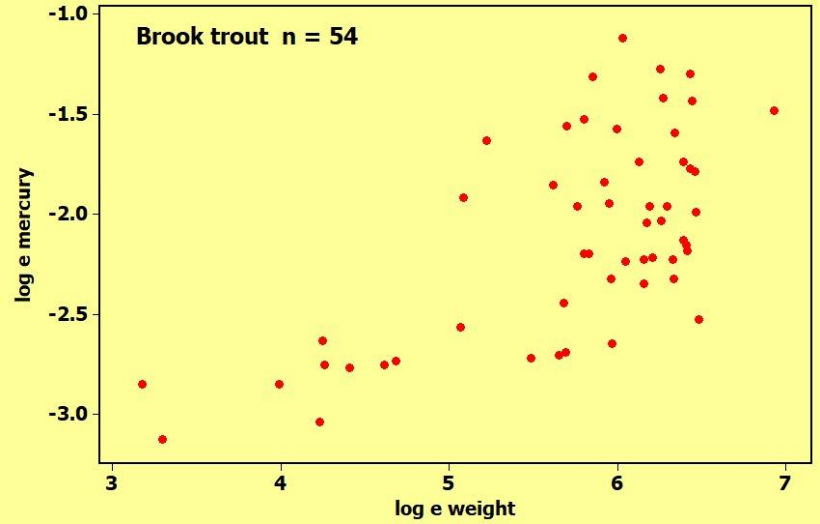
Robust regression modeling using M -estimators assumes a linear relationship between the dependent variable and the predictor variables despite the presence of outliers. In the present case, scatterplots between the dependent variable and predictors of **full** data sets were examined and necessary transformations made in order to ensure largely linear relationships before applying the robust regression modeling. The following linear models were established from full data sets (i.e. without any attempt to remove outliers) of the three fish species and their regression coefficients were determined by robust regression modeling:

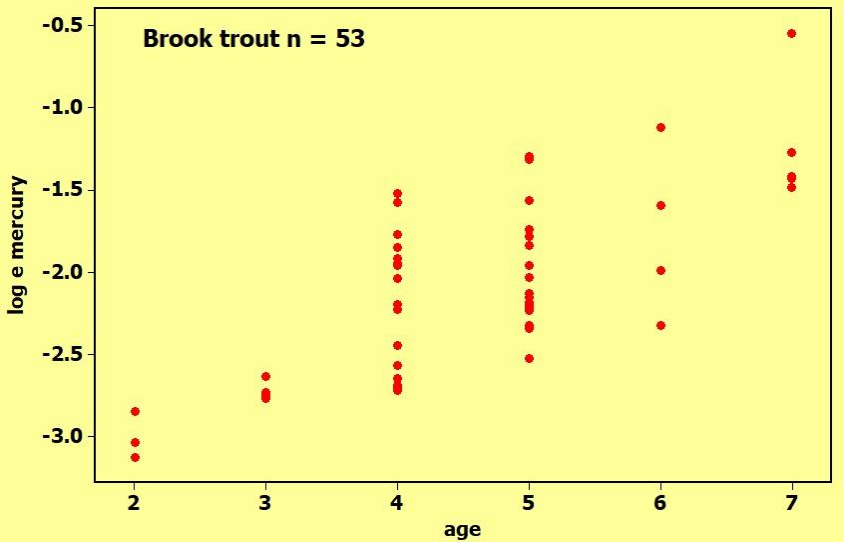
Book trout: $\log_e \text{mercury} = \beta_0 + \beta_1 * \text{fork length} + \beta_2 * \log_e \text{weight} + \beta_3 * \text{age}$

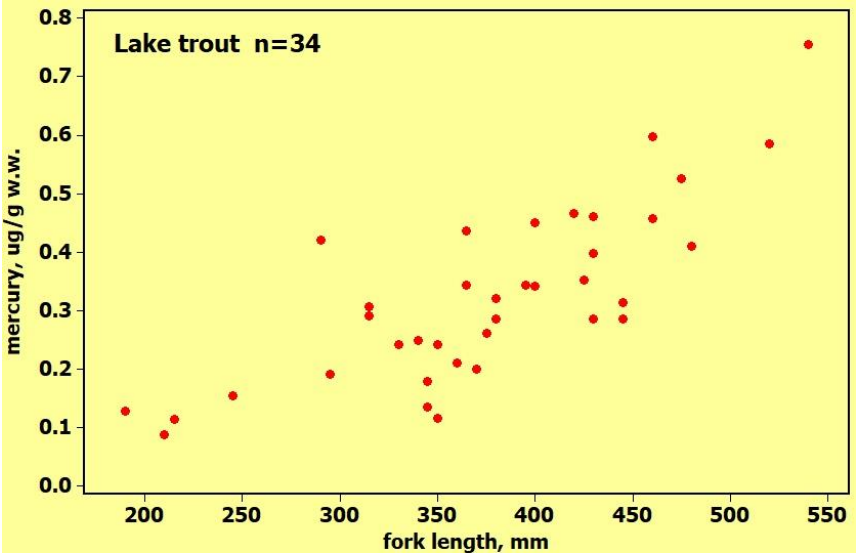
Lake trout: $\text{mercury} = \beta_0 + \beta_1 * \text{fork length} + \beta_2 * \text{weight} + \beta_3 * \text{age}$

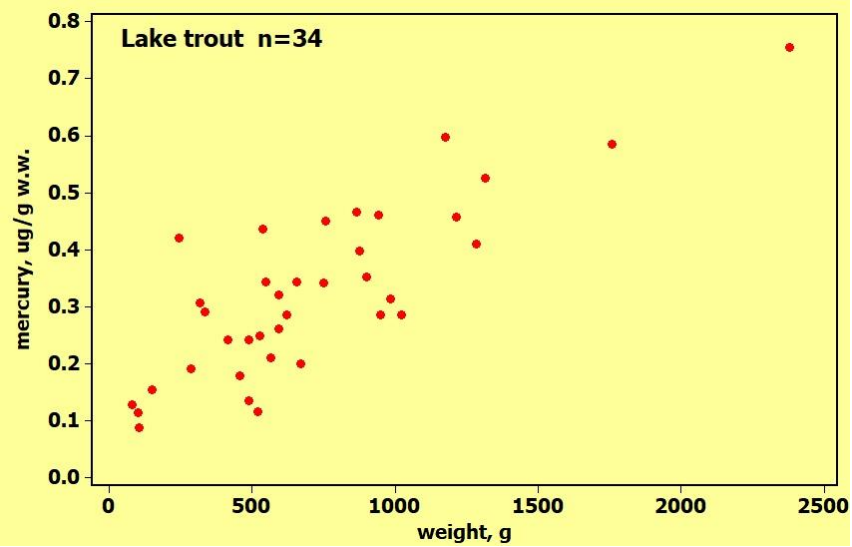
Lake whitefish: $\log_e \text{mercury} = \beta_0 + \beta_1 * (\text{fork length})^2 + \beta_2 * \text{weight} + \beta_3 * \text{age}$

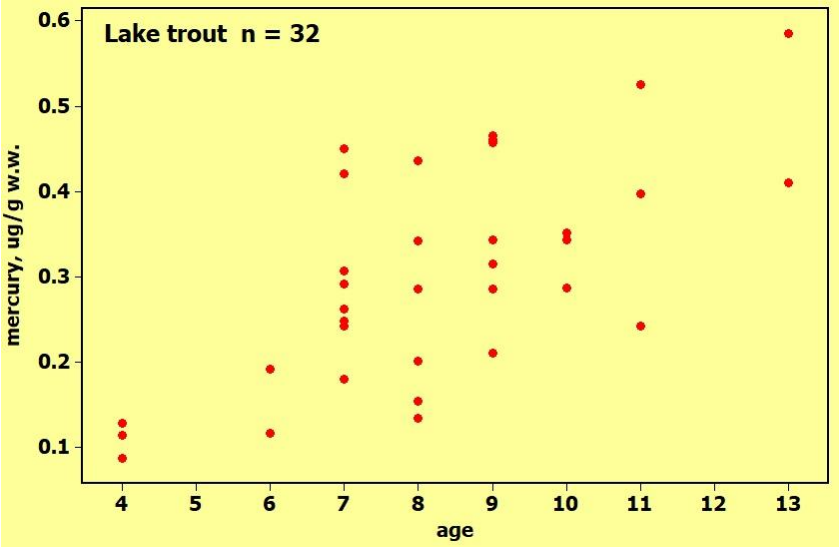


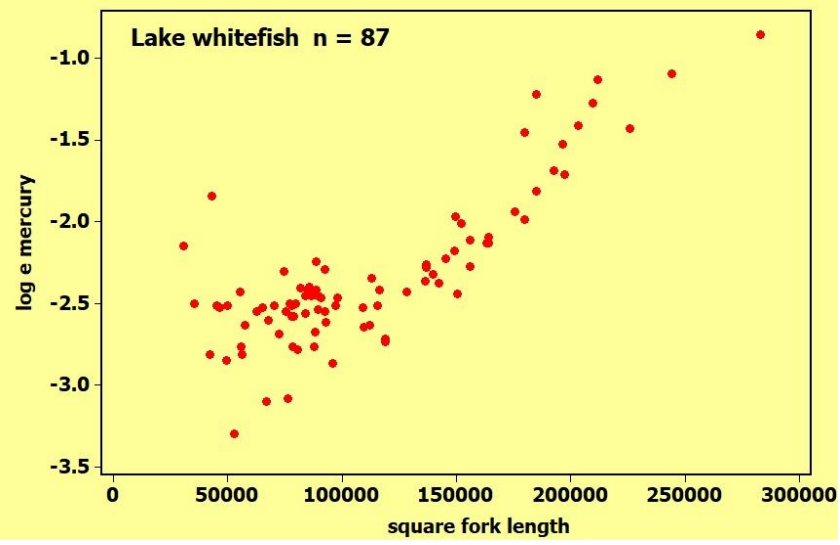

Brook trout n = 54
log e mercury
log e weight
-1.0
-1.5
-2.0
-2.5
-3.0
3
4
5
6
7

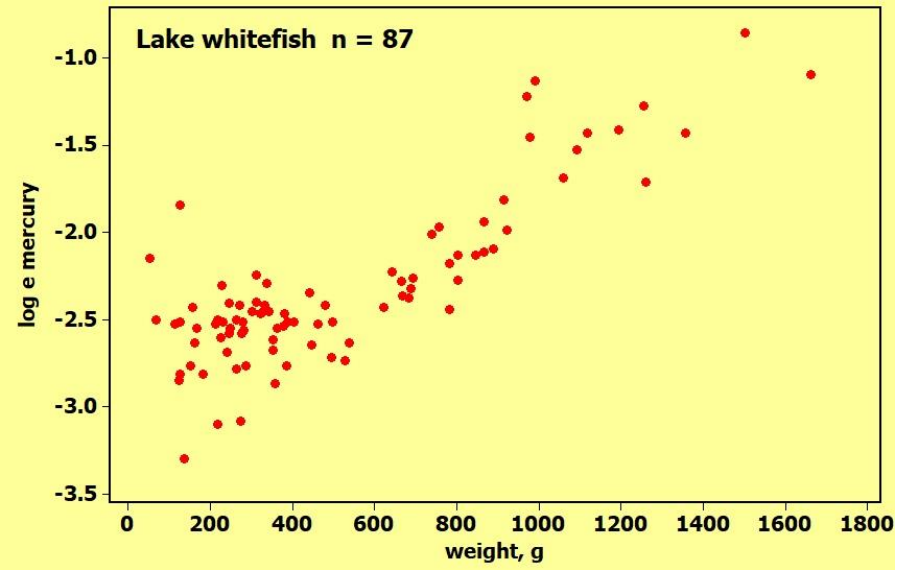

Brook trout n = 53
log e mercury
age
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
2
3
4
5
6
7

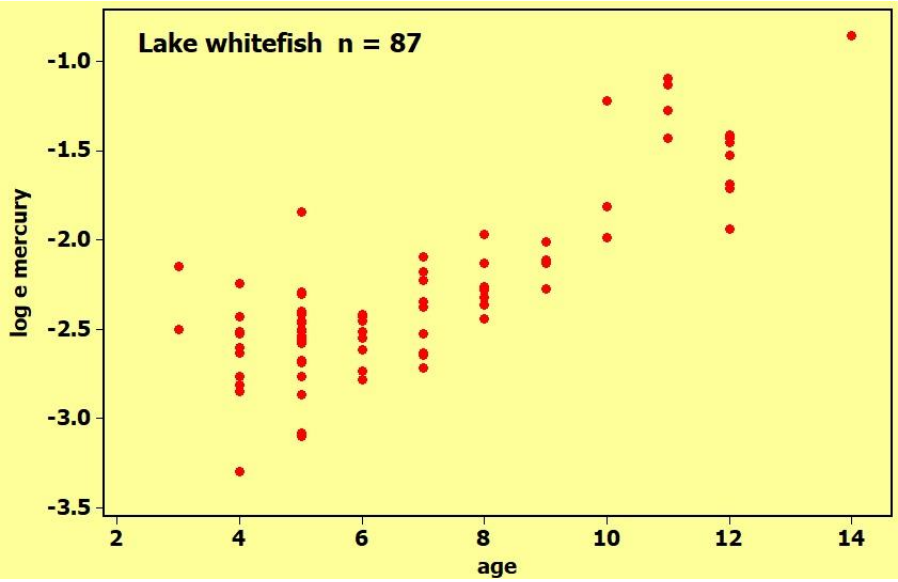

Lake trout n=34
mercury, ug/g w.w.
fork length, mm
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
200
250
300
350
400
450
500
550


Lake trout n=34
mercury, ug/g w.w.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
500
1000
1500
2000
2500
weight, g


Lake trout n = 32
mercury, ug/g w.w.
0.1
0.2
0.3
0.4
0.5
0.6
4
5
6
7
8
9
10
11
12
13
age


Lake whitefish n = 87
log e mercury
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5
0
50000
100000
150000
200000
250000
300000
square fork length





Because of the relatively extensive scattering of data points in each of these plots, it was decided that the more robust Tukey's Bisquare M – estimator was used as the influence function in assigning weight to the observations. The truncation (or tuning) constant used is the default value of 4.685. A maximum of 30 iterations was set to allow for the algorithm to find a solution (i.e. stabilization of coefficients and convergence of residuals). However, if the percentage change in each of the estimated regression coefficients was less than 0.001 (0.1%), the iteration process was terminated.

