Ryan Kladar s141094
Course: 02402 Intro Statistics
Project 2: Skive Fjord 2
4/11/2014

Part I: Analysis

a) The data collected by the Nat'l Maritime Monitoring Program available for this analysis is related to phytoplankton growth via phosphorus concentrations and chlorophyll concentrations. There is one monthly datum for each parameter (total phosphorus concentration in g/m^3 and chlorophyll concentration in g/m^3) for the years 1984 through 2003 (20 years inclusively). This results in a datum count of 720, or 240 points in 4 categories.

The median phosphorus concentration is 0.0704 g/m^3 while the mean is 0.09036 g/m^3. This difference results in a positive density skew, as shown in Figure 1. Other summary statistics are given in Figure 3.
The median chlorophyll concentration is 0.008670 g/m^3 while the mean is 0.012094 g/m^3. This difference results in a positive density skew, as shown in Figure 2. Other summary statistics are given in Figure 3.

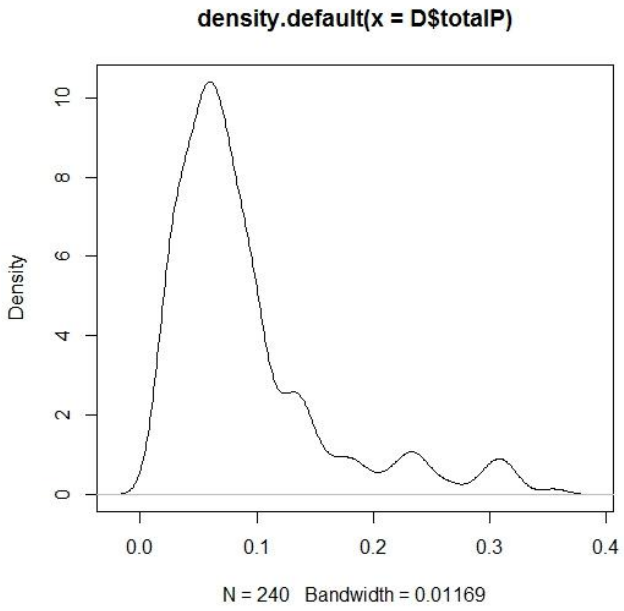

Figure 1

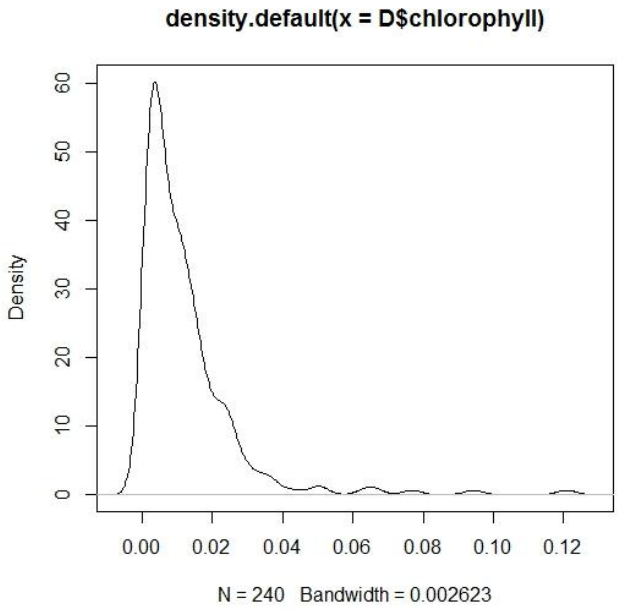

Figure 2

totalP	chlorophyll
Min. :0.01450	Min. :0.000500
1st Qu.:0.04830	1st Qu.:0.003525
Median :0.07040	Median :0.008670
Mean :0.09036	Mean :0.012094
3rd Qu.:0.10038	3rd Qu.:0.015213
Max. :0.35500	Max. :0.120920

Figure 3

b)

