STAT 213: Project II (120 pts)

Colin Reimer Dawson

Due via Blackboard by Monday 4/18 at noon

Goals Use the U.S. liberal arts colleges data from project 1, or another dataset of your choosing (but run it by me first).

- If you use the colleges data, a suggested goal is to find a multiple linear regression model to predict college ranking using a subset of the other variables in the data as predictors.
- Consider models with interaction and quadratic terms, and transform variables as appropriate to satisfy regression conditions and/or to reduce multicollinearity.
- Select a pool of three or four models that satisfy the regression conditions and that yield good fit statistics.
- Once you have narrowed the set of models down to three or four, select one by computing their mean squared prediction error using leave-one-out cross-validation (recall that this requires re-fitting the coefficients for each data subset). If the best performing model according to leave-one-out MSPE is different from the best performing model as evaluated on the full dataset, explain what may have led to the discrepancy.
- Once you have your final model, construct confidence and prediction intervals for the rank of colleges with the same characteristics as Oberlin (do something analogous to this if you are using different data).

Include all code, plots, and results, as well as ample explanatory text guiding the reader through your model construction and selection process.

Data URL: `http://colinreimerdawson.com/data/colleges_2014.csv`
Code book URL: `http://colinreimerdawson.com/data/colleges_2014-codebook.htm`

Honor Code Unlike the labs, this project should be done individually. You are free to consult any materials you would like, but should not collaborate with other students.

Grading Rubric The grade for this project is determined by the following components:

Component	**Weight**
Completeness	90 pts
Clarity/style	30 pts

Deadline and Submission Method Turn in your writeup via Blackboard as you would a lab assignment by **Monday, 4/18 at noon**. I will try to grade all of them on Monday afternoon so that you can get them back before the exam on 4/21. Include a PDF of the compiled report, as well as the source .Rmd. If you used a dataset other than the one provided, include the data and code book as well.