London School of Economics and Political Science

Department of Management

The Health Effects Of Eating Contaminated Fish

Applied Statistics Project 2015/2016

Candidate Number. 13709

Table of Contents

1 Introduction

The aim of this report is to investigate the effects of poisonous metal on human health, and find out if eating contaminated fish does cause chromosome damage to human body. The cases studied specifically involves subjects who ate fish that were contaminated with methylmercury – an organic form of mercury that is primarily responsible for mercury poisoning in humans due to its high toxicity[1].

This paper analyses the information of 39 individuals whose medical history had been provided to our medical research team. 23 of these were exposed individuals who had consumed contaminated fish at a rate of three meals a week at minimum for more than three years, while the remaining 16 did not consume contaminated fish regularly and ate far less fish of all kinds, therefore constituting the control group in this study.

In this report, 5 important hypotheses will be explored and verified through various statistical analyses. This will enable me to systematically derive my conclusion on the correlation between consumption of contaminated fish and chromosome damage. The 5 hypotheses are as follows:

1. Higher levels of mercury in blood of exposed group
2. More chromosome abnormalities in total in exposed subjects
3. More C_u cells in exposed subjects
4. Greatest chromosome damage in subjects with highest levels of mercury in their blood
5. No difference between the level of unrelated conditions in the control group and the exposed group

The statistical software used in this report is Minitab[2].

2 Data exploration and analysis

2.1 Data set

Four outcome measures were given for each of the 39 individuals:

Mercury in blood: Amount of mercury in the individual's blood, recorded in (ng/g)

% abnormal cells: Percentage of cells exhibiting chromosome damage, called abnormal cells

% C_u cells: Percentage of cells exhibiting a particular type of chromosome abnormality called C_u cells, specifically, asymmetrical or

[1] http://www.medicinenet.com/script/main/art.asp?articlekey=31628
[2] https://www.minitab.com/en-us/

incomplete symmetrical chromosome aberrations as recorded in cells cultured between 48 and 120 hours.

Unrelated conditions: Number of unrelated health conditions such as asthma, hypertension, influenza, drugs taken regularly, and diagnostic X-rays, over the same three year period.

Data on these outcome measures are set out with respect to the individual subjects marked 1 to 39 under the ID column, in Table 1 of the Appendix at the end of the report.

2.2 Data analysis

An initial exploratory data analysis is especially crucial because of the small sample size provided; error in value of the variables would be more significant when used to perform statistical test, with a higher tendency of reaching an incorrect conclusion. Table 2 below shows the summary statistics for the 4 variables provided.

Variable	Group	N	N*	Mean	SE Mean	StDev	Minimum	Median	Maximum
Mercury	control	16	0	8.938	0.966	3.863	3.000	8.600	17.000
	exposed	23	0	198.2	46.6	223.3	12.8	150.0	1100.0
Abnormal	control	16	0	4.669	0.837	3.347	1.000	4.350	13.000
	exposed	23	0	8.88	1.03	4.93	0.00	9.00	21.50
Cu	control	16	0	1.075	0.370	1.479	0.000	0.250	5.000
	exposed	23	0	2.778	0.480	2.303	0.000	2.000	9.500
Unrelated	control	16	0	0.625	0.287	1.147	0.000	0.000	4.000
	exposed	23	0	1.174	0.456	2.188	0.000	0.000	9.000

Table 2: Summary statistics

The values in table 2 show that the mean of the exposed group is approximately 22 times higher in mercury in blood, 2 times higher in % abnormal cells and 2.5 times higher in % C_u cells. This is not surprising since the exposed group eats contaminated fish more regularly.

There is no missing value in the data provided.

From the table above, the standard deviation of Mercury in the exposed group, and Cu in the control group have higher values compared to its mean. Besides, the matrix plot of Mercury, Abnormal and Cu (Figure 1 in the Appendix) shows that the linear relationship between Mercury vs Abnormal and Mercury Vs Cu is not that obvious. These 2 points suggest that a log transformation for the Mercury variable may help to linearise the data and account for the high standard deviation at the same time.

A box plot of Mercury for the separate groups as shown in the top left box in Figure 2 depicts that the amount of mercury in the blood for the control group are all near 0. However, in the case of the exposed group, the blood mercury levels seem to lie across the range of values

near 0 to 400, with an outlier observed. Upon further investigation, the outlier belongs to individual 26, with mercury in blood content of 1100ng/g. It is sensible to say that the higher mercury level belongs to an elderly who had accumulatively consumed more contaminated fish over his lifespan, as compared to a younger exposed individuals. However, methylmercury does leave the body after several months[3]. Therefore, this individual has a high mercury in blood content probably because he ate more contaminated fish over the 3 years period.

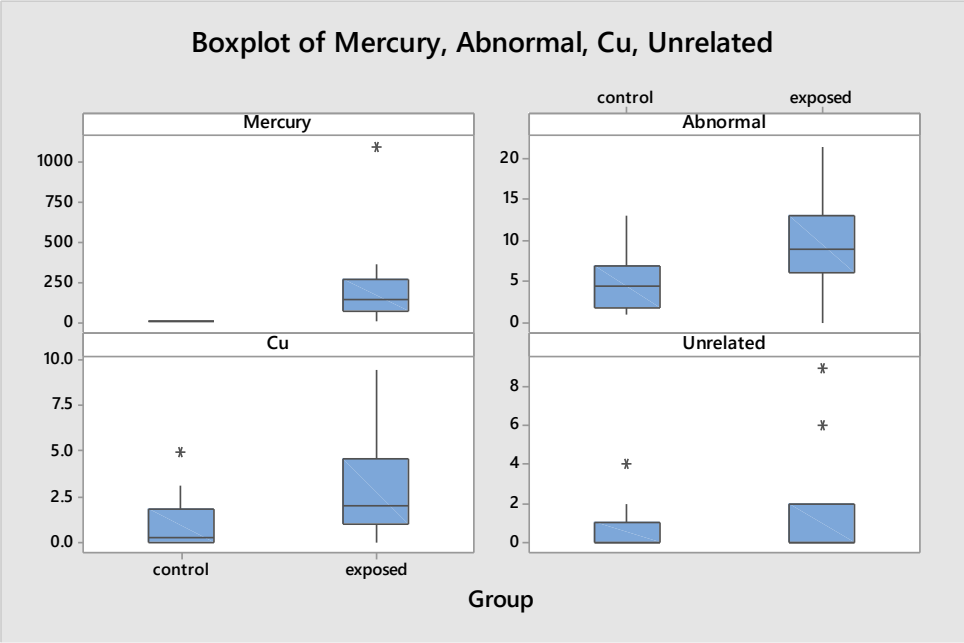

Figure 2: Box plot of Mercuy, Abnormal and Cu

A box plot of % abnormal cells as shown in the top right box in Figure 2 above shows that the median and maximum value in the exposed group is higher than the control group. Although no outlier has been found, what is interesting here is that the minimum of the exposed group is lower that of the control group. Individual 35 and 38 in the exposed group does not exhibit any chromosome damage and further investigation needs to verify if this is true. A box plot of % C_u cells as shown in the bottom left box in Figure 2 above highlights an outlier (individual 5) in the control group. However, given that the mercury in blood and % abnormal cells for individual 5 is the highest amongst the control group, it should not be a major concern that the % C_u cells is the highest. A box plot of Unrelated as shown in the bottom right in Figure 2 is not very insightful because the minimum and median for the control and exposed group is 0. There is 3 outliers in unrelated conditions, with 2 of them coming from the exposed group.

Although t tests are considered robust for violations of normal distributions, I have drawn a probability plot (normal test plot) in Figure 3 below to check if the variables in each group follow a normal distribution. The given p-values for each group in each variables are shown in Table 3 below. The p-values suggest that the data for mercury in the exposed group and Cu in the control group does not show normality, with p value less than 0.005. A log transformation should be performed for these variables.

