Lecture 16:
Linear Model Selection

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 30, 2021

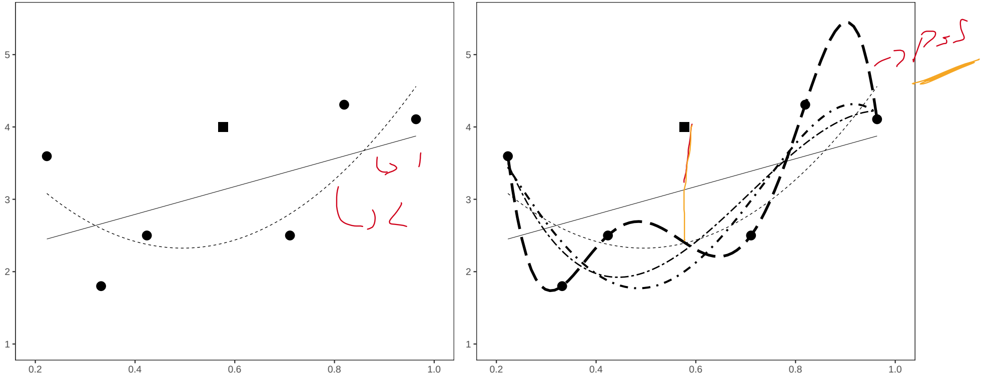
Agenda

1. Motivation
   - Recap: Overfit
2. Classical Framework for Model Selection
   - AIC: Akaike Information Criterion
   - SIC/BIC: Schwarz/Bayesian Information Criterion
   - Connection to Classical Hypothesis Testing: General
   - AIC/SIC in the linear regression model
3. Model Selection in Practice
   - Best Subset Selection
   - Stepwise Selection
4. Review & Next Steps
5. Further Readings
6. Demo

Overfit and out of Sample Prediction

- ML we care about prediction out of sample
- Overfit: complex models predict very well inside a sample but ”bad” outside
- Choose the right complexity level
- How do we measure the out of sample error?
- R^2 doesn’t work: measures prediction in sample, it’s non decreasing in complexity (PS1)

Overfit and out of Sample Prediction



Motivation

- Estimating test error: two approaches

  LOO CV

  1. We can directly estimate the test error, using either a validation set approach, or a cross-validation approach k=5 , k=10
  2. We can indirectly estimate test error by making an adjustment to the training error to account for overfitting.

     ISLR

     - AIC, BIC, C_p and Adjusted R^2

       Mallow's
     - These techniques adjust the training error for the model size, and can be used to select among a set of models with different numbers of variables.
     - I'll focus on AIC and BIC. They are intimately related to more classical notions of hypothesis testing.

       Bayesian / Schwartz

Classical Framework for Model Selection

- The framework for model selection can be described as follows.
- We have a collection of parametric models

$$\underline{f_i(x_i, \theta)} \tag{1}$$

- where $\theta \in \underline{\Theta_j}$ for $j = 1, \ldots, J$.
- Some <u>linear s</u>tructure is usually imposed on the parameter space, so typically $\Theta_j = m_j \cap \theta_J$, where m_j is a linear subspace of $\underline{\mathcal{R}}^{p_J}$ of dimension $\underline{p_j}$ and $\underline{p_1 < p_2 < \cdots < p_J}$.
- e.g.

$$y = \underbrace{X_{n \times p_j}}\beta + u \tag{2}$$

$y = X_1 \beta_1 + \beta_2 X_2 + \;—\; + \beta_J X_J + u$

AIC

- Akaike (1969) was the first to offer a unified approach to the problem of model selection.
- His point of view was to choose a model from the set $\underline{f_i}$ which performed well when evaluated on the basis of forecasting performance.
- His criterion, which has come to be called the Akaike information criterion is

$$\underline{AIC(j)} = \underline{l_j(\hat{\theta})} \boxed{- p_j} \tag{3}$$

- where $l_j(\theta)$ the log likelihood corresponding to the j model maximized over $\theta \in \Theta_j$.

