

## BIOL 458 BIOMETRY

### Lab 10 - Multiple Regression

---

Many problems in biology science involve the analysis of multivariate data sets. For data sets in which there is a single continuous dependent variable, but several continuous independent variables, multiple regression is used. Multiple regression is a method of fitting linear models of the form:

$$\hat{Y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + \varepsilon$$

where  $\hat{Y}$  is the estimated value of  $Y$ , the criterion variable ;  $X_1, X_2, \dots, X_k$  are the  $k$  predictor variables; and  $b_0$  and  $b_1, b_2, \dots, b_k$  are the regression coefficients. The values of the regression coefficients are determined by minimizing the sum of squares of the residuals, *i.e.*, minimizing

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Hypothesis tests about the regression coefficients or about the contribution of particular terms or groups of terms to the fit of the model are then performed to determine the utility of the model. In many studies, regression programs are used to generate a series of models; the "best" of these models is then chosen on the basis of a variety of criteria, as discussed in lecture. These models can be generated using a forward, back-wards, or stepwise regression routine. In forward regression new independent variables are added to the model if they meet a set significance criteria for inclusion (often  $p < 0.05$ , for the partial -  $F$  test for the inclusion of the term in the model). In backwards regression all independent variables are initially entered into the model and sequentially taken out if they do not meet a set significance criterion (often  $p > 0.1$ , for the partial  $F$ -test for removal of a term). Stepwise regression uses both these techniques. A variable is entered if it meets the  $p$ -value to enter. After each variable is added to the equation all other variables in the equation are tested against the  $p$ -value to remove a term and if necessary thrown out of the model. The **SPSS**, SAS, MINITAB, SYSTAT, BMDP and other statistical packages include these routines. The computer output generated by these routines consists of a series of models for estimating the value of  $Y$  and the goodness-of-fit statistics for each model. Each model estimates the value of  $Y$  as a linear combination of values of the predictor variables included in that model.

#### Further Instructions for Lab 10

In **Lab 10** you use the same regression module to fit multiple regression models in **SPSS** that you used in **Lab 9** to fit bivariate regression models. The big difference now

is in entering the multiple independent variables, selecting the algorithm for building the model, and evaluating the fit of each model.

In the **Linear Regression** sub-window, you will see a box with a pull down arrow called "**Method**" that by default is occupied by the word "**Enter**." **Enter** is one of several model building algorithms available in the **Method** box. "**Enter**" in **SPSS** is equivalent of forcing all the variables in the independents variable box to be entered into the model simultaneously. The opposite of **Enter** is "**Remove**" where all variables are removed simultaneously. Other model building algorithms use various criteria to make decisions about which variables are entered (or removed) from the model, and when to stop adding or removing variables from the model. **SPSS** has algorithms named; **Stepwise**, **Backward**, and **Forward**. In the Stepwise algorithm, the variable with the smallest probability of its *F* statistic (if it meets a criteria, such as  $p < 0.05$ ) is entered into the model first. Then this process is repeated for the variables not yet included in the model. The next variable that meets this criterion is added to the model. This process continues to add variables to the model until there are no variables left that have *F* statistics that meet some user specified criteria ( $p < 0.05$  for example). As this process progresses, the *F* statistics for variables already in the model can change. If the significance level of these *F* statistics exceeds the criterion, then these variables are removed from the model. Hence, in a **Stepwise** algorithm, variables can be both added and removed from a model in the model building process. The Forward algorithm is identical to the Stepwise algorithm, except that variables can only be added to the model, not removed. The **Backward** algorithm puts all variables into the model, but then attempts to sequentially remove variables. The variable with the smallest partial correlation with the dependent variable is removed first if it meets the criterion for removal. If this variable is removed, then the variable with the next smallest partial correlation with the dependent variable is considered for removal, and removed if it meets the criterion. Note that in the **Backward** algorithm, variables are removed because their partial correlations exceed the significance criterion ( $p > 0.05$ ), the opposite of the criterion for a **Stepwise** or **Forward** algorithm.

Unfortunately, none of these algorithms are guaranteed to choose the "best" model. I prefer the **Forward** algorithm, but sometimes build models with different algorithms to see if they all choose the same "best" model.

Occasionally you might wish to enter variables in a specific sequence into a model, or to use different algorithms for model building for different groups of independent variables. To do so you need to look at the text and buttons surrounding the **Independent(s)** box in the **Linear Regression** sub-window. Note a light gray line enclosing this region, and blue text that says **Block 1 of 1**. **SPSS** allows you to group variables into "blocks" and specify different variable selection methods for each block. For example, to build the Analysis of Covariance models that I described in class, you would place the variable name for the covariate into the **Independent(s)** box and select **Enter** as the **Method** (since you don't want **SPSS** to do any thinking, just put the variable in the model). Then you would click on the **Next** button. Note that the blue text now says "**Block 2 of 2**." Here you would enter the names of the dummy variables that define your groupings or factors in the covariance analysis. Again use the **Method: Enter**. Finally, you would

click on the “**Next**” button to create the third block of variables. Here you would enter the variable names for the factor-covariate interactions. Once again use the **Method: Enter**.

Assessing model fit involves all the same procedures used in bivariate regression since the same assumptions apply. The dependent variable should be normally distributed, scatter plots should indicate linear relationships between the dependent and independent variables, and residual plots should show homoscedasticity (equality of variances in the residuals throughout the regression line). In addition to these issues, one also needs to check for outliers or overly influential data points, and for high inter-correlations between pairs of independent variables (called multi-collinearity). If two independent variables are highly correlated ( $r > 0.9$ ), then inclusion of both variables in the model causes problems in parameter estimation. You can pre-screen your independent variables by getting a correlation matrix prior to performing the regression and only allowing one variable of a pair of high correlated variables to serve as a candidate variable for model building at a time. You could also examine the **Tolerance** values provided by **SPSS** in the output table named “**Excluded Variables**.” These values also provide you information about whether you have a problem with multi-collinearity. Come to class to find out how to interpret the tolerance values.

## Lab 10 Assignment

---

The exercise to be performed in this lab is to use the **SPSS** stepwise and forward regression routine to generate a series of models, and to select the "best" model from each series, as discussed in lecture. Two data sets will be provided; you are to perform the analysis on either of these two. You must discuss in detail the reasons for choosing the models that you have selected, including showing plots of residuals, information about the distribution of the response variable, examining outliers, and other metrics to demonstrate goodness-of-fit.

### DESCRIPTION OF DATA

The data is stored in a file [MULTR2](#). The variables are as follows (they are in the same order in the data sets):

### VARIABLE (UNITS)

---

Mean elevation (feet)

Mean temperature (degrees F)

Mean annual precipitation (inches)

Vegetative density (percent cover)

Drainage area (miles<sup>2</sup>)

Latitude (degrees)

Longitude (degrees)

Elevation at temperature station (feet)

1-hour, 25-year precipitation. intensity (inches/hour)

Annual water yield (inches) (Dependent Variable)

The data consists of values of these variables measured on all gauged watersheds in the western region of the USA. The dependent variable is underlined. Develop and evaluate a model for estimating water yield from un-gauged basins in the western USA.