[3] http://www.atsdr.cdc.gov/PHS/PHS.asp?id=112&tid=24

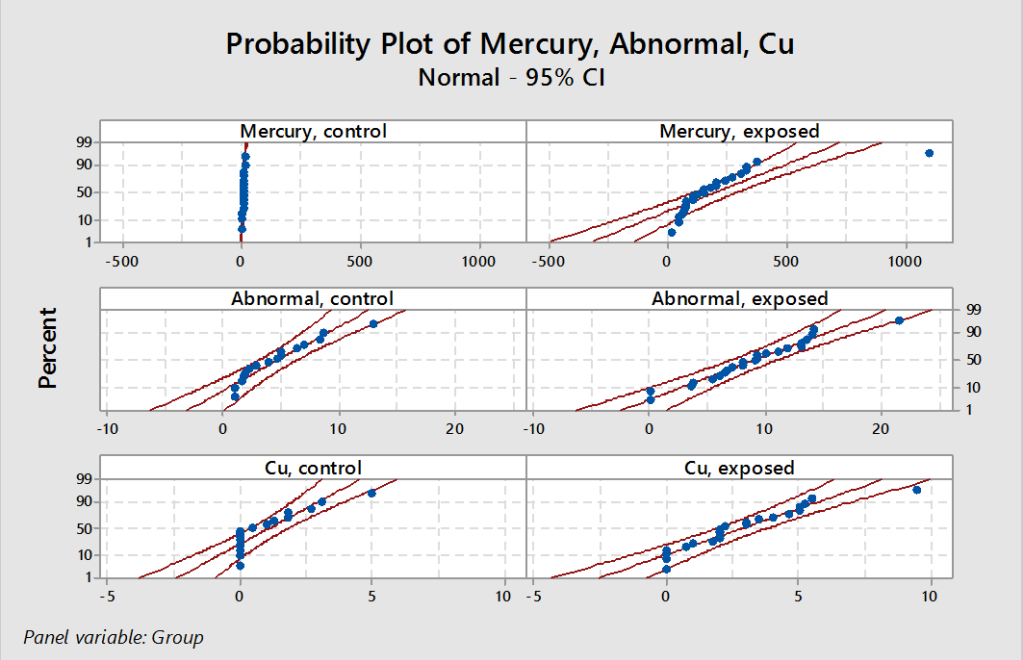

Figure 3: Probability plot (normal test plot) for the different variables

	Mercury	Abnormal	Cu
Control	0.93	0.175	<0.005
Exposed	<0.005	0.754	0.107

Table 3: P-value for each group in each variable

3 Hypothesis 1 – Eating contaminated fish results in higher mercury in blood

To test this hypothesis, a 2 sample test t-test should be carried out since I am interested to find out if higher mercury in blood is present in the exposed group. Bearing in mind that normality does not hold for the exposed group, and the standard deviation for this group is high, it is recommended to carry out a log transformation to address the positively skewed data. The data are independent since an individual appears only in the exposed or control group. Table 4 in the Appendix shows the data with log (mercury) column created.

3.1 Data transformation

A probability plot of log (mercury) as shown in Figure 4 below indicates that the data conforms to normality to a larger extend, where the p-value of the probability plot for the exposed group increased from <0.005 to 0.711 after the log transformation. A box plot of the log transformed data as seen in Figure 5 in the Appendix also show that the outlier has been removed.

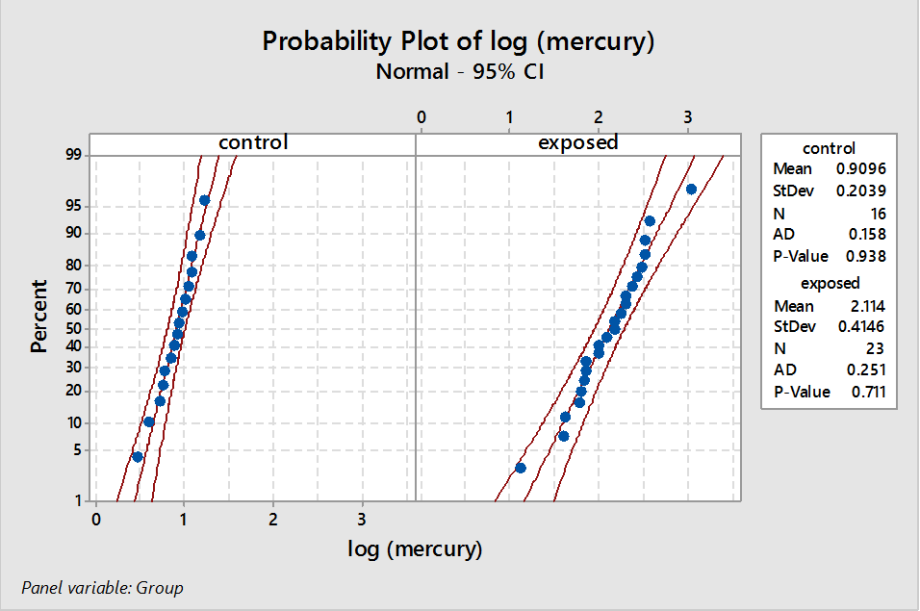

Figure 4: Probability plot of log (mercury)

A Levene's test would then be necessary to check if the control and exposed group have the same variance. This test the null hypothesis that the 2 variances are equal against the hypothesis that the variance are different. The test gives a p-value of 0.022 and is significant; I can conclude that the variance for the 2 samples are different. Details of the Levene's test can be found in Figure 6 in the Appendix.

3.2 Analysis

From here, carrying a 2 sample t-test (95% confidence interval) with null hypothesis that log mercury amount between the 2 groups are the same against the hypothesis that the amount between the 2 groups are different, gives a P-value of 0 when the variances are not assumed to be equal as shown in Table 5 in the Appendix. 95% confidence intervals of the difference between mean of control and exposed group gives (-1.409, -1.001). This is equivalent to the exposed group having a log mercury content of (1.001, 1.409) higher than the control group. It is also estimated that the exposed group has a log mercury level of 1.205 higher than the control group.

One important point to take note is that these values are in the log form. To check how much higher the mercury in blood for the exposed group is compared to the control group, an antilog is required. This is equivalent to ($10^{1.001}$, $10^{1.409}$) for the confidence interval and $10^{1.205}$ for the estimated difference. The exact values for the confidence interval and estimated difference are (10.023, 25.645) and 16.032 respectively.

This is obvious that the blood mercury level for the exposed group is higher than the control group since the 2 sample t-test is significant and the value 0 does not lie between the confidence interval.

4 Hypothesis 2 – Eating contaminated fish increases % abnormal cells

Hypothesis 2 is more straightforward than hypothesis 1. No transformation is needed because data of % abnormal cells in each group does show normal distribution to a certain extent. Using Levene's test to test H_0 that variance of the 2 groups are equal, against H_1 where their variance are not the same (Figure 7 in the Appendix), gives a p-value of 0.206 and the test is not significant. However, I need to take note that the sample size is small and there is a possibility that the Levene's test is not as accurate.

Two-Sample T-Test and CI: Abnormal, Group

```
Two-sample T for Abnormal

Group     N  Mean  StDev  SE Mean
control  16  4.67   3.35     0.84
exposed  23  8.88   4.93      1.0


Difference = μ (control) - μ (exposed)
Estimate for difference:  -4.21
95% CI for difference:  (-6.90, -1.52)
T-Test of difference = 0 (vs ≠): T-Value = -3.18  P-Value = 0.003  DF = 36
```

Table 6: 2 sample t-test of Abnormal by Group

A 2 sample t-test (95% confidence interval) with null hypothesis that % abnormal cells between the 2 groups are the same against the hypothesis that the percentage between the 2 groups are different, gives a P-value of 0.003 when the variances are assumed equal. The 95% confidence interval for the difference in % abnormal cells between exposed and control gives (1.52, 6.90) and the estimated difference is 4.21%. The percentage of abnormal cells in the exposed group is higher since 0 does not lie between this and test is significant.

