ST 307 Project 3 Rainier Sangalang
Group: Liesl Miranda & Vy Tran

Objectives

Our goal in this project was to analyze data taken from a data set containing information on lakes mostly from Maine. We had four main objectives:

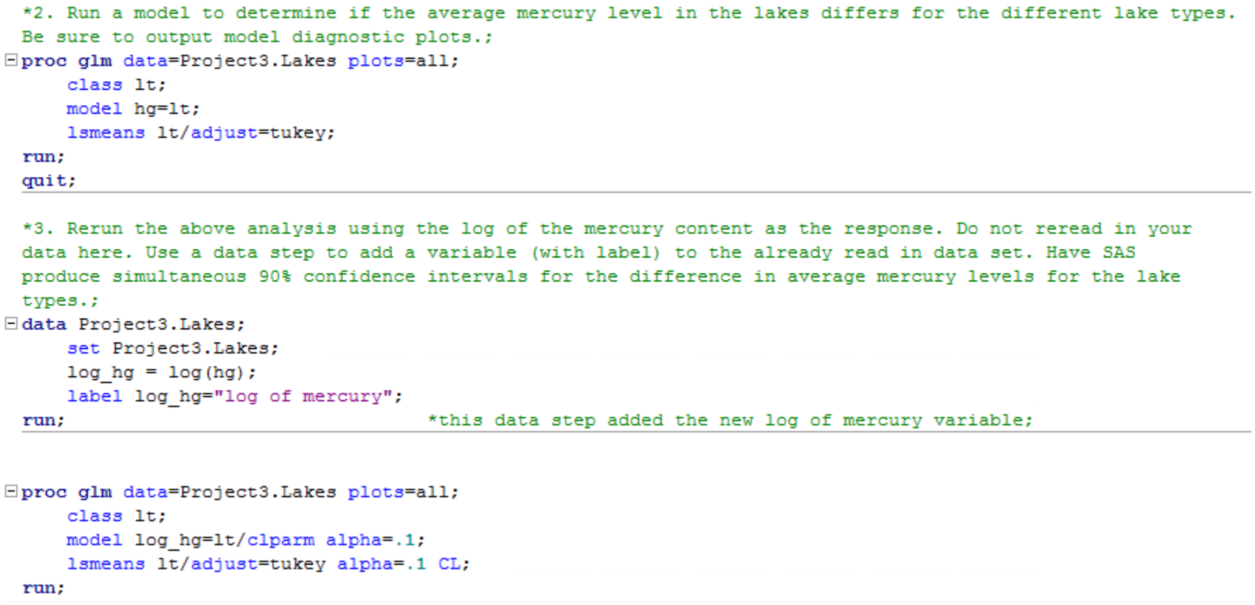
1. Determined if the average mercury level and the log of the mercury levels in the lakes differ for the different lake types. We wanted to produce simultaneous 90% confidence intervals for the difference in average mercury levels for the lake types.
2. Conducted correlation analysis between Mercury, Elevation, Surface Area, Drainage Area, and Flushing Rate
3. Ran a model to investigate the linear relationship between log of Mercury (response) and Flushing Rate (predictor). Had SAS produce a 99% confidence interval for the slope parameter.
4. Produced a 99% confidence interval for the mean value of log mercury at a lake with flushing rate 0.78

Methods

Initially, the data had to be cleaned and formatted to be analyzed through statistical methods. To meet our objectives, we utilized the power of SAS to analyze our data. Our data was stored in a tab delimited text file and read into a database within SAS. We went through each variable individually and designated each variable as one of those two types. We made sure the data was read in correctly by changing variable lengths and formats to improve user readability and functionality within SAS. To conduct regression and correlation tests, we assumed the data set was normally distributed.

Objective 1

To achieve this objective, we ran regression analysis using mercury level as the response variable and lake type as the predictor variable. This was repeated but instead we used the log of the mercury level as the response variable.



