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()`.
- 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.
   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.
- Interpret the results.

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?

Massachusetts Test Score dataset

- Load the `MCAS` dataset from the `car` package into a variable `df` and read about it using `help(MCAS)`.

Cleaning the dataset

- Find the total number of rows.
- Remove the rows with missing values, and compute the number of rows of the resulting data frame.

  - Is the number of rows appreciably smaller?
  - Anticipate some statistical problems that this naive row removal could cause in our analyses.

Preliminary analysis

We'll start out with some more simple linear regressions before moving to a slightly more advanced technique.

- Compute the correlations of `totsc4` and `totsc8` with the other features in the dataset.
  - Why do you think the correlations with `totsc8` tend to be larger than the correlations with totsc4?
- Run a regression of `totsc8` against total expenditure per student, `totday`.
  - What does this say about the effect of spending per student on standardized test scores?
- Form a new data frame `df1` by removing the non-numeric columns and `totsc4`.
- Run a regression of totsc8 against the other features using `lm(totsc8 ~ . , df1)`.
  - Should we be including `code` as a predictor? If not, remove it and see how the results change.
- Run a regression of `totsc8` against the 3 predictors with p-value < 0.01 in the above regression.
  - Is the predictive power appreciably lower?

Stepwise linear regression

In general, the problem of *feature selection* is a difficult one. We ideally want to maximize predictive power using as few features as possible, because adding redundant features actually works *against* interpretability and pulls weight away from the less-redundant ones; however, with n features, we have 2^n possible combinations of features to regress against.

The most simplistic way of solving this problem is with stepwise linear regression. It works like so:

- In *forward* linear regression, we initialize the linear model with no predictors (so the model is just a constant), and we keep adding predictors which, when added, provide the greatest incremental boost to the model quality.

- In *backward* linear regression, we start with all of the predictors added to the model and successively remove predictors which, when removed, are associated with the smallest incremental drop in model quality.
- Forward and backward linear regression can be combined for a method where predictors can both be added or removed based on how doing so affects model quality.
- Eventually, according to some statistical criterion, we reach a stopping point.

The evaluation of "model quality" is often done via the Akaike information criterion, which can intuitively be thought of as being analogous to the *entropy* of a model. (Minimizing the AIC is broadly equivalent to maximizing the entropy in a thermodynamic system.)

Before learning how to run a stepwise regression in R, briefly read about its implementation.

Use stepwise regression by writing:

```
model_init = lm(totsc8 ~ 1, df1);
model = formula(lm(totsc8 ~ ., df1))
step_reg = step(model_init, model, direction = "both")
```

- How does the resulting model differ from the model above?
- Interpret the order in which coefficients are added and removed from the stepwise model.
- What does this say about the effect of educational expenditure on student test scores for schools in the dataset?
  - Reconcile this with your earlier results.
- Repeat the above with `totsc8` replaced by `totsc4` and compare the results.

California Test Score dataset

Explore questions analogous to the ones above for the `Caschool dataset` in the `Ecdat` library, and interpret the results.

Educational effects of smaller class sizes

Open the `Star` dataset from the `Ecdat` library, restricting consideration to those students who were either in regular classes or in small classes. Explore questions analogous to those above, and interpret the results.