

Announcements

- My office hours today will be cut short and end at 2:50 instead of 3:30.
- If you need to pick up old exams or homework, we will have them in our office hours.
- Next Wed Brandon will do my office hours (2 to 3:30)
- Current and upcoming assignments and due dates:

Assigned	Due	Lectures covered
HW #5: Mon, Nov 6	Wed, Nov 15	Mon, Oct 30 & Mon, Nov 6
HW #6: Mon, Nov 13	Wed, Nov 22	Wed, Nov 8 (today) and Mon, Nov 13

- Then back to usual schedule

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 C_p

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 .

$$R_{adj}^2 = 1 - \frac{S_\varepsilon^2}{S_Y^2}$$

Helps factor in the number of predictors in the model

Look for small S_ε

$$S_\varepsilon = \sqrt{\frac{SSE}{n - k - 1}}$$

Look at individual t-tests.

But susceptible to multicollinearity problems

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 1: Go over Minitab example in class

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 .

R: Best Subsets for BlueJays2

```
#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 Mallows's C_p 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
```

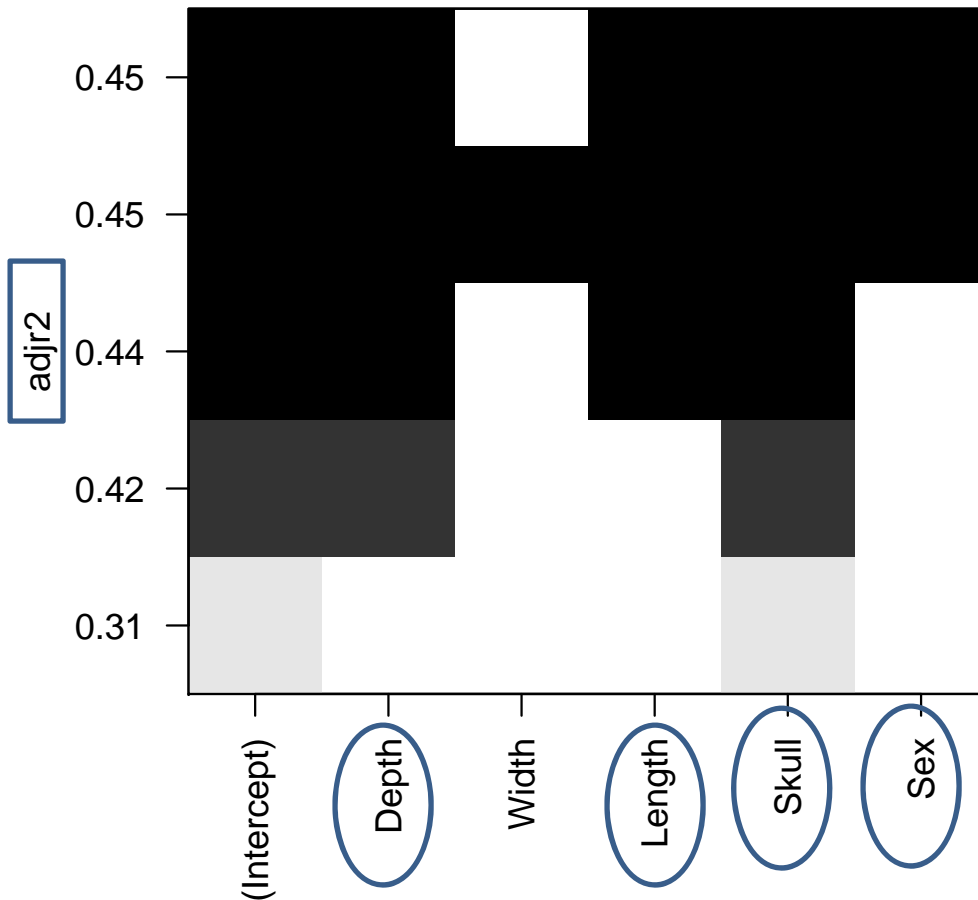
```
> summary(all)$adjr2 #Provides AdjR-sq
```

```
[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_ε , 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),

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

Diagram: A red box labeled "Reduced" has an arrow pointing to SSE_p . A red box labeled "Full" has an arrow pointing to $MSE_{Full(k)}$.

See white board for another version of formula.

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

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 $\text{Adj } R^2$, 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.

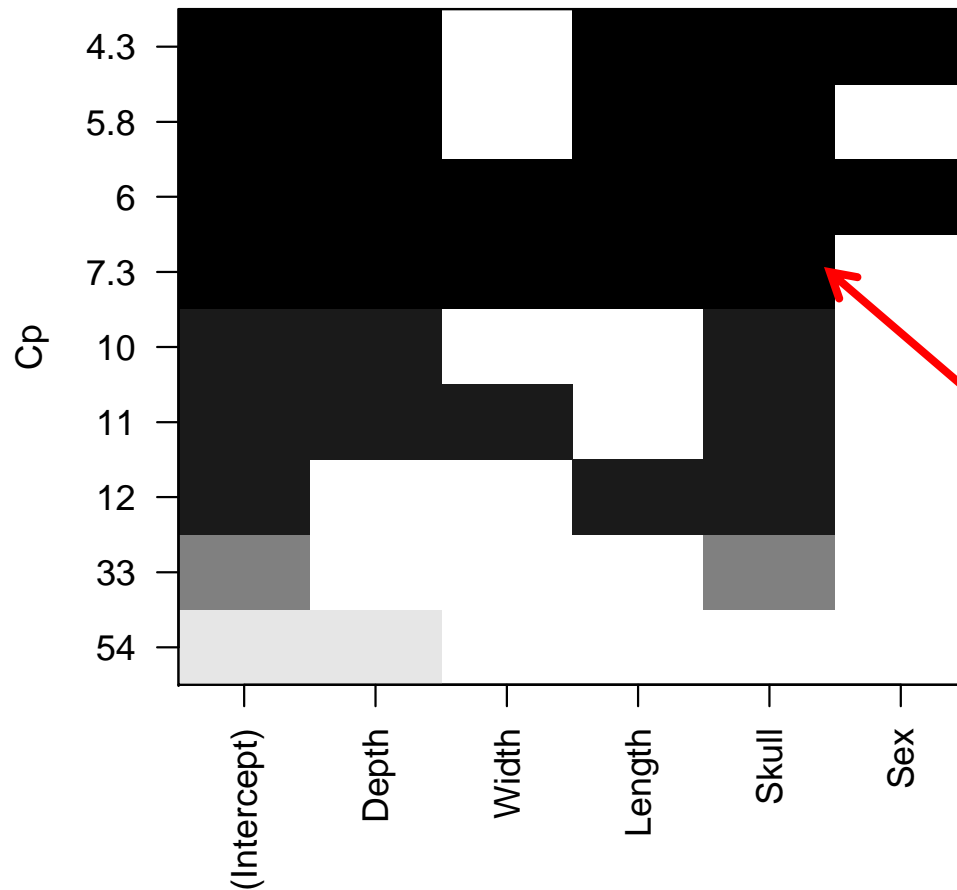
R: Best Subsets for BlueJays2

