A SUGGESTED PROCEDURE FOR
MULTIPLE REGRESSION ANALYSIS

Given data on a dependent variable $Y$ and set of potential explanatory variables $(X_1, ..., X_K)$, the following is a suggested procedure for multiple regression analysis of this data. You need not follow this procedure exactly in every case. Rather it should serve as a rough guideline highlighting the types of analyses you might wish to consider.

(1) First plot histograms of all your data. In examining these histograms, keep the following points in mind:

(i) For your explanatory variables $(X_1, ..., X_K)$, if some of the histograms are very skewed, you should consider transformation like logs or square-roots (for nonnegative data) to achieve a more uniform spread and thereby gain more “leverage” for identifying $\beta$ coefficients. Remember that it is not necessary that your explanatory variables be normally distributed.

(ii) For your $Y$ variable, it is again not essential that this variable appear to be normal. (Remember the example I gave in class where $Y$ was bimodal, and not directly “transformable” into a normal distribution.) Here are some other key points to remember:

(a) In the linear model itself, each $Y_i$ is a sample from a normal distribution with a different mean, depending on the values of $(x_{i1}, ..., x_{ik})$. So the overall $Y$ distribution may not appear to be normal.

(b) What really counts is the distribution of the final regression residuals. It is essential that these be sufficiently normal so that statistical inference can be carried out.

(c) If your $Y$ variable is highly skewed (with a long tail), then this skewness will often persist in the residuals. So this is a case where you need to try some simple transformations to see if you can improve the residual distribution.

(d) If your $Y$ data is strongly multi-modal (as in the bimodal example above), you should try to determine whether this may be reflecting some form of multiple-population structure (age groups, ethnic groups, etc.) that might be controlled for with a suitable categorical variable.

(iii) Finally, do not exclude any variables at this early stage. As a general rule, only drop variables one you have verified (through collinearity checks) that they are indeed insignificant.
(2) Assuming that you have a substantial number of potential explanatory variables (say, \( k \geq 5 \)) you should next do a *stepwise regression* to identify potential sets of significant explanatory variables.

(i) For your initial settings you should try Direction = ‘Mixed’, Entrance and Exit Probabilities = 0.15. But you should also experiment with other settings. For example, if your first stepwise regression produces no significant variables, this only means that no variable *by itself* is significant. A good option to try here is Direction = ‘Backward’ (be sure to start with all variables included by clicking ‘Enter All’). Often this will pick up *combinations* that are significant. This can then serve as a good starting point for a ‘Mixed’ procedure, as above.

(ii) Also remember that these procedures only focus on *P-values*. So to find models with highest *adjusted R-square*, you may wish to try the ‘All Possible Models’ option (described at the end of the web notes on Stepwise Regression).

(3) At this point you should consider whether any key variables are missing from your regression (based on your knowledge and initial hypotheses about the problem). If so, you should check to see whether this might be due to *multicollinearities* with variables selected in the stepwise procedure. This can be accomplished most directly by regressing each such variable, say \( X_{k+1} \), on the included variables \( (X_1, \ldots, X_k) \) to see whether any of these are highly significant predictors of \( X_{k+1} \). You might then try omitting some of these variables from your regression to see if \( X_{k+1} \) is now significant. If so, then you may conclude that \( X_{k+1} \) is indeed involved in a multicollinearity, and that the significance of this variable can only be evaluated by removing one or more of the other variables. More generally, you should also check VIFs of other excluded variables just to be sure that they are not being excluded solely on the basis of multicollinearity effects. Finally, remember that none of these procedures can substitute for your own knowledge about the \( Y \) variable of interest. What you are looking for are sets of predictors that are both statistically significant and intuitively plausible in terms of your own knowledge of \( Y \).

(4) On the other hand, if \( X_{k+1} \) either does not have a high VIF or does not become significant when other variables are removed, then \( X_{k+1} \) may in fact *not* be statistically significant. However, you can also try rerunning your stepwise regression with this variable forced to be included (simply by clicking on it). In some cases, this may generate a new sequence of variables in which \( X_{k+1} \) remains significant. Remember that stepwise regression is only a heuristic tool to find a ‘good’ regression. It is *path dependent*, and may actually lead to a much better regression result by forcing one or more variables to stay in. Try many
alternative starting points, and explore the space of possible regressions. The key point is to use this tool to learn as much about your data as you can.

(5) When you have settled on a set of variables which are both statistically significant and intuitively meaningful (from your knowledge of the problem), you are ready to test the adequacy of your regression assumptions. Be sure to start by clicking ‘Make Model’ to run a multiple regression using your candidate variables. If you have missing data, then these regression results may actually differ from the last step of the stepwise regression (depending on what data rows are included in the final regression).

(6) To analyze these regression results, start by saving the regression residuals. Now check the Normality Assumption on these residuals by using Normal Quantile Plot. If the residuals exhibit significant non-normalities, then your P-values may be misleading, since they are based on the assumption of normally distributed residuals. So you will probably need to consider additional data transformations, and start the analytical process over again. Normality of residuals is crucial.

(7) Finally, use these residuals to check the Gauss-Markov Assumptions for your regression model:

- **Linearity Assumption.** Check for nonlinear trends in the residuals. A good procedure here is to begin by plotting the residuals against each explanatory variable. If some nonlinear trends are evident, construct partial residual plots for these explanatory variables to see if some simple transformations can be found to remove these nonlinearities. [See the web notes “Multiple Regression (k = 2)”]. When you are satisfied with your choices, plot the residuals against each variable again to be sure that no trends remain.

- **Homoscedasticity Assumption.** Check for heteroscedasticity by visual inspection of the residual vs predicted and residual vs variables plots.

  (i) If the variance of the residuals appears to be increasing in Y-predicted (and if Y is a positive random variable), then you can try a variance-stabilizing transformation, such taking the log or square root of Y to reduce this heteroscedasticity. [See the class web notes on Heteroscedasticity.]

  (ii) If Y is non-positive, or if you do not wish to transform Y for some reason (such as ease of interpreting the results) then you should try a Weighted Least-Squares procedure. [See the web notes on “Heteroscedasticity”].

- **Independence Assumption.** This is the most difficult assumption to check. However, there is one important case where it can be checked statistically.
(i) If your data is *time series*, you should always use the Durbin-Watson test to check for (temporal) *autocorrelation*. If this test shows the presence of autocorrelation then you should try the two-stage procedure developed in class for reducing this autocorrelation. If the residuals pass all these diagnostic tests, you can be fairly confident that you have a solid regression result. If they do not, then you should try one or more of the procedures discussed in class for rectifying these problems. [See the web notes on “Autocorrelation”]. Also, if you have *panel data* (observations on the same sample units at multiple time points) then also look at the web notes on “Regression with Panel Data”.

(ii) If your data is *not* time series, then there is little you can do in terms of statistical analysis (based only on what you have learned in class.) However, this does *not* mean that you are free to ignore the problem.

(a) As one example, suppose you are using political-poll data and you know that some of the individuals sampled were from the same household. Then you can be reasonably confident that their political opinions are highly correlated. So if possible, remove all but one sample from each household. If this is not possible, then you should at least mention that this is a possible source of sample dependencies that may detract from your results.

(b) As a second example, if you are studying various attributes of cities around the world, then cities within the same country will exhibit much stronger forms of dependency than those cities in different countries. Even if you try using dummy variables to control for the “country effect”, this is not guaranteed to eliminate such dependencies. So again be sure to at least point out these possible sources of sample dependencies in your data.