Chapter 4  Topic 4.2

Predictor Selection Methods
All Subsets
Backward Elimination
Forward Selection
Stepwise Regression

Lecture on white board first:
Reasons for regression
Assess specific variable(s)
Discover relationships
Predict in future
Mallow’s Cp
Bluejay example
Example 4.2 (book): Predicting First Year GPA

Data: **FirstYearGPA** from Chapter 4

Response: **GPA**

Predictors: (See page 169)

**HSGPA, Male, FirstGen, SATV, HU, White, SATM, SS, CollegeBound**

Find the “best” model for GPA using some or all of these predictors. See text for results!

What determines “best”? 
Criteria to Compare Models?

Look for large $R^2$.

But $R^2$ is always best for the model with all predictors.

Look for large adjusted $R^2$.

Helps factor in the number of predictors in the model

Look for small $S_\varepsilon$

Look at individual t-tests.

But susceptible to multicollinearity problems

\[
R^2_{adj} = 1 - \frac{S^2_\varepsilon}{S^2_Y}
\]

\[
S_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}
\]
Predictor Selection Methods

Choosing an effective set of predictors:

1. *Think, consult, graph*… but if that fails, then:

2. *All subsets*

3. Backward elimination

4. Forward selection

5. Stepwise regression
How to Choose Models to Compare?

Method #1: All Subsets!

Consider all possible combinations of predictors.

How many are there?

Pool of $k$ predictors $\rightarrow 2^k - 1$ subsets

Advantage: Find the best model for your criteria.

Disadvantage: LOTS of computation; brain overload. Also, “data snooping!”
Example 2: BlueJays2

Response variable:

\[ Y = \text{Mass} \]

5 Potential predictors:

- Depth = bill upper surface – lower surface
- Width = width of bill
- Length = length of bill
- Skull = base of bill to back of skull
- Sex = 1 for male, 0 for female

Find model with best adj \( R^2 \).
#find the leaps package (at CRAN on the web)

#load the leaps package
> library(leaps)

#read in BlueJays file using RStudio

#Ask for the best model of each size

> all = regsubsets(Mass ~ BillDepth + BillWidth + BillLength + Skull + Sex, data = BlueJays)

Uses Mallow’s Cp to find good models.
> options(digits=3) # Only display 3 digits

> summary(all)$cp # Provides Cp values

[1] 37.34 11.79  5.93  4.31  6.00

> summary(all)$adjr2 # Provides AdjR-sq

[1] 0.300 0.418 0.449 0.461 0.458

Cp and Adjusted $R^2$ for some good models

But what models are they?
> plot(all,scale="adjr2")

White means that variable isn’t in.

*Top adjR2 model has D, L, Sk, Sx.*

Next model has D, W, L, Sk, Sx.

Next model has D, L, Sk.

These three all look pretty good, and adjusted R-squared is about the same for all, so we may prefer the third. (Simplest)
Mallow’s $C_p$

Note: $R^2$, Adjusted $R^2$, $S_e$, $SSE$, and $MSE$ all depend only on the predictors in the model being evaluated, NOT the other potential predictors in the pool.

Mallow’s $C_p$: When evaluating a subset of predictors from a larger set of $k$ predictors (not including intercept),

$$C_p = \frac{SSE_p}{MSE_{Full(k)}} + 2(p + 1) - n$$

$p$=# predictors (without intercept) in reduced model

See white board for another version of formula.
Notes on $C_p$

• $C_p$ depends on the larger pool of predictors as well as the set being tested.
• For full model, $C_p = p + 1$
• For a “good” set of predictors, $C_p$ should be small.
• Like Adj R$^2$, $C_p$ weighs both the effectiveness of the model ($SSE_p$) and the number of predictors ($p$).
• A model with $C_p$ near $p + 1$ (or smaller) is worth considering.
R: Best Subsets for BlueJays2

#Ask for the best (2) models of each size
> all=regsubsets(Mass~BillDepth+BillWidth+BillLength+Skull+Sex,data=BlueJays,nbest=2)
> summary(all)$cp #get Cp values

Gives nine models in this case:

2 each with 1, 2, 3, 4 variables
1 with all variables

[1] 37.34 55.82 11.79 12.20  5.93 12.51  4.31  7.35     6.00

1 var.   2 vars.  3 vars.  4 vars.  5 vars
Top \( Cp \) model has \( D, L, Sk, Sx \).
Next model has \( D, L, Sk \).
Model with all predictors (\( D, W, L, Sk, Sx \)) has \( C_p = 6 = 5 + 1 \).
Next model has \( D, W, L, Sk \).
> `summary(all)`

<table>
<thead>
<tr>
<th></th>
<th>Depth</th>
<th>Width</th>
<th>Length</th>
<th>Skull</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>4</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Shows best two models of each size.
One variable only: Best is Skull, then Depth
Two variables only: Best is Depth+Skull, then Length+Skull
Three variables: Best is Depth+Length+Skull, etc.
Find the leaps package and the HH package (at CRAN on the web) and multcomp, mvtnorm, RColorBrewer, and latticeExtra packages! The last four all need to be installed for HH to work.