From this, I can say that eating contaminated fish increases the % abnormal cells and % abnormal cell is higher in the exposed group compared to the control group.

5 Hypothesis 3 – Eating contaminated fish increase % C_u cells.

Figure 3 suggest that a transformation of data is recommended for variable % C_u cells since the standard deviation for the control group is higher than its mean as shown in Table 2, and the probability plot shows that this group deviates from normality. However, a log transformation will not work well because there are zero values present in some of the variables, and log of zero is undefined. Therefore, it is not ideal to conduct a 2 sample t-test to test the difference in % C_u cells, between the control and exposed group. However, I will still proceed with a 2 sample t-test since t-test is robust. I will then conduct a non-parametric test to check if this supports the t-test undertaken.

Table 7 below shows the 2 sample t-test conducted and the estimated difference between exposed and control is 1.703, with a p-value of 0.008 which is very significant. The 95% confident interval states that the exposed group is 0.474% to 2.932% higher than the control group.

Two-Sample T-Test and CI: Cu, Group

```
Two-sample T for Cu

Group     N  Mean  StDev  SE Mean
control  16  1.08   1.48     0.37
exposed  23  2.78   2.30     0.48


Difference = μ (control) - μ (exposed)
Estimate for difference:  -1.703
95% CI for difference:  (-2.932, -0.474)
T-Test of difference = 0 (vs ≠): T-Value = -2.81  P-Value = 0.008  DF = 36
```

Table 7: Two sample t-test of Cu by Group

Before conducting a Mann-Whitney test, it is necessary to make sure that both the distributions in the control and exposed group have the same shape. A dot plot of Cu (Figure 8 in the Appendix) indicates that both distributions are rather different, with numerous zero values. Removing the zeros to make the distribution more similar sounds like a good idea but this would leave us with a sample size too small. Table 8 shows the data with the 2 additional columns Cu_control and Cu_group that I have created from Cu, in order to conduct the non-parametric test. I then proceed to conduct the Mann-Whitney test (Table 9 in the Appendix). The results show that a 95.3% confidence interval for the difference in % C_u cells between the 2 group is 0.2% to 2.8%, with the exposed group having a higher value.

Both tests indicates that eating contaminated fish increases % C_u cells and % C_u cells is higher in the exposed group.

6 Hypothesis 4 – Greatest chromosome damage in subjects with the highest levels of mercury in their blood

A matrix plot for the 3 variables (Figure 9 in the Appendix) show that there is a slight positive linear relationship between abnormal vs log (mercury) and Cu vs log (mercury). This suggest that there are more chromosome damage in subjects with the highest levels of mercury. The correlation between log mercury and the variables were calculated using Minitab, as shown in Table 10 below, and the positive correlation also indicates that there is a positive relationship between chromosome damage and levels of mercury in blood. A confidence interval for correlation cannot be constructed because of the small sample size.

```
               log (mercury)       Abnormal
Abnormal               0.536
                       0.000

Cu                     0.410          0.845
                       0.010          0.000


Cell Contents: Pearson correlation
               P-Value
```

Table 10: Correlation between log (mercury), Abnormal and Cu

However, further regression analysis must be conducted to substantiate this. More specifically, I need to obtain an explanatory model where

1) Higher % abnormal cells is caused by higher mercury in blood
2) Higher % C_u cells is caused by higher mercury in blood

At this point, I would need to explore the use of different variables to obtain the best explanatory model and therefore I would need to use a best subset regression to choose the variables needed.

6.1 Investigating the reason for high % abnormal cells

It seems to me that group and unrelated conditions also have an effect on % abnormal cells. In order to check if the group and unrelated conditions have a significant effect on % abnormal cells, I have done a best subset regression using log mercury, group and unrelated conditions as a predictor. A new column named “Group_code” has been added to the data in Minitab. This is a dummy variable required to conduct the analysis. The value 0 and 1 belongs to the control and exposed group respectively, and the addition data can be found in Table 11 in the Appendix.

A best subset regression analyses (Table 12 in the Appendix*)* tells me that I would get the best exploratory model when I only use log mercury as my only variable. This is because its p value of 2 is approximately equal to its Cp value of 1.9, giving minimal bias due to omitted predictor variables. Even though including all 3 predictor variables can be ideal, as shown in the Cp vs p plot (Figure 10 in the Appendix), I chose log mercury only because a larger p increases the prediction error and that is not what we want. A 95% confidence interval regression analysis of Abnormal on log mercury gives the regression equation (Table 13 in the Appendix):

```
Abnormal = 1.14 + 3.713 log (mercury)
```

What this essentially means is that when log mercury increase by one, % abnormal cell increases by 3.713, and 28.7 percent (R-squared) of the variability in % abnormal cell is explained by this model. The F statistic in this regression test the null hypothesis that the coefficient of log mercury is 0. Here, the p-value is 0, indicating that the test is very significant and regression coefficient is not zero. A residual plot of Abnormal vs log (mercury) in Figure 11 in the Appendix shows evidence of heteroscedasticity. However, there is a point on the graph with a large standardised residual and upon further investigation, this belongs to individual 24. Individual 24 has one of the lowest mercury in blood compared to the rest in the exposed group. However, he has the highest % abnormal and C_u cells in his body, with 0 unrelated conditions. This anomaly has to be further investigated to find out why this is so because it goes against the hypothesis that I am testing. On the whole, this test suggest that high % abnormal cell increase when mercury in blood increase.

6.2 Investigating the reason for high % C_u cells

Similarly, I have a done a best subset regression using log mercury, group and unrelated conditions as my predictors. In this particular case, I have also incorporated Abnormal because % C_u cells is a subset of Abnormal and can be affected by this variable. Table 14 in the Appendix gives the best subset analysis that I conducted.

From this analysis, I would eliminate the first, third and fifth model because it does not incorporate log mercury and analysing those model would be meaningless to my hypothesis. I would then reject the second model since Cp value of 75.4 is too high. The next best model I would choose is the fourth model, which incorporates log mercury and abnormal because the Cp and p value is approximately equal and this keeps the prediction error low.
Table 15 below gives the regression analysis I conducted and the regression equation is :

Cu = -0.446 - 0.187 log mercury + 0.3955 Abnormal

Although the R-squared value of this model (71.71%) is high, this model does not make sense because when value of log mercury and abnormal is zero, Cu gives a value of negative 0.446; % C_u cells should not be negative if mercury log mercury and abnormal is 0! This suggest that perhaps Yule-Simpson's paradox[4] takes effect after Abnormal has been introduced. Looking at the residual plot of this regression line (Figure 12 in the Appendix), the standardised residual tends to be concentrated between 1 and 3, and are not heteroscedasticity.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.17878	71.71%	70.14%	65.88%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.446	0.490	-0.91	0.368	
log mercury	-0.187	0.328	-0.57	0.573	1.40
Abnormal	0.3955	0.0473	8.36	0.000	1.40

Regression Equation

Cu = -0.446 - 0.187 log mercury + 0.3955 Abnormal

Table 15: Regression analysis of Cu on log (mercury) and Abnormal

I want to then conduct a best subset regression with Abnormal removed. Upon further analysis (Table 16 in the Appendix), it is obvious that only log mercury should be selected because the Cp and p values are approximately the same and the prediction error would be small since only one variable will be selected, compared to the third and fifth model. This supports Yule-Simpson's paradox because Cu vs log mercury gives a large Cp value (75.4) during the previous best subset regression. A regression analysis using Cu on log mercury as shown in Table 17 in the Appendix gives the equation:

Cu = 0.003 + 1.282 log (mercury)

This model makes more sense because the % C_u cells is negligible (0.003%) when log mercury value is low. As log mercury increases by one, Cu would increase by 1.282% and they have a positive relationship. 16.82 percent (R-squared) of the variability in % C_u cell is explained by this model. The F statistic testing the null hypothesis that the coefficient of log mercury is 0 gives a p-value of 0.01, indicating that the test is very significant and regression coefficient is not zero. The residual plot as shown in Figure 13 in the Appendix suggest the existence of heteroscedasticity as a whole, with an exception of a residual which belongs to individual 24 as mentioned earlier.

