Introduction

In this exercise, we will gain experience conducting transformations of variables in order to use them in regression models and fitting multiple regression models. We will use data collected by graduate student Doug Bruce and undergraduates during summer 2001. In today’s exercise, we will retrace a path I took when I was confronted with this data set, so that you can see the usefulness of data transformations to build a regression model to analyze experimental data.

Objectives

- Learn how to explore relationships among variables using spline fits.
- Compare fits of alternative non-linear models and linearize them using transformations.
- Assess a collection of variables to fit a multiple regression model and avoid problems associated with excessive intercorrelation.

Explanation of the data

This data set comes from the master’s thesis work of Doug Bruce. In a laboratory experiment, Doug assessed effect of enzyme genotype and exposure temperature on mating frequency and fecundity of the leaf beetle *Chrysomela aeneicollis*. Doug collected 180 females and reared them for three weeks with a male during their reproductive period. For three hours each afternoon, beetles were exposed to one of three experimental temperatures (20°C, 26°C, or 32°C). Previous studies indicated that the highest temperature was stressful to the beetles, which live at high elevations in the Sierra Nevada mountains. We were aware that these exposure temperatures caused induction of stress proteins (heat shock proteins, HSP70), which are an indicator of heat stress. Twice each day, Doug and his undergraduate assistants examined each beetle’s petri dish for new egg clutches, determined whether beetles were mating, and counted the number of eggs in each clutch (mean clutch size = 32.6, S.D. = 9.95). When the experiment was complete, Doug weighed each beetle and froze it to determine its enzyme genotype.

In the file ‘labfecunditySum01B345.JMP,’ data taken include:

- Genotype at the enzyme locus *phosphoglucone isomerase* (pgi_fem)
- Exposure temperature (acc temp)
- Female body mass (Mass_Fem)
- Amount of HSP70 stress proteins in body tissue (ng HSP70_Fem)
- Proportion of time spent mating (Mate freq)
- Number of egg clutches laid (# clutches)
- Average clutch size (ave_clu_sz)
- Total number of eggs laid (Egg_number)

We will also examine the number of eggs laid by each female each day. The file ‘lab fecundity indiv fems’ includes the cumulative number off eggs laid by 76 females over the 24 day experimental period.
Exercise 1- Exploring relationships among variables using spline fits

Quinn and Keough (page 108) and the JMP manual (pp 244-245) discuss the use of spline fits to examine bivariate relationships among variables. A spline fit is a type of polynomial regression that fits a line to nearby points rather than to the entire collection of points. Splines are useful as exploratory tools to determine whether two variables are related in a non-linear way.

Opening the data file

Open the file ‘lab fecundity indiv fems.JMP.’ You will see six variables: the female identification number (female), host plant individual (plant), enzyme genotype (pgi_fem), exposure temperature (acc temp), day (day), and cumulative number of eggs (eggnum). Data for each female is coded a different color.

Plotting data

1. Select the ‘Analyze’ and ‘Fit Y by X’ commands, and drag ‘day’ into the ‘X, Factor’ box and ‘eggnum’ into the ‘Y, Response’ box. Select the bottom right corner of the box and drag it down and to the right to see the scatterplot in a larger scale. You should see a triangle of colored points indicating the data for each day for each female.

2. Select the red triangle under ‘Bivariate fit of eggnum by day’ and select ‘Group By..’ A window will open asking you to select the ‘Grouping Column’ and you should place ‘female’ into that box.

Fitting each value

3. Select the red triangle under ‘Bivariate fit of eggnum by day’ and select ‘Fit Each Value.’ You will now see a line connecting each value for each female. The number of eggs laid should resemble a staircase. What does this tell you about the pattern of egg laying over time? Carefully examine the plots to compare egg laying rates among females. You should notice that females differ greatly in the total number of eggs laid over the 24 days. How else do females to vary in egg laying pattern?

