Linear Regressions

Here you will practice calculating simple least-squares regressions and their assumptions on a few data sets. As always, retain and annotate scripts for your future use.

Caterpillar-tannins experiment: In the ongoing war between insects and plants, tannins produced by plants interfere with insect nutrient absorption. Some plants (e.g., oaks) have high tannin content, whereas other plant species do not. But one cannot simply feed different plant species to caterpillars to test for tannin effects – other factors matter, too (leaf toughness, etc.). So here tannins were added to a standard food to answer this question: How much does tannin concentration (mg/g dry weight) depress caterpillar growth (mm length) during the 2-week experiment? Each entry represents an experimental unit containing ad lib food and 30 caterpillars, for which mean growth and tannin level were recorded.

1. Get the tancat.txt from the course web site and attach it.
2. Compute a regression of growth as a function of tannin:

```
model <- lm(growth~tannin)
summary(model)
plot(growth~tannin)  # plots the data and regression model
abline(model)
```

Is there a significant relationship? How strong is the relationship?

3. Let's test assumptions required for this model:
   a) Normality of residuals:

```
plot(model$residuals ~ tancat$tannin) # plots residuals
abline(h = 0, lty = 3)

hist(model$residuals) # makes a histogram of residuals

qqnorm(model$residuals) # makes a qq normality plot of residuals
qqline(model$residuals)

shapiro.test(model$residuals) # runs a Shapiro-Wilk test on residuals
```

   b) Homogeneity of residuals:
      Squint at the first plot (residuals ~ tannin) – are residuals about even left-right?
      Squint at the third plot (QQ plot) – are + & - distances from the line balanced left-right?

Now you will edit the commands above for a new data set.

Mercury in Fish of FL Lakes: Fish are regularly sampled for mercury (Hg) contamination. This data set represents average Hg load in fish (mg/kg) in fish from 53 Florida lakes. Also reported are basic water quality measures – alkalinity (alk), pH, calcium concentration (Ca), and chlorophyll *a* (Chla – a measure of algal concentration in water). Each of these factors may

predict Hg load in fish, related to chemical complexing of Hg and food web uptake. Your mission: find the best single predictor of Hg load in fish based on these four potential predictors.

4. Get the `FLHg.txt` file from the course web site. Attach it.
5. Try this panel of graphs with smoothed model curves to visualize the data set:

```
pairs(FLHg[2:6], panel=panel.smooth) #[2:6] picks the numeric columns
```

6. Repeat the basic commands used above for tannins, but here first evaluate `alk` as a predictor of Hg. Name your model to suit the predictor (e.g., alkmodel).

Does this regression comply with assumptions of normal error variance?
Try a square-root transform of Hg and then repeat the regression and assumptions tests. Better?

7. Now repeat and rinse for the other three predictor variables. Keep renaming your models to suit.

Which ONE predictor variable would you recommend be used to predict Hg load in fish of Florida lakes, *based on these model results*?

Now use model selection (based on AICc) to compare regression models:

7. Install the `bbmle` package (if it is not already in packages), and then load it.
8. Enter this command, where model names are assumed here (edit as needed):

```
AICctab(alkmodel, pHmodel, Camodel, Chlamodel, sort=TRUE, base=TRUE,
delta=TRUE, weights=TRUE)
```

The resulting table shows you a model-selection approach to answer the question (Which ONE...) above. Bottom line = the model with the greatest weight is the most plausible model among those listed. How much better? Delta AICc (`dAICc`) shows you how much better a model is compared to the next one, and weight tells you the probability that a model is most plausible, given the list of models.

How does this approach compare to your evaluation of alternative models based on regression outputs?