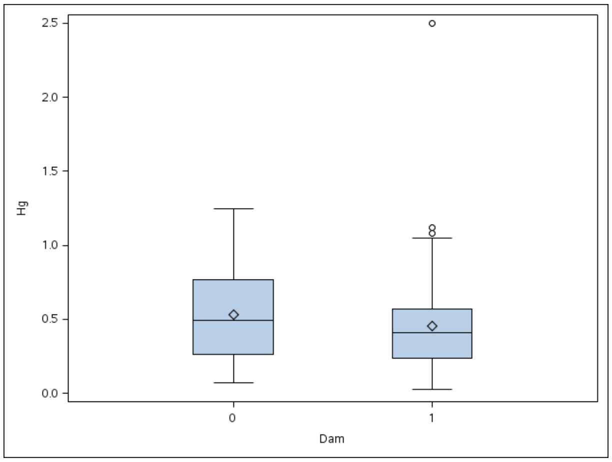
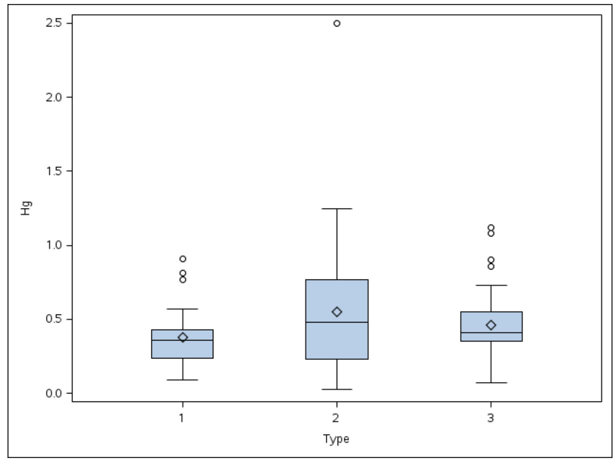
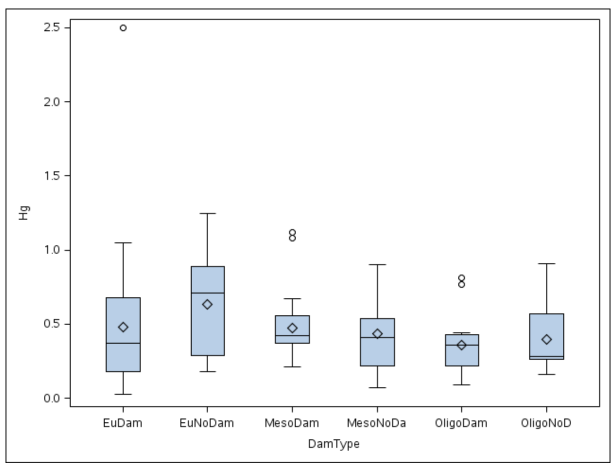
STA 303 H1F / 1002 HF – Winter 2012 – Assignment 1
SOLUTIONS

1. The boxplots below show mercury levels for lakes with and without a functioning dam present, for the three types of lakes, and for lakes classified by both dam status and type.







All plots show one lake (a eutrophic lake with a dam) whose mercury level is quite a bit larger than all other values.

Disregarding this outlier, the distributions, including the means, of mercury levels in lakes with and without dams is similar. Eutrophic lakes have slightly higher mercury levels on average then mesotrophic lakes which in turn have slightly higher mercury levels than oligotrophic lakes and eutrophic lakes seem to have slightly greater spread than the other two types of lake. When the lakes are classified by both dam status and lake type, the most noticeable difference is that eutrophic lakes without a dam appear to have higher mercury levels on average than the other classifications of lakes, and there is some indication that there is, again, greater spread in the observations for the eutrophic lakes.

2. The table below gives the mean mercury levels for lakes with and without a dam and the test statistic and p-value for the two-sided test of equal means under the assumption that the variance is the same in both groups.

	Mean mercury level (ppm)			
	With dam	Without dam	Test statistic	p-value
With Hodgdon Pond	0.46	0.53	1.14	0.26
Without Hodgdon Pond	0.43	0.53	1.93	0.06

On average, lakes without a dam have higher mercury levels than lakes with a functioning dam; lakes without a dam have a mean mercury level of 0.53 p.p.m. while lakes with a dam have a mean of 0.46 p.p.m. (0.43 p.p.m. without Hodgdon Pond). The t-test comparing the means does not result in evidence of a difference between the means when Hodgdon Pond is in the data. When the outlying Hodgdon Pond is removed, there is weak evidence of a difference in the mean mercury levels depending on the presence of a functional dam ($p = 0.06$).

