FEATURE SELECTION

Applied Analytics: Frameworks and Methods 1

Outline

- Motivation behind the use of Feature Selection
- Theory-based approach
- Filter Methods
- Subset selection
- Shrinkage methods
- Dimension Reduction

Motivation for Feature Selection

- Parsimony is a desirable property for models
  - *Predictions from simple (versus complex models) are more stable across samples*
  - *Simple models are easier to interpret and communicate to stakeholders.*
- As number of predictors increases, the chance of finding correlations among a predictor or a set of predictors increases. Such correlations among predictors called *multicollinearity* inflates standard errors of coefficients, potentially leading to erroneous conclusions about the relevance of a predictor.
- In cases where p > n, traditional estimation techniques will not work.

Methods for Feature Selection

1. Theory
2. Filter Methods
3. Subset Selection
4. Shrinkage
5. Dimension reduction
6. Iterative Methods

Methods for Feature Selection

1. Theory: Use published literature or domain knowledge to pick variables.
2. Filter Methods: Keep predictors that are relevant but not redundant.
3. Subset Selection: Identify a subset of p predictors that are related to the outcome. Fit a model with this subset.
4. Shrinkage: Fit a model with all p predictors, but the estimated coefficients are shrunken towards zero relative to least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.
5. Dimension reduction: Group predictors into a reduced number of components based on similarity. Use the components as predictors.
6. Iterative Methods: Repeatedly supply predictor subsets to the model and then use the resulting model performance estimate to guide the selection of the next subset to evaluate.

PS: Above methods do not apply to predictive models (e.g., trees) which automatically pick features. Such models couple the predictor search algorithm with parameter estimation, and are therefore thought to have built-in feature selection.

THEORY

Theory

- Review published literature to determine which variables to include in the model
- Consult experts with extensive domain knowledge
- This theory-based variable selection must be done before examining the data
- This should always be the first line of attack for feature selection. It is unfortunate that with the availability of large datasets and powerful analytical techniques, this method is often overlooked.

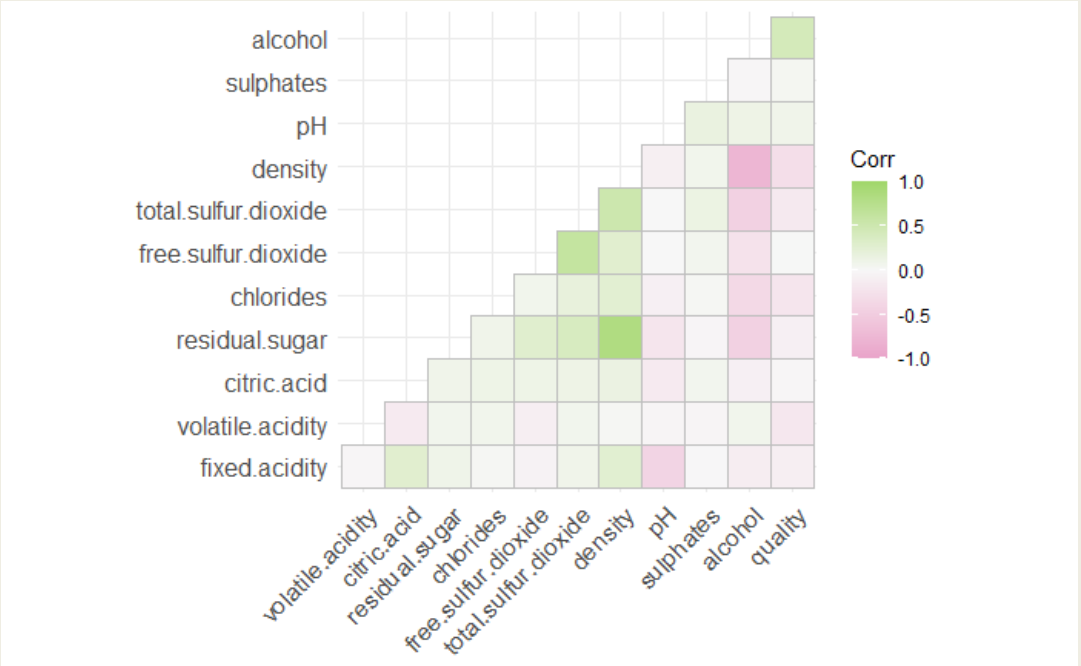
FILTER METHODS

Filter Methods

- Predictors included in a model must be
- Relevant
  - *Related to the outcome variable*
- Non-redundant
  - *Not related to other predictors*