7.2.1 Brook trout

Run Summary Report

Item	Value
Dependent Variable	log_mercury
Number Ind. Variables	3
Weight Variable	None
Robust Method	Tukey's Biweight
Tuning Constant	4.685
MAD Scale Factor	0.6745

Rows	Value
Number Processed	55
Number Used in Estimation	53
Number Filtered Out	0
Number with X's Missing	2
Number with Weight Missing	0
Number with Y Missing	0
Sum of Robust Weights	46.307

Run Information	Value
Iterations	18
Max % Change in any Coef	0.001
R^2 after Robust Weighting	0.7041
S using MAD	0.2670881
S using MSE	0.2613744
Completion Status	Normal Completion

A total of 18 iterations were required for the percentage change in the change of estimated regression coefficients to be stabilized to less than 0.1%. The sum of squares determined using median absolute deviation (MAD) was very similar to the sum of squares determined using mean squared error (MSE); hence, the impact of the outliers was manageable.

Robust Iterations - Coefficients

Robust Iteration	Max Percent Change in Coefficients	b(0)	b(1)	b(2)	b(3)
0		-3.612257	0.001657044	-0.004824827	0.2206958
1	2449.781	-3.392699	0.003164994	-0.1230225	0.2149496
2	66.941	-3.233854	0.004216722	-0.2053745	0.2108608
3	28.152	-3.12208	0.004967087	-0.2631922	0.2075811
4	15.743	-3.042111	0.005511398	-0.3046266	0.2049828
5	19.393	-2.928999	0.006282458	-0.3637031	0.2015011
6	15.421	-2.821163	0.007016605	-0.4197907	0.198133
7	9.360	-2.745466	0.007535489	-0.4590819	0.1956089
8	7.764	-2.6773	0.008005216	-0.4947255	0.1933836
9	3.012	-2.648502	0.008205343	-0.5096287	0.1923051
10	1.035	-2.638367	0.008278891	-0.5149037	0.1918157
11	0.178	-2.636642	0.008293658	-0.5158081	0.1916425
12	0.029	-2.636953	0.008293234	-0.5156604	0.1915881
13	0.050	-2.637465	0.00829051	-0.5154032	0.1915755
14	0.032	-2.637793	0.008288572	-0.5152375	0.191575
15	0.016	-2.637951	0.008287585	-0.5151569	0.1915767
16	0.006	-2.638015	0.008287171	-0.5151244	0.191578
17	0.002	-2.638036	0.008287026	-0.5151135	0.1915786
18	0.001	-2.638041	0.008286986	-0.5151107	0.1915789

The above table shows the largest percentage change in any of the four coefficients: b(0) is the constant (β_0); b(1) is the estimated regression coefficient associated with fork length (β_1); b(2) is the estimated regression coefficient associated with $\log_e$ weight (β_2); b(3) is the estimated regression

coefficient associated with age (β_3). The 0th iteration shows the OLS estimates on the full data set. The coefficients were stabilized after 17 iterations, the percentage change in coefficients reached 0.001 at the 18th iteration.

Robust Iterations - Residuals

Robust Iteration	Max Percent Change in Coefficients	Percentiles of Absolute Residuals 25th	50th	75th	100th
0		0.1234726	0.2193017	0.3931872	0.7221807
1	2449.781	0.1153523	0.2032511	0.3609739	0.801163
2	66.941	0.1131731	0.2080406	0.3356225	0.8658684
3	28.152	0.1147424	0.2141136	0.3465028	0.9142397
4	15.743	0.1181614	0.2111502	0.3407625	0.9501097
5	19.393	0.1242985	0.2081149	0.3302207	0.9995291
6	15.421	0.123421	0.2133148	0.3210711	1.047423
7	9.360	0.1224823	0.2061976	0.3156023	1.082114
8	7.764	0.1212718	0.2010955	0.3100844	1.113043
9	3.012	0.1191329	0.2028633	0.3086561	1.126986
10	1.035	0.1179491	0.2038903	0.3085157	1.132383
11	0.178	0.1173958	0.2044017	0.3088018	1.133688
12	0.029	0.1171845	0.2046059	0.3090202	1.133803
13	0.050	0.1171186	0.2046734	0.3091353	1.133675
14	0.032	0.1171042	0.2046901	0.3091848	1.133561
15	0.016	0.1171043	0.2046914	0.3092026	1.133499
16	0.006	0.1171067	0.2046897	0.3092078	1.133471
17	0.002	0.1171085	0.2046882	0.3092087	1.13346
18	0.001	0.1171094	0.2046874	0.3092085	1.133457

The above table shows that the absolute values of the residuals have converged at the 18th iteration.

The absolute values of the residuals are sorted and the percentiles were calculated. The iteration process was terminated when there was little change in median of the absolute residuals.

Regression Coefficients T-Tests Assuming Random Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Standardized Coefficient	T-Statistic to Test H0: β(i)=0	Prob Level
Intercept	-2.638041	0.6117975	0.0000	-4.312	0.0001
fork_length	0.008286986	0.003289636	1.2464	2.519	0.0151
log_weight	-0.5151107	0.2577036	-0.9075	-1.999	0.0512
age	0.1915789	0.07842185	0.4865	2.443	0.0182

In the above table, the standardized regression coefficients show that fork length is the most important predictor of $\log_e$ mercury. Age also has a positive relationship with $\log_e$ mercury. However, the negative coefficient associated with $\log_e$ weight seem anomalous.

Regression Coefficients Confidence Intervals Assuming Random Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	-2.638041	0.6117975	-3.867494	-1.408588
fork_length	0.008286986	0.003289636	0.001676214	0.01489776
log_weight	-0.5151107	0.2577036	-1.032985	0.002764014
age	0.1915789	0.07842185	0.03398431	0.3491735

Note: The T-Value used to calculate these confidence limits was 2.010.

The Lower and Upper 95% Conf. Limits of β(i) are the lower and upper values of a 100(1 – α)% interval estimate for β_j based on a t – distribution with n – p – 1 degrees of freedom. This interval estimate assumes that the residuals for the regression model are normally distributed which might not be the case in the presence of outliers; hence, bootstrapping technique was used to calculated the confidence intervals. The formulas for these lower and upper confidence limits are: $b_j \pm t_{1-\alpha/2,\, n-p-1}\, s_{bj}$

The T – Value is the value of $t_{1-\alpha/2,\, n-p-1}$ used to construct the confidence limits.

The estimated robust regression model for $\log_e$ mercury on fork length, $\log_e$ weight and age of brook trout is:

$\log_e$ mercury = (-2.6380) +(0.00829) * fork length + (-0.5151) * $\log_e$ weight + (0.1916) * age

The iterative resampling – with – replacement method of bootstrapping was used here to compute the confidence intervals. 3000 draws of bootstrap samples were used; hence, provided 3000 estimates of the coefficients and the confidence limits. The reflection method was used to calculate the bootstrap confidence intervals in which confidence limits were formed by reflecting the percentiles of the bootstrap limits. In the Bootstrap Section Table below, the Original Value is the parameter estimate obtained from the complete original data set **without** bootstrapping. The Bootstrap Mean is the average of the parameter estimates of the 3000 bootstrap samples. The Bias (BM – OV) is the estimate of the bias in the original estimate. It is computed by subtracting the Original Value from the Bootstrap Mean. The Bias Corrected is an estimate of the parameter that has been corrected for its bias. The correction is made by subtracting the estimated bias from the original parameter estimate. The Standard Error term is the bootstrap method's estimate of the standard error of the parameter estimate. It is simply the standard deviation of the parameter estimate computed from the bootstrap estimate. Note that all four

Parameters (Intercept, coefficients for fork length, weight and age) have the same signs as their corresponding upper and lower confidence limits calculated by bootstrapping and these four parameters were all within their corresponding upper and lower confidence limits.

Bootstrap Report for Coefficients

Estimation Results		Bootstrap Confidence Limits	
Parameter	**Estimate**	**Conf. Level**	**Lower**	**Upper**
Intercept				
Original Value	-2.638041	90.000	-3.678781	-0.5841143
Bootstrap Mean	-3.03971	95.000	-4.03103	-0.3638131
Bias (BM - OV)	-0.4016692	99.000	-4.701258	0.1476146
Bias Corrected	-2.236372			
Standard Error	0.9963622			
B(fork_length)				
Original Value	0.008286986	90.000	0.002575874	0.02176121
Bootstrap Mean	0.005304064	95.000	0.0009623758	0.02274342
Bias (BM - OV)	-0.002982922	99.000	-0.001819226	0.02475914
Bias Corrected	0.01126991			
Standard Error	0.006259453			
B(log_weight)				
Original Value	-0.5151107	90.000	-1.5675	-0.03855756
Bootstrap Mean	-0.2964538	95.000	-1.646158	0.07361924
Bias (BM - OV)	0.2186569	99.000	-1.851418	0.3843618
Bias Corrected	-0.7337676			
Standard Error	0.4913193			
B(age)				
Original Value	0.1915789	90.000	0.04611349	0.3166707
Bootstrap Mean	0.2072903	95.000	0.02384769	0.3663127
Bias (BM - OV)	0.01571138	99.000	-0.04182431	0.4698534
Bias Corrected	0.1758676			
Standard Error	0.0842944			

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.

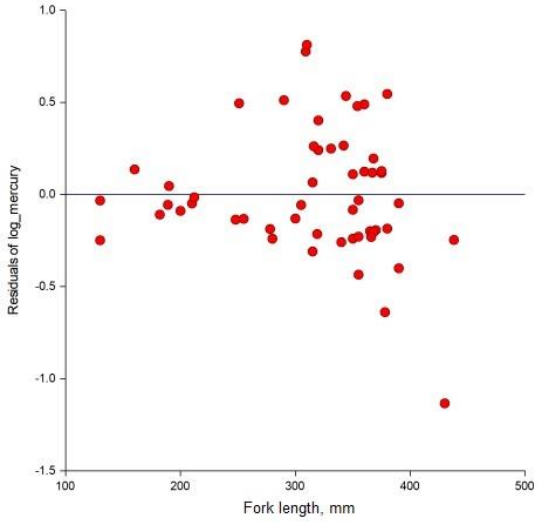
Robust Residuals and Weights

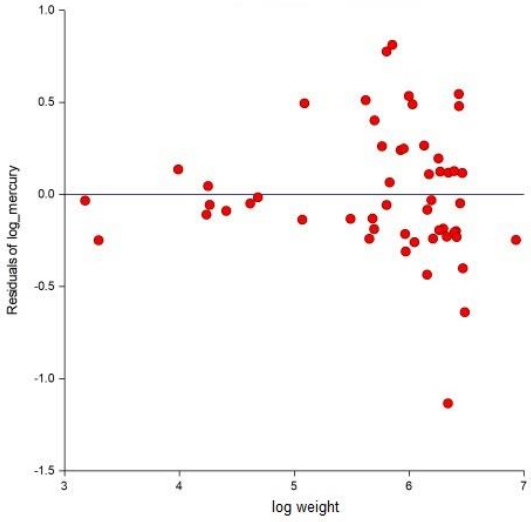
Row	Actual log_mercury	Predicted log_mercury	Residual	Absolute Percent Error	Robust Weight
1	-1.43129	-1.383686	-0.0476043	3.326	0.9974
2	-1.56065	-1.962943	0.4022933	25.777	0.8042
3	-2.56395	-2.427598	-0.1363522	5.318	0.9767
4	-1.78379	-1.900899	0.1171094	6.565	0.9828
5	-1.959	-1.773853	-0.1851469	9.451	0.9570
7	-1.98777	-1.587421	-0.4003488	20.141	0.8060
8	-2.19823	-2.26388	0.06564964	2.986	0.9948
9	-1.41882	-1.542954	0.1241345	8.749	0.9807
10	-1.59455	-1.713231	0.1186813	7.443	0.9824
11	-2.32279	-1.189333	-1.133457	48.797	0.0322
12	-1.52326	-2.299781	0.7765211	50.978	0.3782
13	-2.44185	-2.311546	-0.1303037	5.336	0.9787
14	-1.31304	-2.125712	0.812672	61.892	0.3344
15	-2.15417	-1.955649	-0.1985206	9.216	0.9506
16	-1.959	-1.928018	-0.03098234	1.582	0.9991
17	-2.03256	-1.839322	-0.1932375	9.507	0.9531
18	-2.04022	-2.150389	0.1101693	5.400	0.9848
19	-2.22562	-2.142807	-0.08281309	3.721	0.9915
20	-2.21641	-1.976777	-0.2396325	10.812	0.9283
21	-2.19823	-2.14135	-0.05687985	2.588	0.9962
22	-1.29828	-1.843939	0.545659	42.029	0.6560
23	-1.83885	-2.079961	0.241111	13.112	0.9274
24	-2.63109	-2.677225	0.04613505	1.753	0.9976
25	-2.76462	-2.675856	-0.08876399	3.211	0.9902
26	-1.85151	-2.363628	0.5121179	27.659	0.6933
27	-2.23493	-1.976421	-0.2585091	11.567	0.9167
28	-2.34341	-1.908701	-0.434709	18.550	0.7734
29	-1.12086	-1.610484	0.4896244	43.683	0.7174
30	-1.73727	-1.864208	0.1269381	7.307	0.9798
31	-2.52573	-1.887185	-0.6385446	25.282	0.5472
32	-2.12863	-1.923943	-0.2046874	9.616	0.9475
33	-2.84731	-2.814623	-0.03268715	1.148	0.9989
34	-2.64508	-2.335871	-0.3092085	11.690	0.8819
35	-1.48281	-1.236254	-0.2465558	16.628	0.9241
36	-2.32279	-2.108502	-0.2142879	9.225	0.9425
37	-2.18037	-1.949907	-0.230463	10.570	0.9336
38	-1.77196	-2.251809	0.4798489	27.080	0.7277
39	-2.68825	-2.500845	-0.1874052	6.971	0.9559
40	-3.12357	-2.875298	-0.2482723	7.948	0.9231
41	-2.84731	-2.983732	0.136422	4.791	0.9767
42	-2.22562	-1.996934	-0.2286857	10.275	0.9346
43	-2.7181	-2.585956	-0.132144	4.862	0.9781
44	-2.70306	-2.463028	-0.2400323	8.880	0.9280
45	-2.74887	-2.692816	-0.0560537	2.039	0.9963
46	-1.959	-2.221132	0.2621318	13.381	0.9144
47	-1.57504	-2.109838	0.5347981	33.955	0.6682
49	-1.73727	-2.003137	0.2658668	15.304	0.9120
50	-2.74887	-2.700335	-0.04853477	1.766	0.9973
51	-3.03655	-2.927687	-0.1088626	3.585	0.9852
52	-2.73337	-2.718279	-0.01509118	0.552	1.0000
53	-1.91732	-2.412369	0.4950494	25.820	0.7117
54	-1.94491	-2.193972	0.2490617	12.806	0.9226
55	-1.27297	-1.468793	0.1958228	15.383	0.9519

