More Linear Regression

We'll be doing more simple linear regression, with an open-ended focus toward interpreting the results. It should be assumed that after every step which produces a new result, you should stop and think about what the results *mean*.

If you need a refresher on what linear regression is, refer to yesterday's email on the theory of least squares and skim the relevant sections in *Applied Predictive Modeling*.

For linear regression results in particular,

- What do the coefficients mean, especially when you take into account their p-values?
- Sometimes, when you add or remove variables from a regression, the magnitudes, signs, and p-values of coefficients change significantly. Be sure to interpret these changes.
- Pay attention to how the adjusted R-squared changes (or doesn't change) as you add or remove variables from a regression. You can consider these changes to represent the associated *change in predictive power* as you adjust the model.

`States` dataset

We'll begin by studying the effect of educational expenditures on test scores.

- Load the `States` dataset from the `car` package into a variable `df` and read about it using `help(States)`.
- Try computing the correlations between the columns with `cor()`.

Visualizing correlations

You can display the correlations visually using the library `corrplot`, which you should install and load.

- Set `states_cor = cor[df[-1]]` and pass `states_cor` into `corrplot()`.
  - Why are we omitting the first column of `df`?
- Experiment with different values of the `method` parameter for `corrplot()` until you find one you like. (I like `method="pie"`.) Interpret the results.

Engineering a new feature

Sometimes, it's useful to combine existing dataset features in creative ways to form new ones.

- Add an `SAT` column defined as the sum of `SATV` and `SATM`.
- Run each of the following regressions in sequence, each time using `summary()` to inspect the coefficients, multiple R-squared statistic, and adjusted R-squared statistic. Interpret the results.
  i. SAT against pop, percent, dollars and pay
  ii. SAT against pop, dollars and pay
  iii. SAT against dollars and pay
  iv. SAT against dollars
  v. SAT against pay
  vi. percent against pop, dollars and pay.

Adding interaction terms

We can add *interaction terms* to a linear regression very easily: in the list of predictors we pass in to the `lm()` function, we can include the interaction of `var1` with `var2` by including `var1:var2` or `var1*var2`.

- What's the difference between including `var1:var2` or `var1*var2`? (*Hint:* Try regressing against nothing aside from the interaction term.)
- How much additional predictive power can you get by including well-chosen interaction terms in your regression? Which interaction terms help the most?

Regional-level analysis

We'll also sometimes want to take a step back and group some of our observations together to do data analysis at a different level.

- Aggregate at the level of regions using the `aggregate()` function. (*Hint:* Pass in `FUN=median`.)
- Compute the correlations between the resulting columns.
- How do these compare with the correlations you calculated at the state level? What do you think explains the difference?