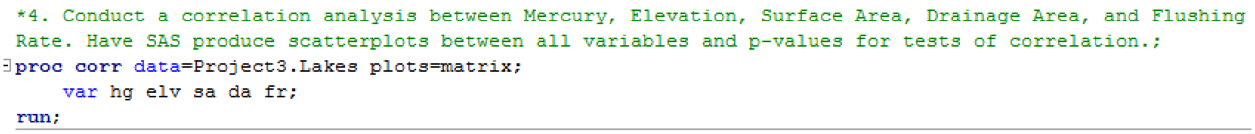
```
*2. Run a model to determine if the average mercury level in the lakes differs for the different lake types.
Be sure to output model diagnostic plots.;
proc glm data=Project3.Lakes plots=all;
    class lt;
    model hg=lt;
    lsmeans lt/adjust=tukey;
run;
quit;

*3. Rerun the above analysis using the log of the mercury content as the response. Do not reread in your
data here. Use a data step to add a variable (with label) to the already read in data set. Have SAS
produce simultaneous 90% confidence intervals for the difference in average mercury levels for the lake
types.;
data Project3.Lakes;
    set Project3.Lakes;
    log_hg = log(hg);
    label log_hg="log of mercury";
run;                                *this data step added the new log of mercury variable;


proc glm data=Project3.Lakes plots=all;
    class lt;
    model log_hg=lt/clparm alpha=.1;
    lsmeans lt/adjust=tukey alpha=.1 CL;
run;
```

Objective 2

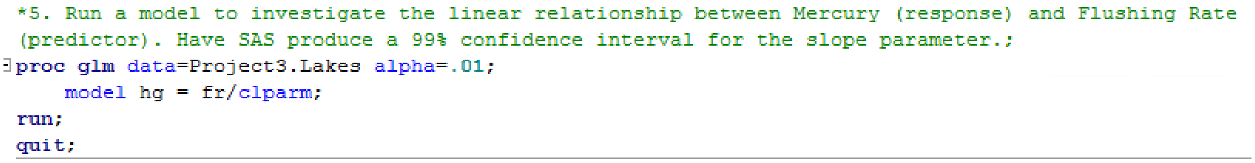
To achieve this objective, we used *proc corr* to perform correlation analysis on the variables Mercury, Elevation, Surface Area, Drainage Area, and Flushing Rate. We had SAS produce scatterplots between all the variables.



```
*4. Conduct a correlation analysis between Mercury, Elevation, Surface Area, Drainage Area, and Flushing
Rate. Have SAS produce scatterplots between all variables and p-values for tests of correlation.;
proc corr data=Project3.Lakes plots=matrix;
    var hg elv sa da fr;
run;
```

Objective 3

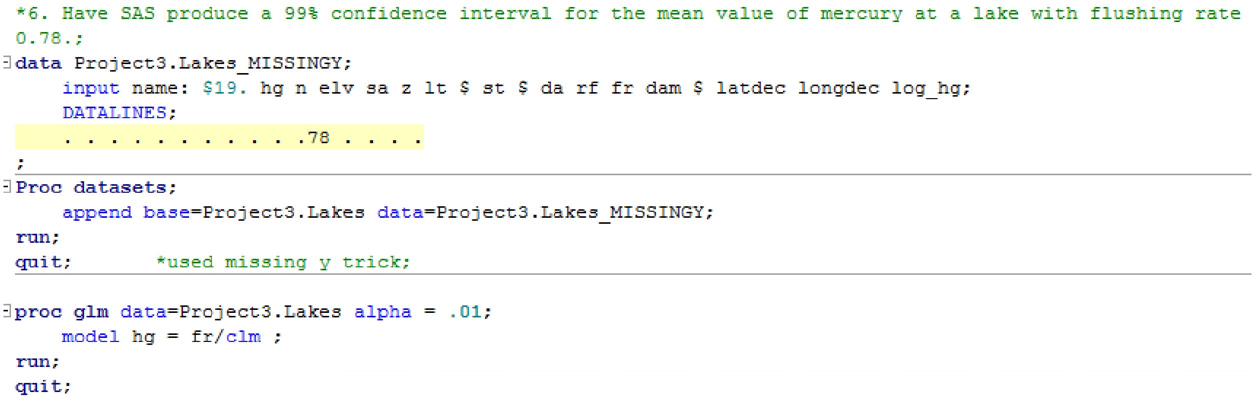
To achieve this objective, we ran a model to investigate the linear relationship between log of Mercury (response) and Flushing Rate (predictor). We had SAS produce a 99% confidence interval for the slope parameter. This was done using regression analysis.



```
*5. Run a model to investigate the linear relationship between Mercury (response) and Flushing Rate
(predictor). Have SAS produce a 99% confidence interval for the slope parameter.;
proc glm data=Project3.Lakes alpha=.01;
    model hg = fr/clparm;
run;
quit;
```

Objective 4

To achieve this objective, we had to introduce a new entry in the data set with a mercury value of .78. The missing y trick was utilized to add the new entry. We then used regression analysis to produce the 99% confidence interval for the mean value of log mercury at a lake with flushing rate 0.78.



