Chapter 9

Selecting Variables

Notation and Example

- Let P – 1 = number of candidate predictors
- Let p – 1 = number in a particular model

Example: Appendix C7, Real estate data

Let Y = log(sales price)

Possible X variables (omit style, year), P – 1 = 9

Square feet/100	Number of bedrooms
Number of bathrooms	Air conditioning? (1 = yes, 2 = no)
Garage size (# cars)	Pool? (1 = yes, 2 = no)
Lot size	Adjacent to highway? (1 = yes, 2 = no)
Quality (1 = high, 2 = medium, 3 = low)	

How to choose predictors

- Note that each of them could be in or out, so there are $2^9 = 512$ possible models!
- For 10 predictors, $2^{10} = 1024$
- For 30 predictors, 2^{30} = approx. 1.07 billion!
- So we need a way to narrow down the possibilities.
- Remember, there isn't one "correct" model, but there are useful models!

Philosophical Issues

1. Will you be trying to *explain* the relationships between Y and the X's, or to *predict* in future?
2. Will it cost extra $$ to measure some X's in the future, if you want to use for prediction?
3. Should certain variable always be in the model, for practical or philosophical reasons?
4. Occam's razor: Simplest is best, if all else is equal!

Practical Issues

1. Unnecessary predictors add "noise."

   $MSE = \frac{SSE}{n-p}$ *both* go down as *p* goes up.

2. Multicollinearity can be a problem if too many variables are included.
3. Too few predictors creates bias (omitting important ones).
4. Be careful about removing two or more correlated variables all at once based on *p*-values. Remember example with left and right foot!

Model Comparison and Selection

- So far, we have only compared models where one is a "reduced" version of the other.
- Now we will compare models without having that requirement.

Model selection methods:

1. All subsets (also called best subsets)
2. Forward selection
3. Backward elimination
4. Stepwise regression (combines methods 2 & 3)

Steps for finding a good model

1. Exploratory analysis on *each* possible predictor: Linear with Y? Outliers? Transform X?
2. Use prior knowledge and plots (residuals vs X_jX_k) to determine what interactions to try.
3. Fit full model, plot residuals versus $\hat{Y}$ to see if Y needs transformation.
4. Reduce number of predictors and compare models (today's topic).
5. Case diagnostics to see if any cases should be corrected or removed (Chapter 10, next time)
6. Repeat above if necessary after fixing or removing cases.
7. If possible, validate the model on new data.

Criteria for Comparing Models

1. R^2 but use only for models with same *p*. High values are good.
2. Adjusted R^2 which is same as comparing MSE, because it's 1 – (MSE/MSTO). Want high adjusted R^2 and low MSE.
3. Mallow's C_p; want it approx equal to *p* (smaller is better).
4. AIC = Akaike's information criterion, low values are good. (See p. 359.)
5. PRESS = Prediction sum of squares, low values are good. (See p. 360.)

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 $p - 1$ predictors from a larger set of $P - 1$ predictors:

$$C_p = \frac{SSE_p}{MSE_P} + 2p - n$$

Reduced

Full

p=# coefficients (including 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, $C_P = P$
- 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 is worth considering.

Example: Real Estate Data, Appendix C7

Y = LogSales = log (Sales Price)

X1 = SqFt100s = Square Feet / 100

X2 = AC = Air conditioning (1 = Yes, 0 = No)

X3 = Bedrms = Number of bedrooms

X4 = Bathrms = Number of bathrooms

X5 = LotSize (in square feet)

X6 = NearHwy = 1 if adjacent to highway, 0 if not

X7 = Garage = number of cars garage will hold

X8 = Pool = 1 if yes, 0 if no

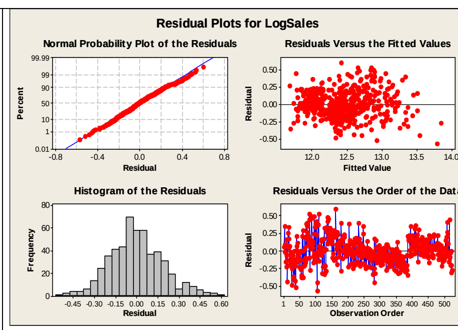
X9 = Quality = 1 (high), 2 (medium) or 3 (low)

EXAMPLE OF "BEST SUBSETS" REGRESSION: Response = Log sales price of house; 9 possible explanatory variables

Best Subsets Regression: LogSales versus SqFt/100, AC, ...