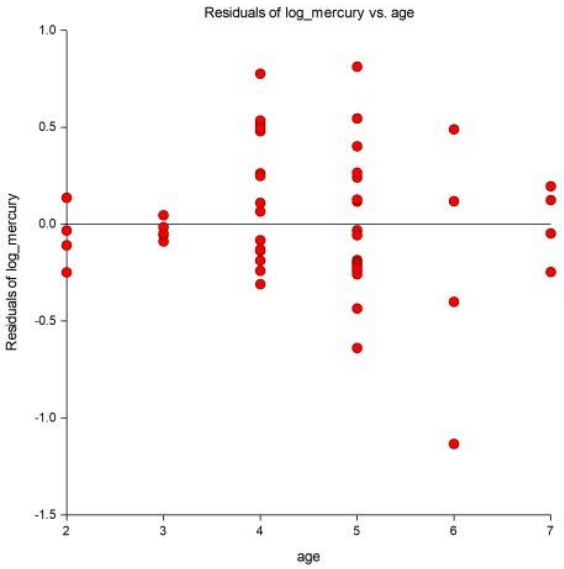
The Robust Residuals and Weights Table above listed out the residuals and the weights assigned by Tukey's Bisquare M – estimator to each of the observations in the entire data set (n = 55). It is apparent that row numbers: 11, 22, 31, and 47 as well as to a lesser extent 28, 29, 38 and 53 were

outliers which received less weight than the residuals of the rest of the data points. These observations with lower weights made only minimum contributions to the determination of the regression coefficients. The table also gives the predicted values of $\log_e$ mercury based on the robust regression equation above from the final iteration for comparing side by side with the actual measured $\log_e$ mercury values. Hence, the residuals were just the difference between the actual measured $\log_e$ mercury and the predicted $\log_e$ mercury. The Absolute Percent Error = (Residual / Actual $\log_e$ mercury) * 100.

The residuals versus predictors plots highlight the presence of outliers in the data set:







7.2.2 Lake trout

Run Summary Report

Item	Value
Dependent Variable	mercury
Number Ind. Variables	3
Weight Variable	None
Robust Method	Tukey's Biweight
Tuning Constant	4.685
MAD Scale Factor	0.6745

Rows	Value
Number Processed	37
Number Used in Estimation	34
Number Filtered Out	0
Number with X's Missing	3
Number with Weight Missing	0
Number with Y Missing	0
Sum of Robust Weights	31.146

Run Information	Value
Iterations	8
Max % Change in any Coef	0.000
R^2 after Robust Weighting	0.6629
S using MAD	0.08150773
S using MSE	0.07440012
Completion Status	Normal Completion

Only 8 iterations were sufficient for the percentage change in the change of estimated regression coefficients to be stabilized to less than 0.1%. The sum of squares determined using median absolute deviation (MAD) was very similar to the sum of squares determined using mean squared error (MSE); hence, the impact of the outliers was manageable.

Robust Iterations - Coefficients

Robust Iteration	Max Percent Change in Coefficients	b(0)	b(1)	b(2)	b(3)
0		-0.02310083	0.0005352143	0.0001182324	0.006198404
1	26.560	-0.02124146	0.0005266641	0.0001365252	0.004552105
2	7.181	-0.01972395	0.0005181392	0.0001418332	0.004225227
3	1.207	-0.01948592	0.0005163095	0.0001420057	0.004257782
4	0.179	-0.01945095	0.0005159886	0.0001420346	0.004264247
5	0.033	-0.01944444	0.0005159267	0.0001420401	0.004265526
6	0.006	-0.01944318	0.0005159146	0.0001420412	0.004265778
7	0.001	-0.01944294	0.0005159123	0.0001420414	0.004265827
8	0.000	-0.01944289	0.0005159118	0.0001420414	0.004265836

The above table shows the largest percentage change in any of the four coefficients: b(0) is the constant (β_0); b(1) is the estimated regression coefficient associated with fork length (β_1); b(2) is the estimated regression coefficient associated with weight (β_2); b(3) is the estimated regression coefficient associated with age (β_3). The 0^{th} iteration shows the OLS estimates on the full data set. The coefficients were stabilized after only 7 iterations.

Robust Iterations - Residuals

Robust Iteration	Max Percent Change in Coefficients	Percentiles of Absolute Residuals 25th	50th	75th	100th
0		0.02046699	0.05132517	0.08902947	0.2164147
1	26.560	0.02104484	0.05398914	0.08975546	0.2240589
2	7.181	0.02202993	0.05496731	0.0897027	0.225996
3	1.207	0.02199552	0.05497459	0.08961066	0.2260183
4	0.179	0.02199336	0.05497649	0.08959322	0.226024
5	0.033	0.02199305	0.05497687	0.08958984	0.2260251
6	0.006	0.02199299	0.05497695	0.08958919	0.2260254
7	0.001	0.02199298	0.05497696	0.08958906	0.2260254
8	0.000	0.02199298	0.05497697	0.08958904	0.2260254

The above table shows that the absolute values of the residuals have converged at the 8th iteration. The absolute values of the residuals were sorted and the percentiles were calculated. The median of the absolute residuals in fact began to converge between the 6th and the 7th iterations.

Regression Coefficients T-Tests Assuming Random Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Standardized Coefficient	T-Statistic to Test H0: β(i)=0	Prob Level
Intercept	-0.01944289	0.1606855	0.0000	-0.121	0.9045
fork_length	0.0005159118	0.0007015544	0.3244	0.735	0.4678
weight	0.0001420414	0.0001349827	0.4349	1.052	0.3011
age	0.004265836	0.01323775	0.0749	0.322	0.7495

In the above table, the standardized regression coefficients show that weight followed by fork length were the most important predictors of $\log_e$ mercury, with age being the least important. All three predictors have a positive relationship with $\log_e$ mercury.

Regression Coefficients Confidence Intervals Assuming Random Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	-0.01944289	0.1606855	-0.3476065	0.3087207
fork_length	0.0005159118	0.0007015544	-0.0009168534	0.001948677
weight	0.0001420414	0.0001349827	-0.00013363	0.0004177128
age	0.004265836	0.01323775	-0.02276925	0.03130092

Note: The T-Value used to calculate these confidence limits was 2.042.

The Lower and Upper 95% Conf. Limits of β(i) were calculated based on the assumption that the residuals for the regression model are normally distributed which might not be the case in the presence of outliers; hence, bootstrapping technique was used to calculated the confidence intervals.

The estimated robust regression model for mercury on fork length, weight and age of lake trout is:

mercury = (-0.01944) +(5.16 * 10^{-4}) * fork length + (1.42 * 10^{-4}) * weight + (4.27 * 10^{-3}) * age

The following Bootstrap Report compared the original confident limits with those generated by 3000 draws using the iterative resampling – with – replacement bootstrapping method:

Bootstrap Report for Coefficients

Estimation Results		Bootstrap Confidence Limits		
Parameter	**Estimate**	**Conf. Level**	**Lower**	**Upper**
Intercept				
Original Value	-0.01944289	90.000	-0.1637438	0.2090507
Bootstrap Mean	-0.03477881	95.000	-0.196568	0.2678696
Bias (BM - OV)	-0.01533592	99.000	-0.3959239	0.4009531
Bias Corrected	-0.004106971			
Standard Error	0.1289335			
B(fork_length)				
Original Value	0.0005159118	90.000	-0.0006500956	0.001425787
Bootstrap Mean	0.000567693	95.000	-0.0009736057	0.001632539
Bias (BM - OV)	5.178111E-05	99.000	-0.001652968	0.002217497
Bias Corrected	0.0004641308			
Standard Error	0.0006671615			
B(weight)				
Original Value	0.0001420414	90.000	-1.689469E-05	0.0004464629
Bootstrap Mean	0.0001190075	95.000	-5.239676E-05	0.000527321
Bias (BM - OV)	-2.303393E-05	99.000	-0.0001371637	0.0006828441
Bias Corrected	0.0001650753			
Standard Error	0.0001416868			
B(age)				
Original Value	0.004265836	90.000	-0.02124089	0.02205884
Bootstrap Mean	0.005434943	95.000	-0.02956825	0.02537567
Bias (BM - OV)	0.001169107	99.000	-0.05181748	0.03590174
Bias Corrected	0.00309673			
Standard Error	0.01390329			

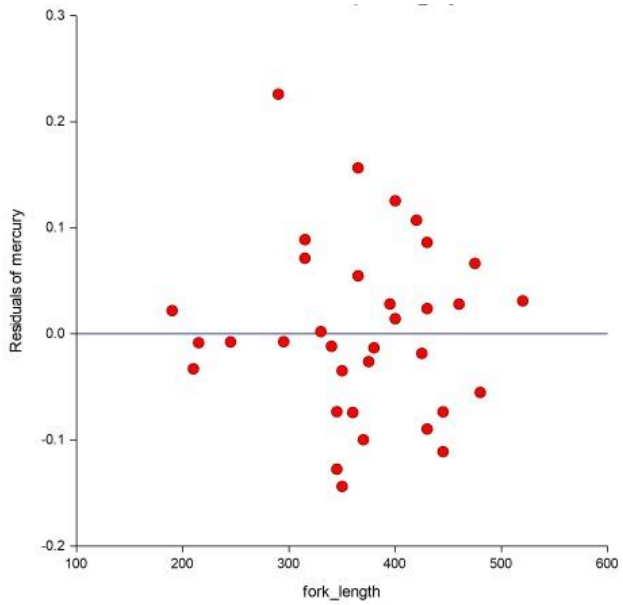
Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.