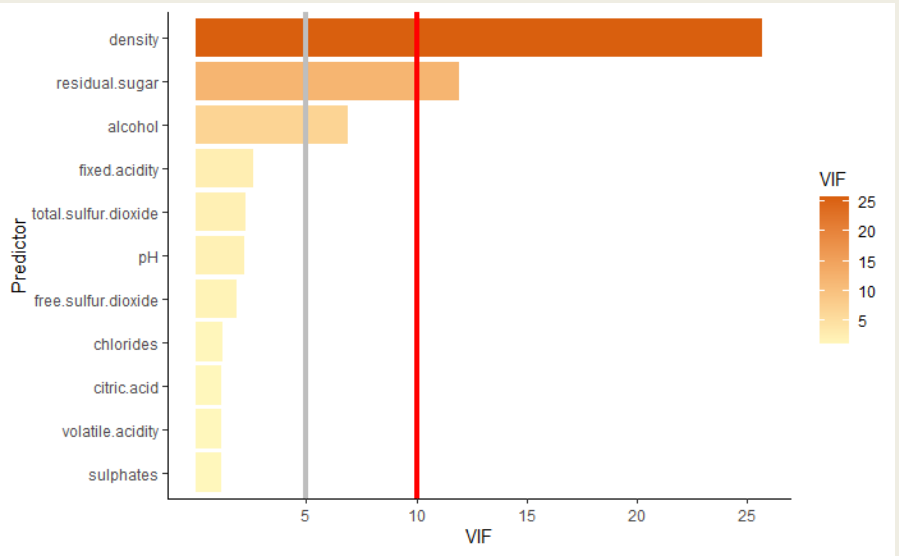
Bivariate Filter

- Examine bivariate correlation to assess relationship between pairs of variables
- Relevant
  - *High bivariate correlation between a predictor and outcome*
- Non-redundant
  - *Low bivariate correlation between a predictor and other predictors*
- In the correlation heatmap, quality is outcome and the rest are predictors.



Multivariate Filter

- Examine relevance and redundancy of all predictors together. Often, what is true in pairwise relationships may not be found when all variables are considered together.
- Relevance
  - *Significant regression coefficients for predictor in a multiple regression*
  - *However, note that this is not a sufficient condition as large sample sizes will render all coefficients statistically significant*
- Non-redundant
  - *Threat of collinearity can also come from linear relationships between sets of variables, known as multicollinearity.*
  - *Multicollinearity is measured by Variance Inflating Factor or Tolerance. 1<VIF<Inf*
    - VIF>10: Serious multicollinearity; VIF>5: Moderate-High multicollinearity.



VIF for a set of predictors of wine quality

See FeatureSelectionMethods.html for an illustration in R

SUBSET SELECTION

Subset Selection

1. Best subset selection
2. Forward
3. Backward
4. Stepwise

Evaluation Criteria

- Before examining these methods, let us look at criteria for evaluating competing models.
- Training set error will decrease as more variables are added, but the test error may not.
- Therefore, training set RSS (residual sum of squares) and training set R2 cannot be used to select from among a set of models with different number of variables.
- In order to select the best model with respect to test error, we need to estimate the test error. There are two approaches to this
  - *Indirectly estimate test error by making an adjustment to the training error to account for bias due to overfitting.*
  - *Directly estimate test error using a validation set approach or cross-validation. (will be discussed in a later topic)*

Evaluation Criteria

- Indirect estimates of test error
  - *apply a penalty to the residual sum of squares for the number of predictors used*
  - *All these indices apply a penalty to the residual sum of squares for the number of predictors used.*
  - *AIC = -2logL + 2d*
  - *BIC = 1/n(sse + log(n)dσ^2)'*
  - *C_p = 1/n(sse + 2dσ^2)*
  - *Adjusted R^2 = 1 – (sse/(n-d-1))/(sst/(n-1))*
    - where L is maximized value of likelihood function, σ^2 is the estimate of the variance of error, d is the number of parameters, sse is sum of squared errors/residuals and sst is total sum of squares.

