Boston University
Department of Mathematics and Statistics

MA 214

Applied Statistics

Lab Session 7: Multiple Regression

In this lab we will use the JMP software to further explore linear regression and gain practice in model selection, and in particular in using the various stepwise methods of deciding on which predictor variables should be included in a multiple linear regression model.

Preparation

This week we will be using a new JMP data set called SurgicalUnit.jmp which is available on the Blackboard site under the Course Documents link. The file contains 54 observations on the time of survival of patients who had liver surgery. There are five covariates: the blood-clotting score, the prognostic index, the enzyme function test score, the liver function test score, and the age in years. SOURCE: Neter, et al., *Applied Linear Statistical Models* (1985).

Variable Descriptions:

Columns

x1	Blood-clotting score
x2	Prognostic index
x3	Enzyme function test score
x4	Liver function test score
x5	Age
y	Survival Time

Questions

1) A very important question at hand is whether or not we can accurately predict the survival of a patient based on some knowledge in the form of the various test scores and other variables such as age. Before fitting a multiple linear regression model to address this question, let us first try to understand the relationship between all of the predictor variables in our dataset.

Go to the “Analyze -> Multivariate Methods -> Multivariate” screen and enter all of the predictor variables, “x1” through “x5”, into the “Y, Columns” role and then press the “OK” button. On the window that pops up, immediately go to the red triangle next to the words “Scatterplot Matrix” and turn off the “Density Ellipses”.

This JMP output now displays the correlations between each of the predictor variables, as well as all of the different possible scatterplots where the value of one predictor variable is plotted against the values of the others, one at a time.

(a) Based on this output, what sort of correlation do you observe among the predictor variables in our data set? Which pairs appear to me most correlated?

The scatterplots appear to show random patterns with no particular direction or trend, this suggests that there is little to no linear correlation between the pairs of predictor variables. In other words, knowing the value of one variable does not give much information about the value of the other variable.

Based on the correlation matrix, the pairs of predictor variables that appear to be most correlated are:

x1 (Blood-clotting score) and x4 (Liver function test score) with a correlation of 0.5024
x2 (Prognostic index) and x4 (Liver function test score) with a correlation of 0.3690
x3 (Enzyme function test score) and x4 (Liver function test score) with a correlation of 0.4164

(b) What effect might this have on a multiple linear regression model?

The lack of correlation between predictor variables is actually a good thing when it comes to multiple linear regression. When predictor variables are highly correlated, which is also called multicollinearity, it can make the model unstable and the estimates of the regression coefficients unreliable. This is because it becomes difficult to disentangle the effects of the correlated variables on the response variable. If the predictor variables are not correlated with each other, this means that each one provides unique information about the response variable, which is what we want in a multiple regression model.

2) Now let us make a first attempt to predict the survival of a patient using all of the knowledge available from the predictor variables. That is, go to the “Analyze -> Fit Model” command and place “Survival” in the “Y” role, and all of the covariates, “x1” through “x5”, in the “Construct Model Effects” box. Change the “Emphasis” to “Minimal Report” and then click the “Run” button to fit the model.

Based on the criteria we have discussed in class, comment on the adequacy of this “full” model in the following steps:

(a) What is the value of the F-test statistic that appears in the JMP output for your fitted model under the “Analysis of Variance” section, and its corresponding p-value? State in words the pair of hypotheses that are being tested by this F-test, and then interpret the outcome of the test using the F-test statistic and p-value.

The F-test statistic in the JMP output is 21.8741 and the corresponding p-value is less than 0.0001.

The pair of hypotheses being tested by this F-test are:

- Null hypothesis (H0): All the regression coefficients are equal to zero. This means that none of the predictor variables (x1 through x5) have any effect on the response variable (Survival Time).
- Alternative hypothesis (H1): At least one regression coefficient is not equal to zero. This means that at least one of the predictor variables has an effect on the response variable.

The F-test statistic of 21.8741 is quite large, and the p-value is less than 0.0001, which is much less than the common significance level of 0.05. This means that we reject the null hypothesis and conclude that at least one of the predictor variables has a significant effect on the response variable. In other words, the model as a whole is statistically significant, and it appears that at least some of the predictor variables are useful for predicting Survival Time.

(b) Scroll down to the “Parameter Estimates” section and perform a t-test to determine which parameter estimates are statistically significant in the model; that is, test whether each parameter estimate equals zero or not. Based on the results of these tests, do you think the full model with all five covariates is necessary? Are there any predictor variables that you would consider dropping from the model? If so, explain why.

Based on the results of these tests, we can see that the predictor variables x1 (Blood-clotting score), x2 (Prognostic index), and x3 (Enzyme function test score) are statistically significant at a common significance level (e.g., 0.05), as their p-values are less than 0.05. This means that these variables have a significant effect on the response variable (Survival Time), assuming that the other variables are held constant.

On the other hand, the predictor variables x4 (Liver function test score) and x5 (Age) are not statistically significant, as their p-values are greater than 0.05. This suggests that these variables do not have a significant effect on the response variable, assuming that the other variables are held constant.

