# Lecture 27

## General Linear Models

General Linear Models (GLMs) is a fancy name for all of the models we have considered so far in this course (and a few we will consider in the next two weeks). Minitab makes working with GLMs relatively easy because it provides them a special dialog box for their analysis. The aspects of preforming a GLM analysis and the difference in the output between running GLM, ANOVA, and Regression will be discussed in this lecture. Also discussed will be the term ANCOVA (Analysis of Covariance).

The following are sketchy notes.

### General Linear Models

A general linear model is any model of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is the response, $\mathbf{X}$ is the matrix of predictors including a constant column for the intercept, $\boldsymbol{\beta}$ lists the intercept and slopes, and $\boldsymbol{\epsilon}$ are the residuals. You might recognize this as the regression model! It's just that some of the predictors might be dichotomous, some might be transformed, the response might be transformed, etc...

We have seen that a one-way ANOVA is a General Linear model. You just include indicator (dichotomous) variables for each group. Continuous linear regression plus groups is also called ANCOVA. The question for ANCOVA is whether there is a difference between groups after controlling for a continuous confounding variable(s). This is precisely what we did for the example of bird heart rates from the last lecture. It is also what we can do in the fish weight data when we have multiple species.

We will run the analysis of fish (log) weight against species and (log) length/height/width in class. We will use the Regression menu and the General Linear Model menu under Stat > ANOVA. We will discuss the output which tells the identical story but differently. The output from the Regression menu, which requires that we create the indicator variables, is as follows:

```
Regression Analysis: log_wt versus species_Abra, species_Esox, ...

The regression equation is
log_wt = - 2.75 - 0.133 species_Abramis_brama(bream)
         - 0.057 species_Esox_lucius(pike) - 0.0255 species_Leuciscus_rutilus(ro
         + 0.109 species_Leusiscus_idus(white
```

```
        - 0.174 species_Osmerus_eperlanus(sm
        + 0.0838 species_Perca_fluviatilis(pe + 1.80 log_len + 0.646 log_ht
        + 0.558 log_wdth


Predictor                         Coef  SE Coef       T      P      VIF
Constant                       -2.7452   0.2474  -11.10  0.000
species_Abramis_brama(bream)  -0.13274  0.03297   -4.03  0.000    4.425
species_Esox_lucius(pike)      -0.0569   0.1274   -0.45  0.656   37.668
species_Leuciscus_rutilus(ro  -0.02554  0.06381   -0.40  0.690   10.403
species_Leusiscus_idus(white   0.10948  0.06602    1.66  0.099    3.849
species_Osmerus_eperlanus(sm   -0.1737   0.1045   -1.66  0.099   21.320
species_Perca_fluviatilis(pe   0.08381  0.06919    1.21  0.228   26.100
log_len                         1.8030   0.1451   12.43  0.000   79.013
log_ht                          0.6460   0.1368    4.72  0.000  137.708
log_wdth                        0.5579   0.1058    5.27  0.000   57.093


S = 0.0808187   R-Sq = 99.6%   R-Sq(adj) = 99.6%

Analysis of Variance

Source           DF       SS      MS        F      P
Regression        9  261.709  29.079  4451.98  0.000
Residual Error  146    0.954   0.007
  Lack of Fit   145    0.953   0.007    13.04  0.218
  Pure Error      1    0.001   0.001
Total           155  262.663


 154 rows with no replicates
```

The output from the General Linear Model is:

```
General Linear Model: log_wt versus species

Factor   Type   Levels  Values
species  fixed       7  Abramis_bjrkna, Abramis_brama(bream),
                        Esox_lucius(pike), Leuciscus_rutilus(ro,
                        Leusiscus_idus(white, Osmerus_eperlanus(sm,
                        Perca_fluviatilis(pe


Analysis of Variance for log_wt, using Adjusted SS for Tests
```

```
Source     DF   Seq SS  Adj SS  Adj MS       F      P
species     6  188.607   0.541   0.090   13.80  0.000
log_len     1   72.483   1.009   1.009  154.41  0.000
log_ht      1    0.437   0.146   0.146   22.29  0.000
log_wdth    1    0.182   0.182   0.182   27.82  0.000
Error     146    0.954   0.954   0.007
Total     155  262.663


S = 0.0808187   R-Sq = 99.64%   R-Sq(adj) = 99.61%


Term         Coef  SE Coef       T      P
Constant  -2.7732   0.2724  -10.18  0.000
log_len    1.8030   0.1451   12.43  0.000
log_ht     0.6460   0.1368    4.72  0.000
log_wdth   0.5579   0.1058    5.27  0.000
```

We will compare the output in class.

### Highly Correlated Input Variables

For highly correlated input variables, there are three (different) standard things to do.

- Take out the variables that are causing the problem.
- Leave the variables in but caution the reader that the input variables were highly correlated and predictions are only valid if new inputs follow the same correlation.
- Replace the dependent variables by a single combination which captures their dependency. This involves computing principal components and can lead to issues of interpreting what the new variable means.

We will explore these options with the fish data in class.

## Exercises for Lecture 27

1. –

2. –