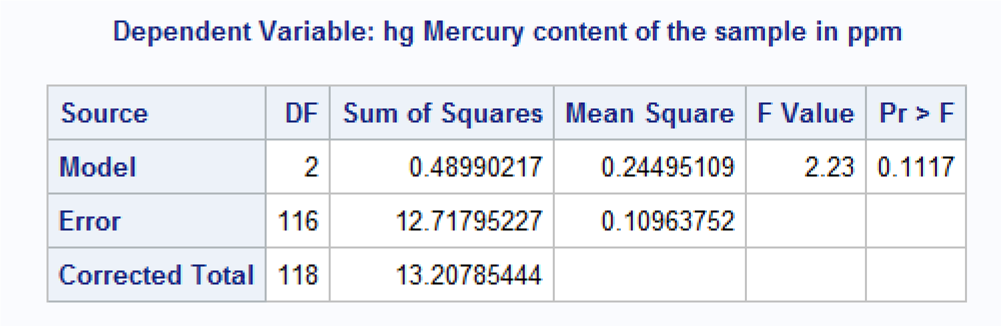
```
*6. Have SAS produce a 99% confidence interval for the mean value of mercury at a lake with flushing rate
0.78.;
data Project3.Lakes_MISSINGY;
    input name: $19. hg n elv sa z lt $ st $ da rf fr dam $ latdec longdec log_hg;
    DATALINES;
    . . . . . . . . . . .78 . . . .
;
Proc datasets;
    append base=Project3.Lakes data=Project3.Lakes_MISSINGY;
run;
quit;      *used missing y trick;

proc glm data=Project3.Lakes alpha = .01;
    model hg = fr/clm ;
run;
quit;
```

Results

Objective 1

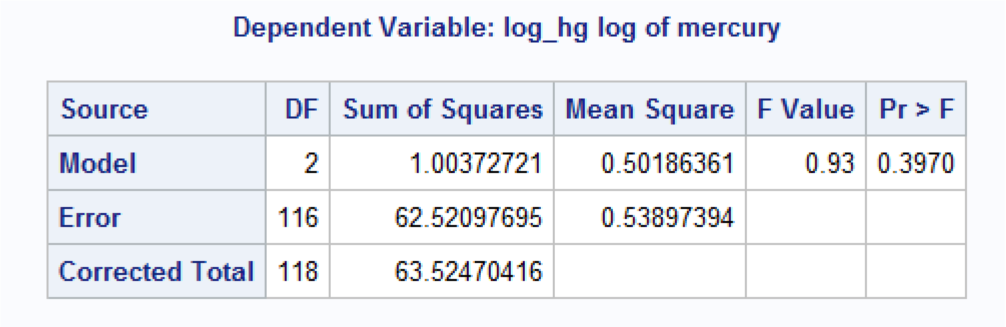
After conducting the regression analysis, we obtained a p-value of .1117 which is greater than the alpha of .05 which means that we do not have evidence that the average mercury level in the lakes differs for the different lake types.



Dependent Variable: hg Mercury content of the sample in ppm

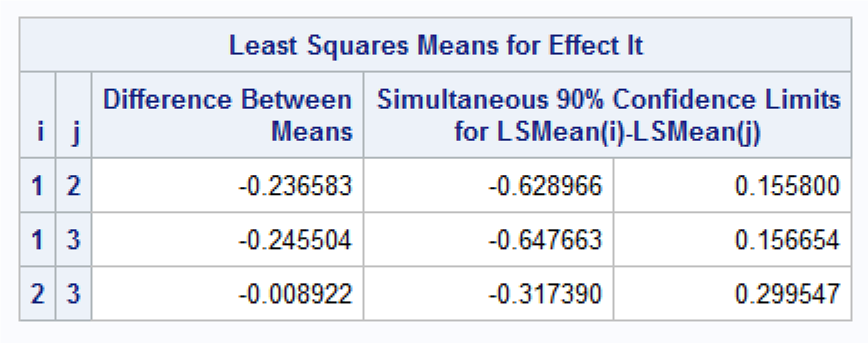
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.48990217	0.24495109	2.23	0.1117
Error	116	12.71795227	0.10963752		
Corrected Total	118	13.20785444			

When running regression analysis for the log of the mercury levels, we obtained a p-value of .3970 which is greater than the alpha of .1 which means that we do not have evidence that the log of the mercury levels in the lakes differs for the different lake types. We also produced 90% confidence intervals.