3. The table below gives the mean mercury levels for the three types of lakes and the test statistic and p-value of the analysis of variance F-test of equal means.

	Mean mercury level (ppm)				
	Oligotrophic	Eutrophic	Mesotrophic	Test statistic	p-value
With Hodgdon Pond	0.38	0.55	0.46	2.23	0.11
Without Hodgdon Pond	0.38	0.51	0.46	1.79	0.17

There is no evidence of differences in the mean mercury levels among the three types of lakes. This is the case whether or not the outlying lake is retained in the data.

4. The table below gives the mean mercury levels for the lakes classified by both type and dam status and the test statistic and p-value of the analysis of variance F-test of equal means.

	Mean mercury level (ppm)							
	Oligotrophic		Eutrophic		Mesotrophic		Test	
	Dam	No dam	Dam	No dam	Dam	No dam	statistic	p-value
With Hodgdon Pond	0.36	0.40	0.48	0.64	0.47	0.44	1.48	0.20
Without Hodgdon Pond	0.36	0.40	0.41	0.64	0.47	0.44	2.55	0.03

When Hodgdon Pond is retained in the data there is no evidence of any differences in the mean mercury level among the 6 classifications of lakes. However when Hodgdon Pond is removed, there is evidence that not all lake classifications have the same mean mercury level ($p = 0.03$).

Since the differences among lake classifications in mean mercury levels are significant only for the data with the outlier removed, post-hoc analysis only needs to be carried out for these data.

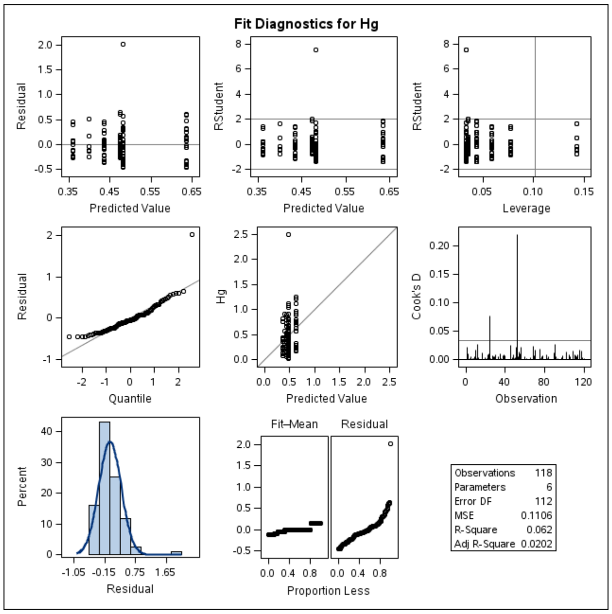
A post-hoc pairwise comparison of the means using Tukey's procedure to control the Type I error rate indicates that there is evidence that eutrophic lakes with no dam have higher mercury levels on average than both eutrophic lakes with a dam ($p = 0.043$) and oligotrophic lakes with a dam ($p = 0.047$). In both cases the mean mercury level is higher in eutrophic lakes with no dam. The p-values for all pairwise comparisons with Tukey's adjustment is given in the table below.