Using the spline fit function to compare females

4. Close the output window and repeat steps 1 and 2 indicated above

5. Select the red triangle under ‘Bivariate fit of eggnum by day’ and select ‘Fit Spline.’ You will see a list of numbers (lambda values) that determine the flexibility of the spline fit. As described in your JMP manual, lower numbers cause a more flexible fit, while higher ones create a fit that conforms more closely to a straight line. You may experiment with different lambda values to get a feel for how it works, but you should eventually select lambda = 10. Examine the resulting plot closely and try to come up with two factors that vary among females and appear to influence the total number of eggs laid by a given female. If you are stumped, ask your instructor for hints. Based on the spline fits, does it look like a linear fit would adequately describe the relationship between time and cumulative fecundity? You can also examine this by opening another Fit Y by X window, grouping by female, and trying a linear fit.
Summarizing data and comparing cumulative fecundity among genotypes

6. Use the ‘Table’ and ‘Group Summary’ commands to obtain daily mean egg numbers for each PGI enzyme genotype. To accomplish this, you need to put ‘pgi_fem’ and ‘day’ into the group window and select ‘Mean’ eggnum in the ‘Statistics’ window. You should obtain a new data table with 72 rows (24 for each genotype). Save this data table as ‘genotype means.jmp.’

7. Select the ‘Analyze’ and ‘Fit Y by X’ commands, and drag ‘day’ into the ‘X, Factor’ box and ‘Mean(eggnum)’ into the ‘Y, Response’ box. You will see three distinct curves corresponding to each genotype. Select the red triangle under ‘Bivariate fit of Mean(eggnum) by day’ and select ‘Group By..’ and ‘pgi_fem.’ Select the red triangle again and select ‘Fit Line.’ You will now see three linear regression lines fit through each group of points. Do the data appear to conform to the assumptions of linear regression? If not, which assumption is not satisfied? Finally, describe each genotype with respect to its cumulative fecundity, and begin filling in a table with 3 rows corresponding to each genotype, and four columns. Record the r square values for the linear fit for each genotype in your table.

Exercise 2- Comparing fits among models and linearizing them using transformations

Finding a transformation to fit the bivariate relationship

JMP has a nice feature that allows you to try out different transformations and examine their fit to the data on the original bivariate plot. If a transformation fits the data more closely than the linear fit, you will see that the points are closer to the regression curve than they would be to a straight line. In addition, the output reports the r square value for the transformed data. If you see that the r square value is distinctly higher than the linear fit, you may infer that you are approximating the relationship between the variables with the transformed data more closely.

8. In the output window from the ‘Analyze’ and ‘Fit Y by X’ commands, with data grouped by ‘pgi_fem,’ select ‘Fit Special.’ You will see a window asking you to specify the transformation. Try different transformations for the X and Y variables and observe whether the lines fit the curves more closely than the linear fit. For relationships that don’t work well, remove the fit and move onto different transformations.

9. When you have found a transformation between the variables that appears to fit the data more closely than a straight line, record the r square values for each genotype in your table, and compare them to those you recorded for the linear fit. Then show your results to your instructor and verify that you arrived at the appropriate transformation. What was your final decision about how each variable should be transformed?

Using the transformation to linearize the relationship between time and fecundity

Transformations are useful to reduce problems associated with variance/mean relationships, yet they can also be used to create a simpler explanation of how two variables are related. A major advantage of linear relationships over others is that they are described by only two parameters: the slope and intercept. More complicated relationships require estimation of more parameters. In the present data, I hope you noticed that genotypes differed not only in total fecundity and also in shape of the fecundity curve. It would be nice to quantify the difference in curve with a single
parameter, i.e. the slope. By transforming the data, we can consider this slope a measure of egg-laying rate.

10. Create two new columns in your data table and apply the same formulas that were indicated in your transformation comparison, following instructions in your JMP manual (pp 57-71). Basically you will start by clicking on the top right side of the column in your summary data table where the observation numbers are located and creating two new columns, then selecting the top of those columns, selecting ‘New Property,’ ‘Edit Formula’ and entering your formula in the ‘Formula Window.’ Ask your instructor if you need help doing this.