```
#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				

```
> plot(all,scale="Cp")
```



*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)
```

		Depth	Width	Length	Skull	Sex
1	(1)	" "	" "	" "	"*"	" "
1	(2)	"*"	" "	" "	" "	" "
2	(1)	"*"	" "	" "	"*"	" "
2	(2)	" "	" "	"*"	"*"	" "
3	(1)	"*"	" "	"*"	"*"	" "
3	(2)	"*"	"*"	" "	"*"	" "
4	(1)	"*"	" "	"*"	"*"	"*"
4	(2)	"*"	"*"	"*"	"*"	" "
5	(1)	"*"	"*"	"*"	"*"	"*"

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.

```
#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)**

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

Model variables with abbreviations

	model
Sk	Skull
D-Sk	Depth-Skull
D-L-Sk	Depth-Length-Skull
D-L-Sk-Sx	Depth-Length-Skull-Sex
D-W-L-Sk-Sx	Depth-Width-Length-Skull-Sex

model with largest adjr2

4

Number of observations

122

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

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

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

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 → Include predictor in the model.
No → Don't include predictor and stop.
3. Find the “most significant” new predictor from among those NOT in the model (use biggest *SSModel*, 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

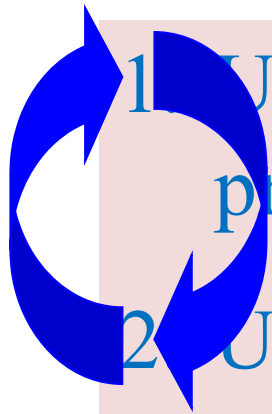
Predictor Selection Methods

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*

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
```

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

```
Start: AIC=100
```

```
Mass ~ 1
```

	Df	Sum of Sq	RSS	Cp	
+ Skull	1	855	1874	32.9	BEST FIRST CHOICE
+ Depth	1	591	2138	54.2	
+ Length	1	427	2301	67.3	
+ Sex	1	340	2389	74.4	
+ Width	1	212	2517	84.7	
<none>			2729	99.8	

```
Step: AIC=33
```

```
Mass ~ Skull
```

	Df	Sum of Sq	RSS	Cp	
+ Depth	1	308	1566	10.2	BEST CHOICE AFTER SKULL IN MODEL
+ Length	1	289	1585	11.7	
+ Width	1	59	1815	30.2	
+ Sex	1	45	1830	31.4	
<none>			1874	32.9	
- Skull	1	855	2729	99.8	

Step: AIC=10

Mass ~ Skull + Depth

	Df	Sum of Sq	RSS	Cp
+ Length	1	78	1488	5.83
<none>			1566	10.15
+ Width	1	16	1550	10.82
+ Sex	1	7	1560	11.60
- Depth	1	308	1874	32.94
- Skull	1	572	2138	54.18

BEST CHOICE WITH SKULL & DEPTH

Step: AIC=5.8

Mass ~ Skull + Depth + Length

	Df	Sum of Sq	RSS	Cp
+ Sex	1	44	1444	4.31
<none>			1488	5.83
+ Width	1	7	1481	7.27
- Length	1	78	1566	10.15
- Depth	1	97	1585	11.68
- Skull	1	576	2064	50.22

BEST CHOICE GIVEN SKULL, DPT, LNTH

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.