

A SUGGESTED PROCEDURE FOR MULTIPLE REGRESSION ANALYSIS

Given data on a dependent variable Y and set of potential explanatory variables (X_1, \dots, 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. If some of the histograms (for nonnegative data) are very skewed, you should consider log or square-root transformations to achieve a more uniform spread. (It is *not* necessary that your individual data be normally distributed.) But highly skewed values tend to produce *heteroscedastic* effects. Also, if some histograms are strongly *multi-modal*, this might indicate a mixture of more than one population in your data. Hence you might want to consider *dummy variables* to reflect these different populations.
- (2) Assuming that you have a significant number of potential explanatory variables (say, $k \geq 5$) you should next do a *stepwise regression* (Direction = 'Mixed', Entrance and Exit Probabilities = 0.15) to identify an initial set of potentially significant explanatory variables. (Start with a 'Step' approach, and check the *adjusted R-square* values at each step to be sure they are increasing. If not, you may wish to stop adding variables.) You may also wish to drop variables with P-values above .10 (i.e., which are not even 'weakly significant'). When finished, click the 'Make Model' option to run a *multiple regression* with the current set of explanatory variables, (X_1, \dots, X_k) , selected by the stepwise procedure.
- (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 by adding each such variable, say X_{k+1} , rerunning the regression, and checking the Variance Inflation Factors (VIFs). If the VIF for X_{k+1} is among the highest values, run a separate regression of X_{k+1} on (X_1, \dots, X_k) to determine which of the current variables (X_1, \dots, X_k) is most significantly related to X_{k+1} (i.e., which have betas with the lowest P-values). 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.

- (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. This is what the procedure was designed to do!)
- (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.
- (6) 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.
- (7) Finally, use these residuals to check the crucial **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 scatter plot.
 - a. 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 web notes on Heteroscedasticity.]
 - b. 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.
 - a. 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 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.]
 - b. 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.