Statistics Project 2 - Water Environment in Skive Fjord

Axel Månson Lokrantz

2023-11-01

A) Statistical Analysis

The data contains observations for the five variables, year, total phosphor concentration in Skive fjord (g/m^3) totalP, temperature of the surface water ($°C$, at 0-1m depth) temp and chlorophyl concentration in Skive fjord (g/m^3) chlorophyl. totalP, temp and chlorophyl are all quantative variables. The observations for chlorophyl are log-transformed to make the data more suitable for analysis. There are 12 observations per year and the data stretches from 1984 to 2003.

	Obs	Mean	Std. Deviation	Median	25% Q	75% Q
totalP	240	0.09036	0.06802911	0.07040	0.04830	0.10038
temp	240	9.494	6.278361	8.290	3.765	15.380
log-chlorophyl	240	-4.902	1.012545	-4.748	-5.648	-4.186

Table 1: Summary Statistics for totalP, temp and log-chlorophyl.

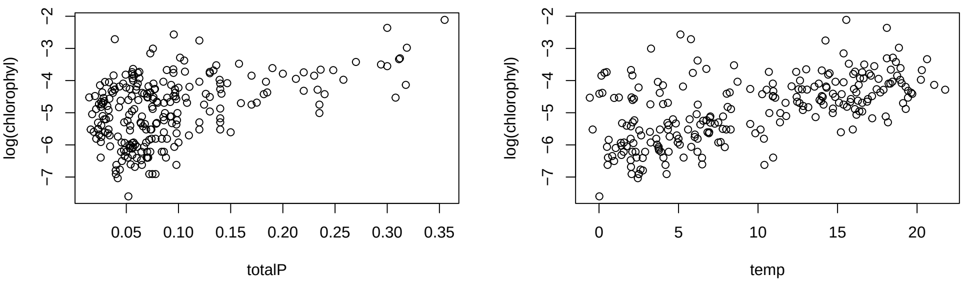

Figure 1: Scatterplots of log-chlorophyl against totalP and temp.

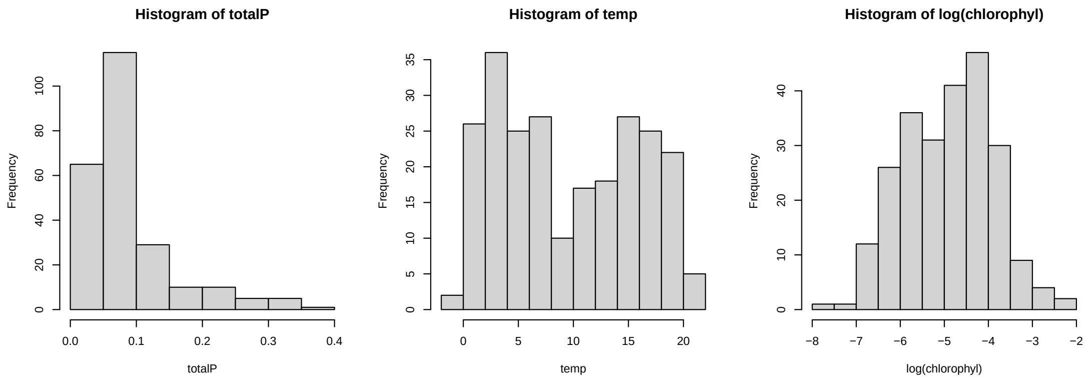

Figure 2: Histograms of all three variables.

