STATS 415: Model selection for linear regression

Prof. Liza Levina

Department of Statistics, University of Michigan

Can we improve on ordinary least squares?

- Least squares is just one way to fit a linear model; there are many others
- Briefly considered robust alternatives, such as minimizing the median absolute deviation instead of sum of squared residuals
- If we wanted a different method, what should we try to improve about OLS?
- Two considerations to keep in mind, as always:
  - Prediction accuracy
  - Model interpretability

Prediction Accuracy

- When the relationship between Y and X is linear and the number of observations n is much bigger than the number of predictors p ($n \gg p$), the OLS work well (low bias, low variance, highly interpretable) prediction accuracy
- When $n \approx p$, then the least squares fit can have high variance and may result in overfitting and poor predictions.
- When $n < p$, the OLS estimates are not unique and their variance is infinite.
- The multi-collinearity problem: unavoidable when $n < p$

  X(n by p)

linear
independent
error

Model Interpretability

- When p is large (a large number of variables X in the model), even OLS is not very interpretable
- There will be some variables that have little or no effect on X; they may mask the effect of the "important variables".
- The model would be easier to interpret if we could remove the unimportant variables, i.e. set their coefficients to zero.

Some solutions

- Subset selection (of variables X)
- Shrinkage (of coefficients $\hat{\beta}$)
- Dimension reduction (of variables X) if p is large

Subset selection

- Identifying a subset of predictors that we believe to be most related to the response Y, and then fitting the model using this subset.
- E.g. best subset selection and stepwise selection
- Advantage: end up with a small linear model
- Disadvantages:
  - Finding the best subset is a combinatorial problem (2^p possible models); greedy search algorithms offer no guarantee of actually finding "the best"
  - Inference may not be valid if you tried many models before choosing "the best".

Shrinkage

- Shrinking the estimated coefficients towards zero (in absolute value)
- Shrinkage reduces variance
- If some of the coefficients are shrunk to exactly zero, those variables can be removed. increase bias
- E.g. ridge regression and the lasso
- Advantages: efficient optimization methods exist; valid inference is being developed
- Disadvanatges: shrinkage increases bias; does not always result in variable selection.

Dimension reduction

- Projecting all p predictors into a k-dimensional space where $k < p$, and then fitting a smaller linear regression model (with k predictors)
- E.g. principle components regression, partial least squares
- Advantage: a much smaller model, faster to fit, coefficients are stable
- Disadvantages: the relationship between y and X is not taken into account when performing dimension reduction; original variables are no longer in the model, therefore interpretation is lost

One-step greedy approaches to subset selection

- General idea: test significance of predictors and either add or eliminate in some principled fashion
- Based on individual p-values
- Multiple testing is not accounted for, but ranking is more important than the absolute size of p-values
- Different methods use different rules to add/delete predictors

Backward Elimination

1. Start with all the predictors in the model
2. Remove the predictor with the highest p-value greater than α
3. Refit the model and go to step 2
4. Stop when all p-values are less than α

$\alpha > 0.05$ may be better if prediction is the goal.

Forward Selection

1. Start with no predictor variables
2. For all predictors not in the model, check the p-value if they are added to the model
3. Add the one with the smallest p-value less than α
4. Refit the model and go to step 2
5. Stop when no new predictors can be added

Stepwise regression is a combination of backward elimination and forward selection (allows to add variables back after they have been removed).

Remarks on one-step methods

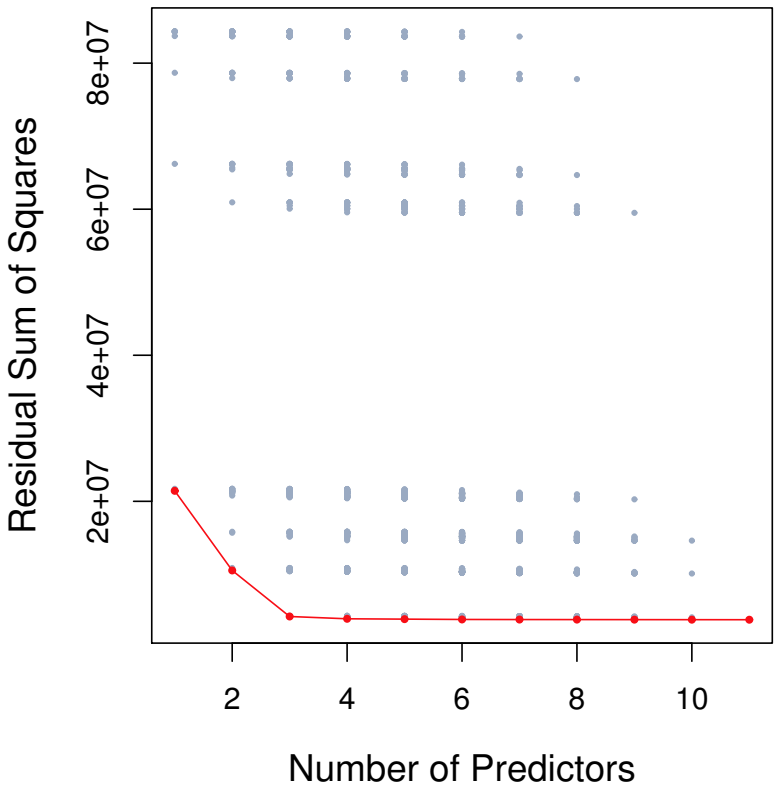
- Greedy. May miss the optimal model.
- Remember not to take p-values at face value (multiple testing).
- Variables not selected can still be correlated with the response, but they do not improve the fit enough to be included.
- Tend to pick smaller models than desirable for prediction purposes; but you can choose your own stopping criterion.