Best Subset Selection

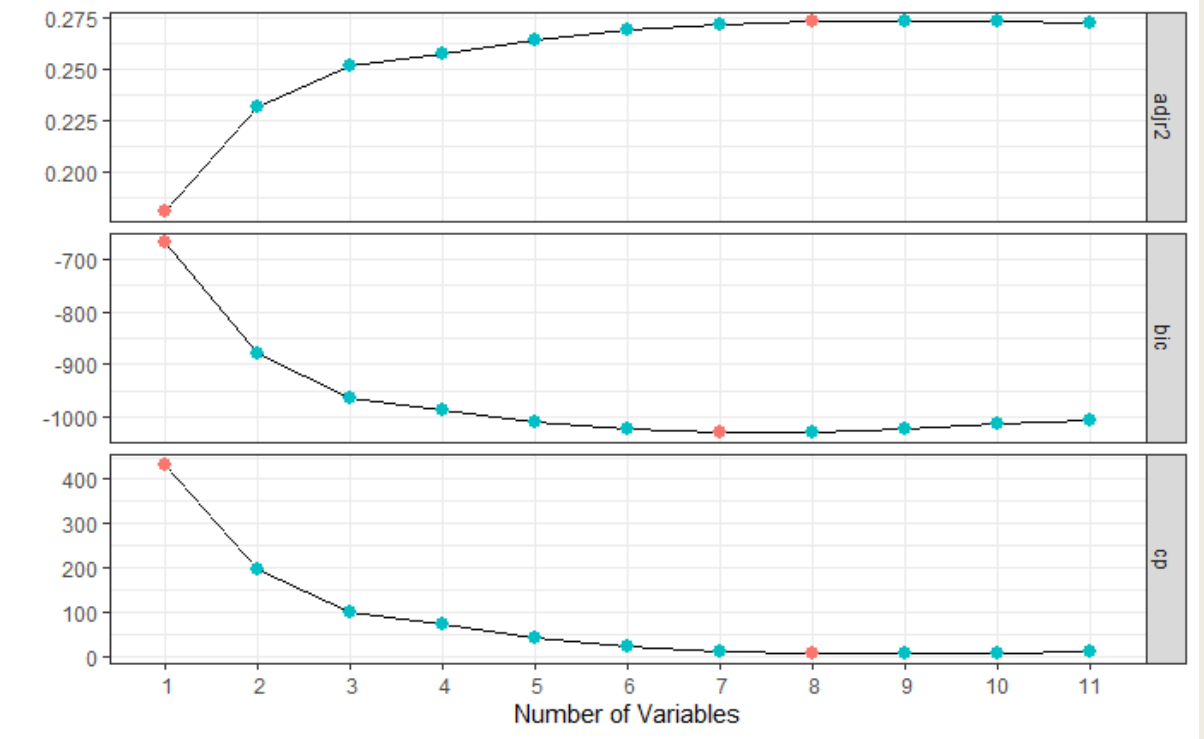
1. Consider all possible subsets of p predictors.
2. Fit a model with each subset.
3. Pick the best performing model

Best Subset Selection: Problems

- As p increases, the computational requirements grow exponentially. For p predictors number of models estimated is given by
  - $^{p}C_1 + {}^{p}C_2 + {}^{p}C_3 + {}^{p}C_4 \ldots {}^{p}C_p = 2^p$
- Larger the value of p, higher the chance of finding models that perform well on training data, even though they might not have any predictive power on future data. In other words, large p may result in overfitting and high variance of coefficient estimates.

Best Subset Selection: Results

- In general, a good model is indicated by a low value of AIC, BIC, and C_p and high value for adjusted R^2.



Forward Selection

1. Forward selection method begins with a model containing no predictors and then adds predictors to the model, one at a time
2. Variables are added in order of the marginal improvement to the model. At each stage, the variable that gives the largest marginal improvement is added.
3. Addition of variables stops when marginal improvement is not significant

Forward Selection

- Compared to Best Subsets method, this method is computationally efficient. For p predictors, number of models estimated is: $1 + p(p+1)/2$

- The intuition behind forward selection method is illustrated using a model with eight predictors. In the following slides, marginal improvement is indicated by a green-red scale where green is significant marginal improvement and red is a non-significant improvement.