Dependent Variable: log_hg log of mercury

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.00372721	0.50186361	0.93	0.3970
Error	116	62.52097695	0.53897394		
Corrected Total	118	63.52470416			

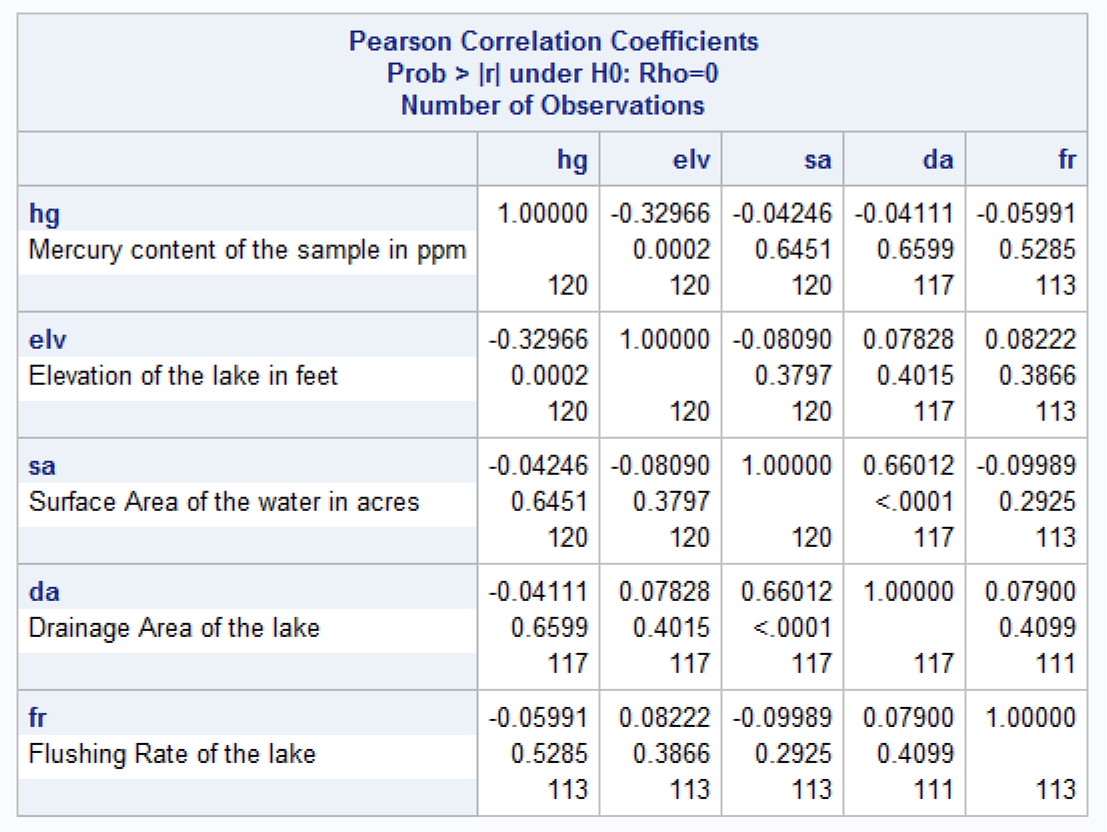


Least Squares Means for Effect lt				
i	j	Difference Between Means	Simultaneous 90% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.236583	-0.628966	0.155800
1	3	-0.245504	-0.647663	0.156654
2	3	-0.008922	-0.317390	0.299547

Objective 2

When we conducted regression analysis on the variables Mercury, Elevation, Surface Area, Drainage Area, and Flushing Rate and produced scatter plots between all variables and p-values for tests of correlation. Below is the output.

According to the results, there exists a linear relationship between drainage area and surface area of the lakes. There also exists a linear relationship between elevation and mercury content.



