Simple Linear Regression

(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*.)

We'll be doing more simple linear regression, with an open-ended focus toward interpreting the results. Below, be sure to interpret your numerical results after every step, paying close attention to the following:

- Sometimes, when you add or remove variables from a regression, the magnitudes, signs, and p-values of coefficients change significantly.
- Similarly, changing the model will also affect the adjusted R-squared statistic, which can be thought of as the approximate predictive power of the model.

Write a report about your interpretation of these results in a separate file and upload it to Github along with the R code.

`States` dataset

We'll begin by studying the effect of educational expenditures on test scores. In particular, we'll look at Math and Verbal subscores on the Scholastic Aptitude Test, a college admissions examination for American high school students.

- 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.

Since we're only provided with subtest scores, we can add them together to form a metric for overall educational performance.

- Add an `SAT` column defined as the sum of `SATV` and `SATM`.
  - Based on the correlation between `SATV` and `SATM`, how much would you expect our results to change if we just used a subtest score instead of the total score?

Running linear regressions

In the following, pay attention to the incremental changes in model quality (as measured by predictive power, *i.e.*, adjusted R-squared) as you add or remove variables to the set of predictors you're regressing against.

- 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`

If you were a state administrator and wanted to increase SAT scores, what would you do?

- Regress `percent` against `pop`, `dollars`, and `pay`.
  - What makes the `percent` value qualitatively different from the other numeric variables in this dataset?
  - What problems arise with predicting `percent` for extreme values of the predictor variables? (What if one of the regression coefficients were negative?)