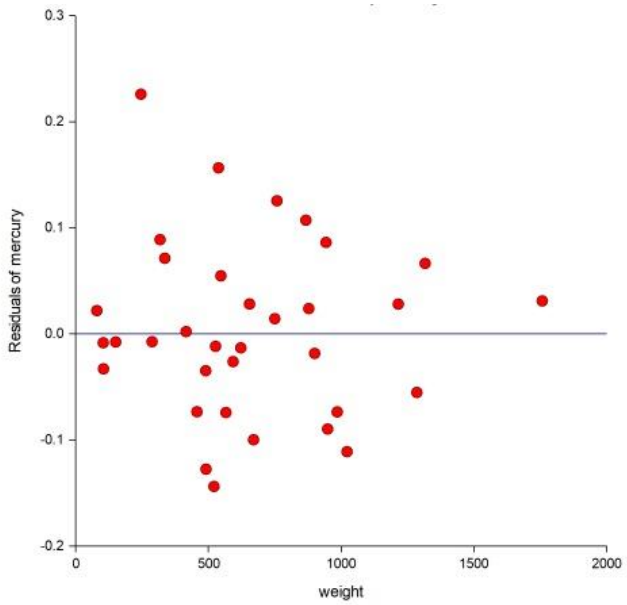
Robust Residuals and Weights

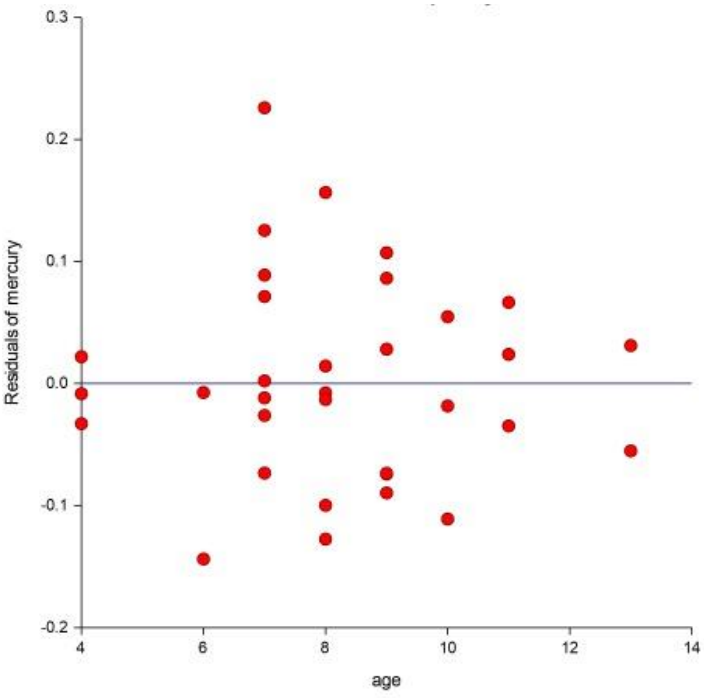
Row	Actual mercury	Predicted mercury	Residual	Absolute Percent Error	Robust Weight
2	0.115	0.1233138	-0.008313815	7.229	0.9991
3	0.421	0.1949746	0.2260254	53.688	0.4221
4	0.286	0.375589	-0.08958904	31.325	0.8930
5	0.585	0.5538539	0.03114609	5.324	0.9868
6	0.45	0.3244501	0.1255499	27.900	0.7955
7	0.211	0.2850733	-0.07407334	35.106	0.9262
8	0.18	0.2533205	-0.07332049	40.734	0.9277
9	0.262	0.2881155	-0.02611547	9.968	0.9907
10	0.242	0.2397581	0.002241892	0.926	1.0000
11	0.457	0.4288494	0.02815059	6.160	0.9892
12	0.155	0.1625305	-0.007530463	4.858	0.9993
13	0.192	0.1992541	-0.007254056	3.778	0.9993
14	0.201	0.3007389	-0.09973893	49.621	0.8683
15	0.461	0.3747368	0.08626321	18.712	0.9006
17	0.135	0.2624157	-0.1274157	94.382	0.7898
18	0.243	0.2776507	-0.03465075	14.260	0.9837
20	0.466	0.3587825	0.1072175	23.008	0.8486
21	0.315	0.3884412	-0.07344121	23.315	0.9275
22	0.398	0.3740358	0.02396423	6.021	0.9922
23	0.344	0.3157719	0.02822805	8.206	0.9892
24	0.286	0.2990801	-0.01308007	4.573	0.9977
25	0.352	0.3703153	-0.01831529	5.203	0.9955
26	0.344	0.2892199	0.05478004	15.924	0.9593
27	0.292	0.2206561	0.07134388	24.433	0.9315
28	0.342	0.3275796	0.01442039	4.216	0.9972
29	0.436	0.2794099	0.1565901	35.915	0.6920
30	0.117	0.2607248	-0.1437249	122.842	0.7368
31	0.249	0.2606838	-0.01168382	4.692	0.9982
32	0.307	0.2180994	0.08890063	28.958	0.8946
33	0.287	0.3979626	-0.1109626	38.663	0.8383
34	0.088	0.1208763	-0.0328763	37.359	0.9853
35	0.526	0.459466	0.06653406	12.649	0.9403
36	0.411	0.4661739	-0.05517389	13.424	0.9587
37	0.129	0.107007	0.02199298	17.049	0.9934

The Robust Residuals and Weights Table above listed out the residuals and the weights assigned by Tukey's Bisquare M – estimator to each of the observations in the entire data set (n = 37). Five possible outliers were identified: rows 3, 29 and to a lesser extent, possibly rows 6, 17 and 30 which received less weight than the residuals of the rest of the data points. These observations with lower weights made only minimum contributions to the determination of the regression coefficients. The table also gives the predicted values of $\log_e$ mercury based on the robust regression equation above from the final iteration for comparing side by side with the actual measured $\log_e$ mercury values.

The residuals versus predictors plots highlight the presence of outliers in the data set:







7.2.3 Lake whitefish

Run Summary Report

Item	Value
Dependent Variable	log_mercury
Number Ind. Variables	3
Weight Variable	None
Robust Method	Tukey's Biweight
Tuning Constant	4.685
MAD Scale Factor	0.6745

Rows	Value
Number Processed	87
Number Used in Estimation	87
Number Filtered Out	0
Number with X's Missing	0
Number with Weight Missing	0
Number with Y Missing	0
Sum of Robust Weights	78.098

Run Information	Value
Iterations	11
Max % Change in any Coef	0.001
R^2 after Robust Weighting	0.7939
S using MAD	0.2197888
S using MSE	0.1965596
Completion Status	Normal Completion

Eleven iterations were required for the percentage change in the change of estimated regression coefficients to be stabilized to less than 0.1%. The sum of squares determined using median absolute deviation (MAD) was not greatly different from the sum of squares determined using mean squared error (MSE); hence, the impact of the outliers was manageable.

Robust Iterations - Coefficients

Robust Iteration	Max Percent Change in Coefficients	b(0)	b(1)	b(2)	b(3)
0		-3.095755	5.584481E-07	0.0005996649	0.06055696
1	223.492	-3.105205	1.806533E-06	0.0005437454	0.04296615
2	12.554	-3.113896	2.033319E-06	0.0005273735	0.04143128
3	3.919	-3.117316	2.112997E-06	0.0005202786	0.04106675
4	1.377	-3.118563	2.142103E-06	0.0005175427	0.04094465
5	0.487	-3.119007	2.152532E-06	0.0005165465	0.04090163
6	0.172	-3.119164	2.156224E-06	0.0005161919	0.04088644
7	0.060	-3.11922	2.157525E-06	0.0005160669	0.04088109
8	0.021	-3.119239	2.157983E-06	0.0005160228	0.0408792
9	0.007	-3.119246	2.158143E-06	0.0005160074	0.04087854
10	0.003	-3.119249	2.1582E-06	0.000516002	0.04087831
11	0.001	-3.11925	2.15822E-06	0.000516	0.04087823

All the estimated regression coefficients stabilized after 11 iterations when the maximum percentage change become 0.1%.

Robust Iterations - Residuals

Robust Iteration	Max Percent Change in Coefficients	Percentiles of Absolute Residuals 25th	50th	75th	100th
0		0.07618278	0.1434869	0.2509609	0.8480917
1	223.492	0.06933996	0.1369131	0.2622249	0.8985443
2	12.554	0.06456183	0.1394394	0.2678245	0.9071613
3	3.919	0.06374979	0.1401639	0.2696272	0.9098507
4	1.377	0.06354098	0.1404093	0.2702567	0.9107932
5	0.487	0.06367098	0.1404918	0.2704792	0.9111272
6	0.172	0.06371715	0.14052	0.2705579	0.9112452
7	0.060	0.06373348	0.1405298	0.2705856	0.9112869
8	0.021	0.06373925	0.1405332	0.2705954	0.9113016
9	0.007	0.06374127	0.1405344	0.2705989	0.9113067
10	0.003	0.06374199	0.1405348	0.2706001	0.9113085
11	0.001	0.06374224	0.1405349	0.2706005	0.9113091

The median of the absolute residuals converged between the 10th and the 11th iterations.

Regression Coefficients T-Tests Assuming Random Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Standardized Coefficient	T-Statistic to Test H0: β(i)=0	Prob Level
Intercept	-3.11925	0.1233988	0.0000	-25.278	0.0000
square_fork_length	2.15822E-06	2.591722E-06	0.2561	0.833	0.4074
weight	0.000516	0.0003782357	0.4125	1.364	0.1762
age	0.04087823	0.0321996	0.2343	1.270	0.2078

The standardized regression coefficients show that weight was the most important predictors of $\log_e$ mercury, whilst square of fork length and age have similar influence on $\log_e$ mercury. All three predictors have a positive relationship with $\log_e$ mercury.

Regression Coefficients Confidence Intervals Assuming Random Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	-3.11925	0.1233988	-3.364685	-2.873814
square_fork_length	2.15822E-06	2.591722E-06	-2.996611E-06	7.313051E-06
weight	0.000516	0.0003782357	-0.0002362956	0.001268296
age	0.04087823	0.0321996	-0.02316548	0.1049219

Note: The T-Value used to calculate these confidence limits was 1.989.

The estimated robust regression model for $\log_e$ mercury on square fork length, weight and age of lake whitefish is:

$\log_e \text{mercury} = (-.3.1193) + (2.16 * 10^{-6}) * \text{fork length}^2 + (5.16 * 10^{-4}) * \text{weight} + (0.0409) * \text{age}$

The following Bootstrap Report compared the original confident limits with those generated by 3000 draws using the iterative resampling – with – replacement bootstrapping method:

Bootstrap Report for Coefficients

Estimation Results		Bootstrap Confidence Limits		
Parameter	**Estimate**	**Conf. Level**	**Lower**	**Upper**
Intercept				
Original Value	-3.11925	90.000	-3.442461	-2.818726
Bootstrap Mean	-3.111678	95.000	-3.573547	-2.754331
Bias (BM - OV)	0.007571935	99.000	-3.790385	-2.584949
Bias Corrected	-3.126822			
Standard Error	0.1976392			
B(square_fork_length)				
Original Value	2.15822E-06	90.000	-3.881521E-06	9.105532E-06
Bootstrap Mean	1.909872E-06	95.000	-5.178678E-06	1.279058E-05
Bias (BM - OV)	-2.483479E-07	99.000	-9.561854E-06	2.271675E-05
Bias Corrected	2.406568E-06			
Standard Error	4.374541E-06			
B(weight)				
Original Value	0.000516	90.000	-0.000333081	0.001482272
Bootstrap Mean	0.000492004	95.000	-0.0007250918	0.001698304
Bias (BM - OV)	-2.399599E-05	99.000	-0.00175607	0.002414419
Bias Corrected	0.000539996			
Standard Error	0.0005864891			
B(age)				
Original Value	0.04087823	90.000	-0.01697551	0.0854281
Bootstrap Mean	0.04536017	95.000	-0.0287133	0.09531674
Bias (BM - OV)	0.004481936	99.000	-0.06004373	0.1124285
Bias Corrected	0.03639629			
Standard Error	0.03164843			

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.

Robust Residuals and Weights

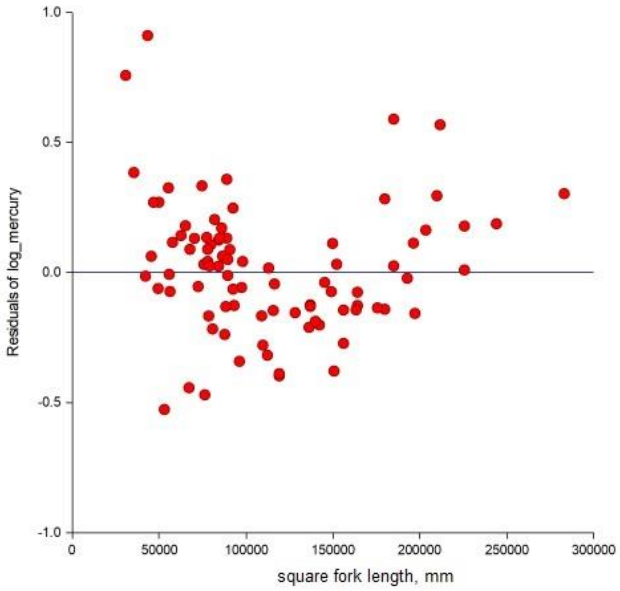
Row	Actual log_mercury	Predicted log_mercury	Residual	Absolute Percent Error	Robust Weight
1	-2.26336	-2.139691	-0.1236686	5.464	0.9714
2	-1.42712	-1.606269	0.1791486	12.553	0.9405
3	-2.09557	-2.020376	-0.07519412	3.588	0.9894
4	-2.55105	-2.582797	0.03174676	1.244	0.9982
5	-2.56395	-2.588872	0.02492207	0.972	0.9989
6	-2.60369	-2.693225	0.08953493	3.439	0.9850
7	-2.51331	-2.368038	-0.1452721	5.780	0.9607
8	-2.52573	-2.35968	-0.1660502	6.574	0.9487
9	-2.00992	-2.042272	0.03235219	1.610	0.9981
10	-2.57702	-2.62044	0.04342037	1.685	0.9965
11	-2.18037	-2.107508	-0.07286223	3.342	0.9901
12	-1.09661	-1.284798	0.1881876	17.161	0.9344
13	-1.98777	-1.847235	-0.1405349	7.070	0.9632
14	-1.70926	-1.552572	-0.1566881	9.167	0.9543
15	-2.22562	-2.188025	-0.03759545	1.689	0.9974
16	-1.96611	-2.078377	0.1122672	5.710	0.9764
17	-2.53831	-2.527379	-0.01093065	0.431	0.9999
18	-2.4651	-2.553685	0.08858548	3.594	0.9853
19	-2.6173	-2.491064	-0.1262363	4.823	0.9702
20	-2.45341	-2.517152	0.06374224	2.598	0.9924
21	-2.45341	-2.504045	0.05063512	2.064	0.9952
22	-2.57702	-2.602543	0.02552316	0.990	0.9989
23	-2.51331	-2.456458	-0.05685163	2.262	0.9940
24	-2.41912	-2.551888	0.1327678	5.488	0.9671
25	-2.52573	-2.706006	0.1802763	7.138	0.9397
26	-2.45341	-2.578552	0.1251421	5.101	0.9708
27	-2.63109	-2.747831	0.1167411	4.437	0.9745
28	-2.43042	-2.756569	0.3261489	13.419	0.8095
29	-0.85802	-1.16161	0.3035903	35.383	0.8338
30	-2.28278	-2.153107	-0.1296726	5.680	0.9686
31	-1.27297	-1.568776	0.2958061	23.237	0.8418
32	-2.76462	-2.758132	-0.00648762	0.235	1.0000
33	-2.24432	-2.603086	0.358766	15.986	0.7720
34	-2.51331	-2.645133	0.1318234	5.245	0.9676
35	-2.50104	-2.608036	0.1069961	4.278	0.9786
36	-2.4651	-2.507856	0.04275572	1.734	0.9966
37	-2.41912	-2.375856	-0.04326383	1.788	0.9966
38	-2.67365	-2.543368	-0.1302821	4.873	0.9683
39	-2.51331	-2.603412	0.09010237	3.585	0.9848
40	-2.3969	-2.568585	0.1716853	7.163	0.9452
41	-2.51331	-2.576372	0.06306186	2.509	0.9926
42	-2.78062	-2.564199	-0.2164212	7.783	0.9137
43	-2.55105	-2.487734	-0.06331591	2.482	0.9925
44	-1.84516	-2.756469	0.9113091	49.389	0.0470
45	-2.68825	-2.634847	-0.05340258	1.987	0.9947
46	-2.36446	-2.154186	-0.2102737	8.893	0.9184
47	-1.1332	-1.701554	0.5683537	50.155	0.4835
48	-2.81341	-2.741116	-0.07229446	2.570	0.9902
49	-2.81341	-2.800021	-0.0133886	0.476	0.9997

