Project 2: Water environment in Skive fjord II

Formalities, structure and expectations – second mandatory project

In this project, we'll formulate and select a suitable multiple linear regression model, which describes the concentration of phytoplankton in Skive fjord. The assignment must be solved using the statistical software R. Some code suggestions are provided but, in addition, it's a good idea to take a look at the R code from project 1, as well as chapter 5 and 6 of the book.

The results of your analysis must be documented in a report with tables, figures, appropriate mathematical notation, and explanatory text. Relevant figures and tables must be included within the text, not in an appendix. Present the results of your analysis as you would when explaining them to one of your peers. Divide the report into subsections, one for each of the questions to be answered.

The report must be handed in as a pdf file. R code should not be included in the report itself but must be handed in as an appendix (a .R-file). The report and appendix must be handed in under Assignments on Learn at: `Projekt 2: Vandmiljø i Skive fjord II`

The report should not exceed 6 pages (excluding figures, tables, and the appendix). A page contains 2400 characters.

It's important that you describe and explain the R output in words – figures and tables cannot stand alone.

When you're asked to state a formula, insert numbers, and then perform certain computations, it's important to show that you've done this by including your intermediate results. (In these cases, it's not enough to report results obtained directly from R). Furthermore, remember that when performing a hypothesis test, you must go through the following steps: State the hypothesis and significance level (α), compute the test statistic and state it's distribution, compute the p-value, and summarize your findings.

Figures and tables are not included in the assessment of the length of the report. However, it's not in itself an advantage to include many figures, if they aren't relevant!

You may work in groups, but the report must be written individually. Questions may be addressed to the teaching assistants, see the guidelines on the *Projects* page of the course website.

Data

Read the dataset `skivefjord2_data.csv` into R. The following code may be used:

```
# Read the dataset 'skivefjord2_data.csv' into R
D <- read.table("skivefjord2_data.csv", header = TRUE, sep = ";")
```

Skive fjord is part of the Limfjord system, which separates the northern part of Jutland from the rest of Jutland. The fjord is monitored intensively via the national maritime monitoring program. This provides a good data source for analysing the effects of initiatives intended for improving the water environment, as well as investigating biological relations in the fjord.

The dataset for this project contains observations of five variables:

- `year`: Year of observation
- `month`: Month of observation
- `totalP`: Total phosphor concentration in Skive fjord (g/m^3)
- `chlorophyl`: Chlorophyl concentration in Skive fjord (g/m^3)
- `temp`: Temperature of the surface water (°C, at 0-1 m depth)

Before you proceed, add the following variable with log-transformed chlorophyl concentrations to the dataset:

```
# Add log-chlorophyl to the dataset
D$logchlorophyl <- log(D$chlorophyl)
```

Phytoplankton is a very visible indicator of the state of the water environment, as the water becomes green and unappealing if the concentration of phytoplankton is high. In this project, we will focus on phytoplankton as an indicator of the water environment. The concentration of phytoplankton is determined using the concentration of chlorophyl (green pigment in plankton). Phytoplankton needs nutrients, such as nitrate and phosphor, as well as light in order to grow. Furthermore, the water temperature also influences its growth.

Statistical analysis

a) Present a short descriptive analysis and summary of the data for the variables `totalP`, `logchlorophyl`, and `temp`. Include scatter plots of the log-transformed chlorophyl concentration against the two other variables, as well as histograms and box plots of all three variables. Present a table containing summary statistics, which includes the number of observations, and the sample mean, standard deviation, median, and 0.25 and 0.75 quantiles for each variable. Furthermore, state which time period the observations cover.

In the following, the statistical model should only be fitted to the first 234 observations in the dataset. Later on, we'll use the six remaining observations (July - December 2003) to evaluate the prediction capabilities of the final model. For example, the following code may be used to split the dataset into two parts, one for estimating the model (`D_model`), and the other for validating prediction accuracy (`D_test`):

```
# Subset containing the first 234 observations (for model estimation)
D_model <- D[1:234, ]

# Subset containing the last 6 observations (for validation)
D_test <- D[235:240,]
```

b) Formulate a multiple linear regression model with the log-transformed chlorophyl concentration as the dependent/outcome variable (Y_i), and the total phosphor concentration and surface temperature as the independent/explanatory variables ($x_{1,i}$ and $x_{2,i}$, respectively). Remember to state the model assumptions. (See Equation (6-1) and Example 6.1).

c) Estimate the parameters of the model. These consist of the regression coefficients, which we denote by β_0, β_1, β_2, and the variance of the residuals, σ^2. You may use the following R code:

```
# Estimate multiple linear regression model
fit <- lm(logchlorophyl ~ totalP + temp, data = D_model)

# Show parameter estimates etc.
summary(fit)
```

Give an interpretation of the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, explaining what they tell us about the relation between the log-transformed chlorophyl concentration and the model's explanatory variables. (See Remark 6.14). Furthermore, present the estimated standard deviations of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, the degrees of freedom used for the estimated residual variance $\hat{\sigma}^2$, and the explained variation, R^2.

d) Perform model validation with the purpose of assessing whether the model assumptions hold. Use the plots, which can be made using the R code below, as a starting point for your assessment. (See section 6.4 on residual analysis).

```
# Plots for model validation

# Observations against fitted values
plot(fit$fitted.values, D_model$logchlorophyl,
     xlab = "Fitted values", ylab = "log(chlorophyl concentration)")

# Residuals against each of the explanatory variables
plot(D_model$EXPLANATORY_VARIABLE, fit$residuals,
        xlab = "INSERT TEXT", ylab = "Residuals")

# Residuals against fitted values
plot(fit$fitted.values, fit$residuals, xlab = "Fitted values",
     ylab = "Residuals")

# Normal QQ-plot of the residuals
qqnorm(fit$residuals, ylab = "Residuals", xlab = "Z-scores",
       main = "")
qqline(fit$residuals)
```

e) State the formula for a 95% confidence interval for the coefficient of the total phosphor concentration, here denoted by β_1. (See Method 6.5). Insert numbers into the formula, and compute the confidence interval. Use the R code below to check your result, and to determine confidence intervals for the two other regression coefficients.

```
# Confidence intervals for the model coefficients
confint(fit, level = 0.95)
```

f) It is of interest whether β_1 might be 5. Formulate the corresponding hypothesis. Use the significance level $\alpha = 0.05$. State the formula for the relevant test statistic (see Method 6.4), insert numbers, and compute the test statistic. State the distribution of the test statistic (including the degrees of freedom), compute the p-value, and write a conclusion.

g) Use backward selection to investigate whether the model can be reduced. (See Example 6.13). Remember to estimate the model again, if it can be reduced. State the final model, including estimates of its parameters.

h) Use your final model from the previous question as a starting point. Determine predictions and 95% prediction intervals for the log-transformed chlorophyl concentration, for each of the six observations in the validation set (`D_test`). See Example 6.8, Method 6.9 and the R code below. Compare the predictions to the observed log-chlorophyl concentrations for the six observations in the validation set and make an assessment of the prediction capabilities of the final model.

```
# Predictions and 95% prediction intervals
pred <- predict(FINAL_MODEL, newdata = D_test,
                interval = "prediction", level = 0.95)

# Observed values and predictions
cbind(year = D_test$year, month = D_test$month,
      logchlorophyl = D_test$logchlorophyl, pred)
```

Hence, don't write the formulas in the report, but instead refer to that the R function `predict` was used for the calculations. The formulas requires a matrix formulation, which are out of the curriculum (to derive the formulas use Equations (6-48) and (6-49) together with the derivations leading to Equations (5-57) and (5-58)).