$\max l(\hat{\theta}) \rightarrow \max R^2$

AIC

$$AIC(j) = \overset{\uparrow}{l_j(\hat{\theta})} \boxed{- p_j}\,\downarrow \tag{4}$$

- Akaike's model selection rule was simply to maximize AIC over the j models, that is to choose the model j^* which maximizes $AIC(j)$.
- This approach seeks to balance improvement in the fit of the model, as measured by the value of the likelihood, with a penalty term, p_j.
- Thus one often sees this and related procedures referred to as penalized likelihood methods.
- The trade-off is simply: does the improvement which comes inevitably from expanding the dimensionality of the model compensate for the increased penalty?

BIC

- Schwarz (1978) showed that while the *AIC* approach may be quite satisfactory for selecting a forecasting model
- However had the unfortunate property that it was inconsistent, in particular, as $n \rightarrow \infty$, it tended to choose too large a model with positive probability.
- Schwarz (1978) formalized the model selection problem from a Bayesian standpoint:

$$\overset{BIC}{SIC}(j) = l_j(\hat{\theta}) - \frac{1}{2}p_j log(n) \tag{5}$$

- It has the property that as $n \rightarrow \infty$, presuming that there was a true model, j^*, then $\hat{j} = argmax\ SIC(j)$, satisfied

$$\boxed{p(\hat{j} = j^*) \rightarrow 1} \tag{6}$$

$\hat{j}$ converge al j^*

AIC vs BIC

$$AIC(j) = l_j(\hat{\theta}) \underbrace{- p_j}_{} \tag{7}$$

$$SIC(j) = l_j(\hat{\theta}) - \underbrace{p_j \frac{1}{2} log(n)}_{} \tag{8}$$

- Note that

$$\underbrace{\frac{1}{2} log(n) > 1}_{} \; for \;\; n > 8 \tag{9}$$

- The SIC penalty is larger than the AIC penalty,
- SIC tends to pick a smaller model.
- In effect, by letting the penalty tend to infinity slowly with n, we eliminate the tendency of AIC to choose too large a model.

Connection to Classical Hypothesis Testing: General

- Recall the likelihood ratio tests, that we classically use to assess goodness of fit/ compare models.

- Suppose that we are comparing a larger model j to a smaller model i

$j > i$

LR

$$T_n = 2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i)) \qquad (10)$$

$$T_1 = 2\left(\frac{\log \mathcal{L}_j}{\log \mathcal{L}_i}\right)$$

- It can be shown that $T_n \to \chi^2_{p_j - p_i}$ for $p_j > p_i = p^*$.

- So classical hypothesis testing would suggest that we should reject an hypothesized smaller model i, in favor of a larger model j iff T_n exceeds an appropriately chosen critical value from the $\chi^2_{p_j - p_i}$ table

H_0 i True

H_A j True

T_n ? CV

Connection to AIC

AIC chooses j over i, iff

j $\quad l_j - p_j$

i $\quad l_i - p_i$

$$l_j(\hat{\theta}) - p_j > l_i(\hat{\theta}) - p_i \tag{11}$$

$$l_j(\hat{\theta}) - l_i(\hat{\theta}) > p_j - p_i \tag{12}$$

$$\frac{T_n}{p_j - p_i}$$

$$2\frac{l_j(\hat{\theta}) - l_i(\hat{\theta})}{p_j - p_i} > 2 \tag{13}$$

$$\frac{T_n}{p_j - p_i} > 2$$

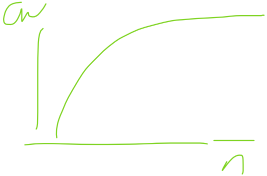
Connection to SIC

$SIC(j) = l_j - p_j \frac{1}{2} \log(n)$

In contrast Schwarz would choose j over i, iff

$$\frac{\overbrace{2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))}^{T_n}}{p_j - p_i} > log(n) \qquad (14)$$

Then $log(n)$ can be interpreted as an implicit critical value for the model selection decision based on SIC