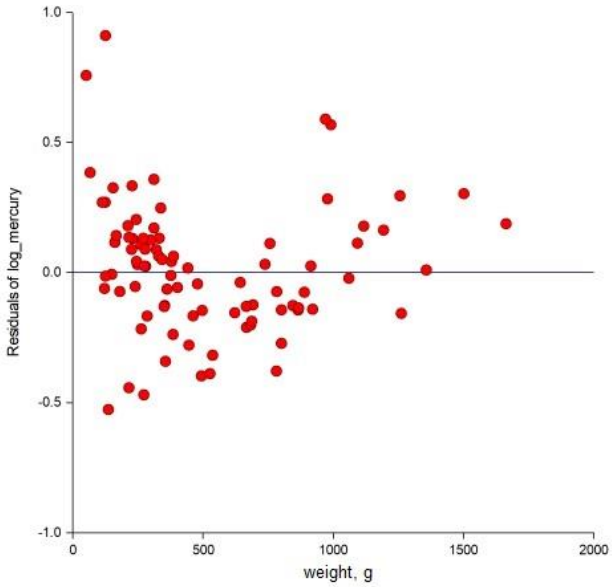
Robust Residuals and Weights

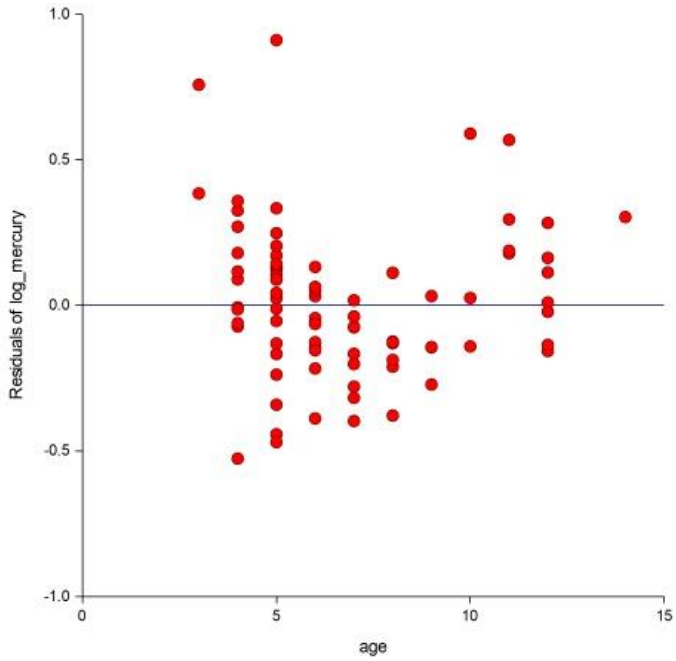
Row	Actual log_mercury	Predicted log_mercury	Residual	Absolute Percent Error	Robust Weight
50	-2.27303	-2.001293	-0.2717368	11.955	0.8656
51	-2.12863	-2.002202	-0.1264284	5.939	0.9702
52	-2.76462	-2.598078	-0.1665421	6.024	0.9484
53	-2.8647	-2.523757	-0.3409426	11.902	0.7928
54	-2.7181	-2.321316	-0.3967842	14.598	0.7251
55	-2.63109	-2.313804	-0.3172863	12.059	0.8192
56	-1.81401	-1.839788	0.02577835	1.421	0.9988
57	-1.43129	-1.44155	0.01026038	0.717	0.9999
58	-2.50104	-2.636606	0.1355665	5.420	0.9657
59	-2.29263	-2.540996	0.2483663	10.833	0.8871
60	-2.30259	-2.63636	0.3337704	14.495	0.8010
61	-2.34341	-2.361376	0.01796551	0.767	0.9995
62	-2.43042	-2.276422	-0.153998	6.336	0.9558
63	-2.37516	-2.173928	-0.2012317	8.472	0.9251
64	-2.14558	-2.903687	0.7581074	35.333	0.2097
65	-2.50104	-2.885763	0.3847227	15.383	0.7404
66	-2.11196	-1.968269	-0.1436908	6.804	0.9615
67	-2.44185	-2.063805	-0.3780454	15.482	0.7486
68	-1.6874	-1.665816	-0.02158355	1.279	0.9992
69	-2.12863	-1.985773	-0.1428566	6.711	0.9620
70	-2.73337	-2.345166	-0.388204	14.202	0.7360
71	-2.40795	-2.612421	0.2044706	8.491	0.9228
72	-2.41912	-2.551384	0.1322639	5.467	0.9674
73	-2.51331	-2.783911	0.2706005	10.767	0.8667
74	-2.52573	-2.796219	0.2704887	10.709	0.8668
75	-3.29684	-2.770875	-0.5259652	15.954	0.5463
76	-2.55105	-2.693798	0.1427476	5.596	0.9620
77	-3.10109	-2.658627	-0.4424632	14.268	0.6649
78	-2.76462	-2.527104	-0.2375162	8.591	0.8965
79	-3.07911	-2.609586	-0.4695242	15.249	0.6274
80	-2.64508	-2.366509	-0.2785708	10.532	0.8590
81	-1.52786	-1.641174	0.1133143	7.417	0.9760
82	-1.41059	-1.574139	0.1635487	11.594	0.9503
83	-1.22078	-1.810892	0.5901123	48.339	0.4510
84	-1.93794	-1.802439	-0.1355005	6.992	0.9657
85	-2.32279	-2.135849	-0.1869415	8.048	0.9352
86	-2.84731	-2.786419	-0.06089111	2.139	0.9931
87	-1.45243	-1.736067	0.2836366	19.528	0.8541

The Robust Residuals and Weights Table above list out the residuals and the weights assigned by Tukey's Bisquare M – estimator to each of the observations in the entire data set (n = 87). Four extreme outliers were identified: rows 44, 64, 75 and 83. Rows 47, 77 and 79 were also significant outliers. To a lesser extent were rows: 33, 54, 65, 67 and 69. These observations with lower weights made only minimum contributions to the determination of the regression coefficients. The table also gives the predicted values of $\log_e$ mercury based on the robust regression equation above from the final iteration for comparing side by side with the actual measured $\log_e$ mercury values.

The residuals versus predictors plots highlight the presence of outliers in the data set. They also suggested a possible pattern of some kind might exist and the potential of a model specification problem should not be dismissed. Despite a possible linear relationship between the dependent variable and the three predictors as shown in the scatterplots, perhaps the relationships were better be described by a quadratic polynomial function (see Chapter 6) which provides a more correct and better fitted model for the lake whitefish data:

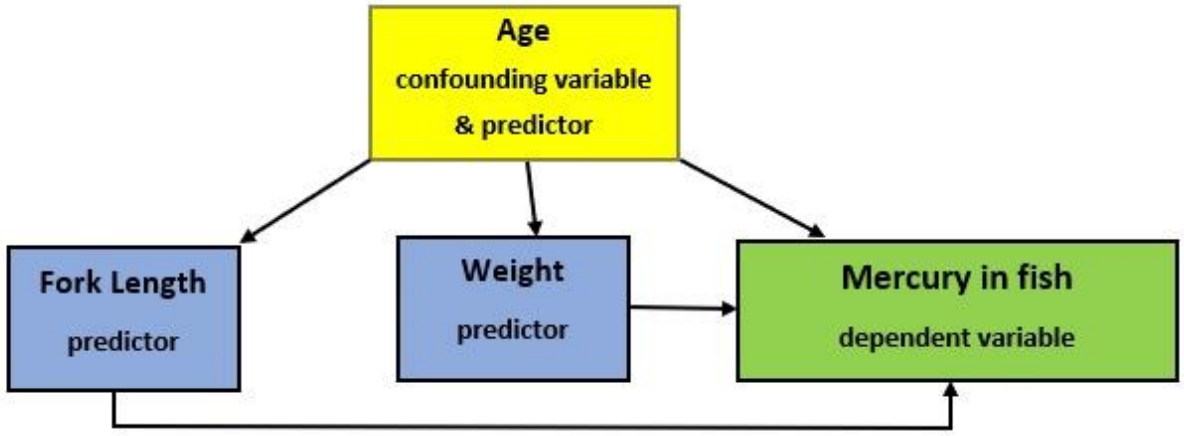






Chapter 8. Hierarchical Multiple Linear Regression Modeling with Bootstrapping using Fish Age as the Confounding Variable.

It is important to realize that theoretical reasoning of the researcher based on sound scientific knowledge and experience has the most important place in data modeling which, in my opinion should always truncate what the computer algorithms and statistical tests tell us otherwise. Afterall, no matter how sophisticated and complex, they are just machines carrying out routines with absolutely no knowledge in the sciences upon which the basic ingredients of a model should be based. In the present case, our previous knowledge regarding bioaccumulation of mercury in fish has given us sound theoretical grounds to the notion that mercury level in fish is most likely increases as the fish grows, (i.e. increase in age and size). Length, weight and age are the obvious candidates as predictors in our model for fish mercury level. Fish age is our confounding variable because the other two predictors (fork length and weight) as well as the dependent variable (fish mercury level) are affected by fish age. Another way of looking at the role of fish age is that for example, in situations when we want to compare mercury level in fish from different locations; we need to include fish age as the covariate in an analysis of covariance (ANCOVA) exercise.



In chapter 6, we used hierarchical multiple linear regression (HMLR) to investigate the quadratic terms of the three predictor variables and worked out the quadratic polynomial models for the lake whitefish data set. In this chapter we used HMLR to control for age as the confounding variable in our regression model but more importantly the process enabled us to assess the impact of fish age on the overall model. HMLR requires the same assumptions as standard multiple linear regression: linearity,

homoscedasticity, normality, independence and randomness of residuals. As seen in chapter 1, after identifying and removing of influential outliers, transformed brook trout data and untransformed lake trout data met these assumptions and were selected for HMLR modeling using fish age as the confounding variable. Collinearity problems might well be an issue here as fork length and weight were shown to have a significant positive correlation. The absence of serious collinearity problems is also an assumption for multiple linear regression and HMR to be valid. Despite the potential collinearity issue with our data sets, we used the IBM™ SPSS™ statistics software to run the HMLR and see what the outcomes suggested. The IBM™ SPSS™ statistics software emphasizes on highlighting the effect of the presence of confounding variables on the overall model was used in our HMLR modeling. The actual procedure of how the HMLR module in IBM™ SPSS™ works has already been touched on in chapter 6.

The bootstrapping algorithm was also used in running HMLR by carrying out 1000 resamplings with replacement to accurately compute standard errors, regression coefficients and confidence intervals. Bootstrapping does not make any assumption on the distribution or homoscedasticity of the data so long as we can assume that the sample approximate the population. It is particularly useful and give more reliable and accurate computation of various regression statistics when regression assumptions are on shaky grounds. Bootstrapping can be used in validating the statistics calculated using conventional frequentist inference in that if the upper and lower confidence limits of the 95% confidence interval computed by bootstrapping both have the same sign as the frequentist inferred statistic, then the p – value associated with that statistic is reliable. On the other hand, if one or both the confident limits calculated by bootstrapping has a sign different from the inferred statistic, that means the inferred statistic is not statistically significant. Bootstrap results are always more reliable than the statistical significance computed by conventional frequentist inference. This is particularly useful when we have non – normally distributed data. In the present case, we draw 1000 bootstrap samples of size n from the original samples with replacement; hence each observation may be selected more than once. For each of the 1000 bootstrap sample, the regression results are computed, stored and used in the final computation of various regression statistics. Bias – corrected and accelerated (BCa) approach was used here for estimating the 95% confidence intervals which is more accurate than the percentile method.

8.1 Brook trout

Descriptive Statistics

			Bootstrap[a]			
					BCa 95% Confidence Interval	
		Statistic	Bias	Std. Error	Lower	Upper
log_mercury	Mean	-2.168619	-.000460	.068175	-2.302552	-2.033321
	Std. Deviation	.4574678	-.0051271	.0419078	.3829081	.5211867
	N	44	0	0	.	.
age	Mean	4.55	.00	.17	4.25	4.84
	Std. Deviation	1.150	-.016	.128	.923	1.336
	N	44	0	0	.	.
log_length	Mean	5.722641	-.001604	.036089	5.649206	5.786212
	Std. Deviation	.2385744	-.0034198	.0279059	.1820767	.2814665
	N	44	0	0	.	.
log_weight	Mean	5.770788	-.005404	.108489	5.559469	5.959708
	Std. Deviation	.7148374	-.0096252	.0832256	.5363844	.8463897
	N	44	0	0	.	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The mean values of the four variables were all have the same signs as and within the upper and lower limits of the BCa 95% confidence intervals computed by bootstrapping which indicated that the normality and homoscedasticity assumptions hold true for the four variables.

