Statistics 191: Introduction to Applied Statistics

Model Selection

Jonathan Taylor
Department of Statistics
Stanford University

March 1, 2010

Topics

Outline

- Goals of model selection.
- Criteria to compare models.
- (Some) model selection.
- Bias- variance trade-off.

Election data

Jonathan Taylor
Department of Statistics
Stanford University

Description

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

Model selection

Statistics 191: Introduction to Applied Statistics

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

Jonathan Taylor
Department of Statistics
Stanford University

General comments

- This is an "unsolved" problem in statistics: there are no magic procedures to get you the "best model."
- In some sense, model selection is "data mining."
- Data miners / machine learners often work with very many predictors.
- Our model selection problem is generally at a much smaller scale than "data mining" problems.

Model selection

Hypothetical example

- Suppose we fit a a model

$$F: \quad Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \varepsilon_{n\times 1}$$

  with predictors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.
- In reality, some of the β's may be zero. Let's suppose that $\beta_{j+1} = \cdots = \beta_{p+1} = 0$.
- Then, any model that includes $\beta_0, \ldots, \beta_j$ is *correct*: which model gives the *best* estimates of $\beta_0, \ldots, \beta_j$?
- Principle of *parsimony* (i.e. Occam's razor) says that the model with *only* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j$ is "best".

Model selection

Hypothetical example: continued

- For simplicity, let's assume that $j = 1$ so there is only one coefficient to estimate.
- Then, because each model gives an *unbiased* estimate of β_1 we can compare models based on

$$\text{Var}(\widehat{\beta}_1).$$

- The best model, in terms of this variance, is the one containing only $\boldsymbol{X}_1$.
- What if we didn't know that only $\widehat{\beta}_1$ was non-zero (which we don't know in general)?

Model selection

Statistics 191: Introduction to Applied Statistics

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 `R`).

Model selection

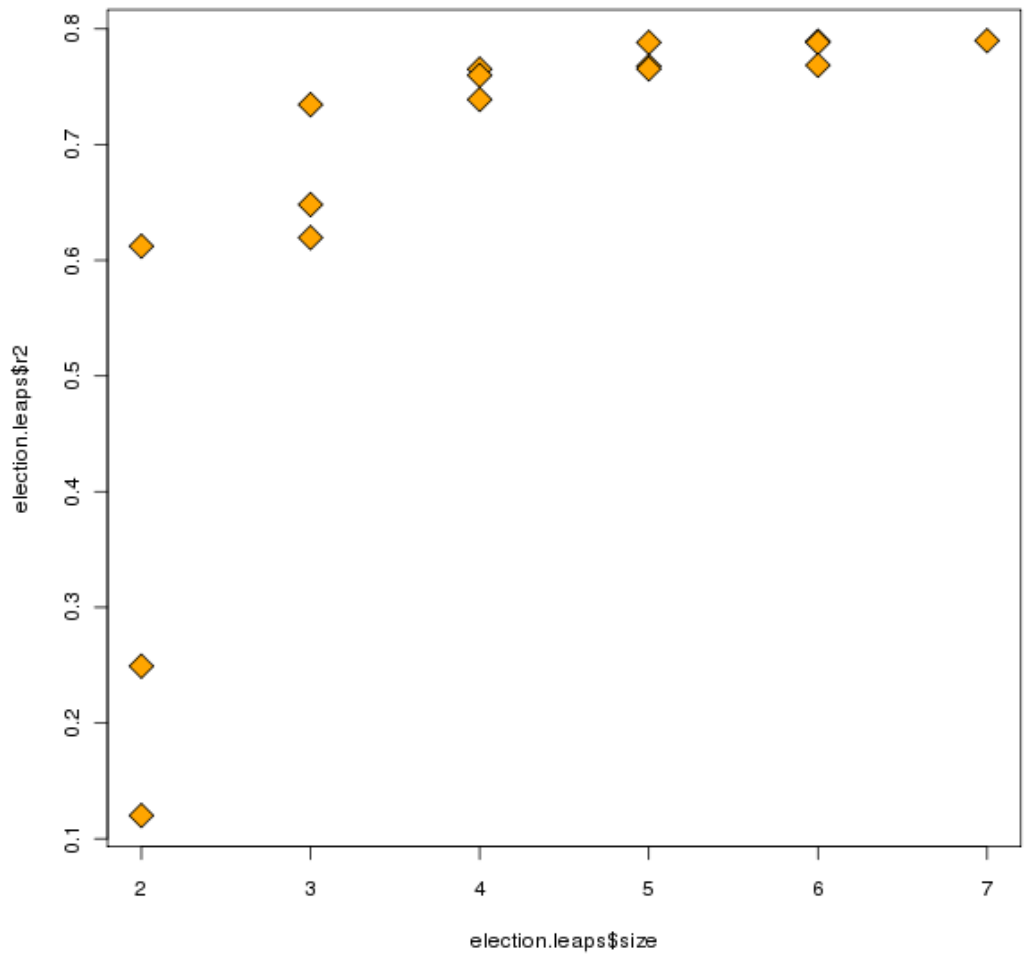
Possible criteria

- R^2: not a good criterion. Always increase with model size $\Longrightarrow$ "optimum" is to take the biggest model.
- Adjusted R^2: better. It "penalized" bigger models. Follows principle of parsimony / Occam's razor.
- Mallow'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 191: Introduction to Applied Statistics