Both explanatory model shows that chromosome damage (both % abnormal cells and % C_u cells) increase as mercury in blood increases.

[4] http://pareonline.net/getvn.asp?v=15&n=15

7 Hypothesis 5 – No difference between the level of unrelated conditions in the control group and the exposed group

In this hypothesis, I want to test the association between level of unrelated conditions and groups (control and exposed). If eating more contaminated fish affects chromosome damage only, then I should see no association between unrelated conditions and the 2 groups.

Since the data provided does not have 2 categorical variables for unrelated conditions and groups, I need to modify the data to make it into 3 columns as shown in Table 18 below, before I can do a cross tabulation using Minitab. Unrelated conditions are split into zero and more than zero (where individuals exhibits one or more unrelated conditions) under the variable "Health conditions".

Group	Health conditions	Frequency
control	Zero	11
control	more than zero	5
exposed	Zero	13
exposed	more than zero	10

Table 18: Modified data

Carrying out a cross tabulation and chi-square test on Minitab, I get the following information.

```
Rows: Health condition   Columns: Group

                  control  exposed  All

more than zero          5       10   15
                    6.154    8.846

zero                   11       13   24
                    9.846   14.154

All                    16       23   39

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 0.596, DF = 1, P-Value = 0.440
Likelihood Ratio Chi-Square = 0.603, DF = 1, P-Value = 0.438
Fisher's exact test: P-Value =  0.516593
```

Table 19: Cross tabulation between Group and Health conditions

All 3 chi-square test the null hypothesis H_0 = no association between Health conditions and Group against H_1 = there is an association between the 2 categorical variables. The p-value of all 3 test are >0.05 suggesting that I cannot reject the null hypothesis. This can be implied that no matter which group an individual belongs to, it does not indicate whether that individual will have zero or more than zero unrelated conditions. Therefore, I can conclude there is no difference in the level of unrelated conditions between the control and exposed group.

8 Conclusion

After conducting the 5 hypothesis, my conclusion is that eating contaminated fish causes chromosome damage, where the percentage of abnormal and Cu cells in human body increase.

This is supported by the statistical test that I have conducted. I have found out that the level of mercury in blood for the exposed group is higher than the control group, from the test done in my first hypothesis. Hypothesis 2 and 3 reveal that higher chromosome abnormalities and C_u cells were found in the exposed subjects, who had higher level of mercury in blood. The exploratory model that I have obtained in hypothesis 4 further supports my conclusion as I have found a positive linear relationship between mercury, and % abnormal cells as well as % C_u cells. Some might argue that higher chromosome damage in the exposed group is a result of other unrelated health conditions. I have therefore showed there is no difference between the level of unrelated conditions and the group an individual comes from.

Although all evidence suggest that eating contaminated fish cause chromosome damage, there are certain limitations in this project analysis. As this is an observational test, small samples were uses for the control and exposed group. This increases the probability of a Type II error, and increasing the probability of measurement bias. Moreover, information on the control group can be collected more easily because from my experience, more people do not consume fish regularly but only 16 individual's data were provided from the control group. Also, the data provided were from the past. Any attempt to verify the accuracy would not be possible unless the individual's blood has been stored. Even if their blood samples were made available, the verification process would need to take time as lab test have to be conducted.

I believe the exposed group who consumed more fish compared to the control tends to eat more seafood as well. We cannot rule out the possibility where the exposed group also ate more food like prawns or crabs contaminated with methylmercury. This would making the conclusion of eating contaminated "fish" on chromosome damage is less significant.

In hypothesis 3, both distributions in the control and exposed group does not have the same shape as required to perform a Mann-Witney test. This would have a significance influence on the results obtained in non-parametric test.

Ultimately, the data given does provide me with enough information to conclude that eating contaminated fish does cause chromosome damage. However, my conclusion can be made with more certainty if there were more data points through a larger sample, lowering the variance in the variables for a more accurate analysis.

9 Appendix:

Table 1: Original data with ID column added

ID	Group	Mercury	Abnormal	Cu	Unrelated
1	control	5.3	8.6	2.7	0
2	control	15	5	0.5	0
3	control	11	8.4	0	0
4	control	5.8	1	0	0
5	control	17	13	5	0
6	control	7	5	0	0
7	control	8.5	1	0	0
8	control	9.4	2.3	1.3	0
9	control	7.8	2	0	2
10	control	12	6.4	1.8	0
11	control	8.7	7	0	0
12	control	4	1.7	0	0
13	control	3	4	1	2
14	control	12.2	1.8	1.8	1
15	control	6.1	2.8	0	4
16	control	10.2	4.7	3.1	1
17	exposed	100	6.4	0.7	0
18	exposed	70	9.2	4.6	0
19	exposed	196	3.6	0	2
20	exposed	69	3.7	1.7	0
21	exposed	370	14.2	5.2	0
22	exposed	270	7	0	2
23	exposed	150	13.5	5	0
24	exposed	60	21.5	9.5	0
25	exposed	330	9	2	1
26	exposed	1100	11	3	9
27	exposed	40	8	1	0
28	exposed	100	9.2	3.5	0
29	exposed	70	8	2	1
30	exposed	150	14	5	0
31	exposed	200	11.9	5.5	0
32	exposed	304	10	2	2
33	exposed	236	6.6	3	0
34	exposed	178	13	4	0
35	exposed	41	0	0	2
36	exposed	120	6	2	0
37	exposed	330	13.1	2.2	6
38	exposed	62	0	0	1
39	exposed	12.8	5.3	2	1