**In R Studio:** Tools -> Install packages, then type HH in the box.

```r
#load the HH package and the leaps package
> library(leaps)
> library(HH)

#Ask for the best model of each size
> all=regsubsets(Mass~Depth+Width+Length+Skull+Sex, data=BlueJays)

#Ask for “nice” output from the regsubsets
> summaryHH(all)
```
> summaryHH(all)

<table>
<thead>
<tr>
<th>model</th>
<th>p</th>
<th>rsq</th>
<th>rss</th>
<th>adjr2</th>
<th>cp</th>
<th>bic</th>
<th>stderr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sk</td>
<td>2</td>
<td>0.313</td>
<td>1874</td>
<td>0.307</td>
<td>32.94</td>
<td>-36.2</td>
<td>3.95</td>
</tr>
<tr>
<td>D-Sk</td>
<td>3</td>
<td>0.426</td>
<td>1566</td>
<td>0.416</td>
<td>10.15</td>
<td>-53.3</td>
<td>3.63</td>
</tr>
<tr>
<td>D-L-Sk</td>
<td>4</td>
<td>0.455</td>
<td>1488</td>
<td>0.441</td>
<td>5.83</td>
<td>-54.8</td>
<td>3.55</td>
</tr>
<tr>
<td>D-L-Sk-Sx</td>
<td>5</td>
<td>0.471</td>
<td>1444</td>
<td>0.453</td>
<td>4.31</td>
<td>-53.6</td>
<td>3.51</td>
</tr>
<tr>
<td>D-W-L-Sk-Sx</td>
<td>6</td>
<td>0.472</td>
<td>1440</td>
<td>0.449</td>
<td>6.00</td>
<td>-49.1</td>
<td>3.52</td>
</tr>
</tbody>
</table>

Model variables with abbreviations

<table>
<thead>
<tr>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sk</td>
</tr>
<tr>
<td>D-Sk</td>
</tr>
<tr>
<td>D-L-Sk</td>
</tr>
<tr>
<td>D-L-Sk-Sx</td>
</tr>
<tr>
<td>D-W-L-Sk-Sx</td>
</tr>
</tbody>
</table>

Number of observations: 122
Choosing an effective set of predictors:

1. *Think, consult, graph*… but if that fails, then:
2. All subsets
3. *Backward elimination*
4. Forward selection
5. Stepwise regression
Backward Elimination

1. Start with the full model (all predictors).

2. Calculate a t-test for each individual predictor.

3. Find the “least significant” predictor (largest p-value or smallest t.s.).

4. Is that predictor significant?
   Yes → Keep the predictor and stop.
   No → Delete the predictor and go back to step 2 with the reduced model.
Backward Elimination

Advantages:
- Removes “worst” predictors early
- Relatively few models to consider
- Leaves only “important” predictors

Disadvantages:
- Most complicated models first
- Individual t-tests may be unstable
- Susceptible to multicollinearity
Choosing an effective set of predictors:

1. *Think, consult, graph* … but if that fails, then:

2. All subsets

3. Backward elimination

4. *Forward selection*

5. Stepwise regression
Forward Selection

1. Start with the best single predictor (fit each predictor or use correlations).

2. Is that predictor significant? (Use individual t-test or partial F-test)
   - Yes $\rightarrow$ Include predictor in the model.
   - No $\rightarrow$ Don’t include predictor and stop.

3. Find the “most significant” new predictor from among those NOT in the model (use biggest $SS_{Model}$, largest $R^2$, or best individual t-test). Return to step 2.
Forward Selection

Advantages:
Uses smaller models early (parsimony)
Less susceptible to multicollinearity
Shows “most important” predictors

Disadvantages:
Need to consider more models
Predictor entered early may become redundant later
Choosing an effective set of predictors:

1. *Think, consult, graph*… but if that fails, then:
   2. All subsets
   3. Backward elimination
   4. Forward selection
   5. *Stepwise regression*
Stepwise Regression

Basic idea: Alternate forward selection and backward elimination.

