# General Approach to Model Building

**by Dr. Woo**

**1. Write your questions of interest. (Convert these into Statistical Language or linear regression language)**

Think about this for our Report for Project One. Are we looking at coefficients.

**2. Define the variables of interest and how to measure them.**

If you are collecting data you can define this. If you data doesn’t fit the regression model framework you may need to change data.

REMEMBER: regression framework requires independence between variables.

**3. Brainstorm which statistical tools you could use to answer your questions of interest. Depending on context, you may have to go back to step 2 and redefine your variables.**

Is your response quant, or categorical? Predictors? Time Sequence (regression models not built for time sequence data).

**4. Design the study and how you're going to collect the data.**

**5. Collect data.**

**6. Enter & clean data.**

7. Run some exploratory data analysis (graphs, basic numerical summaries) on variables of interest that can add insight into your questions of interest in 1.. EDA, as its name suggests, is exploratory. No hypothesis test / confidence intervals performed at this stage.

**NOTES FROM CLASS**: exploratory data analysis is creating the graphical summaries. Basic numerical summaries. Notice interesting things. Observe, think about data.

8. Based on 1. and 7. and perhaps based on automated search procedures, create an initial model. This initial model should be reasonable based on 1. and 7.. Do not expect the initial model to be perfect! Typically, 1st order additive models considered

**NOTES FROM CLASS**: Can start to create the initial model. Use the automated search procedures. Usually only first order additive models. See how models look and act.

9. Carry out diagnostics, especially if the response needs a transformation.

**NOTES FROM CLASS**: Module 8: partial regression plot or added variable plot

10. Check other assumptions.

11. Consider improving your models by dropping terms, adding interaction terms to your initial model.

12. Steps 9., 10., 11. need not be in that specific order, and may need to be repeated. Each step should be based on a reason, rather than trial and error. Make sure what you do is theoretically and contextually sound, and helps answer your questions of interests in 1.

**NOTES FROM CLASS**:

13. Check for outliers in each model that is under considering. Outliers are usually interesting data points and should be investigated further.

**NOTES FROM CLASS**: outliers could be on response, predictors, or both (unusual combination of x and y variables). You probably don’t want to automatically remove outliers. Might be interested in investigating outliers. Create model with and without the outliers.

14. You may have multiple models that are fine in terms of diagnostics. Consider which model best answers your questions of interest. Also compare them using model selection criteria.

**NOTES FROM CLASS**: Use the AIC, BIC, PRESS etc to select model

15. Interpret and write your results.

In some data science settings, steps 1 to 6 are already done as you harvest a database for data, and do not actually plan the study and collect the data in the traditional sense.