

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Statistics 202: Introduction to Data Mining

Model selection

Jonathan Taylor
Department of Statistics
Stanford University

December 2, 2009

Outline

- Goals of model selection.
- Criteria to compare models.
- Model selection techniques.

Election data

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

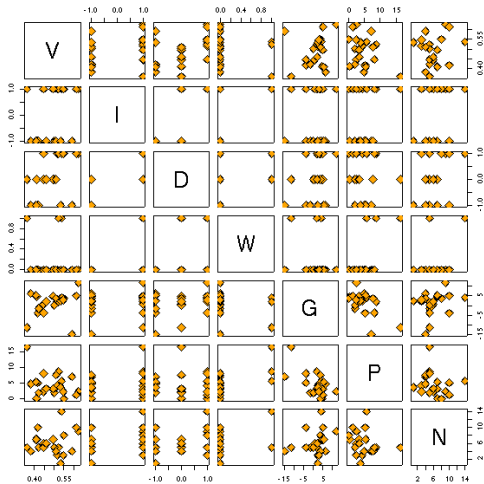
Description

Variable	Description
<i>V</i>	votes for a presidential candidate
<i>I</i>	are they incumbent?
<i>D</i>	Democrat or Republican incumbent?
<i>W</i>	wartime election?
<i>G</i>	GDP growth rate in election year
<i>P</i>	(absolute) GDP deflator growth rate
<i>N</i>	number of quarters in which GDP growth rate $> 3.2\%$

Election data

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University



R code

Model selection

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Problem & Goals

- When we have many predictors (with many possible interactions), it can be difficult to find a good model.
- Which main effects do we include?
- Which interactions do we include?
- Model selection procedures try to *simplify* / *automate* this task.
- Election data has $2^6 = 64$ different models with just main effects!

Model selection

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Strategies

- To “implement” a model selection procedure, we first need a criterion or benchmark to compare two models.
- Given a criterion, we also need a search strategy.
- With a limited number of predictors, it is possible to search all possible models (leaps in \mathbb{R}).

Model selection

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

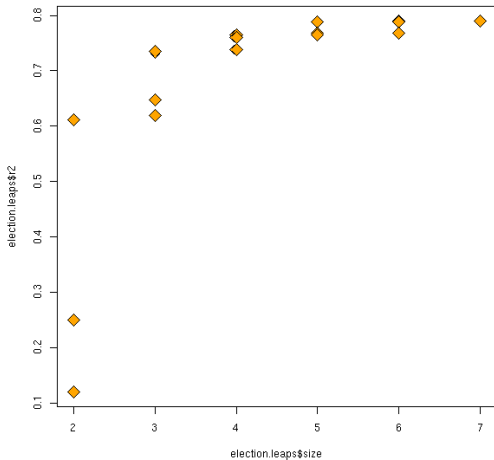
Possible criteria

- R^2 : not a good criterion. Always increase with model size
 \implies “optimum” is to take the biggest model.
- Adjusted R^2 : better. It “penalizes” bigger models.
Follows principle of parsimony / Occam’s razor.
- Mallows’s C_p – attempts to estimate a model’s predictive power, i.e. the power to predict a new observation.

Best subsets, R^2

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

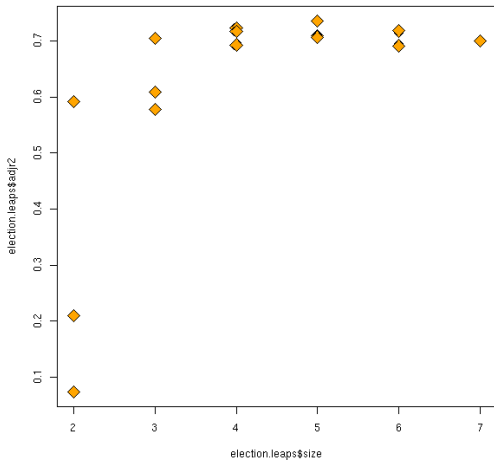


R code

Best subsets, adjusted R^2

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University



Model selection

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

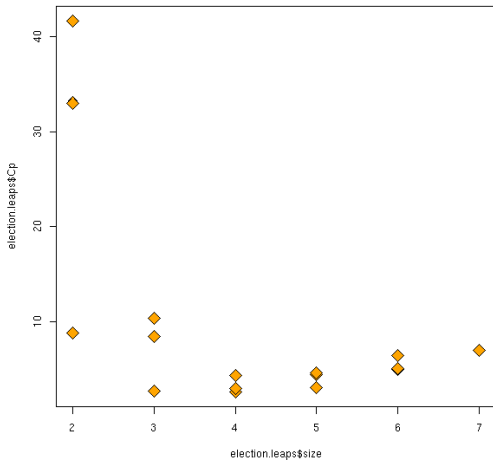
Mallow's C_p

- $$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\hat{\sigma}^2} + 2 \cdot p(\mathcal{M}) - n.$$
- $\hat{\sigma}^2 = SSE(F)/df_F$ is the “best” estimate of σ^2 we have (use the fullest model), i.e. in the election data it uses all 6 main effects.
- $SSE(\mathcal{M})$ is the SSE of the model \mathcal{M} .
- $p(\mathcal{M})$ is the number of predictors in \mathcal{M} .
- This is an estimate of the mean-squared error of $\hat{Y}(\mathcal{M})$, it takes *bias* and *variance* into account.

Best subsets, Mallows's C_p

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University



Model selection

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Search strategies

- Given a criterion, we now have to decide how we are going to search through all possible models.
- “Best subset”: search all possible models and take the one with highest R_a^2 or lowest C_p leaps
- Stepwise (forward, backward or both): useful when the number of predictors is large. Choose an initial model and be “greedy”.
- “Greedy” means always take the biggest jump (up or down) in your selected criterion.

Model selection

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Implementations in R

- “Best subset”: use the function `leaps`. Works only for multiple linear regression models.
- Stepwise: use the function `step`. Works for any model with Akaike Information Criterion (AIC). In multiple linear regression, AIC basically the same as C_p .

Model selection

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

Caveats

- Many other “criteria” have been proposed.
- Some work well for some types of data, others for different data.
- Check diagnostics!
- These criteria are not “direct measures” of predictive power, though Mallows’s C_p is a step in the right direction.
- C_p measures the quality of a model based on both *bias* and *variance* of the model. Why is this important?
- *Bias-variance* tradeoff is ubiquitous in statistics.

Bias-variance tradeoff

Comparing estimators

- When an estimator $\hat{\beta}_1$ of β_1 is unbiased:

$$E((\hat{\beta}_1 - \beta_1)^2) = \text{Var}(\hat{\beta}_1)$$

so it makes sense to compare unbiased estimators in terms of variance.

- Even for biased estimators, the LHS makes sense, called the *mean squared error* of $\hat{\beta}_1$

$$\begin{aligned} \text{MSE}(\hat{\beta}_1) &= E((\hat{\beta}_1 - \beta_1)^2) \\ &= \text{Var}(\hat{\beta}_1) + \text{Bias}(\hat{\beta}_1)^2 \end{aligned}$$

- Paradoxically, it is sometimes possible to reduce *MSE* by *biasing* the estimator.

Shrinking towards zero

Statistics 202:
Introduction
to Data
Mining

Jonathan
Taylor
Department of
Statistics
Stanford
University