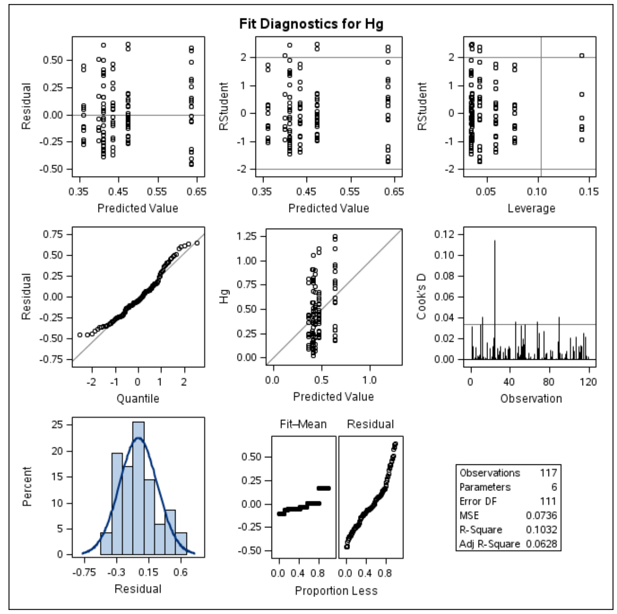
		Oligotrophic		Eutrophic		Mesotrophic	
		Dam	No dam	Dam	No dam	Dam	No dam
Oligotrophic	Dam		0.9996	0.9933	0.0474	0.8147	0.9756
	No dam			1.0000	0.3419	0.9873	0.9997
Eutrophic	Dam				0.0427	0.9533	0.9997
	No dam					0.2854	0.1994
Mesotrophic	Dam						0.9973
	No dam						

Instead of Tukey's procedure, the Bonferroni method could be used to control the Type I error rate. This shows weak evidence that eutrophic lakes with no dam have different mercury levels on average than eutrophic lakes with a dam ($p = 0.057$) and oligotrophic lakes with a dam ($p = 0.064$). In both cases the mean mercury level is higher in eutrophic lakes with no dam. The p-values for all pairwise comparison with Tukey's adjustment is given in the table below.

		Oligotrophic		Eutrophic		Mesotrophic	
		Dam	No dam	Dam	No dam	Dam	No dam
Oligotrophic	Dam		1.0000	1.0000	0.0635	1.0000	1.0000
	No dam			1.0000	0.6987	1.0000	1.0000
Eutrophic	Dam				0.0566	1.0000	1.0000
	No dam					0.5453	0.3422
Mesotrophic	Dam						1.0000
	No dam						

5. The diagnostics panels below give plots of the residuals for the analyses with and without Hodgdon Pond in the data.





The mercury level in Hodgdon Pond is clearly unusual in all plots and this observation is not well-fit by the model as its standardized residual is almost 8. As a result, the inferences for the data with this point are not valid.

Looking at the residual plots for the model fit to the data with Hodgdon Pond removed:

- Looking at the plot of the standardized residuals versus the predicted values, there are no large (positive or negative) outliers of concern. (A few observations have standardized residuals > 2 but a few observations in this range are expected in a data set this large.)
- The residual plots versus the predicted values show approximately equal variance in each of the 6 categories so the assumption of constant variance seems reasonable. Moreover, the standard deviations for observations from the 6 categories of lakes range from 0.216 to 0.340, which is will within the 2:1 ratio suggested by the rule-of-thumb for equal variances.
- The normal quantile plot shows some evidence of a light left tail in the distribution of the residuals. This can also be observed in the histogram of the residuals. However, since we are analyzing means, small departures from normality will still lead to approximately correct inferences.
- Lakes that are close together geographically may have some relationship so it is possible that the observations may not be independent. But we will hope that the lakes chosen for the study are far enough apart that any spatial correlation is negligible and can be ignored.

Since we have no serious concerns about violations of the model assumptions, we can trust the statistical tests carried out on the data without the outlier.

6. (a) The model fit in question 4 has 5 predictor variables (which are all indicator variables) since there are 6 categories of lakes. The predictor variables in the two-way model would be: 1 indicator variable for dam status (since it has 2 categories), 2 indicator variables

for lake type (since it has 3 categories), and thus $2(= 1 \times 2)$ interaction terms, for a total of 5. So both models have the same number of predictor variables.

(b) In question 4, our results showed that for eutrophic lakes there is a significant difference in the mean mercury level between those with and without dams. But for the other types of lakes, there is no significant difference between those that do and do not have a dam. Thus there is an interaction between dam status and type of lake so the F-test for the interaction in the two-way model will likely be statistically significant.

7. From the combined samples, we expect there will be lower variation for samples comprised of more fish. This violates the assumption that the error term in the model has the same variance for all observations. So we should be concerned about this.

   Since there are fewer fish in eutrophic lakes, it is reasonable to speculate that these are the lakes in which only 2 or 3 fish were caught. And in question 1 we noted that there seems to be more variability in observations from eutrophic lakes.

   (Weighted least squares with weights proportional to the number of fish used in the measurement for the observation may be appropriate.)