Therefore, we might consider dropping the variables x4 and x5 from the model, as they do not appear to provide any additional predictive power beyond that provided by the other variables.

(c) From the red triangle next to “Response y” at the top of the JMP output, select the “Plot Residual by Predicted” and “Plot Residual by Row” commands from the “Row Diagnostics” option. Based on these plots and the criteria discussed in class, comment on the adequacy of the model. Do the residuals appear to be independent and normally distributed around zero with constant variance?

The "Residual by Row" plot showing randomness around the center line 0 indicates that the residuals are independent, which is a good sign. However, the "Residual by Predicted" plot showing a fan-out pattern is a concern. This pattern suggests that the variance of the residuals is not constant, violating the assumption of homoscedasticity in linear regression. This could potentially bias the standard errors and thus the statistical inference (like confidence intervals and p-values).

(d) Does the residual plot suggest any transformation on Y may be necessary? If so, apply the transformation for the rest of the analysis. State the transformation you have selected.

The fan-out pattern in the "Residual by Predicted" plot suggests that a transformation might be necessary to stabilize the variance. A common choice in such situations is a log transformation on the response variable Y, which can often help in reducing the heteroscedasticity and making the model more appropriate for the data.

3) Let us now use the tools within JMP to implement the techniques of stepwise variable selection discussed in class. To do this, go to the “Analyze -> Fit Model” menu and setting up the full model as usual, but this time, before clicking the “Run” button, change the “Personality” to “Stepwise”.

You will be taken to a new JMP screen that allows you to control the various parameters involved in the decision-making process of the stepwise regression. As discussed in class, we will use the “P-value Threshold” stopping rule, so first change the stopping rule accordingly.

(a) First, we will move in the “Forward” direction, using the default initial values for the “Prob to Enter” and “Prob to Leave”. Make sure that all of the variables have been removed from the model by clicking the “Remove All” button, and then click the “Go” button to have JMP perform the stepwise regression. At the end of the process, the predictor variables that are checked in the “Current Estimates” section will have been the ones chosen to be in the final model. At this point, you can click the “Run Model” button to run and inspect the adequacy of that current model. Take a few minutes to familiarize yourself with this new JMP screen by trying different “Prob to Enter” and “Prob to Leave” values and trying the different directions of stepwise regression such as “Forward”, “Backward”, and “Mixed”. Remember that for “Forward” stepwise regression you must remove all predictor variables from the model to start fresh, and that for “Backward” stepwise regression you must instead enter all predictor variables to the model to start fresh.

Although we have not discussed them in much detail in the class, the AICc and BIC values offer criteria similar to the P-value Threshold that help us determine which predictor variables to include in our model. Try the stepwise regression procedure using those information criteria to see how

they compare to the P-value Threshold.

Note that different criteria may lead to different models and all of them may be adequate. In the table below, list all the different model selection procedures that you have tried and

Method	**Prob to enter/leave**	**Fitted model**	**R^2**	**Notes on the fit**
Forward Selection	0.25/0.1	x1, x2, x3, x4	0.691041	/
Forward Selection (Use different prob to Enter)	0.05/0.1	x1, x2, x3, x4	0.691041	/
Backward elimination	0.25/0.1	x1, x2, x3	0.684128	/
Mixed	0.05/0.05	x1, x2, x3, x4	0.691041	/
AICC	/	x1, x2, x3, x4	0.691041	/
BIC	/	x1, x2, x3, x4	0.691041	/

(b) Based on your answers in part (a), choose your final model and write down the fitted model in the space below.

It seems that the models selected by Forward Selection (with both sets of probabilities), Mixed Selection, AIC, and BIC all include the variables x1, x2, x3, and x4 and have the same R^2 value of 0.691041. The model selected by Backward Elimination includes the variables x1, x2, and x3 and has a slightly lower R^2 value of 0.684128.

Given that the majority of the methods selected the model with x1, x2, x3,

and x4 and that this model has the highest R^2 value, it seems reasonable to choose this as the final model.

So we have:

$y = \beta0 + \beta1*x1 + \beta2*x2 + \beta3*x3 + \beta4*x4 + \varepsilon$

where y is the survival time, x1 is the blood-clotting score, x2 is the prognostic index, x3 is the enzyme function test score, x4 is the liver function test score, β0 is the intercept, β1 through β4 are the coefficients for the predictor variables, and ε is the error term.

(c) Suppose that you received the following information on two new patients:

Patient 1
x1 5.5
x2 75
x3 101
x4 3.14
x5 38

Patient 2
x1 7.6
x2 35
x3 71
x4 2.71
x5 59

Enter these patients as two new rows in the SurgicalUnits data table, and then fit your model chosen in part **(b)** again and use it to predict the survival of each patient as usual by using the 95% Indiv Confidence Interval command under the “Save Columns” option in the red triangle on the JMP output for your fitted model. Write down your answers for each patient. If your fitted model involves transformation on Y, remember to transform it back to the original scale.

	95% Prediction Interval for Survival Time
Patient 1	(605.46937513, 1575.2599407)
Patient 2	(44.063286732, 1021.5407263)