The Correlation Table below shows that he Pearson correlations for all the four variables in all combinations were of the same signs as the upper and lower limits of the bootstrap confidence intervals corresponded to that correlation. Linearity between the dependent variable and the three predictor variables was confirmed by the significant positive correlations between the dependent variable and each of the three predictor variables. Significant positive correlations amongst the three predictor variables suggested potential collinearity problems.

Pearson Correlation

log_mercury	log_mercury	1.000
	age	.817
	log_length	.708
	log_weight	.647
age	log_mercury	.817
	age	1.000
	log_length	.840
	log_weight	.819
log_length	log_mercury	.708
	age	.840
	log_length	1.000
	log_weight	.990
log_weight	log_mercury	.647
	age	.819
	log_length	.990
	log_weight	1.000

Bootsrtap for Pearson Correlation (Bias)

log_mercury	log_mercury	.000
	age	-.005
	log_length	-.002
	log_weight	-.003
age	log_mercury	-.005
	age	.000
	log_length	.000
	log_weight	.000
log_length	log_mercury	-.002
	age	.000
	log_length	.000
	log_weight	.000
log_weight	log_mercury	-.003
	age	.000
	log_length	.000
	log_weight	.000

Bootstrap for Pearson Correlation (BCa 95% Confidence Interval)

Lower	log_mercury	log_mercury	.
		age	.705
		log_length	.561
		log_weight	.463
	age	log_mercury	.705
		age	.
		log_length	.719
		log_weight	.689
	log_length	log_mercury	.561
		age	.719
		log_length	.
		log_weight	.981
	log_weight	log_mercury	.463
		age	.689
		log_length	.981
		log_weight	.
Upper	log_mercury	log_mercury	.
		age	.891
		log_length	.813
		log_weight	.785
	age	log_mercury	.891
		age	.
		log_length	.920
		log_weight	.903
	log_length	log_mercury	.813
		age	.920
		log_length	.
		log_weight	.996
	log_weight	log_mercury	.785
		age	.903
		log_length	.996
		log_weight	.

Sig. (1 -tailed)

log_mercury	log_mercury	.
	age	.000
	log_length	.000
	log_weight	.000
age	log_mercury	.000
	age	.
	log_length	.000
	log_weight	.000
log_length	log_mercury	.000
	age	.000
	log_length	.
	log_weight	.000
log_weight	log_mercury	.000
	age	.000
	log_length	.000
	log_weight	.

Bootstrap for Pearson Correlation (Std. Error)

log_mercury	log_mercury	.000
	age	.051
	log_length	.065
	log_weight	.083
age	log_mercury	.051
	age	.000
	log_length	.045
	log_weight	.049
log_length	log_mercury	.065
	age	.045
	log_length	.000
	log_weight	.004
log_weight	log_mercury	.083
	age	.049
	log_length	.004
	log_weight	.000

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.817[a]	.667	.659	.2671513	.667	84.088	1	42	.000
2	.880[b]	.774	.757	.2255480	.107	9.462	2	40	.000

a. Predictors: (Constant), age

b. Predictors: (Constant), age, log_weight, log_length

c. Dependent Variable: log_mercury

The Model Summary Table showed that 77.4% of the variance in the dependent variable was explained by Model 2 ($\log_e$ fork length + $\log_e$ weight + age). Age alone (Model 1) accounted for a huge 66.7% of the variance in $\log_e$ mercury. After the effect of age has been controlled (i.e. removed), only

10.7% of the variance in $\log_e$ mercury was explained by $\log_e$ fork length + $\log_e$ weight; however, the change in R – Square was statistically significant (p = 0.000). Hence, adding $\log_e$ fork length and $\log_e$ weight to the age model increased the model's predictive capacity at predicting $\log_e$ mercury in a significant way (p = 0.000), it increased the percentage of variance accounted for by 10.7%. Both Models 1 and 2 were statistically significant in predicting $\log_e$ mercury (p = 0.000).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.001	1	6.001	84.088	.000[b]
	Residual	2.998	42	.071		
	Total	8.999	43			
2	Regression	6.964	3	2.321	45.631	.000[c]
	Residual	2.035	40	.051		
	Total	8.999	43			

a. Dependent Variable: log_mercury

b. Predictors: (Constant), age

c. Predictors: (Constant), age, log_weight, log_length

The ANOVA Table showed that both models were significant (p = 0.000); hence, we rejected the null hypothesis that the slopes of the lines were zero. The F value of 45.631 associated with Model 2 tested the hypothesis that the R – Square value (0.774) associated with Model 2 was statistically significant.

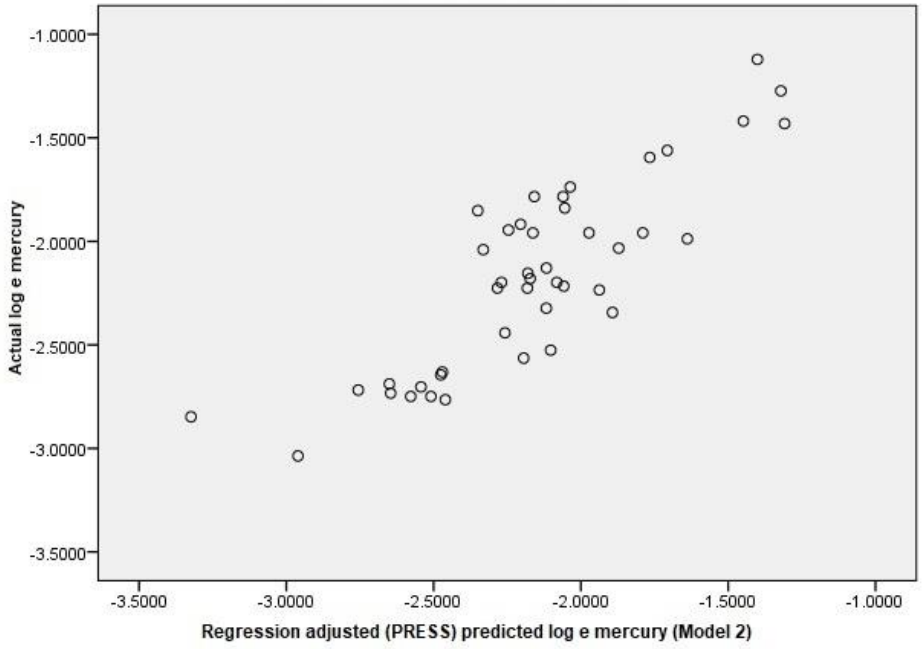
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-3.645	.166		-21.965	.000					
	age	.325	.035	.817	9.170	.000	.817	.817	.817	1.000	1.000
2	(Constant)	-22.340	4.370		-5.112	.000					
	age	.256	.056	.643	4.560	.000	.817	.585	.343	.284	3.518
	log_length	4.866	1.126	2.537	4.322	.000	.708	.564	.325	.016	60.967
	log_weight	-1.531	.355	-2.392	-4.316	.000	.647	-.564	-.325	.018	54.346

a. Dependent Variable: log_mercury

The collinearity statistics in the Coefficients Table confirmed serious collinearity problems existed for $\log_e$ fork length and $\log_e$ weight (VIF > 10 and Tolerance < 0.1). An important issue of multicollinearity is that the estimated regression coefficient may change erratically and significantly in response to small changes in the model or the input data. Multicollinearity impacts calculations

regarding "individual" predictors hence gives erratic results about the relative importance of individual predictors or about which predictors are statistically redundant with respect to others. The estimate of one predictor's impact on the dependent variable while controlling for other predictors is no longer accurate. Hence both the standardized and unstandardized regression coefficients as well as their t – values for both $\log_e$ fork length and $\log_e$ weight were uninterpretable. However, it is important to point out that multicollinearity does not reduce the predictive power of the model "as a whole", it just that we should not and cannot make interpretation on individual predictors based on their estimated regression coefficients. The plot of actual $\log_e$ mercury versus adjusted (PRESS) predicted $\log_e$ mercury of Model 2 actually showed a nice positive linear relationship:



One feature of multicollinearity is that the standard errors of the affected coefficients tend to be large. In the present case, the standard errors of the coefficients associated with $\log_e$ fork length and $\log_e$ weight were 1.126 and 0.355 respectively; whereas the standard error of the coefficient associated with age is only 0.056.

The equation that HMLR computed for the overall model with all three predictor variables is:

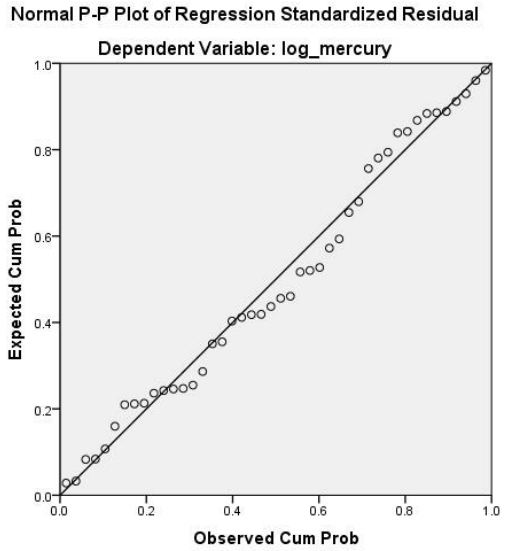
$\log_e$ mercury = (-22.340) + (4.866) * $\log_e$ fork length + (-1.531) * $\log_e$ weight + (0.256) * age

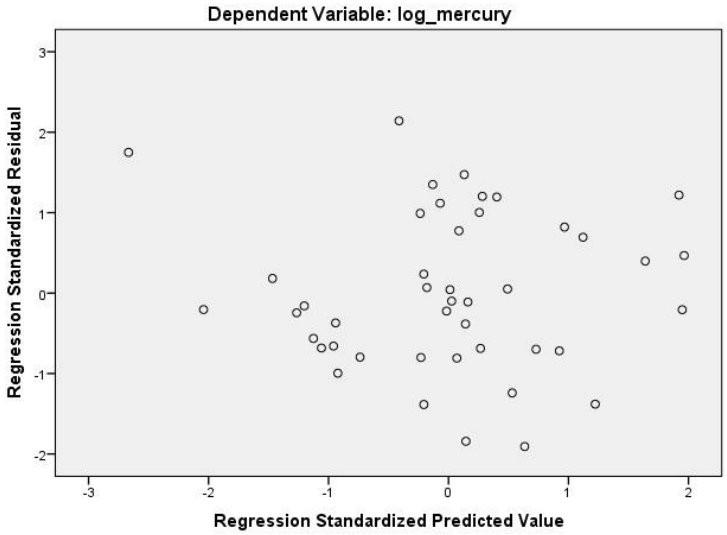
Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	-3.645	.001	.119	.001	-3.844	-3.428
	age	.325	-7.147E-005	.026	.001	.273	.376
2	(Constant)	-22.340	-.403	4.305	.001	-29.856	-14.809
	age	.256	-.005	.045	.001	.174	.323
	log_length	4.866	.099	1.102	.001	2.450	7.120
	log_weight	-1.531	-.025	.351	.001	-2.114	-.934

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

All the estimated regression coefficients were of the same signs as their corresponding lower and upper limits of the BCa (bias – corrected and accelerated) 95% confidence intervals computed by bootstrapping. All the estimated regression coefficients were within their corresponding bootstrapped 95% confidence intervals. And they were all statistically significant (p = 0.001) estimated using bootstrapping.





The Normal P – P plot confirmed the normal distribution of the dependent variable. The standardized residuals plot all the residuals were within ± 2 which confirmed homoscedasticity of residuals.

Residuals Statistics[a]

			Bootstrap[b]			
					BCa 95% Confidence Interval	
		Statistic	Bias	Std. Error	Lower	Upper
Predicted Value	Minimum	-3.241919				
	Maximum	-1.378242				
	Mean	-2.168619	-.000460	.068175	-2.302552	-2.033321
	Std. Deviation	.4024352	-.0022597	.0493192	.3120347	.4839767
	N	44	0	0	.	.
Std. Predicted Value	Minimum	-2.667				
	Maximum	1.964				
	Mean	.000	.000	.000	.000	.000
	Std. Deviation	1.000	.000	.000	.	.
	N	44	0	0	.	.
Standard Error of Predicted Value	Minimum	.039				
	Maximum	.111				
	Mean	.065	-.003	.006	.056	.068
	Std. Deviation	.020	-.001	.003	.016	.022
	N	44	0	0	.	.
Adjusted Predicted Value	Minimum	-3.369401				
	Maximum	-1.374639				
	Mean	-2.172217	.000267	.068807	-2.308518	-2.034700
	Std. Deviation	.4046944	-.0035361	.0502197	.3130133	.4827977
	N	44	0	0	.	.
Residual	Minimum	-.4303178				
	Maximum	.4830552				
	Mean	0E-7	0E-7	0E-7	0E-7	0E-7
	Std. Deviation	.2175378	-.0092217	.0200260	.1884411	.2283701
	N	44	0	0	.	.
Std. Residual	Minimum	-1.908				
	Maximum	2.142				
	Mean	.000	.000	.000	.000	.000
	Std. Deviation	.964	.000	.000	.	.
	N	44	0	0	.	.
Stud. Residual	Minimum	-1.951				
	Maximum	2.175				
	Mean	.007	-.001	.004	.000	.012
	Std. Deviation	1.007	-.001	.008	.995	1.019
	N	44	0	0	.	.
Deleted Residual	Minimum	-.4497979				
	Maximum	.5220910				
	Mean	.0035979	-.0007270	.0020831	.0003621	.0051990
	Std. Deviation	.2378967	-.0104587	.0218939	.2067692	.2491054
	N	44	0	0	.	.
Stud. Deleted Residual	Minimum	-2.025				
	Maximum	2.287				
	Mean	.010	-.002	.007	-.001	.017
	Std. Deviation	1.026	.002	.011	1.013	1.060
	N	44	0	0	.	.
Mahal. Distance	Minimum	.290				
	Maximum	9.522				
	Mean	2.932	.000	.000	2.932	2.932
	Std. Deviation	2.485	.024	.352	1.880	3.321
	N	44	0	0	.	.
Cook's Distance	Minimum	.000				
	Maximum	.327				
	Mean	.024	.000	.005	.016	.033
	Std. Deviation	.050	-.007	.026	.022	.062
	N	44	0	0	.	.
Centered Leverage Value	Minimum	.007				
	Maximum	.221				
	Mean	.068	.000	.000	.068	.068
	Std. Deviation	.058	.001	.008	.044	.077
	N	44	0	0	.	.