1. Use forward selection to choose a new predictor and check its significance.
2. Use backward elimination to see if predictors already in the model can be dropped.
Backward elimination in R

#1 Fit the full model
> full=lm(Mass~Depth+Width+Length+Skull+Sex)

#2 Find the MSE for the full model
> MSE=(summary(full)$sigma)^2

#3 Use the `step()` command for backward
> step(full, scale=MSE, direction="backward")
Stepwise in \( R \)

#For stepwise start with a model with no predictors

> none=lm(Mass~1)

> step(none,scope=list(upper=full),scale=MSE)

#R uses \( Cp \) (AIC) to pick next model
```r
> step(none, scope=list(upper=full), scale=MSE)
Start:  AIC=100
Mass ~ 1

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>Skull</td>
<td>1</td>
<td>855</td>
<td>1874</td>
</tr>
<tr>
<td>+</td>
<td>Depth</td>
<td>1</td>
<td>591</td>
<td>2138</td>
</tr>
<tr>
<td>+</td>
<td>Length</td>
<td>1</td>
<td>427</td>
<td>2301</td>
</tr>
<tr>
<td>+</td>
<td>Sex</td>
<td>1</td>
<td>340</td>
<td>2389</td>
</tr>
<tr>
<td>+</td>
<td>Width</td>
<td>1</td>
<td>212</td>
<td>2517</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>2729</td>
<td>99.8</td>
<td></td>
</tr>
</tbody>
</table>

Step:  AIC=33
Mass ~ Skull

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>Depth</td>
<td>1</td>
<td>308</td>
<td>1566</td>
</tr>
<tr>
<td>+</td>
<td>Length</td>
<td>1</td>
<td>289</td>
<td>1585</td>
</tr>
<tr>
<td>+</td>
<td>Width</td>
<td>1</td>
<td>59</td>
<td>1815</td>
</tr>
<tr>
<td>+</td>
<td>Sex</td>
<td>1</td>
<td>45</td>
<td>1830</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>1874</td>
<td>32.9</td>
<td></td>
</tr>
<tr>
<td>-  Skull</td>
<td>1</td>
<td>855</td>
<td>2729</td>
<td>99.8</td>
</tr>
</tbody>
</table>
```
Step: AIC=10
Mass ~ Skull + Depth

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Length</td>
<td>1</td>
<td>78</td>
<td>1488</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>1566</td>
<td>10.15</td>
</tr>
<tr>
<td>+ Width</td>
<td>1</td>
<td>16</td>
<td>1550</td>
</tr>
<tr>
<td>+ Sex</td>
<td>1</td>
<td>7</td>
<td>1560</td>
</tr>
<tr>
<td>- Depth</td>
<td>1</td>
<td>308</td>
<td>1874</td>
</tr>
<tr>
<td>- Skull</td>
<td>1</td>
<td>572</td>
<td>2138</td>
</tr>
</tbody>
</table>

Step: AIC=5.8
Mass ~ Skull + Depth + Length

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>RSS</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Sex</td>
<td>1</td>
<td>44</td>
<td>1444</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>1488</td>
<td>5.83</td>
</tr>
<tr>
<td>+ Width</td>
<td>1</td>
<td>7</td>
<td>1481</td>
</tr>
<tr>
<td>- Length</td>
<td>1</td>
<td>78</td>
<td>1566</td>
</tr>
<tr>
<td>- Depth</td>
<td>1</td>
<td>97</td>
<td>1585</td>
</tr>
<tr>
<td>- Skull</td>
<td>1</td>
<td>576</td>
<td>2064</td>
</tr>
</tbody>
</table>
Step:  AIC=4.3
Mass ~ Skull + Depth + Length + Sex

Df  Sum of Sq  RSS    Cp
<none>            1444  4.31  BEST CHOICE NOW IS “NONE”
- Sex     1     44 1488  5.83
+ Width   1      4 1440  6.00
- Length  1    115 1560  11.60
- Depth   1    129 1574  12.72
- Skull   1    610 2054  51.44

Call:
lm(formula = Mass ~ Skull + Depth + Length + Sex)

Coefficients:
(Intercept)        Skull        Depth       Length          Sex
   -69.38         2.73         3.42        1.18        -1.67
Missing Values

Warning! If data are missing for *any* of the predictors in the pool, “Stepwise” and “Best Subsets” procedures will eliminate the data case from *all* models.

Thus, running the model for the selected subset of predictors alone may produce different results than within the stepwise or best subsets procedures.