Agenda

1. Variable selection methods: Best subsets, backward elimination, forward selection, bidirectional
2. Criteria: R^2, C_p, AIC, BIC

Variable Selection Recall the problem of finding the "best" model, given a set of potential explanatory variables. How do you know what terms to include in your model? The fundamental difficulty is that if you have k possible explanatory variables, then there are 2^k possible models. Thus, while it is straightforward to simply check all the models and pick the best one, the number of models to check grows *exponentially* with respect to the number of variables, and thus this algorithm is unfeasible.

We'll be thinking about the Zagat data, with information about *Price* of a dinner in Italian restaurants in New York City, as well as the restaurant's customer ratings (measured on a scale of 0 to 30) of the *Food*, *Decor*, and *Service*, as well as whether the restaurant is located to the *East* or west of 5th Avenue.

```
nyc <- read.csv("http://www.math.smith.edu/~bbaumer/mth247/sheather/nyc.csv")
```

Four related methodologies, under the umbrella of stepwise regression:

- Best subsets: Try all the combinations and find the best model for each fixed number of explanatory variables

  ```
  # install.packages("leaps")
  require(leaps)
  best <- summary(regsubsets(Price ~ Food + Service + Decor + East, data = nyc, nbest = 1))
  with(best, data.frame(rsq, adjr2, cp, rss, bic, outmat))
  ```

  This code uses the `regsubsets()` command to find all the best subsets of different sizes. The parameter `nbest=1` tells the function to only show the single best subset of each size (best model with one predictor, best model with two, etc.).

  The second line pulls out some information from the summary of that process, to allow us to compare between the models of different sizes.

- Backward elimination: Start with the full model and iteratively remove terms that are not significant
- Forward selection: Start with nothing, and add terms in order of significance
- Stepwise or bidirectional elimininination: Start with forward selection, but prune using backward elimination

Criteria In order to stepwise algorithms to make sense, we need a criteria for determining which of two models is "better". As you might suspect, there is no universally agreed upon criteria for evaluating models.

- Adjusted R^2 We've talked about this one quite a bit in class already.
- Mallow's C_p The book prefers Mallow's C_p, probably because it can be defined in terms we have been thinking about. C_p is defined as

$$C_p = \frac{SSE_m}{MSE_k} + 2(m+1) - n$$

  Where m is the number of predictors in the model you are considering and k is the number of predictors in the full model. The smaller the C_p, the better.

- Akaike Information Criteria (AIC) AIC is equivalent to Mallow's C_p for linear regression, but is more general. AIC is the default criteria in R, so that's what we will focus on. AIC is defined as

$$AIC = 2k - 2 \cdot \ln(L)$$

where k is the number of explanatory variables, and L is the maximized value of the likelihood function for the estimated model. We're not talking explicitly about likelihood in this class, but it gets bigger the more predictors you add. So, we're trying to penalize for model size. In the linear regression case you can think of it like this,

$$AIC = 2k + n \ln(RSS)$$

The smaller the AIC, the better.

- Bayesian information criterion (BIC) BIC is similar to AIC, but puts a larger penalty on additional terms in the model. BIC is my favorite criteria. It is defined as

$$BIC = k \cdot \ln(n) - 2 \cdot \ln(L)$$

Again, for the case of linear regression we can substitute in for the likelihood and get this expression,

$$BIC = k \cdot \ln(n) + n \cdot \ln(RSS/n)$$

Again, smaller values are better.

Caveats:

- Backward Elimination: All terms in result are significant, and few models are constructed. But can throw out important variables if multicollinearity is an issue
- Forward selection: Consider more models, but can add a predictor early that could be replaced by other variables
- Word of Caution: automated methods are *not* substitutes for careful analysis. Are assumptions met? Are measurements efficient? Is it worth squeezing out a few drops of R^2? What about transformations?

Model selection lab Some code for your convenience.

```
require(mosaic)
require(Stat2Data)
data(FirstYearGPA)
head(FirstYearGPA)
# install.packages("leaps")
require(leaps)

# Reports the two best models for each number of predictors
best <- regsubsets(GPA ~ ., data=FirstYearGPA, nbest=2)
with(summary(best), data.frame(rsq, adjr2, cp, rss, outmat))

backward <- regsubsets(GPA ~ ., data=FirstYearGPA, nbest = 1, nvmax = 6, method = "backward")
summary(backward)
with(summary(backward), data.frame(cp, outmat))

forward <- regsubsets(GPA ~ ., data=FirstYearGPA, nbest = 1, nvmax = 6, method = "forward")
with(summary(forward), data.frame(cp, outmat))

stepwise <- regsubsets(GPA ~ ., data=FirstYearGPA, nbest = 1, nvmax = 6, method = "seqrep")
with(summary(stepwise), data.frame(cp, outmat))
```