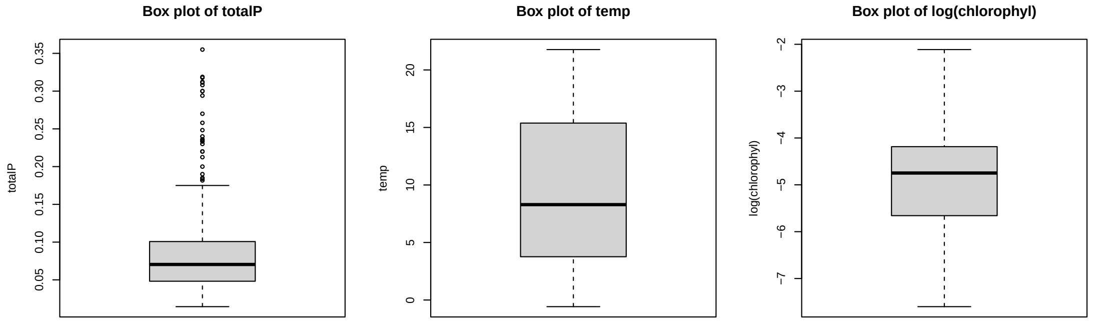

Figure 3: Boxplots of all three variables.

B) Multiple Linear Regression Model

The formula for multiple linear regression model used can be seen below.

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

The symbol εi represents the residual or error term for the ith observation. It is commonly used to account for variation or noise in the relationship between the dependent variable Yi and the independent variables $x1, i$ and $x2, i$. The symbol explains the difference between the actual observed value Yi and the value predicted by the regression model based on the coefficients $\beta 0$, $\beta 1$, and $\beta 2$. The model assumes that the residuals are independent and identically distributed following a normal distribution. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The depended variable Yi represents log-chlorophyll and the independent variables $x1$

and $x2$ are phosphorous concentration (totalP) and surface temperature (temp). These independent variables, $x1$ and $x2$, are utilized to predict the dependent variable Yi.

C) Parameter Estimation

	Estimate	**Std. Error**
$\hat{\beta_0}$	-5.84158	0.10173
$\hat{\beta_1}$	2.57796	0.97795
$\hat{\beta_2}$	0.07533	0.01050

Table 2: Interpretation of the estimates and standard deviations for the coefficents $\hat{\beta_0}$, $\hat{\beta_1}$, $\hat{\beta_2}$

$\hat{\beta_0}$ (Intercept): Represents the y-intercept of the regression line, when all prediction variables are zero. The estimated value is -5.84158 which means that if the independent variables totalP and temp were zero the predicted log-chlorophyl concentration would be -5.84158.

$\hat{\beta_1}$ totalP: Represents the coefficient for the totalP variable. For each unit increase in totalP, the predicted log-transformed chlorophyl concentration increase by 2.57796 units, assuming temp remains constant.

$\hat{\beta_2}$ temp: Represents the coefficient for the temp variable. For each unit increase in temp, the predicted log-transformed chlorophyl concentration increase by 0.07533 units, assuming totalP remains constant.

The standard error which can be seen in table 2 provides a measure of uncertainty in the estimates, For $\hat{\beta_0}$ the standard error is 0.10173 which means that if we were to take many samples from the population and estimate the model for each of them, the average distance between those estimates and the true intercept would be around 0.10173.

The degrees of freedom for a multiple linear regression model is calculated as below where n represents the number of observations and p the number of variables. Through this calculation we can conclude that the degrees of freedom is equal to 231.

$$df = n - (p + 1)$$

The explained variance is given through R when preforming a summary of our model. R^2 is 0.3414, which means about 34.14% of the variation in log-transformed chlorophyl concentration can be explained by totalP and temp.