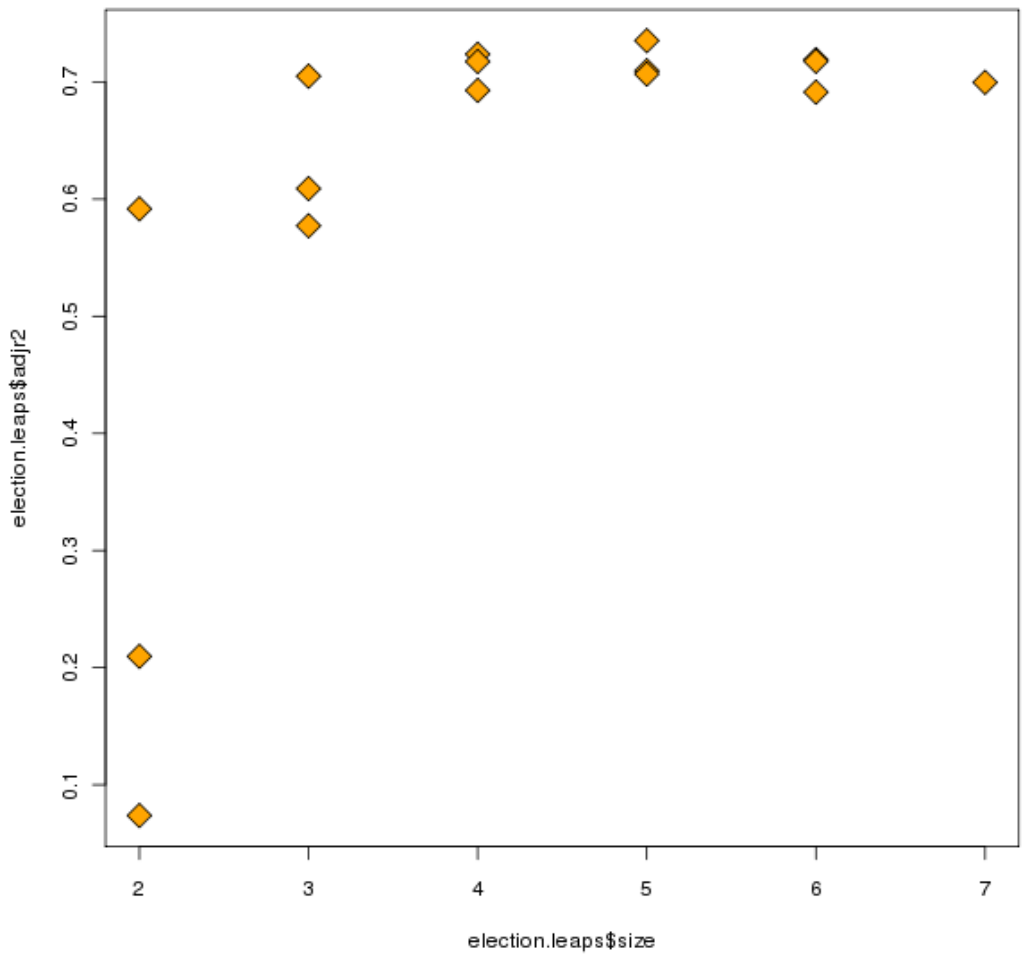
Jonathan Taylor
Department of Statistics
Stanford University