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.3

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.3
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.3
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.4

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.3
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.4
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	

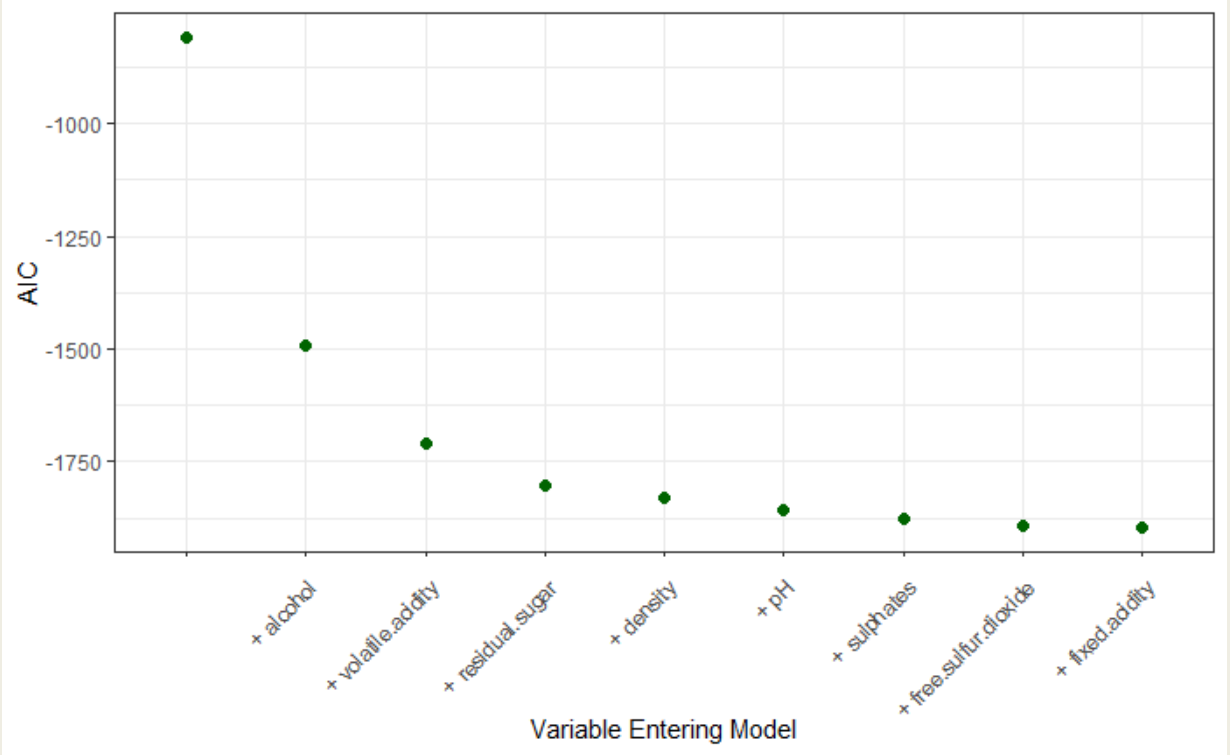
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.3
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.4
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.45

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.3
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.4
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.45
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Adj R^2
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.3
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.4
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.45
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	
Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	0.45

Forward Selection: Results

- For a set of 11 predictors, forward method added variables until marginal improvement in AIC was not significant.