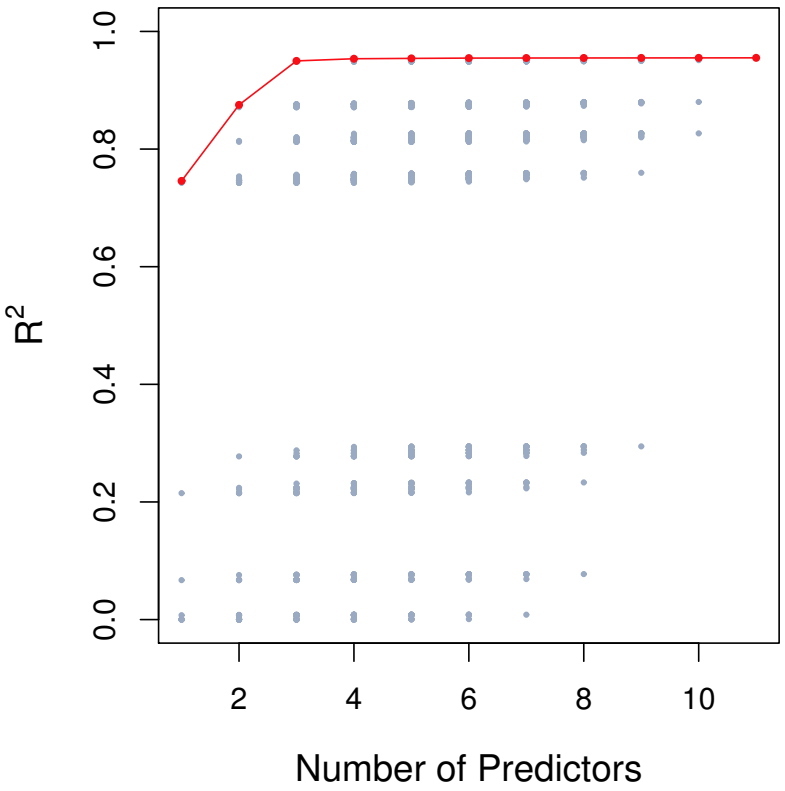
Best Subset Selection

- Suppose we can in fact run a linear regression for each possible combination of the X predictors (p is not so large that this is computationally infeasible).
- How do we judge which subset is the "best"?
- Standard measures of fit: Residual Sum of Squares (RSS), R^2.
- But: the model that includes all the variables will always have the largest R^2 and smallest RSS - they are based on the training data!

Credit Data: R^2 vs. Subset Size

- The RSS will always decrease (and R^2 increase) as the number of variables increases
- The red line tracks the best model for a given number of predictors, according to RSS and R^2





Criterion-based Model Selection

- General idea: choose the model that optimizes a criterion which balances goodness-of-fit and model complexity
- No p-values involved
- Some theoretical guarantees are available
- Different methods use different goodness-of-fit measures and different penalties for complexity (for linear regression, complexity is the number of predictors)

Criteria for Model Comparison

- Adjusted R^2
- AIC (Akaike information criterion)
- BIC (Bayesian information criterion)
- Mallows C_p
- These methods add a penalty to RSS which increases with the number of variables; i.e., they penalize model complexity.
- None are perfect.

AIC and BIC

- Consider a candidate model with d predictors, $d \leq p$
- Akaike information criterion (AIC)

σ_p^2=1/(n-p)*RSS_p

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2d\hat{\sigma}^2)$$

where $\hat{\sigma}^2$ is estimated from the model with all p predictors.

