Topic # 03: Data Analytics

Model Selection and Cross Validation

Instructor: Prof. Arnab Bisi, Ph.D.

Johns Hopkins Carey Business School

Model Selection and Extensions

Session 3:

Agenda

- Part 1: Variable Selection
- Break
- Part 2: Working with code

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Model Selection

- Options for choosing variables
  - Domain-specific knowledge
  - Exploratory analysis
  - Statistical Selection
    - Principal Components Regression
- Subset Selection (predictor selection)
  - Best subset selection
  - Stepwise: Forward and Backward selection

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Why subset selection?

- Choose a smaller number of predictors from a complete set of candidate predictors
  - Improve prediction accuracy: too large a model leads to smaller bias, but larger variance
  - Appropriately smaller model can reduce the prediction error
  - Easy interpretation
    - Fewer variables is almost always easier to interpret, explain, and implement

What is best-subset selection?

- Considers the entire set of candidate predictors: X_1, X_2, …X_p
  - A given model can include/omit any particular predictor
  - Number of possible combinations is 2^p
  - Best-subset selection chooses the "best" subset of predictors from all 2^p possible combinations
    - Must define what "best" means

Linear Regression

- Linear regression is an approach to model the relationship between a scalar dependent variable y and one or more explanatory variables denoted as $\mathbf{X}$
- The case with one explanatory variable is called simple linear regression
- For more than one explanatory variable, it is called multiple regression

Questions Related to Regression Models

- Prediction accuracy
  - Least squares estimates have Low Bias (B-L-U-E)
    - LSE using all available variables minimizes bias
      - If the true relationship involves all variables, this can only be found using a model with all variables. Forcing one variable out of the model introduces a bias.
    - On the other hand, adding variables always increases variance
      - $Var[a + bx + E] = b^2 * Var[x] + Var[E]$
      - $Var[a + bx_1 + cx_2 + E] = b^2 * Var[x_1] + c^2 * Var[x_2] + Var[E]$

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Questions Related to Regression Models

- If *n* is much larger than *p*, LSE have low variance (Deviance) and perform better on test observations
- If *n* is not much larger than *p*, LSE has a lot of variability and may result in over-fitting and poor prediction
  - Over-fitting refers to using idiosyncrasies in a data set to fit a model to that set
    - Results in a model that seems to fit one set but will not work well on others

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Questions Related to Regression Models

- When minimizing RSS, if $p > n$ there is no unique solution for the FOC
  - No unique LSE coefficient estimates
  - We address this case later
- Two statistical approaches come to mind
  - Constrain the choice set by forcing variables out of consideration
    - By definition adding a constraint introduces a type of bias
  - Shrinking the coefficients (make them smaller) until some become 0
    - This effectively drops them from consideration
    - Shrinking also adds a type of bias

What we really care about
is the quality of our predictions!!!

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Best Subset Selection

- We fit separate regression models for each possible combination of the p predictors
- Look at all of the resulting models and select the best one
- Algorithm for Best Subset Selection

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

C_p, AIC, BIC, Adjusted R^2

- For a fitted least squares model containing *d* predictors C_p estimate of test MSE is:

$$C_p = \frac{1}{n}\left(RSS + 2d\hat{s}^2\right)$$

- Akaike information criteria (AIC) is defined as:

$$AIC = \frac{1}{n\hat{s}^2}\left(RSS + 2d\hat{s}^2\right)$$

- The Bayesian information criterion (BIC) is given by:

$$BIC = \frac{1}{n}\left(RSS + \log(n)d\hat{s}^2\right)$$

- For each of these measures **smaller** is **better**

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Best Criteria

- "We shall see that choosing the subset of the predictors that has the highest value of adjusted R^2 tends towards over-fitting."
  - From *A Modern Approach to Regression with R* by Simon J Sheather

- AIC has the desirable property that it is an efficient model selection criteria. What this means is that as the sample gets larger, the error obtained in making predictions using the model chosen becomes indistinguishable from the error obtained using the best possible model.. Other criteria such as BIC do not have this property.
  - From Simonoff JS (2003) *Analyzing categorical data*. Springer, New York.

- For model selection, there is no clear choice between AIC and BIC. BIC is asymptotically consistent as a selection criteria. What this means is that given a family of models, including the true model, the probability that BIC will select the correct model approaches one as the sample size approaches infinity. This is not the case for AIC.
  - From Hastie T, Tibshirani R, and Freedman J. 2001. *The elements of statistical learning*. Springer, New York.