Backward Selection

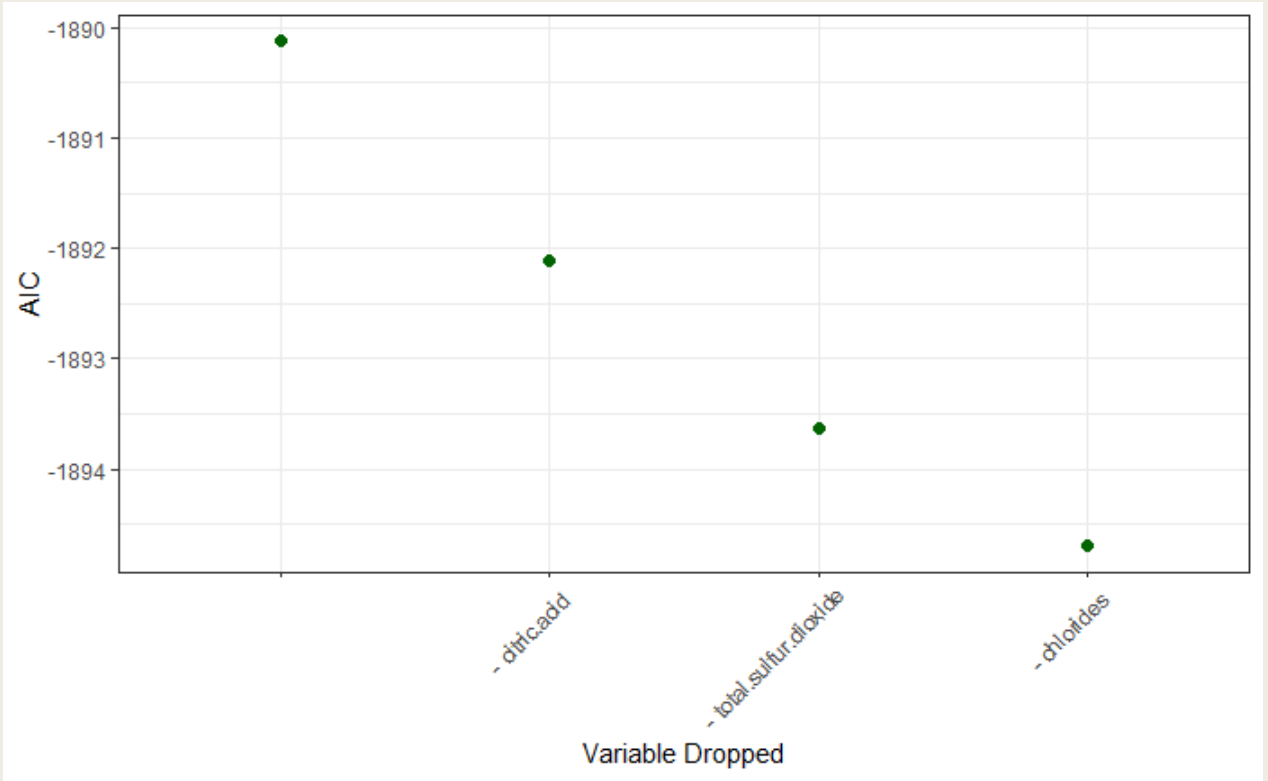
1. Backward selection method begins with a model containing all predictors and then removes predictors from the model, one at a time
2. Variables are removed iteratively with the least useful predictor being dropped first.
3. Removal of variables stops when elimination of a predictor significantly worsens the model.

Backward Selection

- Backward selection is just forward selection in reverse. Like Forward selection, For p predictors, number of models estimated is: $1 + p(p+1)/2$
- However, one limitation of backward selection is that it cannot be used for models where $p > n$. On the other hand, forward selection can handle such high-dimensional models.

Backward Selection: Results

- For a set of 11 predictors, backward method removes variables from full model until elimination of a predictor significantly worsens the model.