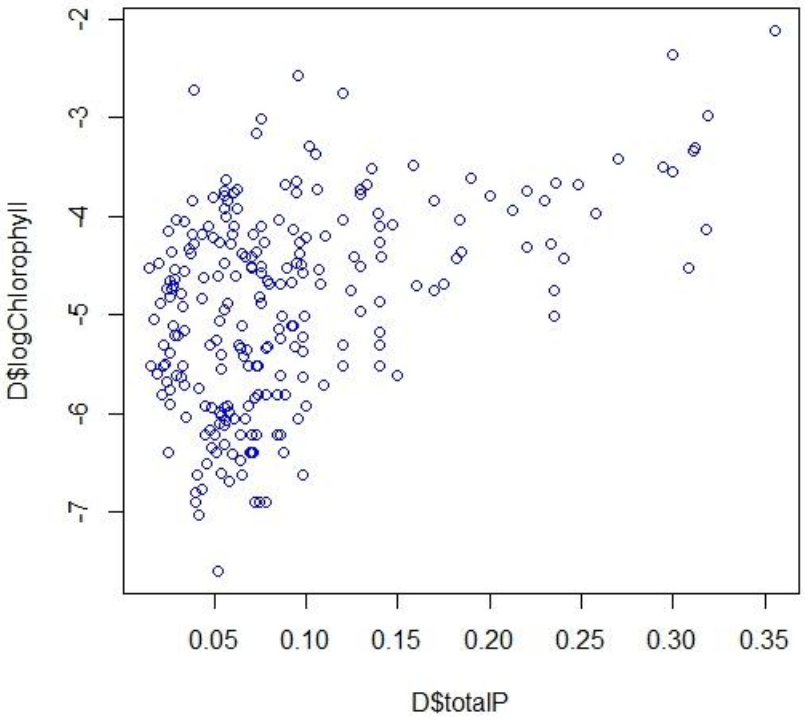

Figure 4: Shown is a scatter plot with the log transformation of the chlorophyll concentration vs. the concentration of phosphorus (both in g/m^3). It is clear that as the concentration of chlorophyll (or rather, the order of magnitude of chlorophyll concentration) increases, so does the concentration of total phosphorus in the fjord.

c) A linear regression model of two variables, in this case the log of the chlorophyll concentrations and the total phosphorus concentrations, requires a calculation of an intercept term (typically β_0), a single regression coefficient term (since we have one independent variable; typically β_1), and an error, or residuals term (because it's not perfect, it's just a model; typically ε_i). The best model will minimize error, or residuals, based on the least squares sum method. To determine the least square estimates, we also use the regression terms β_0 and β_1. Each can be calculated using Equation 1, Equation 2, Equation 3, and Equation 4 below. For a linear regression model to be valid, we must make several assumptions about the data and variables. First, the dependent variable Y has some sort of linear relationship to X (Equation 1), which we observe in Figure 4. Next, for each value X (x_i), the probability distribution of Y ($P(a<Y_{xi}<b)$) has the same standard deviation, σ_y. We can assume this in most cases, but can also check this by plotting the residuals and observing that they are random. Finally, for any given X, Y values are independent (again checked in the random residuals plot, if not assumed) and roughly normally distributed (in this case, we have a large sample size, so skewness, like that observed above in Figures 1 and 2, is alright).

$$\beta_0 = \bar{Y} - \beta_1 \bar{x} \qquad Eq\ 1$$

$$\beta_1 = r \times \left(\frac{s_y}{s_x}\right) \qquad Eq\ 2$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}} \qquad Eq\ 3$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad Eq\ 4$$

$$S_{xx} = s_x^2 \times (n-1) \qquad Eq\ 5$$

In this case, our independent variable, *x*, is total phosphorus concentration, and our dependent variable, *Y*, is the log of the chlorophyll concentration. *R* is the correlation (square root of the mean multiple R-squared). S_y and s_x are the standard deviations of Y and X, respectively.

d) We can estimate the coefficients in the model using Equation 1, 2 (or 3), 4 (or 5) and the known means and individual values of totalP and logChlorophyll.