Figure 1: Matrix plot of Mercury, Abnormal, Cu

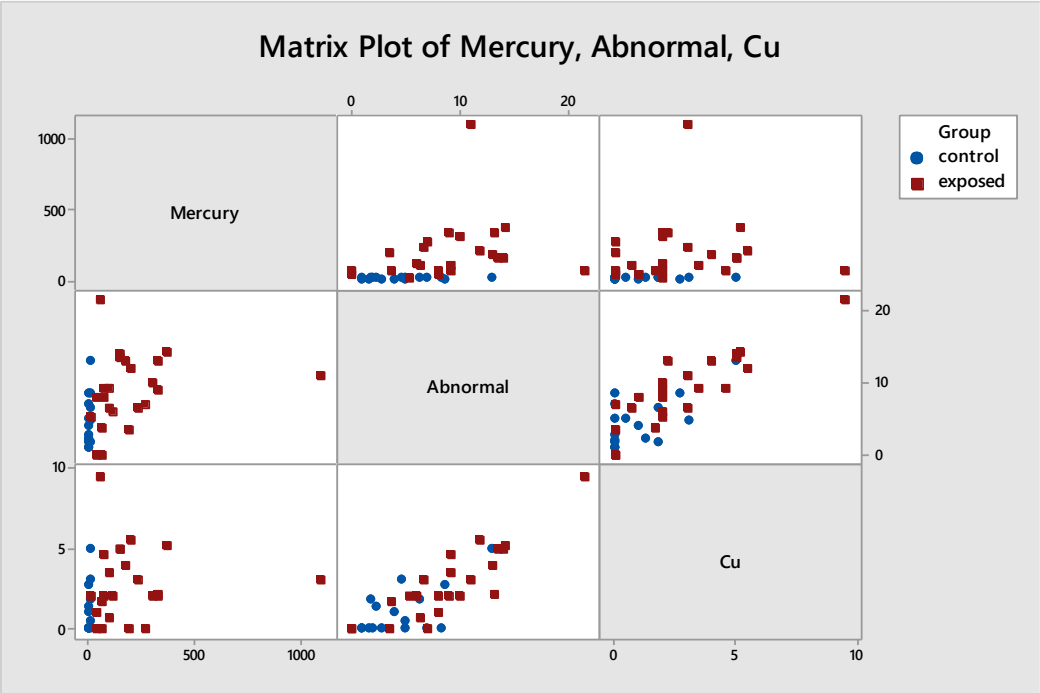

Table 4: Data with log (mercury) column added

ID	Group	Mercury	Abnormal	Cu	Unrelated	log (mercury)
1	control	5.3	8.6	2.7	0	0.72427587
2	control	15	5	0.5	0	1.176091259
3	control	11	8.4	0	0	1.041392685
4	control	5.8	1	0	0	0.763427994
5	control	17	13	5	0	1.230448921
6	control	7	5	0	0	0.84509804
7	control	8.5	1	0	0	0.929418926
8	control	9.4	2.3	1.3	0	0.973127854
9	control	7.8	2	0	2	0.892094603
10	control	12	6.4	1.8	0	1.079181246
11	control	8.7	7	0	0	0.939519253
12	control	4	1.7	0	0	0.602059991
13	control	3	4	1	2	0.477121255
14	control	12.2	1.8	1.8	1	1.086359831
15	control	6.1	2.8	0	4	0.785329835
16	control	10.2	4.7	3.1	1	1.008600172
17	exposed	100	6.4	0.7	0	2
18	exposed	70	9.2	4.6	0	1.84509804

19	exposed	196	3.6	0	2	2.292256071
20	exposed	69	3.7	1.7	0	1.838849091
21	exposed	370	14.2	5.2	0	2.568201724
22	exposed	270	7	0	2	2.431363764
23	exposed	150	13.5	5	0	2.176091259
24	exposed	60	21.5	9.5	0	1.77815125
25	exposed	330	9	2	1	2.51851394
26	exposed	1100	11	3	9	3.041392685
27	exposed	40	8	1	0	1.602059991
28	exposed	100	9.2	3.5	0	2
29	exposed	70	8	2	1	1.84509804
30	exposed	150	14	5	0	2.176091259
31	exposed	200	11.9	5.5	0	2.301029996
32	exposed	304	10	2	2	2.482873584
33	exposed	236	6.6	3	0	2.372912003
34	exposed	178	13	4	0	2.250420002
35	exposed	41	0	0	2	1.612783857
36	exposed	120	6	2	0	2.079181246
37	exposed	330	13.1	2.2	6	2.51851394
38	exposed	62	0	0	1	1.792391689
39	exposed	12.8	5.3	2	1	1.10720997

Figure 5: Box plot of log (mercury)

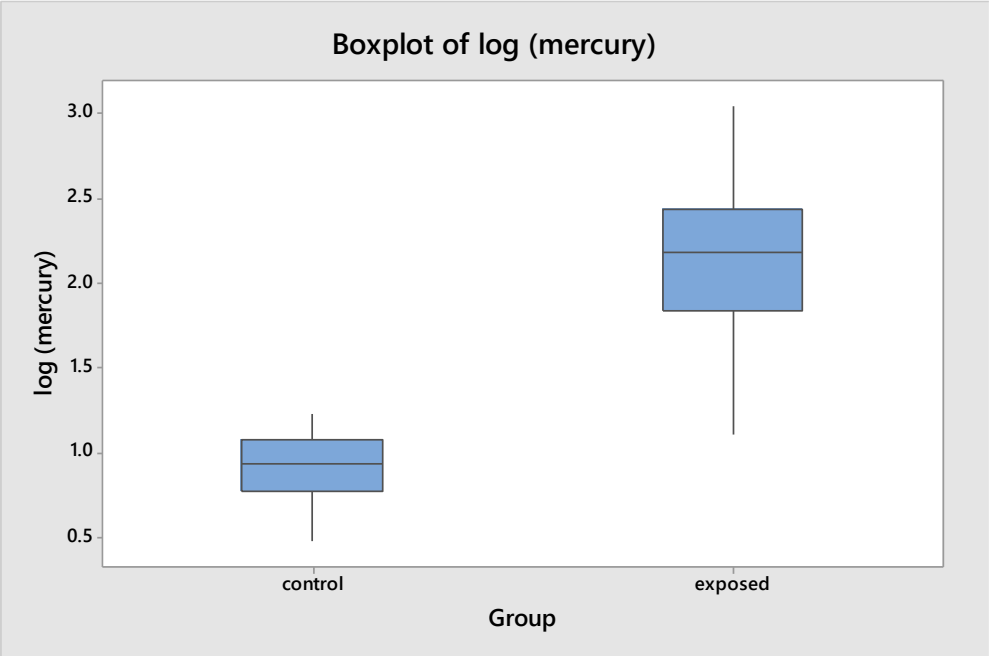

Figure 6: 2 variance test of log (mercury) vs Group

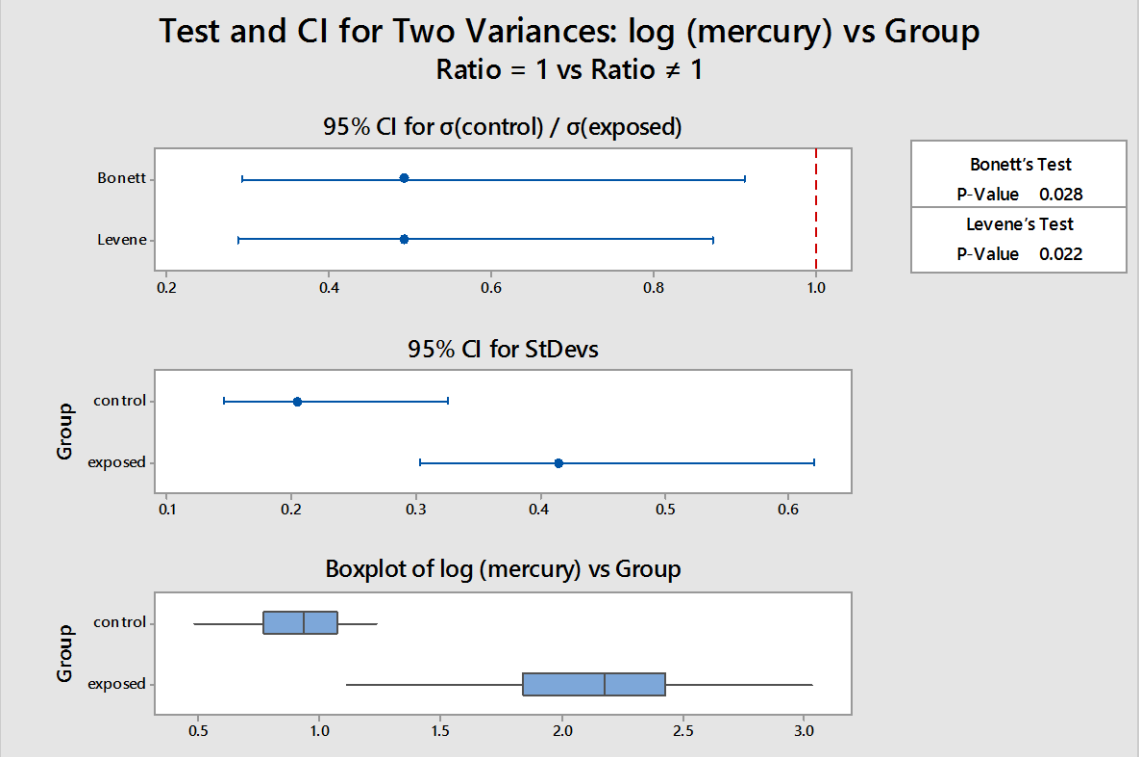

Table 5: Two sample t-test of log (mercuy) by Group

Two-Sample T-Test and CI: log (mercury), Group

```
Two-sample T for log mercury

Group      N   Mean  StDev  SE Mean
control   16  0.910  0.204    0.051
exposed   23  2.114  0.415    0.086


Difference = μ (control) - μ (exposed)
Estimate for difference:  -1.205
95% CI for difference:  (-1.409, -1.001)
T-Test of difference = 0 (vs ≠): T-Value = -12.00  P-Value = 0.000  DF = 33
```