AIC/SIC in the linear regression model

Recall that for the for the Normal/Gaussian linear regression model the log likelihood function is

$$l(\beta,\sigma^2) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\overset{u'}{(y - X\beta)'}\overset{u}{(y - X\beta)} \tag{15}$$

evaluating at $\hat{\beta}$, and $\hat{\sigma}^2 = \underset{\hat{u}'}{(y - X\hat{\beta})'}\underset{\hat{u}}{(y - X\hat{\beta})}$ we get the concentrated/profile log-likelihood

$\hat{\beta} = (X'X)^{-1}X'y$

$$l(\hat{\beta},\hat{\sigma}^2) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\hat{\sigma}^2) - \frac{n}{2} \tag{16}$$

$$= cons - \frac{1}{2} \log(\underline{\hat{\sigma}^2})$$

AIC/SIC in the linear regression model

Thus maximizing SIC

$$l_i - \frac{1}{2}p_i log(n) \tag{17}$$

is equivalent to minimize

$$\frac{n}{2}log(\hat{\sigma}_i^2) + \frac{1}{2}p_i log(n) \tag{18}$$

Penalty

or minimizing

$$log(\hat{\sigma}_i^2) + \boxed{\frac{p_i}{n}log(n)} \tag{19}$$

- Similarity for AIC
- When using software is important to check what is being computed. In `R`, the function `AIC` minimizes and not maximizes, and defines AIC as $-2l_i + kp_i$ with $k = 2$ as default that can be changed,e.g. $k = log(n)$ gives SIC

Comparison LR, t, AIC, BIC in the linear regression model

$y_i = X_1\beta_1 + u$

$y_j = X_1\beta_1 + X_2\beta_2 + u$

Example of adding one more covariate $p_j - p_i = 1$

$$\underline{T_n} = 2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i)) \underset{\frown}{\to} \underline{\chi^2_{p_j-p_i}} \qquad p_j - p_i = 1 \tag{20}$$

$$\frac{T_n}{p_j - p_i} \to \frac{\chi^2_{p_j-p_i}}{p_j - p_i} \approx F_{p_j-p_i,\infty} \tag{21}$$

$\chi^2_{d_1}/d_1 \; / \; \underline{\chi^2_\infty/\infty}$

$\frac{\chi^2_\infty}{\infty} = 1$

$\lim_{n\to\infty} \frac{1}{n}\chi^2_n = 1$

$\frac{T_n}{p_j - p_i} \approx F_{p_j-p,\underline{\infty}}$

$\chi^2 \to N^2$

$F_{d_1,d_2} = \frac{\chi^2_{d_1}/d_1}{\chi^2_{d_2}/d_2}$

HW

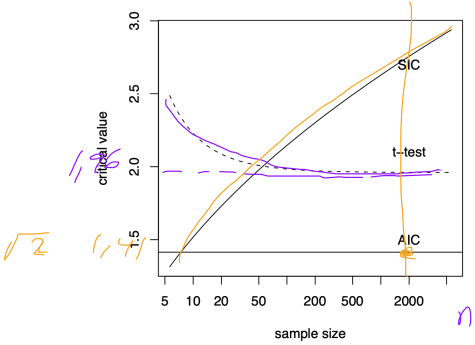
Comparison LR, t, AIC, BIC in the linear regression model

$P_j - P_i = 1$

$$\sqrt{2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))} \to \sqrt{F_{1,\infty}} = \underline{t_\infty} \quad \mathcal{N}$$

$$\underline{\sqrt{2l_j(\hat{\theta}) - l_i(\hat{\theta})} > \sqrt{2}}$$

$$\sqrt{2(l_j(\hat{\theta}_j) - l_i(\hat{\theta}_i))} > \sqrt{log(n)}$$



Model Selection in ~~Practice~~ ML Practice

- We have M_k models
- We want to find the model that best predicts out of sample
- We have a number of ways to go about it
  - Best Subset Selection
  - Stepwise Selection
    - Forward selection
    - Backward selection