11. When you have made your two new columns with transformations of ‘days’ and ‘eggnum,’ use **Fit Y by X** again, group by genotype, and examine scatterplot with the three regression lines. Based on this graph, which one of the three genotypes differs, with respect to egg laying rate? Write down slope estimates for each genotype and add them to your table. Examine your data and the regression carefully to determine why the slope values are negative!

**Interpretation**

By examining the relationships between time and cumulative fecundity for a range of females, we discovered two factors that influence the total egg output of females during the experiment. One of these factors concerned the rate of egg laying, as seen by differences in the curves of cumulative fecundity versus time. Our data transformation allowed us to quantify this rate using the slope of the linearized data. When I analyzed these data, I calculated this rate for each female and included it as a variable in the multiple regression model below.

**Exercise 3- Fitting a multiple regression model of fecundity**

Quinn and Keough (pages 137-141) provide very good suggestions for finding the ‘best’ multiple regression model to fit a set of data. Researchers are gradually coming to a consensus that traditional stepwise regression methods for determining which variables to include in a model have conceptual flaws and sometimes lead to faulty conclusions. The best approach is to separate the process of model selection from statistical hypothesis testing. However, there is no universally agreed upon method for choosing among models. JMP offers a traditional, easy to use stepwise regression procedure, which we will modify to follow Quinn and Keough’s general recommendations about how to select among models.

**Opening the data file**

Open the file ‘labfecunditySum01B345.JMP.’ Inspect the list of variables and compare it to the list on page 1 of this handout. Which two variables have been added?
**Deciding whether to transform the dependent variable**

Certain variables tend to require transformation because their variance is related to the mean. Most often, a log transformation is used. I tend to prefer base 10 over natural logarithms because the numbers can be more easily interpreted \( \log(10) = 1 \), \( \log(100) = 2 \), etc.\]. Before we conduct our multiple regression, we will decide whether total egg number should be transformed by examining residual plots.

12. Use the **Fit Y by X** platform with number of clutches as the X variable and egg number and log_egg_num as Y variables. After fitting the regression line, examine the two regressions closely, then plot the residuals. Which variable seems to conform to the homogeneity of variance assumption more closely? Choose this variable for all subsequent analyses.

**Assessing degree of correlation between the independent variables and Y**

It is always recommended to examine the simple correlations and scatterplots among variables before fitting a multiple regression model. The simple correlations can detect non-linear relationships and intercorrelation among independent variables.

13. Use the **Fit Multivariate** platform with all continuous potential X variables and your chosen Y variable to obtain the correlation matrix and scatterplot matrix. Which two X variables show the highest intercorrelation? Which variable is most closely correlated with Y?

14. Use the ‘**Distribution**’ platform to examine the distribution of the two intercorrelated X variables. Pick the variable whose distribution is closest to normal (which you can assess by visual inspection) for further analysis—leave the other variable out of your regression model fitting analysis.

**Conducting the multiple regression**

We will use the ‘Stepwise’ feature found in the **Fit Model** platform to explore the fit of alternative models to the fecundity data.

15. Use the **Fit Model** platform and enter the continuous X variables (without one you omitted in #14, above) into the ‘Construct Model Effects’ box, and your chosen Y variable into the Y box. At the top right, change the option under ‘Personality:’ from ‘Standard Least Squares’ to ‘Stepwise.’ Then press ‘Run Model.’ You will see a new, unfamiliar output window, which includes a list of the X variables as columns and a row of numbers beneath the ‘Current Estimates’ box.

The row of numbers includes the r square, adjusted r square, Mallow’s Cp value, and the Akaike Information Criterion (AIC). Each of these represent different measures of model fit. Quinn and Keough object to the use of r square, because it only increases as more variables are added to a model, and prefer adjusted r square, which incorporates the number of variables and can decrease if adding variables yields little useful information. The AIC value is widely used for model
fitting in analyses of frequencies, and is based on assessment of the information content of a particular model (see Quinn and Keough, page 139). Although Quinn and Keough advocate the use of another information criterion (BIC), which penalizes complicated models more heavily than AIC, we will practice model fitting using AIC. You will compare AIC values for models with different numbers of variables, and choose to include a variable in the model only if it lowers the AIC value.