Figure 7: 2 variance test of Abnormal vs Group

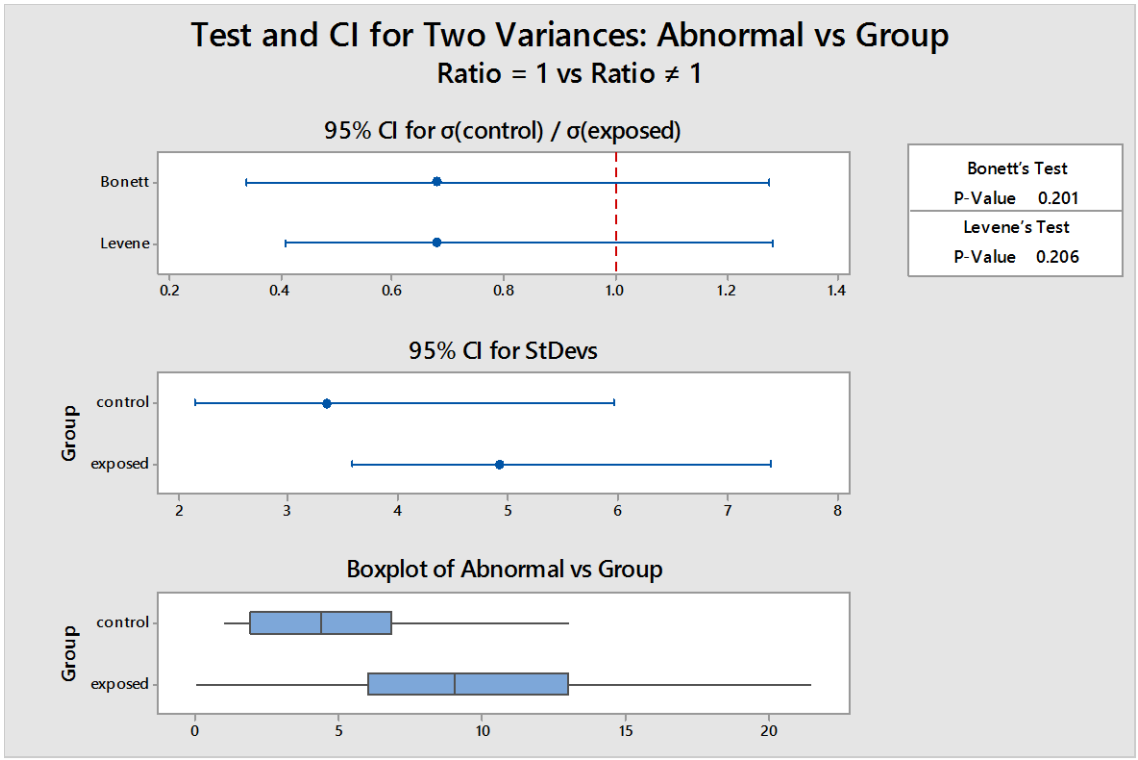

Figure 8: Dot plot of Cu

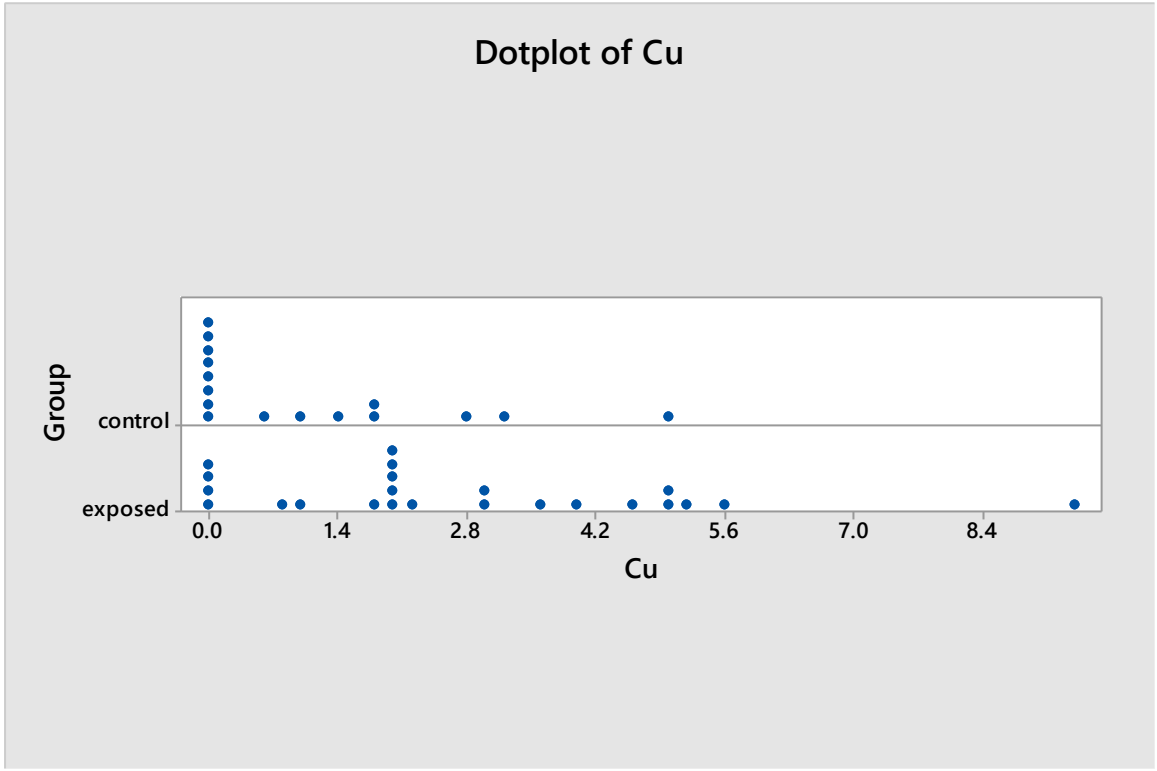

Table 8: Data with Cu_control and Cu_exposed added (arranged from Cu)

ID	Group	Mercury	Abnormal	Cu	Unrelated	log (mercury)	Cu_control	Cu_exposed
1	control	5.3	8.6	2.7	0	0.72427587	2.7	0.7
2	control	15	5	0.5	0	1.176091259	0.5	4.6
3	control	11	8.4	0	0	1.041392685	0	0
4	control	5.8	1	0	0	0.763427994	0	1.7
5	control	17	13	5	0	1.230448921	5	5.2
6	control	7	5	0	0	0.84509804	0	0
7	control	8.5	1	0	0	0.929418926	0	5
8	control	9.4	2.3	1.3	0	0.973127854	1.3	9.5
9	control	7.8	2	0	2	0.892094603	0	2
10	control	12	6.4	1.8	0	1.079181246	1.8	3
11	control	8.7	7	0	0	0.939519253	0	1
12	control	4	1.7	0	0	0.602059991	0	3.5
13	control	3	4	1	2	0.477121255	1	2
14	control	12.2	1.8	1.8	1	1.086359831	1.8	5
15	control	6.1	2.8	0	4	0.785329835	0	5.5
16	control	10.2	4.7	3.1	1	1.008600172	3.1	2
17	exposed	100	6.4	0.7	0	2		3
18	exposed	70	9.2	4.6	0	1.84509804		4
19	exposed	196	3.6	0	2	2.292256071		0
20	exposed	69	3.7	1.7	0	1.838849091		2
21	exposed	370	14.2	5.2	0	2.568201724		2.2
22	exposed	270	7	0	2	2.431363764		0
23	exposed	150	13.5	5	0	2.176091259		2
24	exposed	60	21.5	9.5	0	1.77815125		
25	exposed	330	9	2	1	2.51851394		
26	exposed	1100	11	3	9	3.041392685		
27	exposed	40	8	1	0	1.602059991		
28	exposed	100	9.2	3.5	0	2		
29	exposed	70	8	2	1	1.84509804		
30	exposed	150	14	5	0	2.176091259		
31	exposed	200	11.9	5.5	0	2.301029996		
32	exposed	304	10	2	2	2.482873584		
33	exposed	236	6.6	3	0	2.372912003		
34	exposed	178	13	4	0	2.250420002		
35	exposed	41	0	0	2	1.612783857		
36	exposed	120	6	2	0	2.079181246		
37	exposed	330	13.1	2.2	6	2.51851394		
38	exposed	62	0	0	1	1.792391689		
39	exposed	12.8	5.3	2	1	1.10720997		