$$y_1 = \beta_1 x_1 + \beta_2 x_2 + u$$

$$y = \beta_1 x_1 + u$$

$$y_1 = \beta_1 x_1 + \beta_2 x_1 + \beta_3 x_3 + u$$

$$y_1 = \beta_1 x_1 + \beta_2 x_2 + u$$

$$y_1 = \beta_1 x_1 + \beta_3 x_3 + u$$

Best Subset Selection

$y^{(0)} = \beta + u \Rightarrow \hat{\beta} = \frac{\sum y_i}{n}$

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

$\binom{10}{1} = 10 \quad \binom{10}{2} = 45 \quad 2^p = 1,024$

$2^{20} = 1$ millon $\quad 2^{40} = 1$ billon

$p = 10$

$p = 20$

2. For $k = 1, 2, \ldots, p$:
   1. Fit all $\binom{p}{k}$ models that contain exactly k predictors
   2. Pick the best among these $\binom{p}{k}$ models, and call it M_k. Where *best* is the one with the smallest SSR

R^2 max

$M_0, M_1, M_2, \ldots, M_p$

3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error. (ISLR suggest AIC (C_p), BIC, or adjusted R^2)

Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large p.
- Best subset selection may also suffer from statistical problems when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates.
- For both of these reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

Forward Stepwise Selection

$M_0 \rightarrow 0$ predict $\quad y = \beta_0 + u$

$M_1 \rightarrow 1$ predicto $\quad y = \beta_0 + \beta_1 X_1 + u$ ✓

$M_2 \rightarrow 2$ predict $\quad y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.

- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.

Forward Stepwise Selection

$p = 20$ → 1 millon

$p = 20$ → $1 + \frac{(p+1)p}{2} = 211$

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 0, 1, \ldots, p-1$:

   1. Consider all $p - k$ models that augment the predictors in M_k with one additional predictor.

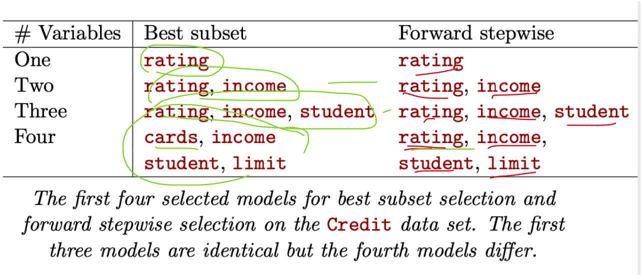
   2. Choose the best among these $p - k$ models, and call it M_{k+1}. Where *best* is the one with the smallest SSR max R^2

3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error. (ISLR suggest AIC (C_p), BIC, or adjusted R^2)

$M_0, \ldots, M_p$ → M^*

Forward Stepwise Selection

- Computational advantage over best subset selection is clear.
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
- ISLR Example



# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income student, limit	rating, income, student, limit

The first four selected models for best subset selection and forward stepwise selection on the Credit *data set. The first three models are identical but the fourth models differ.*

Backward Stepwise Selection

- Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection.

- However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Backward Stepwise Selection

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = p, p-1, \ldots, 1$:

   1. Consider all k models that contains all but one of the predictors in M_k, for a total of $k-1$ predictors

   2. Choose the best among these k models, and call it M_{k-1}. Where *best* is the one with the smallest SSR

      not R^2

3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error. (ISLR suggest AIC (C_p), BIC, or adjusted R^2)

Backward Stepwise Selection

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection.
- Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors.
- Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

FSLR

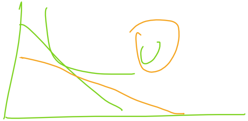
$n < p \rightarrow$? $X_{n x p}$

$\rightarrow$ Regularización

Validation and Cross-Validation

- Each of the procedures returns a sequence of models M_k indexed by model size $k = 0, 1, 2, \dots$.
- Our job here is to select $\hat{k}$. Once selected, we will return model $M_{\hat{k}}$
- We compute the validation set error or the cross-validation error for each model M_k under consideration, and then select the k for which the resulting estimated test error is smallest.
- This procedure has an advantage relative to AIC (C_p), BIC, and adjusted R^2, in that it provides a direct estimate of the test error, and doesn't require an estimate of the error variance σ^2
- It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2

Review & Next Steps



- Today:
  - Basic Classical Framework for Model Selection AIC, SIC/BIC
  - Model Selection in Practice
    - Best Subset Selection
    - Stepwise Selection
- Next class after the break: Regularization/Shrinkage Methods
- Remember to work on your proposals

