STAT 215: Mini-Project 1 (Multiple Linear Regression)

Colin Reimer Dawson

Due via Blackboard by Friday 11/17

Goals

The goal of this project is to use multiple linear regression to create a predictive model in a domain of your choice, using any of the tools and methods you have learned about in class. You can choose your own topic, or use one of the suggestions below. As with Mini-Projec 0, you will document the process and write about the results in a reproducible format (i.e., RMarkdown). We have more tools at our disposal in the MLR context compared to the SLR context of MP0, so use them.

Here are a few possible directions you could go (though again, you can choose any dataset and response variable you want, provided there are at least a half dozen or so possible predictors to consider):

1. Build on MP0, predicting Life Expectancy using multiple predictors in a single model.

   (data: `CountryData.csv`, code book: `CountryData-CodeBook.txt`)

2. Use data on liberal arts colleges to "reverse-engineer" a ranking equation — i.e., try to use various variables about schools to predict their rank (or model a different response variable!).

   (data: `colleges_2014.csv`, code book: `colleges_2014-codebook.txt`)

3. Model birth weights of babies using variables that are knowable during pregnancy, to help predict when a mother and fetus are at risk of dangerously low birth weights.

(data: `low_birth_rate.csv`, code book: `low_birth_rate_codebook.txt`).

4. Anything else of interest to you!

If choosing your dataset, start with the cases and response variable that you want to be able to model (the response must be quantitative for MLR to apply), and try to find data that contains information about that response variable for a sample of cases, as well as several other potentially useful predictor variables. You do not have to gather the data yourself, although you can if you want to. If you want to do this, you should talk to me about your plans first before carrying out any data collection.

What to Turn In

You should turn in a Markdown report documenting your data exploration, the CHOOSE, FIT, ASSESS cycle. You can make use of any metrics, tools, code, etc. that we have used in class. There is not a set path of model-fitting and assessing that you must follow, but here are some things that would be good to include:

1. A brief introduction of why your chosen topic is of interest.
2. Before fitting any models, set aside a random subset of your dataset (perhaps 20% of cases) that you will use only to evaluate prediction error for your handful of "finalist" models. It should not be used for any parameter fitting or hypothesis tests.
3. Consider some quadratic (or higher-order) polynomial terms, interaction terms, etc., and interpreting these, even if you do not wind up keeping them in your final model. Don't go overboard considering all sorts of polynomials and interactions: just consider a couple that might make sense in context.
4. Some assessment of multicollinearity in your (nearly?) final model, and if appropriate, doing something about it
5. Once you have settled on a final model, report and interpret confidence and prediction intervals at a couple of representative values of the predictors.

Throughout, discuss what you are doing in text! Where possible, interpret the components of the models you are considering (particularly the "finalists"). It may not always be possible to give an intuitive interpretation of every coefficient, but try to do this when you can.