Homework 3

Gather data where there is a numerical response variable and at least four numerical potential predicting variables. An indicator variable defining membership in one of two groups counts as a numerical variable, but at least two of your numerical predictors should not be indicator variables. Any variables you might create during the analysis, such as product or squared variables, do not count to the minimum of four. There should be at least 30 observations in your sample. **Do a full and complete analysis of what is going on in the data, being sure to build your chosen model using all of the methods and principles we have discussed in class, and being sure to check assumptions.** Use corrective procedures where necessary. Discuss the implications of your results.

Your criterion in choosing a data set should be that it is reasonable to expect that the target variable might be related to the predicting variables. If it turns out that the target is not related to the predictors (after careful analysis), that’s fine — you’ve learned something (you might start out with four potential predictors in your model, and ultimately decide that you only need three, two, one, or even none of them). In this assignment, **and in all future assignments**, one of your goals should be to find what you consider the best model(s) to summarize your data; in particular, if that means excluding effects that you feel are not needed, then that is what you should do. **The number of predictors in your final model of choice has nothing to do with the quality of your analysis.** Further, in this and all future assignments, **even if you believe that there are violations of assumptions in your final model(s), you should still discuss in detail all of the implications of those models, while also (of course) noting the potential limitations of those implications.**

*Note:* It’s okay if you want to build on your analysis from Homework 2 here. If you decide to do that, you **must** treat this assignment as if it is based on new data — don’t assume that I will remember results or discussion from Homework 2. The report **must** stand alone as your analysis of the data. Also, in this and any future assignment, please do not send me a “flowchart” of the steps you plan to take (or have already taken) in your analysis so that I can confirm or correct it; it is precisely how you have decided to do your analysis, and what you think it implies, that is the point of an assignment.

You can analyze a data set that is a time series if you wish, but I don’t recommend it; if you do so you should recognize that such data have the possibility of the occurrence of
additional problems related to autocorrelation. While I would not expect you to be able to address such problems (yet), their presence could make your analysis more challenging, so I recommend that you stick with cross-sectional data for this assignment.

Remember that you may not, in this homework or any future homework, take your data from a textbook, a journal article that includes a regression analysis of the data, an online digest of data sets that have been put together for educational or research purposes (such as by faculty at a university or college) or use as regression analysis examples, including (but not limited to) the Data and Story Library (DASL), the UCLA Statistics Online Computational Resource (SOCR), StatCrunch, StatLib, RPubs, Kaggle, crowdflower, or the UC Irvine Machine Learning Repository, or a data analysis competition like the KDD Cup or the Yelp Academic data sets, and you should not use data sets that are included with implementations of software, such as being included as part of an R package. Thus, you should not try to find data by doing a Google search of “regression data” or use data from a web page with the title “Regression data sets”).

Please be sure to include a cover page for your homework that has your name on it. It needn’t have anything else on it, but it should definitely not contain any of the text of your homework.

Homework due date: November 5