Further Readings

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Koenker, R. (2013) Economics 508: Lecture 4. Model Selection and Fishing for Significance. Mimeo

Demo: Prelogomenon

```r
#Load the required packages
library("McSpatial") #loads the package
library("dplyr") #for data wrangling
library("caret") #ML
data(matchdata) #loads the data set
matchdata <- matchdata %>% mutate(price=exp(lnprice)) %>% select(-lnprice)
#transforms log prices to standard prices
set.seed(123) #set the seed for replication purposes
str(matchdata) #compact display
```

```
## 'data.frame':    3204 obs. of  18 variables:
##  $ year     : num  1995 1995 1995 2005 1995 ...
##  $ lnland   : num  8.23 8.63 8.7 8.63 8.63 ...
##  $ lnbldg   : num  6.98 7.02 7.22 6.87 7.2 ...
##  $ rooms    : int  5 5 5 4 6 7 7 6 4 6 ...
##  $ bedrooms : int  3 3 3 2 3 3 4 3 2 3 ...
##  $ bathrooms: num  1 1.5 1.5 1 1 2 1 2 1.5 1.5 ...
##  $ centair  : int  0 1 1 0 1 0 1 1 1 1 ...
##  $ fireplace: int  0 0 0 0 0 1 0 0 0 0 ...
##  $ brick    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ garage1  : num  1 0 0 1 0 1 0 0 1 0 ...
##  $ garage2  : num  0 1 1 0 1 0 1 0 0 0 ...
##  $ dcbd     : num  13.6 13.5 13.6 13.5 13.6 ...
##  $ rr       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ yrbuilt  : num  1953 1952 1952 1949 1953 ...
##  $ carea    : Factor w/ 11 levels "Albany Park",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ latitude : num  42 42 42 42 42 ...
```

Demo: K-fold CV

```
model2 <- train(price ~ bedrooms,   # model to fit
                data = matchdata,
                trControl = trainControl(method = "cv", number = 10),
                # Method: crossvalidation, 10 folds
                method = "lm")# specifying regression model

model2
```

```
## Linear Regression
##
## 3204 samples
##    1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2883, 2884, 2884, 2883, 2884, 2885, ...
## Resampling results:
##
##   RMSE      Rsquared    MAE
##   147308.8  0.01385314  122117.8
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Demo: K-fold CV

```r
model2 <- train(price ~ bedrooms+bathrooms+ rooms+centair+fireplace+brick+
                        lnland+lnbldg+garage1+garage2+
                        dcbd+ rr +
                        yrbuilt+ factor(year) +
                        carea+ latitude+longitude,
                        data = matchdata,
                        trControl = trainControl(method = "cv", number = 10),
                        method = "lm")

model2
```

```
## Linear Regression
##
## 3204 samples
##   17 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2883, 2884, 2884, 2883, 2883, 2884, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   74297.54  0.7490674  48170.37
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Model Subset Selection

- We have M_k models
- We want to find the model that best predicts out of sample
- We have a number of ways to go about it
  - Best Subset Selection
  - Stepwise Selection
    - Forward selection
    - Backward selection

Best Subset Selection

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + u \tag{22}$$

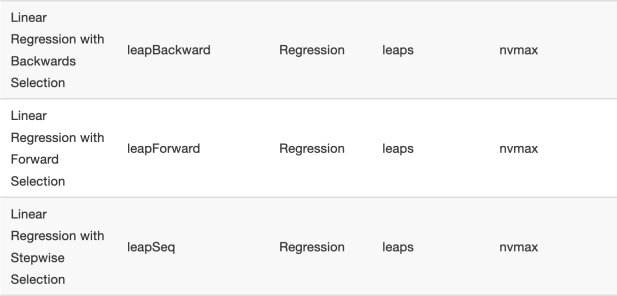
1. Estimate **all** possible models with $k = 0, 1, ..., p$ predictors.