$$\beta_1 = \frac{\sum_{i=1}^{240}(Y_i+4.902)(x_i-0.0904)}{S_{xx}} = 6.7506 \quad \text{or} \quad \beta_1 = \sqrt{0.2057} \times \left(\frac{1.01255}{0.06803}\right) = 6.7506$$

$$S_{xx} = \sum_{i=1}^{240}(x_i - 0.0904)^2 = \quad 1.1061 \quad \text{or} \quad S_{xx} = s_x^2 \times (n-1) = 1.1061$$

$$\beta_0 = \ -4.902 - \beta_1 \times 0.0904 = \ -5.5123$$

We can calculate the standard error of our coefficients with Equation 6 and Equation 7:

$$SE_{\beta 1} = \frac{s_{y/x}}{\sqrt{\sum(x_i - \bar{x})^2}} = 0.85985 \qquad Eq\ 6$$

$$SE_{\beta 0} = \sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}} \ = 0.09718 \quad Eq\ 7$$

We can also calculate our residual standard error using Equation 8:

$$SE_{residual} = \sqrt{\frac{\sum \widehat{\epsilon_i^2}}{n-2}} = 0.9043 \qquad Eq\ 8$$

For degrees of freedom, we must subtract two from our number of observations because we estimate both β_0 and β_1, limiting our freedom somewhat. Thus we get Equation 9:

$$DOF = n - 2 = 240 - 2 = 238 \qquad Eq\ 9$$

e) We can check our calculations using the built in simple linear regression function in R (see R code included in supplementary materials.) Summarized below, we have two estimated coefficients, each with their own standard deviation, an estimate of the standard deviation of the errors, the number of degrees of freedom used in the estimate of the standard deviation of errors, and the variance of the explanatory variable.

$\hat{\beta}_0 = \ -5.51215$
$\hat{\beta}_1 = 6.7506$
$Standard\ Error_{\beta 0} = 0.09718$
$Standard\ Error_{\beta 1} = 0.85985$
$Residual\ Standard\ Error = 0.9043$
$Degrees\ of\ Freedom = 238$
$Variance\ of\ Explanatory\ Variable\ (Var(totalP)) = \ 0.004628$
$RMSE = 0.2057$

f)

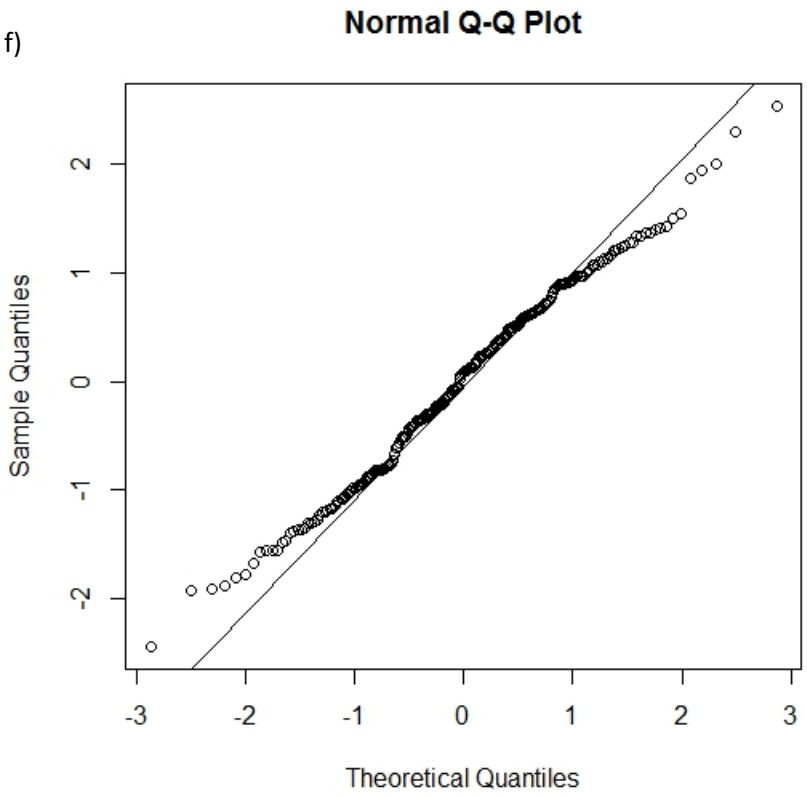

Figure 5. Shown to the left is the Q-Q plot of sample quantiles given with the data for the fjord vs. theoretical quantiles as calculated by our linear regression model. Included is the line of best fit for the plot, which shows good agreement along several deviations from the mean en either direction. The model performs more poorly as we get further from the mean, into the ends of the distribution, which is to be expected as point to point variation increases in these regions. The departure from normality challenges our assumption of normally distributed dependent variables, but verifies our observation of skew in the data. The large sample size is working against the skew of the data, decreasing the departure from normality. Had we not received as large a sample size, the departure from normality would be even more apparent and our model may need adjustment.