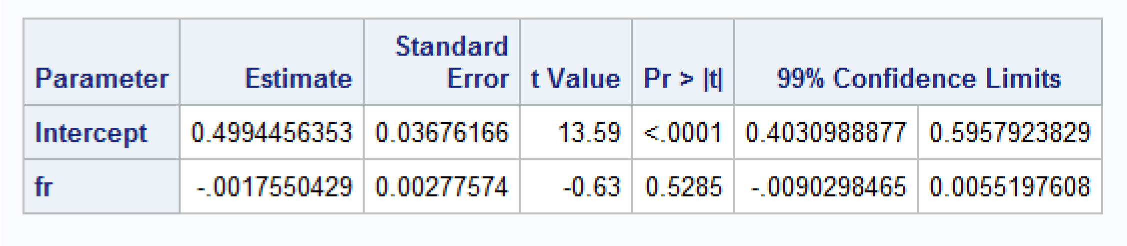
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	hg	elv	sa	da	fr
hg Mercury content of the sample in ppm	1.00000 120	-0.32966 0.0002 120	-0.04246 0.6451 120	-0.04111 0.6599 117	-0.05991 0.5285 113
elv Elevation of the lake in feet	-0.32966 0.0002 120	1.00000 120	-0.08090 0.3797 120	0.07828 0.4015 117	0.08222 0.3866 113
sa Surface Area of the water in acres	-0.04246 0.6451 120	-0.08090 0.3797 120	1.00000 120	0.66012 <.0001 117	-0.09989 0.2925 113
da Drainage Area of the lake	-0.04111 0.6599 117	0.07828 0.4015 117	0.66012 <.0001 117	1.00000 117	0.07900 0.4099 111
fr Flushing Rate of the lake	-0.05991 0.5285 113	0.08222 0.3866 113	-0.09989 0.2925 113	0.07900 0.4099 111	1.00000 113

Objective 3

We investigated the linear relationship between the log of Mercury and Flushing Rate. 99% confidence interval for the slope parameter was produced. Below is the output.

The p-value for the flush rate in this case is well above the alpha of .01 which means that we do not have evidence of a linear relationship between the log of Mercury and Flushing rate.

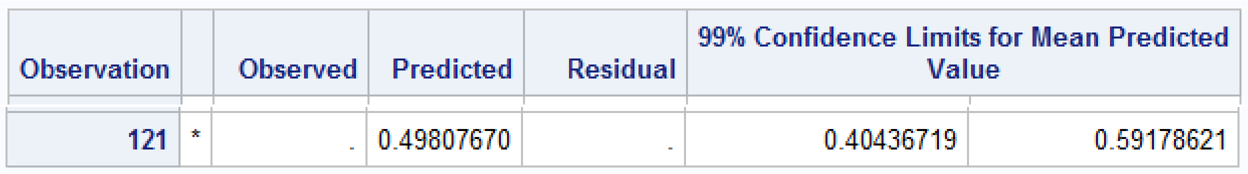


Parameter	Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	0.4994456353	0.03676166	13.59	<.0001	0.4030988877	0.5957923829
fr	-.0017550429	0.00277574	-0.63	0.5285	-.0090298465	0.0055197608

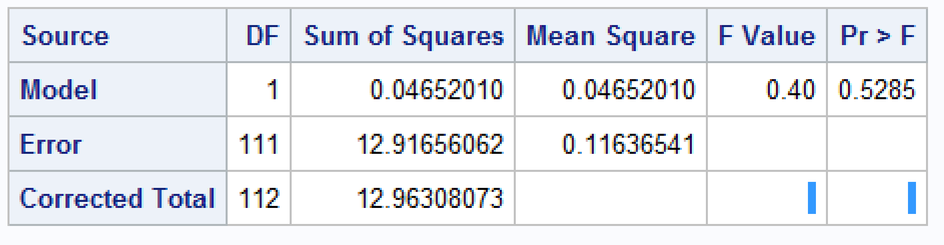
Objective 4

We used regression analysis to produce the 99% confidence interval for the mean value of log mercury at a lake with flushing rate 0.78. Below is the output.

Again, the p-value is well above the alpha level of .01 which means that we do not have evidence of a linear relationship between the mean value of the log of mercury and the flushing rate of .78.



Observation		Observed	Predicted	Residual	99% Confidence Limits for Mean Predicted Value	
121	*	.	0.49807670	.	0.40436719	0.59178621



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.04652010	0.04652010	0.40	0.5285
Error	111	12.91656062	0.11636541		
Corrected Total	112	12.96308073			

Summary

The majority of the tests performed in this project showed that there was not enough evidence to show any correlation between variables. The only correlation present was between drainage area and surface area, and with elevation and mercury content. It is important to remember that correlation does not imply causation.