2. Compute the prediction error using cross validation

3. Pick the one with the smallest prediction error

Demo: Best Subset Selection

- `Caret` is usually the solution but it doesn’t have everything =(



Linear Regression with Backwards Selection	leapBackward	Regression	leaps	nvmax
Linear Regression with Forward Selection	leapForward	Regression	leaps	nvmax
Linear Regression with Stepwise Selection	leapSeq	Regression	leaps	nvmax

Note: https://topepo.github.io/caret/available-models.html

Demo: Best Subset Selection

```
require("leaps")
```

```
class(matchdata$carea)
```

```
## [1] "factor"
```

```
class(matchdata$year)
```

```
## [1] "numeric"
```

```
matchdata$year<-factor(matchdata$year)
```

```
best<-regsubsets(price ~ ., method="exhaustive",data = matchdata)
summary(best)
```

Demo: Best Subset Selection

```
## Subset selection object
## Call: regsubsets.formula(price ~ ., method = "exhaustive", data = matchdata)
## 26 Variables  (and intercept)
## ...
## Selection Algorithm: exhaustive
##          year2005 lnland lnbldg rooms bedrooms bathrooms centair fireplace
## 1  ( 1 ) "*"      " "    " "    " "   " "      " "       " "     " "      
## 2  ( 1 ) "*"      " "    "*"    " "   " "      " "       " "     " "      
## 3  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "      
## 4  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "      
## 5  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "      
## 6  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "      
## 7  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "      
## 8  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     "*"      
##          brick garage1 garage2 dcbd rr  yrbuilt careaEdgewater careaEdison Park
## 1  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "             
## 2  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "             
## 3  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "             
## 4  ( 1 ) " "   " "     " "     " "  " " " "     "*"            " "             
## 5  ( 1 ) " "   " "     " "     " "  " " " "     "*"            " "             
## 6  ( 1 ) " "   " "     " "     " "  " " " "     "*"            " "             
## 7  ( 1 ) " "   " "     " "     " "  " " " "     "*"            " "             
## 8  ( 1 ) " "   " "     " "     " "  " " " "     "*"            " "             
```

Demo: Best Subset Selection

```
##          careaForest Glen careaJefferson Park careaLincoln Square
## 1  ( 1 ) " "              " "                 " "                
## 2  ( 1 ) " "              " "                 " "                
## 3  ( 1 ) " "              " "                 " "                
## 4  ( 1 ) " "              " "                 " "                
## 5  ( 1 ) "*"              " "                 " "                
## 6  ( 1 ) "*"              " "                 " "                
## 7  ( 1 ) "*"              " "                 "*"                
## 8  ( 1 ) "*"              " "                 "*"                
##          careaNorth Park careaNorwood Park careaRogers Park careaUptown
## 1  ( 1 ) " "             " "               " "              " "        
## 2  ( 1 ) " "             " "               " "              " "        
## 3  ( 1 ) " "             " "               " "              " "        
## 4  ( 1 ) " "             " "               " "              " "        
## 5  ( 1 ) " "             " "               " "              " "        
## 6  ( 1 ) " "             " "               " "              "*"        
## 7  ( 1 ) " "             " "               " "              "*"        
## 8  ( 1 ) " "             " "               " "              "*"        
##          careaWest Ridge latitude longitude
## 1  ( 1 ) " "             " "      " "      
## 2  ( 1 ) " "             " "      " "      
## 3  ( 1 ) " "             " "      " "      
## 4  ( 1 ) " "             " "      " "      
## 5  ( 1 ) " "             " "      " "      
## 6  ( 1 ) " "             " "      " "      
## 7  ( 1 ) " "             " "      " "      
## 8  ( 1 ) " "             " "      " "      
```

Stepwise Selection

1. Forward Stepwise Selection
   - Start with no predictors
   - Test all models with 1 predictor. Choose the one with smallest prediction error using cross validation
   - Add 1 predictor at a time, without taking away.
   - Of the p+1 models, choose the one with smallest prediction error using cross validation
2. Backward Stepwise Selection
   - Same idea but start with a complete model and go backwards, taking one at a time.