- Bayesian information criterion (BIC)

$$\text{BIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + \ln(n)d\hat{\sigma}^2)$$

- Find a model to minimize AIC or BIC, over all values of d and all subsets of variables

Mallows' C_p

- Definition:

$$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$$

- For linear regression, C_p and AIC are equivalent
- In general, they are not the same; C_p aims to estimate the test MSE of the candidate model

$$\frac{1}{n}\sum_i E(\hat{y}_i - Ey_i)^2$$

Adjusted R^2

- Recall

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

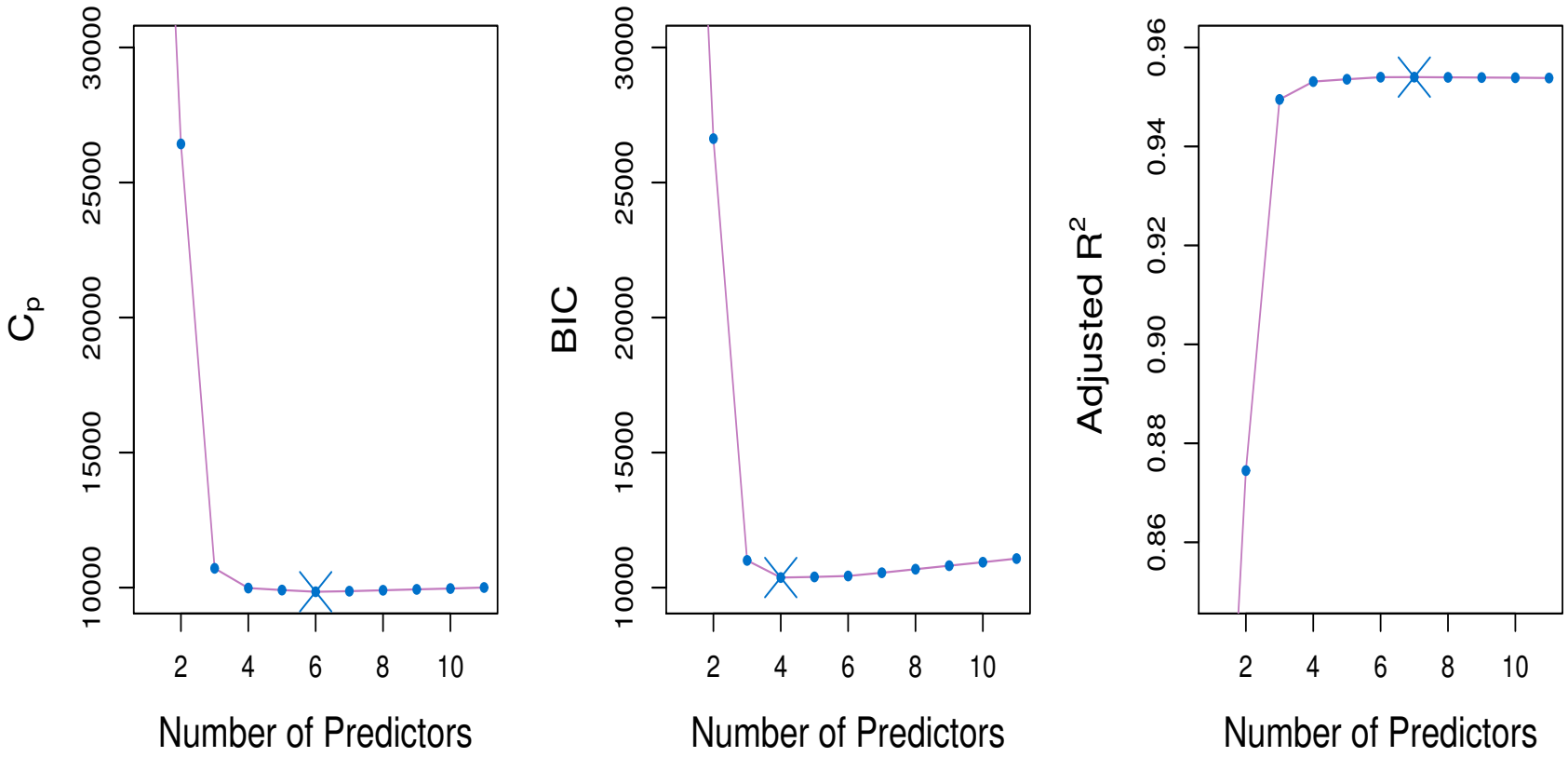
- Definition of adjusted R^2:

$$\begin{aligned} R_a^2 &= 1 - \frac{\text{RSS}/(n-(d+1))}{\text{TSS}/(n-1)} \\ &= 1 - \left(\frac{n-1}{n-(d+1)}\right)(1-R^2) \end{aligned}$$

- Adding a predictor will only increase R_a^2 only if it has predictive value.

Credit Data: AIC (C_p), BIC, and Adjusted R^2

- A small value of AIC, BIC and C_p indicates a better model.
- A large value of Adjusted R^2 indicates a better model.



Subset Selection Summary

- Generally, criterion-based methods are preferred; but they are computationally intensive for large p and so not always feasible
- In practice, a smaller range of values of d may be considered instead of all possible model sizes from 0 to p
- It may happen that several models provide very similar fit
- If models with similar fit lead to very different conclusions, the data are ambiguous
- If conclusions are similar, choose a simpler model and/or predictors that are easier to measure or interpret