The Residual Statistics Table above showed that the mean of all the residual statistics were of the same signs as and within the lower and upper limits of the BCa (bias – corrected and accelerated) 95% confidence intervals computed by bootstrapping. The maximum values of the Mahalanobis distance, the Cook's distance and the centred leverage values all confirmed the absence of influential outliers in the data set.

8.2 Lake trout

Descriptive Statistics

			Bootstrap[a]			
					BCa 95% Confidence Interval	
		Statistic	Bias	Std. Error	Lower	Upper
mercury	Mean	.2853	.0006	.0214	.2462	.3283
	Std. Deviation	.11842	-.00194	.01144	.09774	.13566
	N	31	0	0	.	.
age	Mean	8.13	.00	.37	7.42	8.81
	Std. Deviation	2.109	-.055	.276	1.648	2.480
	N	31	0	0	.	.
fork_length	Mean	365.4839	.4587	13.7282	336.4394	392.5837
	Std. Deviation	76.31464	-1.83828	9.52444	58.25680	88.52550
	N	31	0	0	.	.
weight	Mean	638.8710	2.4450	61.1902	518.7630	764.2463
	Std. Deviation	340.43832	-6.55462	35.46586	276.61080	388.91126
	N	31	0	0	.	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The mean values of the four variables were all have the same signs as and within the upper and lower limits of the BCa 95% confidence intervals computed by bootstrapping which indicated that the normality and homoscedasticity assumptions hold true for the four variables.

The Pearson correlations for all the four variables in all combinations were of the same signs as the upper and lower limits of the bootstrap confidence intervals corresponded to that correlation. Linearity between the dependent variable and the three predictor variables was confirmed by the significant positive correlations between the dependent variable and each of the three predictor variables. Significant positive correlations amongst the three predictor variables suggested potential collinearity problems.

Pearson Correlation

mercury	mercury	1.000
	age	.666
	fork_length	.799
	weight	.803
age	mercury	.666
	age	1.000
	fork_length	.816
	weight	.783
fork_length	mercury	.799
	age	.816
	fork_length	1.000
	weight	.962
weight	mercury	.803
	age	.783
	fork_length	.962
	weight	1.000

Bootstrap for Pearson Correlation (Bias)

mercury	mercury	.000
	age	-.004
	fork_length	-.002
	weight	-.005
age	mercury	-.004
	age	.000
	fork_length	-.005
	weight	-.003
fork_length	mercury	-.002
	age	-.005
	fork_length	.000
	weight	.003
weight	mercury	-.005
	age	-.003
	fork_length	.003
	weight	.000

Bootstrap BCa 95% Confidence Intervals

Lower	mercury	mercury	.
		age	.392
		fork_length	.674
		weight	.656
	age	mercury	.392
		age	.
		fork_length	.615
		weight	.567
	fork_length	mercury	.674
		age	.615
		fork_length	.
		weight	.
	weight	mercury	.656
		age	.567
		fork_length	.
		weight	.
Upper	mercury	mercury	.
		age	.842
		fork_length	.888
		weight	.900
	age	mercury	.842
		age	.
		fork_length	.931
		weight	.913
	fork_length	mercury	.888
		age	.931
		fork_length	.
		weight	.
	weight	mercury	.900
		age	.913
		fork_length	.
		weight	.

Sig. (1 - tailed)

mercury	mercury	.
	age	.000
	fork_length	.000
	weight	.000
age	mercury	.000
	age	.
	fork_length	.000
	weight	.000
fork_length	mercury	.000
	age	.000
	fork_length	.
	weight	.000
weight	mercury	.000
	age	.000
	fork_length	.000
	weight	.

Bootstrap for Pearson Correlation (Std. Error)

mercury	mercury	.000
	age	.099
	fork_length	.051
	weight	.060
age	mercury	.099
	age	.000
	fork_length	.078
	weight	.083
fork_length	mercury	.051
	age	.078
	fork_length	.000
	weight	.007
weight	mercury	.060
	age	.083
	fork_length	.007
	weight	.000

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.666[a]	.444	.425	.08982	.444	23.150	1	29	.000
2	.809[b]	.655	.616	.07336	.211	8.236	2	27	.002

a. Predictors: (Constant), age

b. Predictors: (Constant), age, weight, fork_length

c. Dependent Variable: mercury

d. Predictors: (Constant), age, fork_length, weight

The Model Summary Table showed that 65.5% of the variance in the dependent variable was attributed to Model 2 (fork length + weight + age). Age alone (Model 1) accounted for a huge 44.4% of the variance in the dependent variable. After the effect of age has been removed, there is still 21.1% of the variance in mercury was explained by fork length + weight; this change in R – Square was statistically significant (p = 0.002). Hence, adding fork length and weight to the age model significantly increased the model's predictive capacity at predicting mercury (p = 0.002). Both Models 1 and 2 were statistically significant in predicting mercury (p < 0.005). The F change for Model 1 (F = 23.150) was exactly the same as the F value associated with Model 1 in the ANOVA Table below, since this F value only associated with a model with age as the sole predictor variable. The addition of age into the model caused a change of F value of 8.236 and the F value associated with the Model 2 which contains all three predictor variables was 17.059. The ANOVA Table showed that both models were significant (p = 0.000); hence, we rejected the null hypothesis that the slopes of the lines were zero. The F value of 17.059 associated with Model 2 tested the hypothesis that the R – Square value (0.655) associated with Model 2 was statistically significant.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.187	1	.187	23.150	.000[b]
	Residual	.234	29	.008		
	Total	.421	30			
2	Regression	.275	3	.092	17.059	.000[c]
	Residual	.145	27	.005		
	Total	.421	30			

a. Dependent Variable: mercury

b. Predictors: (Constant), age

c. Predictors: (Constant), age, weight, fork_length

d. Predictors: (Constant), age, fork_length, weight

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.019	.065		-.288	.775		
	age	.037	.008	.666	4.811	.000	1.000	1.000
2	(Constant)	-.019	.150		-.126	.900		
	age	.003	.011	.046	.234	.816	.333	3.001
	fork_length	.000	.001	.319	.711	.483	.063	15.792
	weight	.000	.000	.459	1.102	.280	.074	13.586

a. Dependent Variable: mercury

The Collinearity Statistics above showed that some degrees of collinearity problems were detected for fork length and weight (Tolerance < 0.1, VIF > 10), although it was not as severe as in the case for brook trout. It is very interesting to note that the unstandardized regression coefficients associated with fork length and weight were zero and that associated with age was very low (0.003) for Model 2 when all three predictors were present. The unstandardized regression coefficient for age on its own in Model 1 was much higher than that in Model 2 (0.037) and this was the only regression coefficient that was statistically significant (p = 0.000). The SPSS™ algorithm literally throwed out Model 2 and suggested that a model with age alone as the only predictor was most appropriate (i.e. Model 1). For Model 1, the unstandardized regression coefficient indicated that as age increases by one year, fish mercury increases by 0.037 unit, in this case 0.037µg/g. The standard error of the unstandardized regression coefficient associated with age was very low (0.008).

The equation that HMLR computed for the model with age as the sole predictor variable is:

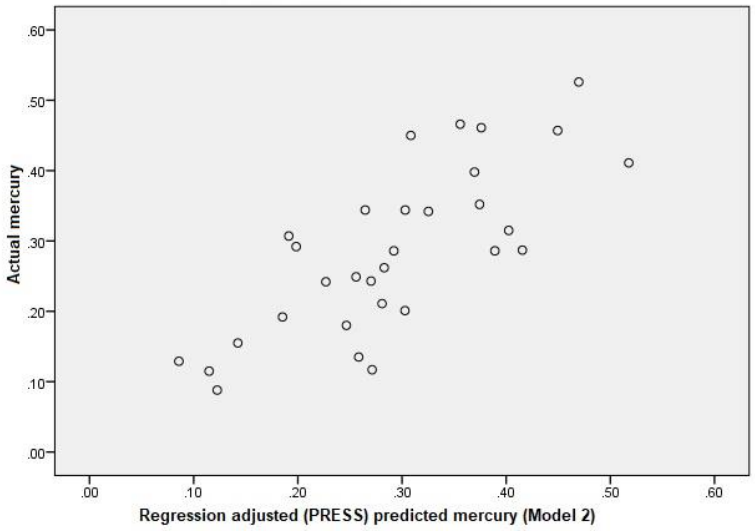
mercury = -0.19 + (0.037) * age

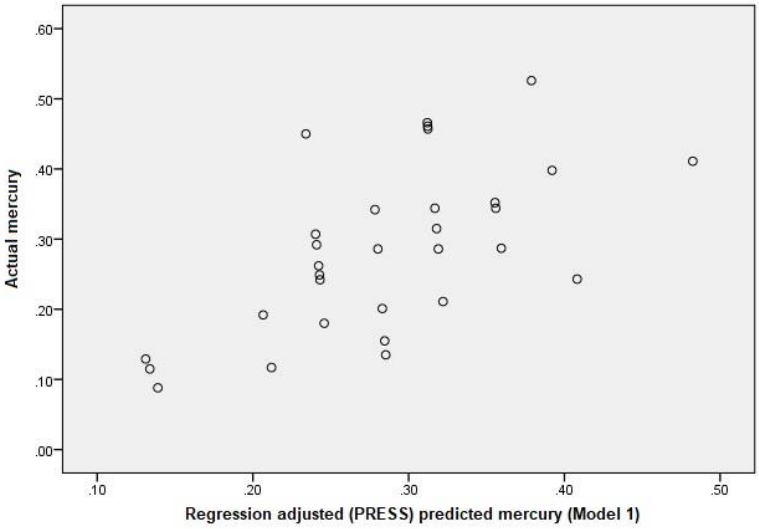
Bootstrap for Coefficients

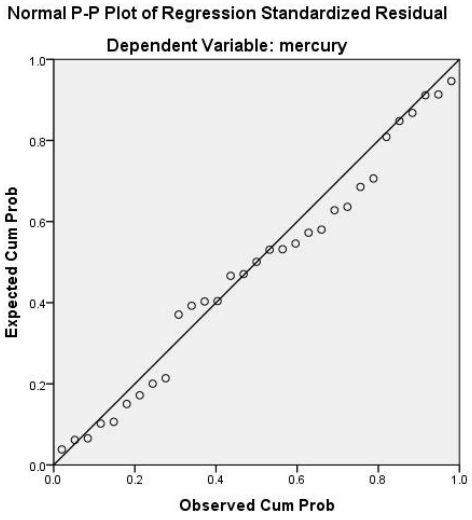
Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	-.019	-.002	.050	.699	-.096	.058
	age	.037	.000	.006	.001	.022	.053
2	(Constant)	-.019	-.017	.116	.853	-.261	.140
	age	.003	.000	.011	.824	-.020	.027
	fork_length	.000	6.572E-005	.001	.406	-.001	.002
	weight	.000	-1.550E-005	.000	.207	.000	.000

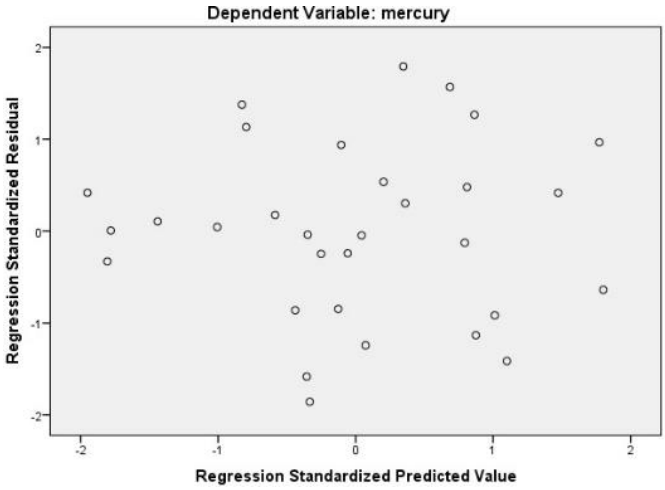
a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The bootstrapping run of the coefficients led us to the same conclusion. The unstandardized regression coefficient associated with age in Model 1 (0.037) was of the same sign and within the upper and lower 95% confidence limits computed by bootstrapping; and it was the only coefficient that was statistically significant (p = 0.001). It is interesting to note that despite SPSS throwed out Model 2, the plot of actual mercury versus adjusted (PRESS) predicted mercury of Model 2 actually showed a nicer (less scattered) positive linear relationship than the same plot with only age as the predictor (Model 1):


Actual mercury
.00
.10
.20
.30
.40
.50
.60
Regression adjusted (PRESS) predicted mercury (Model 2)


Actual mercury
.00
.10
.20
.30
.40
.50
.60
Regression adjusted (PRESS) predicted mercury (Model 1)


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: mercury
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Dependent Variable: mercury
Regression Standardized Residual
Regression Standardized Predicted Value
-2
-1
0
1
2

The Normal P – P plot confirmed the normal distribution of the dependent variable. The standardized residuals plot all the residuals were within ± 2 which confirmed homoscedasticity of residuals.