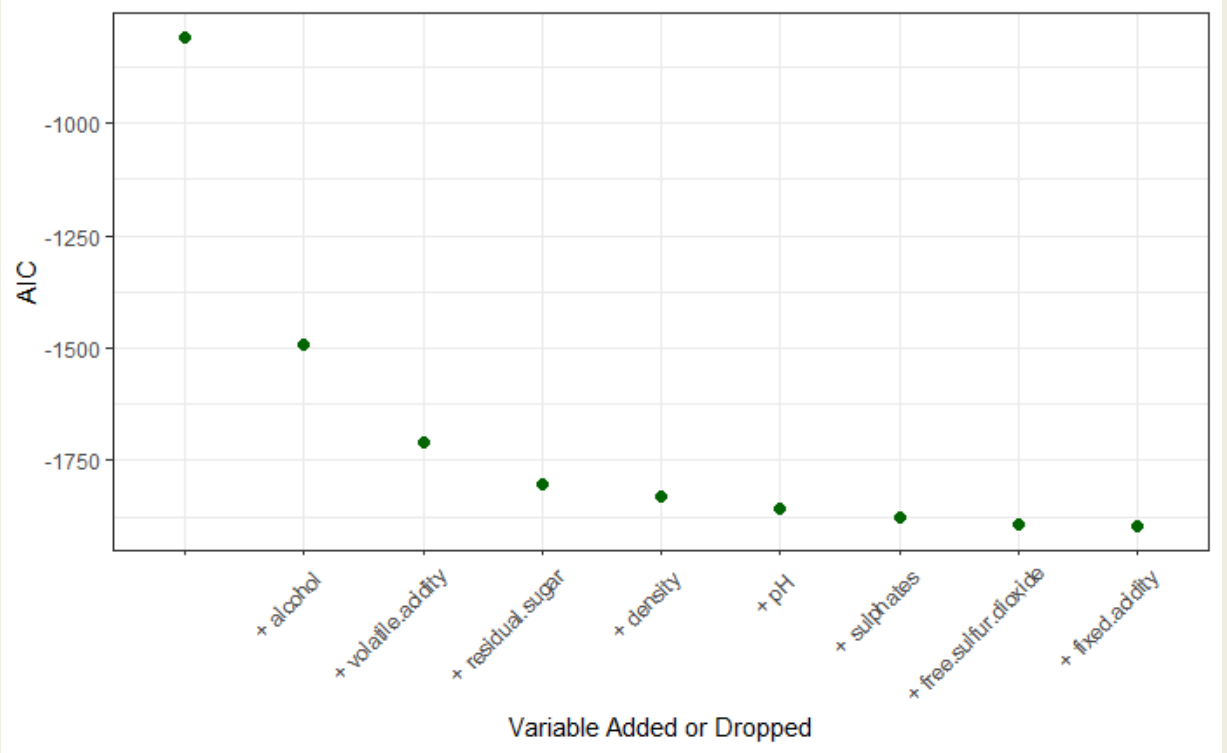


Stepwise

- Forward and Backward methods are run simultaneously
- Predictors may be added or removed at each stage until the optimal model is reached
- Stepwise method attempts to more closely mimic benefits of best subsets while retaining the computational advantages of forward and backward selection methods.

Stepwise Selection: Results

- For a set of 11 predictors, stepwise method added variables but did not drop any variables.



See FeatureSelectionMethods.html for an illustration in R

SHRINKAGE

Shrinkage Methods

- Shrinkage methods constrain or regularize coefficient estimates, or equivalently, shrink the coefficient estimates towards zero. They do so by adding a shrinkage penalty to punish a model for complexity.
- While OLS minimizes the residual sum of squares (sse) to estimate coefficients, shrinkage methods minimize the sum of squared errors (SSE), also known as residual sum of squares and a shrinkage penalty.
  - *OLS: min(SSE)*
  - *Shrinkage methods: min(SSE + λ* shrinkage penalty)*