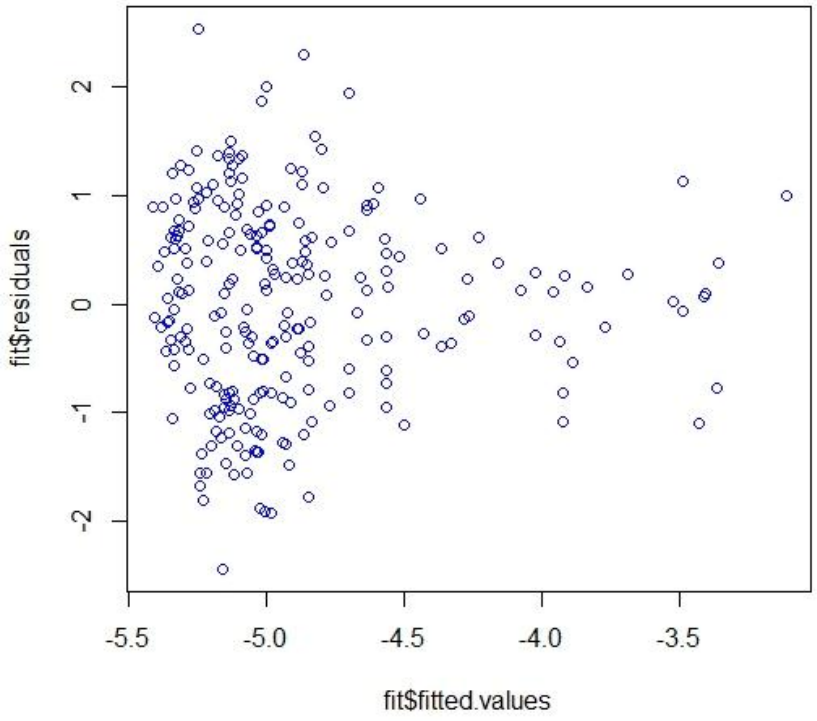

Figure 6. Shown to the left are the residuals vs. our fitted values. While there doesn't seem to be a strong pattern in the residuals as a function of some part of the fitted values, we can clean up the image a bit, manipulating the residuals just to be sure. We can see a sort of trumpet shape to the data, hinting at maybe some sort of relationship. See Figure 7.