Table 9: Mann-Witney test

Mann-Whitney Test and CI: control, exposed

```
          N  Median
control  16   0.250
exposed  23   2.000

Point estimate for η1 - η2 is -1.700
95.3 Percent CI for η1 - η2 is (-2.800,-0.200)
W = 228.5
Test of η1 = η2 vs η1 ≠ η2 is significant at 0.0094
The test is significant at 0.0083 (adjusted for ties)
```

Figure 9 : Matrix plot of log (mercury), Abnormal and Cu

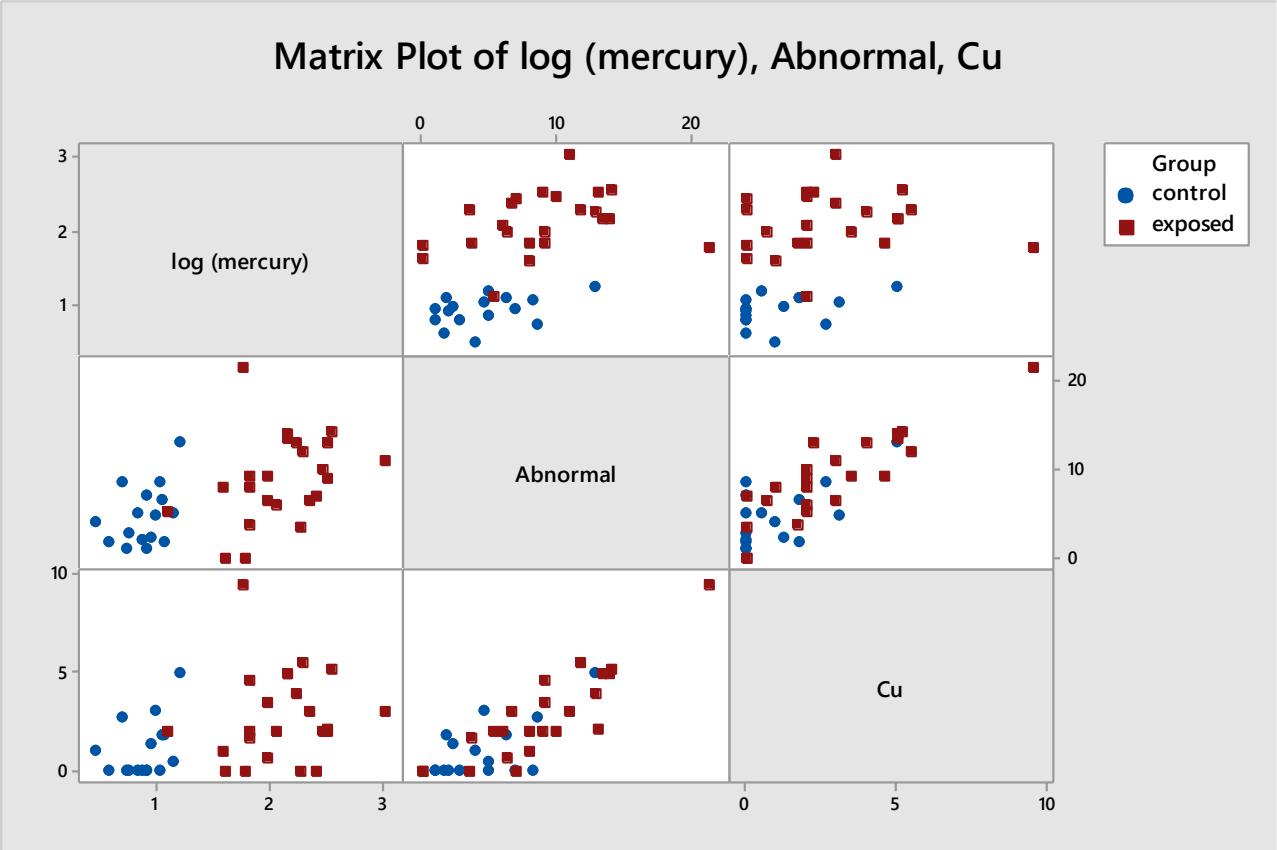

Table 11: Additional Column added

ID	Group_code
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
31	1
32	1
33	1
34	1
35	1
36	1
37	1
38	1
39	1

Table 12: Best subset regression using Abnormal as response

Best Subsets Regression: Abnormal versus log (mercury), Group_code, ...

Response is Abnormal

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	log (mercury)	Group_code	Untreated
1	28.7	26.8	22.8	1.9	4.0957	X		
1	19.2	17.0	10.8	6.8	4.3590		X	
2	31.3	27.5	22.0	2.6	4.0756	X		X
2	29.0	25.0	17.8	3.8	4.1432	X	X	
3	32.4	26.6	18.4	4.0	4.1003	X	X	X

Figure 10: Cp v p plot for Table 12

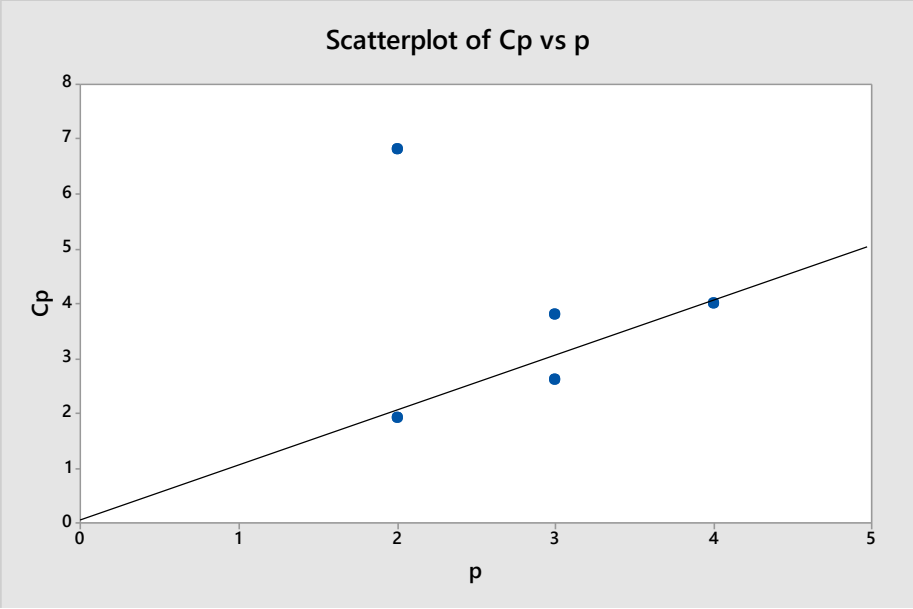

Table 13: Regression analysis of Abnormal vs log mercury

Regression Analysis: Abnormal versus log (mercury)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	249.56	249.564	14.88	0.000
log (mercury)	1	249.56	249.564	14.88	0.000
Error	37	620.67	16.775		
Lack-of-Fit	33	607.50	18.409	5.59	0.052
Pure Error	4	13.17	3.292		
Total	38	870.24			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.09572	28.68%	26.75%	22.81%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.14	1.69	0.67	0.506	
log (mercury)	3.713	0.963	3.86	0.000	1.00

