Computer lab, non-linear multiple regression
Course: Statistics using R

March 3, 2017

This is the lab for the morning of Friday March 10. You will not need to hand in anything at the end of the class.

1. We will continue to work with the “Finnish fish catch data” we looked at in an earlier lab. Locate the dataset, called “fishcatch.dat”, via the link on the course home page and read it into R, calling the resulting object “fish”.

2. We will look at how well we can predict fish weight from fish length. First we will consider the Perch species; take out a subset of “fish” consisting of the observations for Perch, and call it “perch”. Attach “perch”, and start by doing a simple scatter plot of Length1 versus Height. Now proceed to do a polynomial regression of Weight on Length1. Hint: you will use formulas like

```
Weight ~ Length1 + I(Length1^2)
```

Determine how many powers of Length1 you need to get a good fit (i.e., if you need a second degree polynomial, a third degree polynomial, a fourth degree polynomial, etc.): Look at ANOVA results or the R-squared output from regression to make your decision. Hint, your final model should include the linear term and one higher order term and restrict your search to a fourth order polynomial. Why would it be a bad idea to include multiple higher order terms?

3. Can you get a better fit by using Length2, or Length3, instead of Length1? Maybe you get a better prediction by looking at a combination of the length measurements? Try this out, and make a decision for the model you believe is best for prediction. Considering the descriptions of the three length measurements is this a good modeling approach?

4. Decide upon one model from question 2. If you feel uncertain ask the teachers which models could be useful. Make a plot showing the Length and the Weight, together with the polynomial regression line and the
prediction lines, using commands as those in section 12.1 of Dalgaard. An assumption of the linear model we are using is that the variances of the differences between the model line and the actual observations are the same for all observations. Does the plot give you reasons to doubt if this assumption is reasonable for these data?

5. R has several built in diagnostic plots to examine regression models. If you called your chosen model “model1”, investigate the diagnostics of the model with

```r
plot(model1, which=1:2)
```

Interpret the plots.

6. You may by how have observed that there is more variability, compared to the prediction, of large fish compared to smaller fish, and that this is quite natural, concerning the biology. It is also quite natural that weight, which is connected to the volume of the fish, has some some kind of power-law relationship with the length: \( W = aL^b \cdot c \) where \( a \) and \( b \) are constants and \( c \) has a random variation on a log scale. In short, it may be a good idea to log-transform both predictor and response, and try a linear regression on the result (as the formula above corresponds to \( \log(W) = \log(a) + b \cdot \log(L) + \log(c) \)). Compute log-transformations of your Weight and Length variables, and add them to the data frame perch (remember to do first detach and afterwards attach for the data frame). Then, re-do the analysis above on the log-transformed variables. Investigate the R squared results and compare them with the earlier models to compare the fits of the models. Also, re-do the diagnostics plots of the previous question, to compare the new model with the old. Finally, make a plot similar to that you made for original model above. (Hint: Even if the linear modeling is done on log scales, the plot should be on the original scale. To achieve this, you need to compute the predictions using log scale, and then convert to the non-log scale by applying the exponential function to the values before plotting).

7. OPTIONAL: Go back to the original fish dataset and try out similar models for the other species, using log-transformed data. Can you use the same regression coefficients, or do you need to have different regression coefficients for the different species, coded as an interaction between the log-transformed length values and the Species variable?