The Basic Regression Modeling Process

1) Examine your data to understand what kinds of variables are present, and check for any that may have a significant number of missing values. (There are methods for dealing with missing values, but they are beyond the scope of this course.)
2) Plot histograms of each variable to understand the overall distribution as well as the ranges (for quantitative) and levels (for qualitative) for each one.
3) Examine the correlations between all quantitative variables to see how they are related to each other, keeping in mind that variables that are strongly correlated with each other (either positively *or* negatively) will introduce multicollinearity into your model if they are both included as predictor variables.
4) Decide what your **outcome** variable is (i.e., what it is you want to predict or explain); this will be your Y variable. All others will be potential predictor variables.
5) Fit a model with all potential predictor variables included, including any interactions.
6) Once the model has been built, evaluate it using the following criteria:
    a. Review the Analysis of Variance and ensure that the F statistic is significant
    b. Check the *p*-values for each predictor variable to determine significance
    c. Check the sign of each parameter estimate (i.e., coefficient) to ensure the direction of the relationship makes sense (for example, if you are trying to predict annual income based on education level, you would not expect the coefficient for education level to be *negative*, as this suggests that income *decreases* as education level increases)
    d. Examine the residual plot for any indication of non-normality or non-constant variance, and also perform formal tests (such as the Shapiro-Wilk test for normality and the `ncvTest` for constant variance); if either of these issues are identified, a transformation of the outcome variable may be necessary (such as a log)
    e. Review the Variance Inflation Factors (VIFs) for evidence of multicollinearity (a VIF of 10 or greater is generally an indicator of this problem)
    f. Review and interpret the Adjusted R^2 value; if it is rather low, this suggests that you are missing key predictor variables in your model (which may not be in your dataset)
7) After your evaluation of the model, do the following:
    a. Drop any non-significant predictor variables (i.e., those with a non-significant p-value) from the model
    b. If multicollinearity is present, drop one of the variables that is causing it
    c. Transform the outcome variable if the residual plot looks suspicious
    d. Fit a new model

8) Repeat steps 6 and 7 as necessary until you have built a statistically satisfactory model
9) If desired, also employ either stepwise variable selection ("backward" is recommended) and/or all subsets regression to help identify potential predictor variables that may not have been included in your initial model. You should evaluate any resulting models using the same criteria listed in step 6, and compare these models to your original model (as well as to each other) to determine which one is the most appropriate from both a statistical as well as a practical perspective.

Once you have built the final model, it must be evaluated in terms of its usefulness and relevance to the business problem, as well as its interpretability to others who may be interacting with it. ***Always ensure that it makes sense from the standpoint of the real-world dynamics it is intended to explain.***

Extended steps in the process (not discussed in this course)

- **Data partitioning** — It is good practice to partition your data into two parts: 1) a training set that is used to build the model, and 2) a test set that is used to validate the results of the model built from the training set. This helps prevent overfitting, which can occur when you build a model to all of the data you have available. Partitioning is often done by selecting a random percentage of the data (typically 30-40%) and excluding it ("setting it aside") during the model building process; this is your test set. Once you've built a model on the training set (the 60-70% of the data you didn't exclude), you should apply it to the test set and compare the output of the model (for the outcome variable) to the actual value for each record.

  When the "partitioning-model building" process is repeated multiple times (sometimes hundreds or thousands of times, typically through the aid of a computer program), the practice is known as **bootstrapping** or **case resampling**.