Demo Stepwise Selection

```r
forward <- train(price ~ ., data = matchdata,
                 method = "leapForward",
                 trControl = trainControl(method = "cv", number = 10))
forward
```

```
## Linear Regression with Forward Selection
##
## 3204 samples
##   17 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2884, 2884, 2884, 2883, 2883, 2883, ...
## Resampling results across tuning parameters:
##
##   nvmax  RMSE      Rsquared   MAE
##   2      84219.28  0.6768962  55184.58
##   3      79212.49  0.7147519  51009.82
##   4      78595.00  0.7193113  50631.17
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was nvmax = 4.
```

Demo Stepwise Selection

```r
summary(forward$finalModel)
```

```
## Subset selection object
## 26 Variables  (and intercept)
##                      Forced in Forced out
## ...

## 1 subsets of each size up to 4
## Selection Algorithm: forward
##          year2005 lnland lnbldg rooms bedrooms bathrooms centair fireplace
## 1  ( 1 ) "*"      " "    " "    " "   " "      " "       " "     " "
## 2  ( 1 ) "*"      " "    "*"    " "   " "      " "       " "     " "
## 3  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "
## 4  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "
##          brick garage1 garage2 dcbd rr  yrbuilt careaEdgewater careaEdison Park
## 1  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "
## 2  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "
## 3  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "
## 4  ( 1 ) " "   " "     " "     " "  " " " "     "*"            " "
##          careaForest Glen careaJefferson Park careaLincoln Square
## 1  ( 1 ) " "              " "                 " "
## 2  ( 1 ) " "              " "                 " "
## 3  ( 1 ) " "              " "                 " "
## 4  ( 1 ) " "              " "                 " "
##          careaNorth Park careaNorwood Park careaRogers Park careaUptown
## 1  ( 1 ) " "             " "               " "              " "
## 2  ( 1 ) " "             " "               " "              " "
## 3  ( 1 ) " "             " "               " "              " "
## 4  ( 1 ) " "             " "               " "              " "
##          careaWest Ridge latitude longitude
```

Demo Stepwise Selection

```r
backwards <- train(price ~ ., data = matchdata,
                   method = "leapBackward",
                   trControl = trainControl(method = "cv", number = 10))
backwards
```

```
## Linear Regression with Backwards Selection
##
## 3204 samples
##   17 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2884, 2882, 2884, 2885, 2883, 2884, ...
## Resampling results across tuning parameters:
##
##   nvmax  RMSE      Rsquared   MAE
##   2      84353.39  0.6769755  55217.19
##   3      79280.53  0.7153318  51000.22
##   4      78137.63  0.7235894  49820.72
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was nvmax = 4.
```

Demo Stepwise Selection

```r
summary(backwards$finalModel)
```

```
## Subset selection object
## 26 Variables  (and intercept)
## ...
## Selection Algorithm: backward
##          year2005 lnland lnbldg rooms bedrooms bathrooms centair fireplace
## 1  ( 1 ) "*"      " "    " "    " "   " "      " "       " "     " "      
## 2  ( 1 ) "*"      " "    "*"    " "   " "      " "       " "     " "      
## 3  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "      
## 4  ( 1 ) "*"      "*"    "*"    " "   " "      " "       " "     " "      
##          brick garage1 garage2 dcbd rr  yrbuilt careaEdgewater careaEdison Park
## 1  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "             
## 2  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "             
## 3  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "             
## 4  ( 1 ) " "   " "     " "     " "  " " " "     " "            " "             
##          careaForest Glen careaJefferson Park careaLincoln Square
## 1  ( 1 ) " "              " "                 " "                
## 2  ( 1 ) " "              " "                 " "                
## 3  ( 1 ) " "              " "                 " "                
## 4  ( 1 ) "*"              " "                 " "                
##          careaNorth Park careaNorwood Park careaRogers Park careaUptown
## 1  ( 1 ) " "             " "               " "              " "        
## 2  ( 1 ) " "             " "               " "              " "        
## 3  ( 1 ) " "             " "               " "              " "        
## 4  ( 1 ) " "             " "               " "              " "        
##          careaWest Ridge latitude longitude
## 1  ( 1 ) " "             " "      " "      
## 2  ( 1 ) " "             " "      " "      
## 3  ( 1 ) " "             " "      " "      
## 4  ( 1 ) " "             " "      " "      
```