Your book and JMP manual describe the problems that occur in multiple regression when predictor variables are intercorrelated, but I have not yet found a description of a related problem- inclusion of a predictor variable in the model that is highly correlated to the dependent variable. When this occurs, slope estimates for other predictor variables fluctuate wildly, depending on whether the variable that is closely correlated to the dependent variable is included in the model. One must examine such correlations carefully, because it may mean that the correlated predictor variable simply represented another way of measuring Y. In our case, number of clutches is closely correlated to number of eggs because they are two ways of measuring egg output. It is often best to leave such variables out of a multiple regression analysis.

16. If ‘# clutches’ is on your list of variables in the output window, click in the ‘Lock’ box on the left side next to this variable name so that you do not accidentally include it in your model.

17. The next step is to compare alternative one-variable regression models. One by one, check the ‘Entered’ box next to each variable and watch the AIC values change. Deselect the variable before proceeding to the next one. Once you have identified the variable with the lowest AIC value, check the ‘Lock’ box next to it. Now create a table with five columns: Step, Variables included, r square, adjusted r square, and AIC. Fill in your values for the best one variable model into the table.

18. Now you can begin to select the best two-variable model to fit the data. Choose the variable that lowers the AIC value most by checking the enter box for all of them. When you have chosen your variable, enter the values into your table, and proceed to the third variable. Stop adding variables when you find that all choices cause an increase in AIC.

19. When you have chosen the best model, click on the ‘Make Model’ box, which will open a new ‘Fit Model’ box. Here, simply click ‘Run Model’ and examine the leverage plots in the output. Inspect the leverage plots to make sure that there are nice even clouds of points (see pages 320-323 of your JMP manual for a description of the effects of collinearity on leverage plots). Write a brief paragraph summarizing the results of this model, using declarative sentences as described in the ‘laboratory introduction’ handout.

20. Return to the ‘Stepwise Fit’ window and search for the ‘next’ best model by choosing the one that increases AIC the least. Add this model to your table and make three graphs, using graph paper, showing the change in r square, adjusted r square, and AIC on the Y axis, and the number of predictor variables in the model on the X axis. Do the different measures change in a similar way when AIC is used as the measure for model selection?
21. Now click on the ‘Remove All’ box to omit all variables from the model, and try the automatic stepwise model selection procedure implemented by JMP by clicking on the ‘GO’ button. How does this model compare to the one you selected by AIC?

22. Remove all variables from your model again and uncheck the ‘Lock’ box by the number of clutches. Repeat steps 17-19 without making a new table, and examine the new ‘best’ model. Write a brief paragraph comparing the ‘best’ model found when the number of clutches included to the best model found without that variable. Which ‘best’ model seems to make the most biological sense?

Tasks

1. Identify factors that influence total egg output by examining cumulative fecundity over time for individual females using 'Fit Values' and spline fits.

2. Construct a table comparing r square values for a linear fit of time versus cumulative fecundity to a fit using transformed variables. You will also indicate what the most appropriate transformation you used was and create transformed variables that you will compare using linear regression.

3. Decide whether an original or transformed variable is the most appropriate variable for use in a regression, based on examination of residual plots.

4. Construct a table showing your process of fitting a multiple regression model by adding variables to the model based on comparisons among AIC values for different models.

5. Write a brief paragraph summarizing the results found in the 'best' multiple regression model that you found.

6. Plot the change in r square, adjusted r square, and AIC with increasing numbers of variables, including one variable more than the 'best' model found above. Describe how the pattern of change in values relates to the topics discussed on pages 137-139 of Quinn and Keough.

7. Interpret the difference in results between a model that includes a variable (number of clutches) that shows high correlation to the dependent variable (number of eggs) to results found when that variable is not included in the analysis (constructed for tasks 4 and 5).