Shrinkage Methods

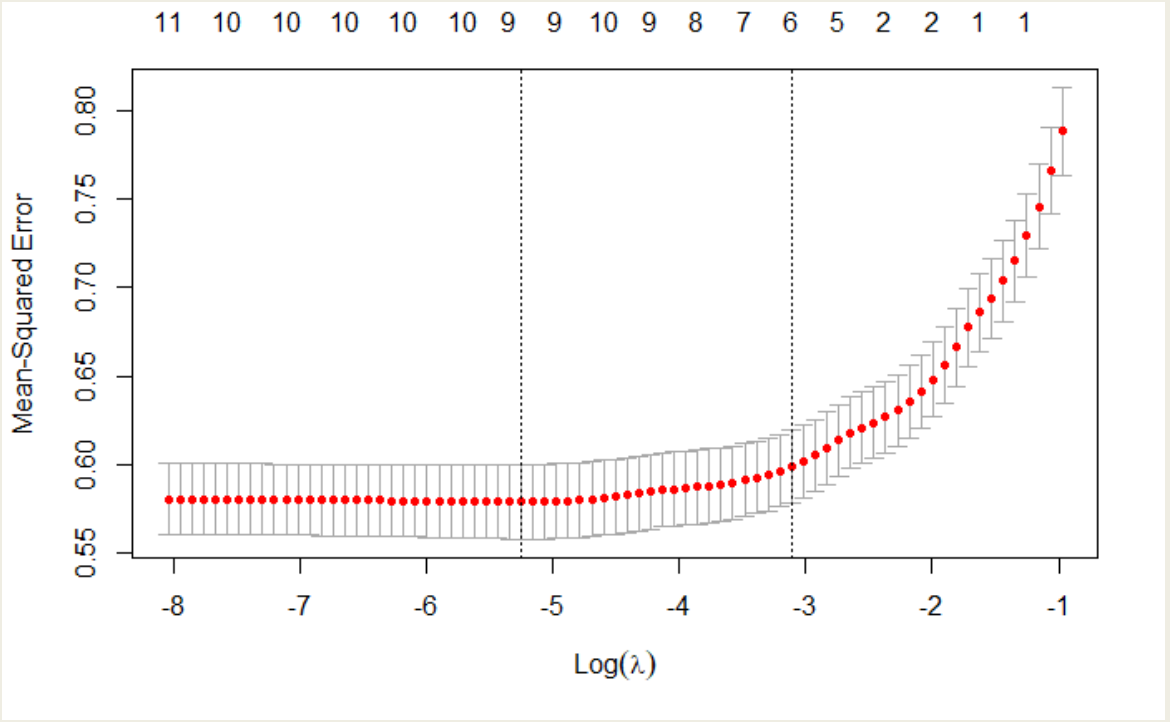
- Based on the shrinkage penalty used, there are two shrinkage methods
  - *Ridge Regression (Shrinkage penalty is $\lambda\Sigma\beta_j^2$): Shrinking coefficient estimates can significantly reduce their variance and lead to better fit. However, ridge regression will include all predictors in the model.*
  - *Lasso (Shrinkage penalty is $\lambda\Sigma|\beta_j|$): With a small modification to the shrinkage penalty from ridge regression, Lasso is able to force some of the coefficients to equal zero.*
  - *The shrinkage penalty in Ridge regression suppresses all coefficients, on the other hand, the penalty in Lasso has the effect of forcing some, not all, coefficient estimates to be exactly zero. By forcing only some coefficient estimate to zero, Lasso in essence performs feature selection.*

Lasso

- The extent of shrinkage penalty depends on the value of the tuning parameter, λ.
- When λ = 0, the penalty has no effect and Lasso will produce least square estimates.
- As λ gets larger, the penalty grows and coefficient estimates are forced to zero.
- The optimal value of λ is typically determined by examining cross-validation error of a set of λs using k-fold cross-validation.

Advantages of Lasso

- Performs better than stepwise selection
- Performs variable selection to any number of variables
- Can create models where p > n
- Statistically well founded
- Relatively fast



See FeatureSelectionMethods.html for an illustration in R

DIMENSION REDUCTION

Dimension Reduction

- p predictors are reduced to a smaller number of components based on a measure of similarity (e.g., correlation). Two such techniques are
  - *Principal Components Analysis (PCA)*
  - *Partial Least Squares (PCR)*

Principal Components Regression (PCR)

- PCA generates linear combinations of original p predictors
- A matrix with p variables will generate p components such that the first component captures the most variance, followed by the second, and so on.
- Generally, the number of components to be retained is based on eigen-value of extracted components, a scree plot, and capturing a certain amount of variance (e.g., 70%).
- The reduced number of components are used to predict the outcome instead of the original set of predictors.

Partial Least Squares

- Like PCR, PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \ldots, Z_M$ that are linear combinations of the original features, and then fits a linear model via OLS using these M new features.
- But unlike PCR, PLS identifies these new features in a supervised way – that is, it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response.
- Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

See FeatureSelectionMethods.html for an illustration in R

Summary

- We discussed the Motivation behind the use of Feature Selection, and examined six ways of carrying out feature selection
  1. *Theory-based approaches*
  2. *Manual approach*
  3. *Subset selection*
  4. *Shrinkage methods*
  5. *Dimensionality Reduction*