D) Model Validation

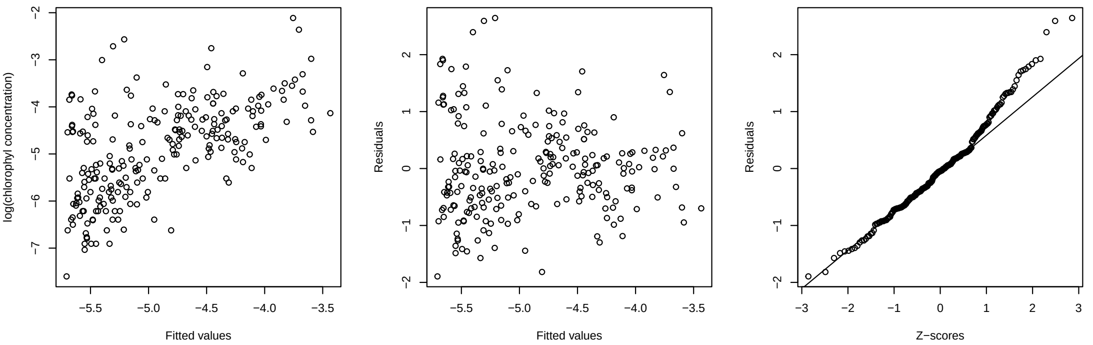

Figure 4:

To validate a multiple linear regression model we use a QQ plot of residuals. The residuals of the data has to follow the QQ line, which suggest that the plot is normally distributed and that the residuals meet the assumption of normality. Through observation of the QQ plot and the QQ line (figure 4) we can conclude that this is the case of our data.

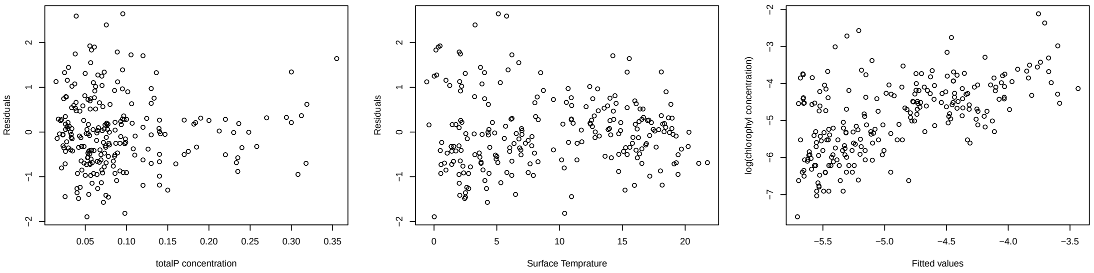

Figure 5:

The plots in figure 5 helps us check for linearity. If the residuals are evenly scattered with no clear patterns, it suggest that the spread of residuals is consistent across different values for the independent variables. If a clear pattern is visible in the plot it may indicate that the variance of errors is not constant. Upon observing the plots of the residuals it appears that totalP could potentially have a quadratic relationship with the dependent variable. Therefore, in an experiment, the term x^2 was added to the model. The addition resulted in a lower p-value from 0.00896 to 0.00245, however upon plotting the residuals the pattern became even more apparent, suggesting that the quadratic term might no the the best fit for our data and was therefor removed from the model.

The p-values for all three coefficients (Intercept, totalP, and temp) are less than 0.05, which suggests that these predictors are statistically significant at the 5% significance level. In other words, there is strong evidence to reject the null hypothesis that these coefficients are zero (i.e., they have no effect on log-chlorophyl).

E) Coefficient Confidence Interval

The 95% confidence interval for the coefficient of total phosphor concentration (totalP) in the linear regression model is given by:

$$\hat{\beta}_i \pm t_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\beta_i}$$

Where: $\hat{\beta}_{\mathrm{i}}$ is the estimate of the coefficient, $t_{1-\frac{\alpha}{2}}$ is the critical t-value for a 95% confidence interval with $n-(p+1)$ degrees of freedom, $\hat{\sigma}_{\beta_i}$ is the standard error of the coefficient estimate. The values for the coefficient of total phosphor concentration (totalP) in the linear regression model is given by R through the summary command where fit is the name of the model.

```
(summary(fit))
```

$$2.57796 \pm\ 1.970287 \times 0.8201$$

Which gives us the interval [4.193792369 ; 0.9621276313]. To verify the result the following command was run using R.

```
confint(fit, level = 0.95)
```

F) Test Statistic

To formulate a hypothesis with a significance level of 5% whether $\beta 1 = 5$ we formulate a null-hypothesis and an alternative hypothesis. The test statistic for a hypothesis test about regression coefficient follows a t-distribution.

