Stepwise Linear Regression Terms and Implementation

Rose Wieland

November 2, 2007

Math 410, Dr. Lambert
Abstract

Scientists in all fields are constantly performing experiments to explain how the world works. Finding the significant relationships among pieces of data is often the most challenging aspect of a project. There are myriads of statistical tools available that will help but only if one is familiar with the statistical jargon. Stepwise and Stepwisefit are two Matlab functions which contain an algorithm for determining which variables have a significant relationship with the output. The statistical analysis performed behind the Matlab function is explained, and the outcomes of random and correlated data are displayed.

Introduction

Every experiment attempts to establish a relationship between two or more variables. After the data is collected experimenters must make meaningful conclusions from the data they have collected; the data must be analyzed and presented. The presentation must establish credibility and do it in a way that will be understood by others. To read or write scientific papers it is necessary to understand the terms and theory, understand the methods of implementation, and have confidence in those methods.

Background

Linear regression is a technique used to fit data points. When more than one variable is thought to affect the response multiple linear regression is used to model the data. If there are k independent inputs it is assumed the data follows the form:

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \]

In this model, \( Y \) is a linear function of all \( k \) input variables. It is also sometimes necessary to account for interaction between variables. This is the case if \( x_i \cdot x_j \) is included in the regression model. Then with two variables the model is of the form:

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \]

It is also possible to include higher order terms of any of the variables, \( x_i^2 \) or even \( x_i^3 \). This allows for a quadratic or cubic relationship between the input and the outcome. It is also technically possible to include \( x_i \cdot x_j \cdot x_k \) terms (a three way interaction) but this is rarely done.

When interaction is considered in a model the coefficients no longer represent the strength of effect a variable has. Because the variables are modeled as interacting, an increase in \( x_i \) is going to affect the \( Y \) value and will also affect the \( x_i x_2 \) variable.

Never in real life does the data perfectly predict an outcome. Instead statistics help us to make predictions with an acceptable degree of accuracy. To find the coefficients in the regression a Least Squares Fit is preformed to solve the system. A Least Squares Fit of the data has the property that the squared distance between the predicted and actual values will be minimized.
The difference between the solution predicted by the model and the actual solution is called the residual.

The value of the relationships predicted by the model can be calculated from the residuals. The residual sum of squares (the sum of the squares of the residuals), the regression sum of squares (the sum of the square of the difference between the predicted value and the mean) and the total corrected sum of squares all help establish statistical significance.

Significance of both the regression as a whole and of each coefficient must be established. Significance is established using a null hypothesis. The hypothesis is that unless a test statistic can be established with a predetermined level of confidence it should be assumed to be zero. The p value is the probability that a statistic is insignificant.

Just because a model has been established as significant does not mean it is the best, or even a good model. One way to verify a model is to plot the residuals. Randomly distributed residuals with a small standard variation in comparison to the data suggest a linear model is a good choice.

An $R^2$ value is an estimate of what percent of the variation in Y is explained by the data. $R^2$ values range from 0 to 1. The larger $R^2$ is the more the variation in Y is explained. However, adding new variables always makes $R^2$ bigger but not necessarily, does adding more data make the model a better predictor. In the case of multiple variables an adjusted $R^2$ can also be calculated which accounts for the number of considered variables. The adjusted $R^2$ value will only increase if the reduction in the residuals is large enough to compensate for the loss of one residual degree of freedom.

A stepwise regression model takes advantage of this property of the Adjusted $R^2$ value to determine significance. The $R^2$ value of each variable considered alone is computed and then the variables are ranked according to these values. The variable with the highest $R^2$ can be assumed to account for the largest percent of variation in Y. A stepwise regression starts with the variable with the highest $R^2$ and calculates an adjusted $R^2$ value. The variable with the next highest $R^2$ is then added to the model and a new adjusted $R^2$ value is calculated. This pattern is followed until the adjusted $R^2$ value decreases rather than increases. At this point the addition of variables is no longer statistically justified.

A stepwise regression can also be performed backwards. In this case all of the variables are included and the variables with the lowest $R^2$ values are sequentially removed until the adjusted $R^2$ value stops increasing. At that point all of the remaining variables are considered to significantly contribute to the model.

**Methods**

The Matlab stepwisefit and stepwise functions perform a stepwise regression of the input data to uncover statistically significant relationships. Stepwise is an interactive tool, while stepwisefit is a command line version of the stepwise function. Both functions allow the user to input:
- the independent variables
- the dependent variable
- the required level of significance for a variable to be kept or removed from the model
- the maximum number of iterations
- any variables desired to be considered in the initial calculation of adjusted $R^2$

The tools also allow the user to keep or scale each variable. When one variable has values ranging from 10,000 to 20,000 and another variable ranges from .0001 to .0002, it is necessary to scale the variables to have a well conditioned matrix.

The interactive tool (stepwise function) provides the user with a graphical interface to step through each step in the regression. Any variable that is determined should be kept in the final model is colored blue while those variables that do not add to the model are colored red. The coefficients for all variables if they were contained in the model are shown as well as their student’s $t$ distribution value which establishes the $p$ (or probability of insignificance) value.

The entire regressions statistics, $R^2$, adjusted $R^2$, F and p values which are measures of statistical significance as well as the intercept of the data and the root mean squared error of the data are all displayed. The intercept is the $y$ value if all input variables were zero. Using the stepwise function a stepwise regression was performed of random data and perfectly correlated data generated in Excel, as well as example data from a statistics text.

**Results and Discussion**

As expected the stepwise regression found no statistically significant relationships among any of the random data trials. In Figure 1 you can see the results of one trial. The recommendation is to move none of the variables into the model because none of them achieve a high enough $p$ value to be significant. The lack of correlation is exemplified in Figure 2, where it can be seen the data is distributed randomly about the predicted line. The other random trials produced similar insignificant results.

![Stepwise Regression](image)

*Figure 1. The output of stepwise for first trial of random data.*
Performing a stepwise regression on artificially perfectly correlated data found the correct coefficients. The p values were accurately calculated to be 0 and the $R^2$ and adjusted $R^2$ values were appropriately 1. The perfect correlation is illustrated in Figure 3. From the Model History plot you can see the error in the analysis decrease with each additional variable added. Figure 4 shows a plot of perfectly correlated data.

As a last implementation of the stepwise regression, data from a statistics text was used. The results obtained from this trial exactly matched the results obtained in the text. A plot of realistic data exhibiting partial correlation is presented in Figure 5.
As a final note of interest, the stepwise regression will also consider interactive models if the terms are appended to the original data. The code to do this is included in the appendix. It is clearly indicated which part of the code includes the effect of interaction. With the random data no interaction was found. With contrived data the interaction between variables was found exactly. In the partial example if the interaction variable was included it overrode the effect of either variable independently. Clearly, stepwise regression could be used to find interaction between variables but more analysis is necessary to interpret the results.

The stepwise function is a thorough tool for experimenters who are looking to analyze data. It will compute values necessary to be confident in the results of an analysis, as well as, provide plots for visual confirmation of the outcome.
Appendix

clear all;
%initialize variables
counter1 = 1;
counter2 = 1;
%get data
data = load('file.txt');
%normalize data
[m n] = size(data);
numvar = n-1;

y = data(:,n);
x = data(:,1:numvar);

% %Check for interaction
% for i = 1:numvar
%     for j = counter1:numvar
%         x2(:,counter2) = x(:,i).*x(:,j);
%         counter2 = counter2+1;
%     end
%     counter1 = counter1+1;
% end
% %Stepwise Regression
% x = [x x2];

stepwise(x,y)
References