Regression Modeling

The list below summarizes steps which should be taken after you've preliminarily explored a regression model. The steps can be taken in any order, and can be tried repeatedly as you continue to improve your model.

1. Use sets of dummy variables to represent qualitative variables in your mode. An analysis of variance (ANOVA) will tell you if the data supports inclusion of these variables.
2. Plot the residuals against each explanatory variable. If you see a "U" (bending upwards or downwards), try adding the square of that explanatory variable to your model. Then look at "c" and "-b/(2c)" to see the nature of the nonlinearity you've captured.
3. For each explanatory variable in turn, ask yourself whether its impact on the dependent variable might vary as some other explanatory variable varies. If so, try adding the product of those two explanatory variables to your model (in order to capture a possible interaction). Interpret the regression results in terms of the "conceptual" model in which the coefficient of the first variable explicitly incorporates the second.
4. Find the sample observations with the largest positive residuals, and those with the largest (in magnitude) negative residuals. If some as-yet-not-in-your-model factor seems to differentiate the two groups, collect data on that factor and try including it as a new explanatory variable in your model.
5. Do a "model analysis," and examine any outliers that turn up. Check that the data was entered correctly. If it was, see if you can identify something "special" about the outliers (ideally, new explanatory variables which will yield a model where the observations are no longer outliers).
6. [Plot the residuals against the predicted values of the dependent variable. If the "scatter" of the residuals grows as the predicted values grow, consider using the logarithm of the dependent variable as the dependent variable in a new model.]
7. [If you suspect that the effects of the explanatory variables are "scale" effects (for example, if you think that changes in an explanatory variable are associated with percentage changes in the dependent variable, rather than additive changes), consider using the logarithms of the explanatory variables in a new model, instead of the original explanatory variables themselves.]

(The last two steps reach a bit beyond our course material. They're the next two modeling issues I'd discuss if we had more time.)