Multiple Linear Regression Analysis

Statistics is not an exact science, especially the data analysis part. Most times statisticians do not have the luxury or option of meeting all the assumptions related to a data analysis problem. So do we stop the analysis – NO. We do the complete analysis of the problem and we point out the assumptions that are not verified. At the end, we remind that caution should be used when using or interpreting the results.

The analysis of any data set should include the steps below.

1) Identify the correct '*y*' and '*x*' variables (read the question carefully for clues).

2) Verify a linear relationship between *y* and *x* variables (scatterplots, correlation, etc.).

3) Check for high correlation among the '*x*' variables. Remember, high correlation between '*y*' and each of the '*x*' variables are desirable (a good thing). [Look at the correlation matrix on the R window, MS Word usually distorts the columns of correlation matrix]

4) Fit the full model.

5) Investigate whether variables need to be removed from the full model.

   a) Check the significance of the coefficients and SEs.

   b) Check for high correlation from step 3.

   c) If needed then remove variables from the full model. Remember removing variable from a model results in loss of information. Due diligence is required.

6) Fit the reduced model. Compare R^2 values of the full and reduced models.

7) Perform residual analysis of the final model. Discuss all 4 assumptions, comment on whether they are valid or not. Investigate if transformation is needed.

8) Choose the final model.

   a) If drop in R^2 value is not big (less than 5%) and all the coefficients are significant in the reduced model, then choose the reduced model.

   b) Otherwise select the full model, but mention why you fit the reduced model.

9) Discuss all the coefficient values and their interpretations. Pay close attention to the coefficient standard errors. If some of the standard errors are large (greater than one fourth of the value of the coefficient), coefficients could be unstable.

10) Conclusion for the analysis. Restate violations of assumptions (if any) and caution about estimation and prediction.