Class Notes

Statistical Computing & Machine Learning

Class 4

Review

- The linear model (e.g. what `lm()` does)
- A variety of questions relevant to different purposes, e.g.
  - how good will a prediction be?
  - what's the strength of an effect?
  - is there synergy between different factors?

ISL book's statement on why to study linear regression

> "Though it may seem somewhat dull compared to some of the more modern statistical learning approaches described . . . later . . . , linear regression is still a useful and widely used statistical learning method. Moreover, it serves as a good jumping-off point for newer approaches. . . . Consequently, the importance of having a good understanding of linear regression before studying more complex learning methods cannot be overstated."

Concepts from linear regression:

- Choice of explanatory variables and model term (such as interaction).
- "Degrees of freedom"
- Ease of interpretability of coefficients and their standard errors.

Small data

The regression techniques were developed in an era of small data, such as that that might be written in a lab notebook or field journal. As a result:

1. Emphasis on very simple descriptions, such as means, differences between means, simple regression.
2. Theoretical concern with details of distributions, such as the t-distribution.
3. No division into training and testing data. Data are too valuable to test! (Ironic, given the importance of replicability in the theory of the scientific method.)

As a consequence of (3), there's a great deal of concern about *assumptions*, e.g.

- linearity of $f(\mathbf{X})$
- structure of ϵ: IID — Independent and Identically Distributed
  - uncorrelated between cases
  - each is a draw from the same distribution.

Selecting model terms

The regression techniques

- Heirarchical principal
- Increase in R^2

Theory of whole-model ANOVA.

Standard measure: $\frac{\text{Explained amount}}{\text{Unexplained amount}}$

Examples:

- Standard error of mean: $\frac{\hat{\mu}}{\sigma/n}$ – note the n.
- t statistic on difference between two means: $\frac{\hat{\mu}_1-\hat{\mu}_2}{\sigma/(n-1)}$
- F statistic: $\frac{SS/df1}{SSR/df2}$
  - df1 is the number of degrees of freedom involved by the model or model term under consideration.
  - df2 is $n-(p-1)$ where p is the total degrees of freedom in the model. (I called this m in the Math 155 book.) The intercept is what the -1 is about: the intercept *can never* account for case-to-case variation.

Trade-off between eating variance and consuming degrees of freedom.

The R^2 versus p picture.

- Adjusted R^2
- Whole model ANOVA.
- ANOVA on model parts

Forward, backward and mixed selection

Create a whole bunch of model terms

- "main" effects
- "interaction" effects
- nonlinear transformations: powers, logs, sqrt, steps, ...
- categorical variables

Result: a set of k vectors that we're interested to use in our model.
Considerations:

- not all of the k vectors may pull their weight
- two or more vectors may overlap in how they eat up variance

Algorithmic approaches:

- Try all combinations, pick the best one.
  - computationally expensive/impossible 2^k possibilities
  - what's the sensitivity of the process to the choice of training data?
- "Greedy" approaches

Programming basics: Graphics

Basic functions:

1. Create a frame: `plot()`. Blank frame: `plot( , type="n")`
   - set axis limits,
2. Dots: `points(x, y)`, `pch=20`
3. Lines: `lines(x, y)` — with `NA` for line breaks
4. Polygons: `polygon(x, y)` — like lines but connects first to last.
   - fill
5. Color, size, ... `rgb(r, g, b, alpha)`, "tomato"

In-class programming activity

Day 4 activity
Drawing a histogram.