$$H_{0,i} : \beta_i = 5$$
$$H_{1,i} : \beta_i \neq 5$$

Next, we compute the test statistic with the the formula for a level α t-tests for parameters.

$$t_{\text{obs},\beta_i} = \frac{\hat{\beta}_i - \beta_{0,i}}{\hat{\sigma}_{\beta_i}}$$

where: $\hat{\beta}_i = 2.57796$ (estimated slope) $\beta_{0,i} = 5$ (null hypothesis value) $\hat{\sigma}_{\beta_i} = 0.97795$ (standard error). When plugging in the values our t_{obs,β_i} becomes -2.476650135.

To compute the evidence against the null hypothesis we must first find the corresponding p_{value_i} which can be found using the formula below.

$$p_{\text{value}_i} = 2P\left(T > |t_{\text{obs},\beta_i}|\right)$$

This can be done using R with the following line of code.

```
p_value <- 2 * (1 - pt(abs(tobs), 231))
```

The resulting p_{value_i} is 0.01397964 which is smaller than our significance level of 0.05 therefor we reject the null-hypothesis $\beta_i = 5$. Furthermore, 5 is not part of the confidence interval which was calculated in assignment E which further strengthen our evidence against the null hypothesis.

G) Backward Selection

Backward selection is a method used in statistics for model selection, in particular in the context of regression. The goal of the method is to build a model that includes only the parameters which are statistically significantly related to the dependent variable y. First we need to determine if the parameters are significant on a 5% level.

This can be done using R with the following line of code.

```
confint(lm(y ~ totalP + temp))
```

	2.5%	97.5%
totalP	0.65111334	4.50480706
temp	0.05464581	0.09601646

Table 3: Confidence Intervals

The totalP parameter ranges from 0.65 to 4.50. The range does not include zero, so it is significant at the 5% level.

The temp parameter ranges from 0.05 to 0.09. This range does not include zero, so it is also significant at the 5% level.

Since both of the variables are significant neither of them should be removed through backward selection. The final model therfore becomes:

$$Y_i = -5.760993 + 2.57796x_{1,i} + 0.091237x_{2,i} + \varepsilon_i$$

H) Prediction Interval

Year	Month	-chlorophyl	Fit	Lwr	Upr	Differences
2003	7	-4.378035	-4.113287	-5.745301	-2.481273	0.264748
2003	8	-3.335129	-3.484984	-5.141718	-1.828249	0.149855
2003	9	-3.476676	-4.250814	-5.874621	-2.627007	0.774138
2003	10	-5.637995	-4.847923	-6.467270	-3.228577	0.790072
2003	11	-6.400938	-5.172466	-6.792513	-3.552418	1.228472
2003	12	-6.165818	-5.434359	-7.056511	-3.812207	0.731459

Table 4: Prediction interval for log-chlorophyl

In table 4 the differences represent the residuals of the model, i.e., the discrepancy between observed and predicted values. Generally, smaller residuals indicate a better fit of the model to the data. In the case of our model it perform relatively well for month 7, and 8 with smaller residuals. However, for moth 9, 10, 11 and 12 the residuals are larger which indicates that the model's predictions deviate more from the observed values.

The 'LWr' and 'Upr' columns represent the lower and upper bounds of the 95% prediction intervals, respectively. These intervals give and estimated range where we expect to see the true log-chlorophyl concentration 95% of the time. All six values fall within this interval, which is a good sign for the model.

The R-squared value measure how well the model explains the variation of log-chlorophyl. A higher value closer to 1 indicates a better fit, while a lower value closer to 0 indicate a poor fit. The R-squared for the model is calculated in R through the following command.

```
fit <- lm(logchlorophyl ~ totalP + temp, data = D_model)
summary(fit)
```

The R-squared value in this model is 0.34, which means that the models explains approximately 34% of the variation of log-chlorophyl around its mean. This could indicate that there is a lot of inherent variability in the data that is not captured by the independent variables and that the model has room for improvements.