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Best Subset Selection in R

- Use Hitters data set
- Use data to predict a baseball player's salary on the basis of his stats
- Simple R code
  - library(ISLR)
  - dim(Hitters)
    - This set has 20 variables
  - library(leaps)
    - regfits.full=regsubsets(Salary~.,Hitters)
    - summary(regfits.full)

Best Subset Selection in R

- Load data set of interest
- Load "leaps" library
- Use "regsubsets" function
  - "regsubsets" needs all columns of data frame to be relevant
    - Eliminate irrelevant vectors
    - Rows with missing info
    - Columns with labels or values not to be analyzed
  - Issues
    - regsubsets() reports the 1 Best model that includes 1-8 variables
      - Best means lowest value of RSS
      - **R** stores the resulting values of "rsq" "adjr2" "cp" and "bic"
      - Use the option "nvmax" to change the "8" to another size
      - Use the option "best=" to change the "1" to another size
    - **R** does not store the actual models, but it does tell you which variables are in the models

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Problem with Best Subset Approach

- The number of possible models that must be considered grows rapidly as p increases
- There are 2^p different models that can be built using p predictors
  - For $p = 10$, there are roughly 1000 models
  - For $p = 20$, there are roughly 1,000,000 models
  - For $p > 40$, it is computationally infeasible
- Best subset selection can lead to over-fitting
  - "Best" can be determined by very small improvements based on test data

Forward Stepwise Selection

- Begins with a model containing no predictors
  - Null model
  - Mean value of Y
- Add predictors one at a time
  - The variable that gives the greatest additional improvement to the fit is added.
  - Once a variable is added to the list it stays on the list

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Backward Stepwise Selection

- Begins with a model that includes all *p* predictors
- Remove the least useful predictor one at a time
  - Once a predictor is dropped from the list it stays dropped
- Use regsubset() to perform Forward or Backward stepwise selection
  - Add argument method="forward" or method="backward"

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

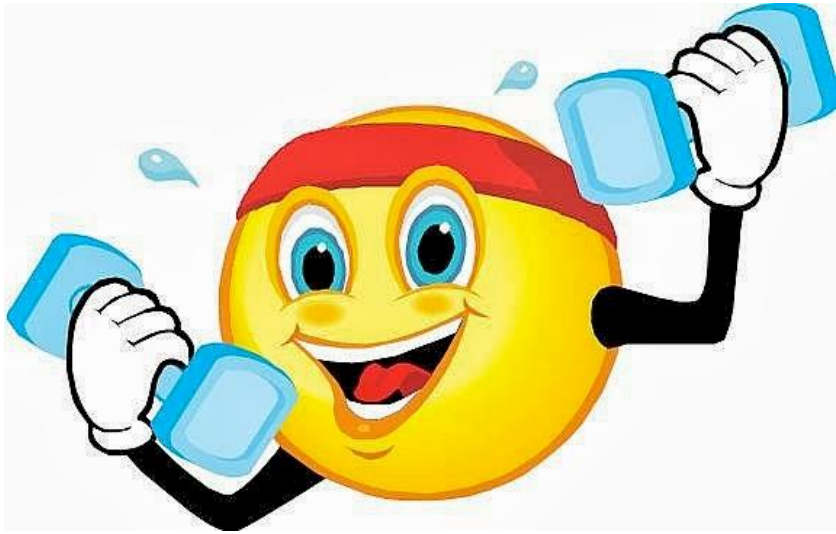
JOHNS HOPKINS CAREY BUSINESS SCHOOL

Stepwise Selection Sample Code

```
## Best subsect selection
library(ISLR)
names(Hitters)
Hitters2=na.omit(Hitters)
library(leaps)
regfit.full=regsubsets(Salary~.,data=Hitters2,nvmax=15)
reg.summary=summary(regfit.full)
### use different approaches: R^2, adjusted R^2, CP, BIC
names(reg.summary)
reg.summary$rsq
which.max(reg.summary$adjr2)
coef(regfit.full,11)
which.min(reg.summary$cp)
which.min(reg.summary$bic)
plot(reg.summary$bic,xlab="Number of Variables",type="l")
#### stepwise selection: forward or backward
regfit.fwd=regsubsets(Salary~.,data=Hitters,nvmax=19, method="forward")
summary(regfit.fwd)
regfit.bwd=regsubsets(Salary~.,data=Hitters,nvmax=19, method="backward")
summary(regfit.bwd)
coef(regfit.full,7)
coef(regfit.fwd,7)
coef(regfit.bwd,7)
```

Exercise

- Use Auto.csv data set
- Subset selection: Choose best 3 predictors
  - Best subset selection
  - Forward subset selection
  - Backward subset selection
- 15 mins



Summary of Model Selection

- Subset selection
  - Best subset
  - Forward step approach
  - Backward step approach

What we really care about is prediction!