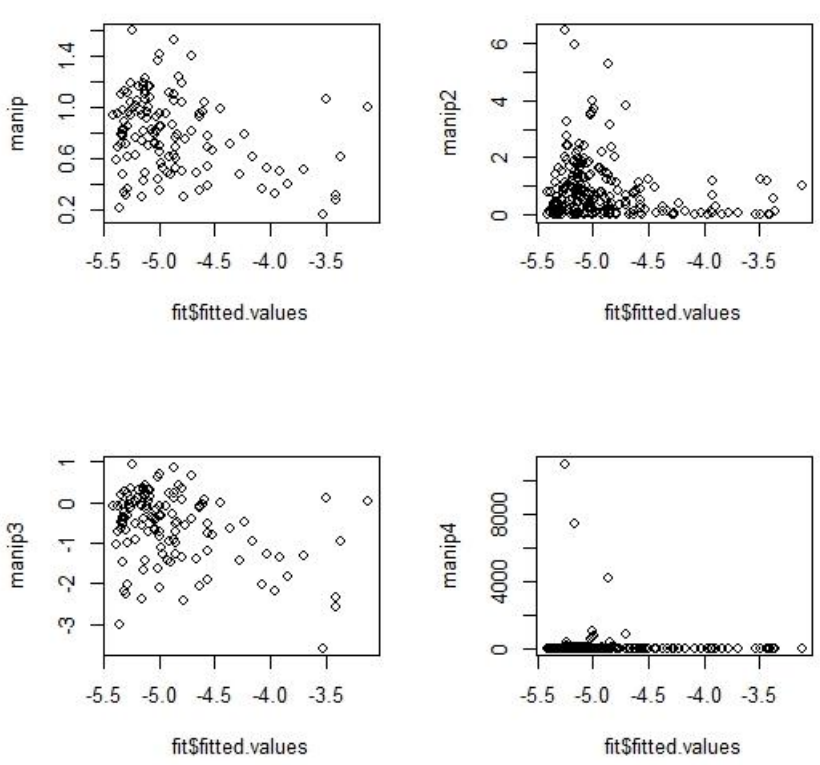

Figure 7. Shown is a series of manipulations done to the residuals and plotted against the fitted values, in an attempt to coax out any hidden dependencies not seen in the plot above. While there are an infinite number of transformations that could be performed on the residuals, here are basic ones often used to look at varying relationships. Clockwise from the top left, the manipulations to the residuals are: square root, squared, log, and antilog. We can see no discernible pattern in the residuals in any of them, so there are no hidden dependencies. Our earlier assumption that Y is a linear function of one variable, x, is correct.

.

g) For calculating confidence intervals of the coefficients in the model, the general approach for calculating confidence intervals is used, with slight modifications for simplicity (Equation 10):

$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \qquad Eq\ 10.1$$

$$\beta_1 \pm t_{\alpha/2} \times SE \times \frac{1}{\sqrt{S_{xx}}} \qquad Eq\ 10.2$$

$$\beta_1 \pm t_{0.975} \times 0.85985 \times \frac{1}{\sqrt{1.1061}} = [5.0567, 8.4444]$$

$$\beta_0 \pm t_{0.975} \times 0.09718 \times \frac{1}{\sqrt{1.1061}} = [-5.7036, -5.3207]$$

h) It seems that the independent variable has an impact on the dependent variable. To be sure, we must test it:

$$H_0 : Y_{mean} = \beta_0$$
$$H_1 : Y_{mean} \neq \beta_0$$

That is to say that if the mean of the dependent variable, Y, the log of Chlorophyll concentration, is not statistically significantly different than the intercept of the model (what Y would be equal to if our independent variable, X, the total phosphorus concentration in the fjord were equal to zero, or if it were not important, as is the case here), then the null

hypothesis is accepted. If there is a statistically significant difference between the dependent variable and the independent variable term, the null hypothesis can be rejected. For this we need a test statistic, t_{obs}, calculated with Equation 11:

$$t_{obs} = \frac{\bar{Y} - \beta_0}{s_y/\sqrt{n}} = \frac{-4.9021 + 5.51215}{1.0125/\sqrt{240}} = 9.3385 \qquad Eq\ 11$$

Using this t_{obs} and our confidence interval equations, we can calculate a 95% interval for the null hypothesis:

$$\bar{Y} = [-5.0309, -4.7733]$$
$$P_{value} = 2.2x10^{-16}$$

With at least 95% confidence, we can say that the true mean of the log of chlorophyll concentrations is not equal to the intercept value of -5.5125, β_0. We reject the null hypothesis and can state that phosphorus concentrations do play a significant part in chlorophyll concentrations in the fjord.

i) In December 2006, more measurements were taken, with $totalP_{dec2006}$ = 0.0703 g/m^3. Using our linear regression model, we can estimate the chlorophyll concentration at that point.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = -5.51215 + 6.7506(0.0703) = -5.0379$$

To find the chlorophyll concentration, we must undo the log transformation:

$$e^y = chlorophyll\ concentration \qquad \ln(chlorophyll\ concentration) = y$$
$$e^{-5.0379} = 0.006489$$

We can find a confidence interval for this new concentration, which must involve both the error from the fitted model and the error associated with future observations. This confidence interval is also referred to as the prediction interval, and takes into account the uncertainty of the system (Equation 12):

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{1-\alpha/2} \times \sigma \times \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}} \qquad Eq\ 12$$

$$-5.51215 + (6.7506 \times 0.0703) \pm t_{0.975} \times .9045 \times \sqrt{1 + \frac{1}{240} + \frac{(0.0703 - 0.0904)^2}{1.1061}}$$
$$= [-3.2521, -6.8231]$$

Finally, undoing the log transformation:

$$e^{-3.2521} = 0.038693$$
$$e^{-6.8231} = 0.00109$$