Homework 3

Due: Friday March 6 at 2pm
See general homework tips and submit your files via the course website.

Code for creating the **anovafish** data set is in **HW3Data.sas** in the course website. The data set is built upon the **sashelp.fish** data set included in SAS and described here online (and in the SAS help files). The data set contains the original variables and a categorical **widthgroup** variable for a subset of the **sashelp.fish** data. **Species** and **widthgroup** are classification variables. **Widthgroup** indicates whether the fish is wider or thinner than the average width of fish in the original data set.

Exercise 1

a) For **weight**, create a cross-tabulation of the mean, standard deviation and counts by **species** and **widthgroup**. Comment on any interesting features (e.g. apparent differences between species and/or width groups).

b) Obtain your best ANOVA model for **weight** as a function of **species** and **widthgroup** and possibly the interaction between them. Show and explain your selection process for the model, and be sure to use the correct SAS procedure. Comment on significance of the model and the individual terms in the model.

c) For the model chosen in part **b**, comment on variation explained by the model, and any significant group differences (main effects and interactions if the interaction term is still in your model). What does this tell us about **weight** differences between species and/or wider or thinner fish?

Exercise 2

We might expect that fish would have a certain amount of symmetry. Consider modeling the **weight** as a function of the length measurement **length1**.

a) Fit a linear regression model for **weight** as a function of **length1** ignoring species. If any points are unduly influential, note those points, then remove them and refit the model.

b) Comment on the quality of the final model, significance of the parameters, and any remaining issues noted in the diagnostics. What does this model tell us about the relationship between weight and length for the types of fish included in this data?

Exercise 3

Now consider modeling the other possible continuous predictors for weight as well.

a) Perform and explain model selection for a linear regression model of **weight** as a function of the other continuous variables in the data set. Consider the additional contribution to the model R^2 when determining whether to keep significant terms with high variance inflation. If any points are unduly influential, note those points, then remove them and refit the model.

b) Comment on the quality of the final model, significance of the parameters, and any remaining issues noted in the diagnostics. What does this model tell us about the relationship between weight and the dimensions for the types of fish included in this data?

Exercise 4

It might be expected that volume (product of length, width, and height) might be more related to weight than a linear combination of the 3 dimensions. A log-linear model with log(**weight**) as the response, and logs of length, width and height measurements would be consistent with such a model.

Add variables for these log measurements to the data set and repeat exercise 3 using the log of weight as the response and the other log variables as possible predictors. Also comment on whether this model or the one obtained in exercise 3 is a better model