Regression Equation

Abnormal = 1.14 + 3.713 log (mercury)

Fits and Diagnostics for Unusual Observations

Obs	Abnormal	Fit	Resid	Std Resid	
24	21.50	7.74	13.76	3.41	R

R Large residual

Figure 11: Residual plots for abnormal vs log (mercury)

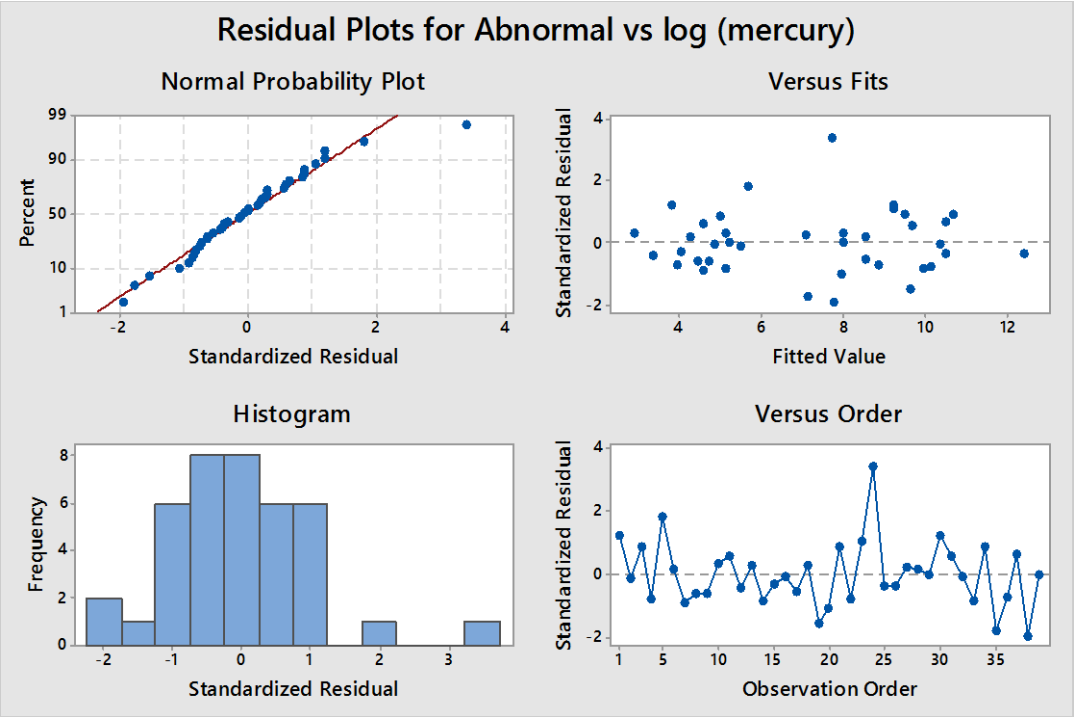

Table 14: **Best Subsets Regression: Cu versus log (mercury), Abnormal, ...**

Response is Cu

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	log (mercury)	Abnormal	Group_cat	Unrelated
1	71.5	70.7	67.5	2.9	1.1680		X		
1	16.8	14.6	9.6	75.4	1.9939	X			
2	73.6	72.2	67.7	2.0	1.1379		X		X
2	71.7	70.1	65.9	4.5	1.1788	X	X		
3	73.9	71.6	66.1	3.7	1.1488		X	X	X
3	73.6	71.4	65.8	4.0	1.1540	X	X		X
4	74.4	71.4	64.4	5.0	1.1542	X	X	X	X

Figure 12: Residual plot for Cu vs log (mercury) and Abnormal

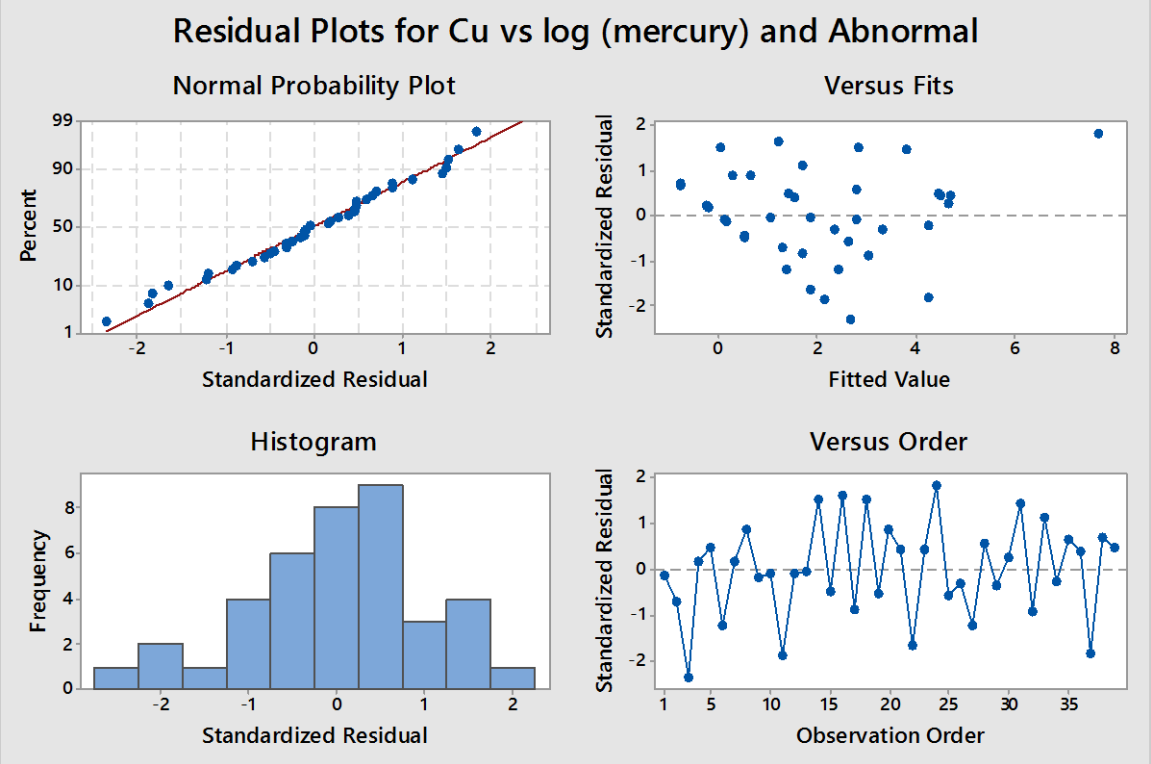

Table 16: Best subset regression using Cu as response, with Abnormal removed

Best Subsets Regression: Cu versus log (mercury), Group_code, Unrelated

Response is Cu

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	log (mercury)	Group_code	Unrelated
1	16.8	14.6	9.6	3.6	1.9939	X		
1	15.5	13.2	6.8	4.2	2.0099		X	
2	24.5	20.4	12.2	2.0	1.9252	X		X
2	19.2	14.7	1.3	4.5	1.9923		X	X
3	24.5	18.1	6.5	4.0	1.9525	X	X	X

Table 17: Regression analysis of Cu vs log (mercury)

Regression Analysis: Cu versus log (mercury)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	29.742	29.742	7.48	0.010
log (mercury)	1	29.742	29.742	7.48	0.010
Error	37	147.101	3.976		
Lack-of-Fit	33	139.781	4.236	2.31	0.216
Pure Error	4	7.320	1.830		
Total	38	176.844			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.99392	16.82%	14.57%	9.62%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.003	0.824	0.00	0.997	
log (mercury)	1.282	0.469	2.74	0.010	1.00

Regression Equation

Cu = 0.003 + 1.282 log (mercury)

Fits and Diagnostics for Unusual Observations

Obs	Cu	Fit	Resid	Std Resid	
24	9.500	2.282	7.218	3.67	R

R Large residual

Figure 13: Residual plot for Cu vs log (mercury)