Model Selection Recipe

- Find manageable set of candidate models
  - Fitting all models is fast enough
  - Choose candidate that does the BEST job of making predictions using data NOT used to select the model
- Two issues here
  - What data to use for our test
  - How to define BEST

Finding the BEST Model

- BEST may be defined directly in terms of statistical measures (AIC, BIC, etc.)
  - Deals only with the Fitted model
  - This is not really what we care about
- LOOCV – Leave-One-Out Cross-Validation
  - Given *n* observations leave out observation 1
  - Create model using observations 2:*n*
  - Use this model to make prediction about observation 1 and calculate the error
  - Now leave out observation 2
  - Create a new model using observations 1, 3:*n*
  - Repeat for all *n* observations

JOHNS HOPKINS CAREY BUSINESS SCHOOL

LOOCV

- Average error can be calculated across the *n* models
  - For each model there is an error regarding the one held out
  - Consider the average of these values

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

- This can be done for each model being considered and the BEST can be selected
- This will work, however;
  - Very time consuming if *n* is large and we wish to consider a large number of candidate models

k-Fold Cross Validation

- LOOCV may be modified to increase speed
- k-Fold Cross Validation
  - Split data set into k non-overlapping subsets of about equal size (folds)
    - k is typically set to 5 or 10
  - Given *k* folds leave out fold 1
  - Create model using folds 2:*k*
  - Use this model to make prediction about fold1 and calculate the average error
  - Now leave out fold 2
  - Create model using folds 1, 3:*n*
  - Repeat for all *k* folds

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

JOHNS HOPKINS CAREY BUSINESS SCHOOL

Validation Set Approach

- Big data set
- Split data set into 2 pieces
  - Fit model using first piece (Training data)
  - Test its ability to make predictions on second piece (Test data)
- Error rates (MSE, Deviance, etc) can be different for the two sets
  - What really matters is the errors involving the test data

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

How to Make the Split

- Want distributions to be the same across the two sets
  - Random selection without replacement
    - Even vs odd rows
  - Split based on time if the world does not change over time
    - nrow(Data.Frame)/2
- Avoid Splits that are clearly different populations
  - Models fit using young customers may not make good predictions about old customers
  - Models fit using urban customers may not make good predictions about rural customers
  - Models fit using male customers may not make good predictions about female customers

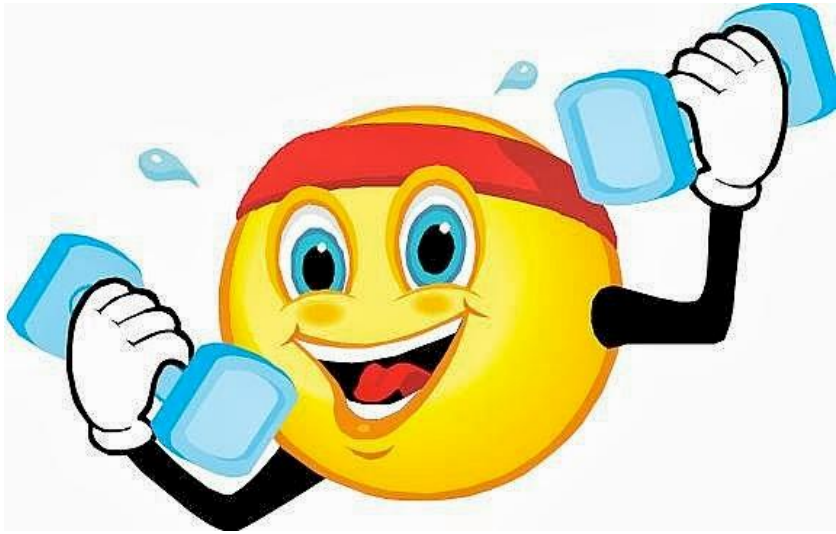
Issues With Validation Set Approach

- Pros
  - Easy to explain and execute
  - Easy to generalize across many types of models

- Cons
  - Can be hard to tell if a split alters outcomes
  - If splits are random the test error rate can be highly variable depending on precisely which observations are included in each set
  - Statistical methods tend to perform worse when trained on fewer observations. Therefore, the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set

In-Class Exercise

- Use Auto.csv data set
- Split data set into 2 pieces using sample function
- Find candidates using “training” set
  - Best subset selection with 1, 2, and 3 variables
- Use these models to make predictions involving “test” set
- Calculate MSE of predictions to select BEST model
- 15 mins



Questions, Comments?

Let's move to the Code.