Best subsets, adjusted R^2

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Model selection

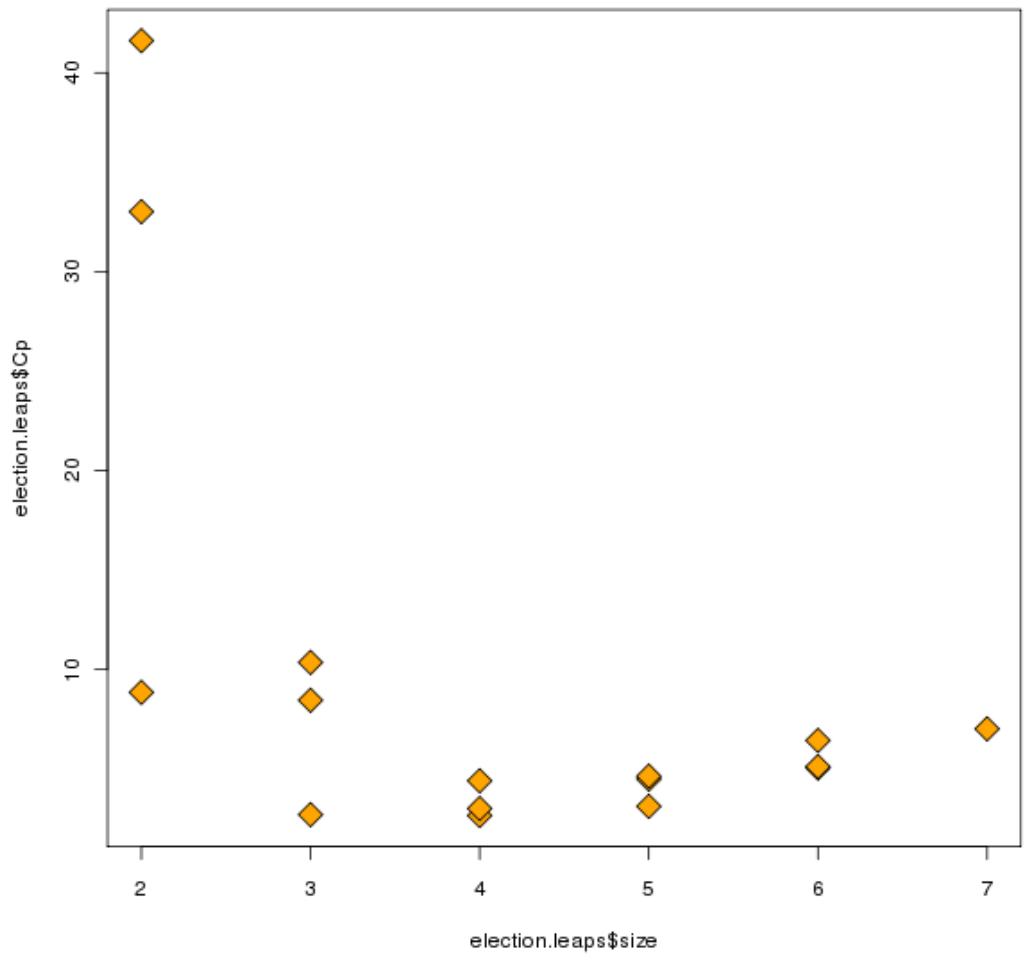
Mallow's C_p

- $$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\widehat{\sigma}^2} + 2 \cdot p(\mathcal{M}) - n.$$
- $\widehat{\sigma}^2 = SSE(F)/df_F$ 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 expected mean-squared error of $\widehat{Y}(\mathcal{M})$, it takes *bias* and *variance* into account.

Best subsets, adjusted R^2

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Model selection

Statistics 191: Introduction to Applied Statistics

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^2_a 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

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 is (almost) a linear function of C_p.

Model selection

Akaike / Bayes Information Criterion

- Akaike (AIC) defined as

$$AIC(\mathcal{M}) = -2 \log L(\mathcal{M}) + 2p(\mathcal{M})$$

  where $L(\mathcal{M})$ is the maximized likelihood of the model.
- Bayes (BIC) defined as

$$BIC(\mathcal{M}) = -2 \log L(\mathcal{M}) + \log np(\mathcal{M})$$

- Strategy can be used for whenever we have a likelihood, so this generalizes to many statistical models.
- In linear regression with unknown σ^2

$$-2 \log L(\mathcal{M}) = n \log(2\pi\widehat{\sigma}^2_{MLE}) + n$$

Model selection

Akaike / Bayes Information Criterion

- In linear regression with unknown σ^2

$$-2 \log L(\mathcal{M}) = n \log(2\pi \widehat{\sigma}^2_{MLE}) + n$$

where

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n} SSE(\widehat{\beta})$$

- In linear regression with known σ^2

$$-2 \log L(\mathcal{M}) = n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} SSE(\widehat{\beta})$$

so AIC is very much like Mallow's C_p.

Model selection

Akaike / Bayes Information Criterion

- BIC will always choose a model as small or smaller than AIC.
- As our sample size grows, we can show that
  - AIC will (asymptotically) always choose a model that contains the true model, i.e. it won't leave any variables out.
  - BIC will (asymptotically) choose exactly the right model.

Model selection

Statistics 191: Introduction to Applied Statistics

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 Mallow'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 $\widehat{\beta}_1$ of β_1 is unbiased:

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

  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 $\widehat{\beta}_1$

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

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

Bias-variance tradeoff

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

Shrinking toward zero

- Suppose we observe

$$Y_i \sim N(\mu_i, 1), 1 \leq i \leq n$$

and our goal is to estimate the entire vector μ.

- Minimum variance unbiased estimator is

$$\widehat{\mu}_i = Y_i, \qquad 1 \leq i \leq n.$$

Bias-variance tradeoff

Shrinking toward zero

- How good an estimator is $\widehat{\mu}$?

$$MSE(\widehat{\mu}, \mu) = \frac{1}{n} E(\sum_{i=1}^{n} (\widehat{\mu}_i - \mu_i)^2) = 1.$$

- *However*, we can improve on the MSE very simply by *shrinking* $\widehat{\mu}$ toward 0.
- Define

$$\widehat{\mu}_i^{\alpha} = \alpha \cdot Y_i, \qquad 1 \leq i \leq n, 0 \leq \alpha \leq 1.$$

Shrinking an estimator

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University