Residuals Statistics[a]

			Bootstrap[b]			
					95% Confidence Interval	
		Statistic	Bias	Std. Error	Lower	Upper
Predicted Value	Minimum	.0983				
	Maximum	.4578				
	Mean	.2853	-.0006	.0215	.2427	.3268
	Std. Deviation	.09581	-.00123	.01387	.06694	.12141
	N	31	0	0	31	31
Std. Predicted Value	Minimum	-1.951				
	Maximum	1.801				
	Mean	.000	.000	.000	.000	.000
	Std. Deviation	1.000	.000	.000	1.000	1.000
	N	31	0	0	31	31
Standard Error of Predicted Value	Minimum	.014				
	Maximum	.044				
	Mean	.025	-.001	.003	.018	.028
	Std. Deviation	.009	-.001	.001	.005	.011
	N	31	0	0	31	31
Adjusted Predicted Value	Minimum	.0813				
	Maximum	.4801				
	Mean	.2844	-.0008	.0216	.2411	.3264
	Std. Deviation	.09652	-.00078	.01392	.06914	.12258
	N	31	0	0	31	31
Residual	Minimum	-.13626				
	Maximum	.13151				
	Mean	.00000	.00000	.00000	.00000	.00000
	Std. Deviation	.06960	-.00426	.00757	.05093	.07973
	N	31	0	0	31	31
Std. Residual	Minimum	-1.857				
	Maximum	1.793				
	Mean	.000	.000	.000	.000	.000
	Std. Deviation	.949	.000	.000	.949	.949
	N	31	0	0	31	31
Stud. Residual	Minimum	-1.976				
	Maximum	1.916				
	Mean	.006	.001	.008	-.009	.023
	Std. Deviation	1.001	.004	.010	.991	1.029
	N	31	0	0	31	31
Deleted Residual	Minimum	-.15418				
	Maximum	.15028				
	Mean	.00088	.00016	.00134	-.00125	.00386
	Std. Deviation	.07774	-.00401	.00864	.05734	.09054
	N	31	0	0	31	31
Stud. Deleted Residual	Minimum	-2.096				
	Maximum	2.023				
	Mean	.005	.002	.012	-.016	.033
	Std. Deviation	1.026	.010	.020	1.010	1.090
	N	31	0	0	31	31
Mahal. Distance	Minimum	.055				
	Maximum	9.733				
	Mean	2.903	.000	.000	2.903	2.903
	Std. Deviation	2.818	.069	.531	1.932	4.067
	N	31	0	0	31	31
Cook's Distance	Minimum	.000				
	Maximum	.131				
	Mean	.029	.004	.013	.022	.058
	Std. Deviation	.037	.020	.062	.023	.165
	N	31	0	0	31	31
Centered Leverage Value	Minimum	.002				
	Maximum	.324				
	Mean	.097	.000	.000	.097	.097
	Std. Deviation	.094	.002	.018	.064	.136
	N	31	0	0	31	31

The Residual Statistics Table above showed that the mean of all the residual statistics were within the lower and upper limits of the BCa (bias – corrected and accelerated) 95% confidence intervals computed by bootstrapping. The maximum values of the Mahalanobis distance, the Cook's distance and the centred leverage values all confirmed the absence of influential outliers in the data set.

Epilogue

What do all these different models actually mean? Not much. At least not at this stage. Ultimately each of these models needed to be subjected to validation with more data before they mean anything. More fish are needed to be collected from the same watersheds from which the fish data used in constructing the models came from for used in testing these models. The three growth parameters (fork length, weight and age) of these new fish samples will be measured and the data fed into these models to calculate the predicted fish mercury levels. These predicted fish mercury levels will then be compared with the "actual" mercury levels determined by laboratory analysis of the fish samples to assess the performance of each model.

Predictive modeling has important potentials. So long as we apply a validated model to the same fish species collected from the same watershed, in the same season and within the same ranges of measurements of these growth parameters of fish used in establishing the models (i.e. no extrapolation); we should be able to predict mercury level in newly collected fish samples with acceptable accuracy just by knowing these growth parameters. The predictive modeling approach can be transplanted to other fish species in other watersheds: First, collect a fair number of a chosen species of fish from that watershed for mercury determination and growth parameter measurement in order to obtain data for establishing the model(s). Then, more fish of that species are collected from the same watershed to validate the models for their reliability. The models developed are location – and species – specific until demonstrated otherwise.

There are two amongst many quotations I particularly like when it comes to data science. One is by the American humorist Evan Esar:

"Statistics: The only science that enables different experts using the same figures to draw different conclusions."

The other one is from the McGill University mathematical statistician Phillip I. Good:

"In our research efforts, the only statements we can make with God – like certainty are of the form "our conclusions fit the data". The true nature of the real world is unknowable. We can speculate, but never conclude."

****

References

Bayes, T. (1763). "Essay towards solving a problem in the doctrine of chances." *Philosophical Transactions. Royal Society, London*. 53: 370 – 418.

Belsley, D.A., E. Kuh and R. E. Welsch (2013). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley and Sons Inc. ISBN: 9780471691174.

Cawley, G. C. and Talbot, N. L. C. (2010*).* "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation" *Journal of Machine Learning Research* **11**: 2079–2107.

Clyde, M., Ghosh, J. and Littman, M. (2010). "Bayesian adaptive sampling for variable selection and model averaging." *Journal of Computational Graphics and Statistics* **20**: 80 – 101.

Goldfeld, S. M. and R. E. Quandt (1965). "Some tests for homoscedasticity." *Journal of the Royal Statistical Society Series D*. **45** (1): 49 – 56.

Hoerl, A.E. and Kennard, R.W. (1970). "Ridge Regression: Biased estimation for nonorthogonal problems." *Technometrics* **12**: 55-82

Huber, P. J. (1981). *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons. New York.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann.* **2** (12): 1137–1143.

Kwan, M. K. H. (2019). "*Mercury concentration in fish from the Inukjuak River watershed and nearby lakes collected in the summer of 2019. A technical report on laboratory determination of total mercury in fish and statistical analysis*." Makivik Corporation, Nunavik Quebec. October 2019.

Magel, R. C. and S. H. Wibowo (1997). "Comparing the powers of the Wald – Wolfowitz and Kolmogorov – Smirnov tests." *Biometrical Journal*. **39** (6): 655 – 675.

Montgomery, D.C. and E.A. Peck (1992). *Introduction to Linear Regression Analysis*. 2nd edition. John Wiley and Sons inc. New York.

O'Brien, R. M. (2007). "A caution regarding rules of thumb for variance inflation factors." *Quality & Quantity.* **41** (5): 673.

Thursby, J. (1982). "Misspecification, Heteroscedasticity, and the Chow and Goldfeld - Quandt Tests". *The Review of Economics and Statistics*. **64** (2): 314–321.

Wilkinson, L., G. Blank and C. Gruber (1996). *Desktop Data Analysis with SYSTAT*. New Jersey, Prentice Hall. pp 168.

Wold, H. (1975). "Soft modeling by latent variables; the nonlinear iterative partial least squares approach." In *Perspectives in Probability and Statistics. Paper in Honour of M.S. Bartlett*, ed. J. Gani, Academic Press.

Zimek, A. and P. Filzmoser (2018). "*There and back again: Outlier detection between statistical reasoning and data mining algorithms*". Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. **8** (6): e1280.

List of Software and Algorithms

BayES Bayesian Econometric Software©, 2019, version 2.4

IBM™ SPSS™ Statistical Software, version 20.0

JASP open – source statistical project, version 0.11.1.0 and various modules in R

MathType© version 5.0

Minitab™, version 16.2.3

NCSS™ Statistical Software, 2019 version 19.0.3

SAS™ JMP™ 2019, version 15.0

BAS algorithm (Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling) by Merlise Clyde, Duke University

The Metropolis – Hastings algorithm for Markov – chain Monte Carlo simulation

Appendix (outliers are in red, see Chapter 1)

Lake whitefish collected from the Inukjuak River watersheds in summer 2019

# Fish ID	fork length, mm	weight, g	mercury, µg/g w.w.	age
26	370	692	0.104	8
27	475	1117	0.240	11
28	405	889	0.123	7
29	275	248	0.078	6
30	290	280	0.077	5
31	260	226	0.074	4
32	340	497	0.081	6
34	330	462	0.080	7
35	390	738	0.134	9
36	279	245	0.076	5
74	386	783	0.113	7
75	494	1663	0.334	11
76	424	921	0.137	10
77	444	1261	0.181	12
78	381	643	0.108	7
79	387	757	0.140	8
80	299	377	0.079	5
81	301	321	0.085	5
82	305	353	0.073	6
83	294	330	0.086	6
84	299	343	0.086	6
85	281	275	0.076	5
86	312	402	0.081	6
87	298	332	0.089	5
88	255	212	0.080	4
89	290	300	0.086	5
90	240	162	0.072	4
38	235	155	0.088	4
39	532	1501	0.424	14
40	370	666	0.102	8
41	458	1256	0.280	11
42	236	150	0.063	4
44	298	312	0.106	4
45	265	229	0.081	5
118	282	262	0.082	5
119	313	379	0.085	5
120	341	479	0.089	6
121	297	351	0.069	5
122	279	278	0.081	5
123	293	312	0.091	5
124	213	387	0.081	6
125	284	263	0.062	6
126	304	362	0.078	6
127	208	126	0.158	5

Lake whitefish collected from the Inukjuak River watersheds in summer 2019

(continue…)

# Fish ID	fork length, mm	weight, g	mercury, µg/g w.w.	age
128	269	240	0.068	5
129	369	667	0.094	8
130	460	991	0.322	11
238	237	181	0.060	4
239	205	126	0.060	4
240	395	801	0.103	9
242	405	845	0.119	8
244	280	286	0.063	5
245	310	356	0.057	5
246	345	494	0.066	7
247	335	537	0.072	7
248	430	914	0.163	10
249	475	1357	0.239	12
58	278	216	0.082	5
59	304	338	0.101	5
60	273	228	0.100	5
61	336	442	0.096	7
62	358	622	0.088	6
63	377	683	0.093	7
138	175	52	0.117	3
139	188	67	0.082	3
140	395	865	0.121	9
141	388	782	0.087	8
142	439	1060	0.185	12
143	404	801	0.119	9
144	345	527	0.065	6
145	286	244	0.090	5
146	291	271	0.089	6
147	223	125	0.081	4
168	216	114	0.080	4
169	230	137	0.037	4
170	250	167	0.078	5
171	259	216	0.045	5
172	296	385	0.063	5
173	276	273	0.046	5
174	331	446	0.071	7
175	443	1093	0.217	12
176	451	1193	0.244	12
177	430	970	0.295	10
178	419	867	0.144	12
179	374	687	0.098	8
180	222	122	0.058	4
181	424	978	0.234	12

Brook trout collected from the Inukjuak River watersheds in summer 2019

# Fish ID	fork length, mm	weight, g	mercury, µg/g w.w.	age
2	390	628	0.239	7
3	320	298	0.210	5
4	248	159	0.077	4
5	375	640	0.168	5
6	380	542	0.141	5
7	260	185	0.196	
8	390	643	0.137	6
9	315	340	0.111	4
10	360	528	0.242	7
11	367	567	0.203	6
12	430	565	0.098	6
13	309	331	0.218	4
14	300	293	0.087	4
15	310	348	0.269	5
16	365	606	0.116	5
17	355	489	0.141	5
18	370	524	0.131	5
19	350	479	0.130	4
20	350	472	0.108	4
22	350	496	0.109	5
23	305	331	0.111	5
24	380	621	0.273	5
25	320	374	0.159	5
65	190	70	0.072	3
66	200	82	0.063	3
67	290	276	0.157	4
68	340	422	0.107	5
69	355	471	0.096	5
70	360	415	0.326	6
71	375	596	0.176	5
72	378	654	0.080	5
73	368	598	0.119	5
92	130	24	0.058	2
43	315	391	0.071	4
132	438	1021	0.227	7
133	319	389	0.098	5
134	366	609	0.113	5
135	354	622	0.170	4
232	278	297	0.068	4
233	130	27	0.044	2
234	160	54	0.058	2
235	355	559	0.108	5
236	255	242	0.066	4
237	280	285	0.067	4

Brook trout collected from the Inukjuak River watersheds in summer 2019

(continue ….)

# Fish ID	fork length, mm	weight, g	mercury, µg/g w.w.	age
54	189	71	0.064	3
55	316	318	0.141	4
56	344	402	0.207	4
57	396		0.577	7
149	342	459	0.176	5
153	210	101	0.064	3
154	182	69	0.048	2
155	212	108	0.065	3
156	251	162	0.147	4
157	331	384	0.143	4
158	368	520	0.280	7

Lake trout collected from the Lake Tasialuk in summer 2019

# Fish ID	fork length, mm	weight, g	mercury, µg/g w.w.	age
187	380	595	0.322	
188	215	104	0.115	4
189	290	246	0.421	7
190	430	949	0.286	9
191	520	1757	0.585	13
201	400	758	0.450	7
202	360	566	0.211	9
203	345	457	0.180	7
204	375	593	0.262	7
205	330	416	0.242	7
206	460	1215	0.457	9
207	245	151	0.155	8
208	295	288	0.192	6
209	370	670	0.201	8
210	430	943	0.461	9
212	540	2377	0.755	12
213	345	491	0.135	8
214	350	490	0.243	11
215	460	1176	0.598	
216	420	867	0.466	9
217	445	985	0.315	9
218	430	878	0.398	11
219	395	655	0.344	9
221	380	622	0.286	8
223	425	900	0.352	10
225	365	547	0.344	10
226	315	336	0.292	7
227	400	750	0.342	8
228	365	538	0.436	8
229	350	521	0.117	6
230	340	527	0.249	7
272	315	318	0.307	7
273	445	1022	0.287	10
274	210	105	0.088	4
275	475	1316	0.526	11
276	480	1285	0.411	13
277	190	80	0.129	4