Response is LogSales

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	**SqFt/100**	**AC**	Bedrooms	**Bathrooms**	**LotSize**	NearHwy	**GarageSize**	Pool	**Quality**
1	70.5	70.4	285.5	0.23472	X								
1	62.0	61.9	517.3	0.26644									X
2	78.5	78.4	69.5	0.20056	X								X
2	73.6	73.5	203.7	0.22235	X						X		
3	79.7	79.6	39.3	0.19515	X				X				X
3	79.5	79.3	45.5	0.19624	X						X		X
4	80.5	80.3	19.7	0.19149	X				X		X		X
4	80.2	80.1	26.8	0.19276	X			X	X				X
5	80.9	80.7	9.3	0.18942	X			X	X		X		X
5	80.7	80.5	15.3	0.19051	X	X			X		X		X
6	81.1	80.9	6.4	0.18871	**X**	**X**		**X**	**X**		**X**		**X**
6	81.0	80.8	9.5	0.18928	X			X	X		X	X	X
7	81.2	80.9	6.8	0.18861	X	X		X	X		X	X	X
7	81.1	80.9	7.8	0.18879	X	X		X	X	X	X		X
8	81.2	80.9	8.3	0.18869	X	X		X	X	X	X	X	X
8	81.2	80.9	8.5	0.18873	X	X	X	X	X		X	X	X
9	81.2	80.9	10.0	0.18882	X	X	X	X	X	X	X	X	X



Above are the diagnostic plots for the model chosen, which is the one shown in bold on the left. The "residuals versus order of the data" plot isn't useful in this example, but the other three plots are. See note #3 below.

NOTES:

1. All of the highlighted models have acceptable Mallow's Cp. I chose the model (in bold) with good Cp and smallest number of variables to get best R-Sq(adj), which stays the same for the rest of the models, at 80.9%.
2. That model has the variables in bold as predictors. They include SqFt/100, AC, Bathrooms, Lot size, Garage size and Quality. Bedrooms, near highway and pool are not included.
3. The diagnostic plots for the chosen model are shown on the right. They look good. The normal probability plot and the histogram of residuals show that the residuals are approximately normal, and the plot of residuals versus fitted values looks like random scatter, as it should.
4. The final model is:

LogSales = 11.9 + 0.0283 SqFt/100 + 0.0552 AC + 0.0418 Bathrooms + 0.000004 LotSize + 0.0643 GarageSize - 0.206 Quality

Same example using R

```
> AppendixC7 <- read.table("AppendixC7.txt", header=T)
> AppendixC7<-cbind(AppendixC7, log(AppendixC7$SalesPrice))
> names(AppendixC7)[14]<-"LogSales"
> AppendixC7<-cbind(AppendixC7, AppendixC7$SqFt/100)
> names(AppendixC7)[15]<-"SqFt100s"
> library(leaps) #NEED TO LOAD THIS PACKAGE
> Best<-regsubsets(LogSales ~ SqFt100s + AC + Bedrms + Bathrms +
LotSize + NearHwy + Garage + Pool + Quality, data=AppendixC7)
> summary(Best)
```

Partial output from "Summary"

```
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         SqFt100s AC  Bedrms Bathrms LotSize NearHwy Garage Pool Quality
1 ( 1 )  "*"      " " " "    " "     " "     " "     " "    " "  " "
2 ( 1 )  "*"      " " " "    " "     " "     " "     " "    " "  "*"
3 ( 1 )  "*"      " " " "    " "     "*"     " "     " "    " "  "*"
4 ( 1 )  "*"      " " " "    " "     "*"     " "     "*"    " "  "*"
5 ( 1 )  "*"      " " " "    "*"     "*"     " "     "*"    " "  "*"
6 ( 1 )  "*"      "*" " "    "*"     "*"     " "     "*"    " "  "*"
7 ( 1 )  "*"      "*" " "    "*"     "*"     " "     "*"    "*"  "*"
8 ( 1 )  "*"      "*" " "    "*"     "*"     "*"     "*"    "*"  "*"
> summary(Best)$cp
[1] 285.50 69.54 39.31 19.69 9.29 6.41 6.82 8.29
```

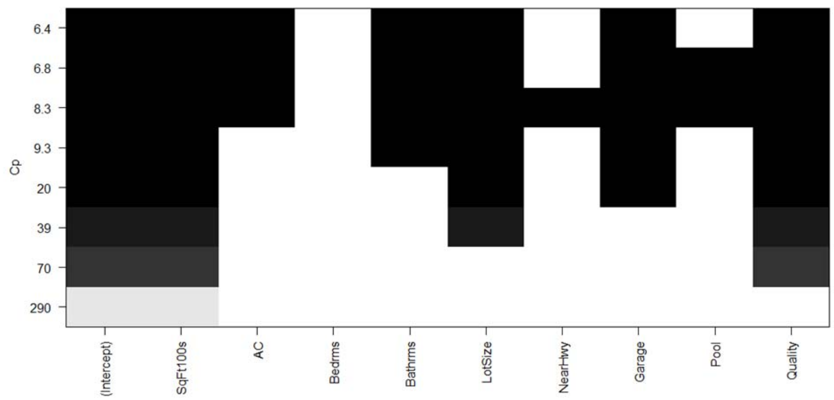
NOTES:

1. What was added at each step is shown in red. (Not part of the R output.)
2. It won't always be the case that each size contains everything from the previous size.
3. Most reasonable Cp value is 6.41, which is for 6 explanatory variables (so p = 7). Next best is 6.82, which is for 7 explanatory variables (so p = 8).

Graphical display of models, with Cp

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

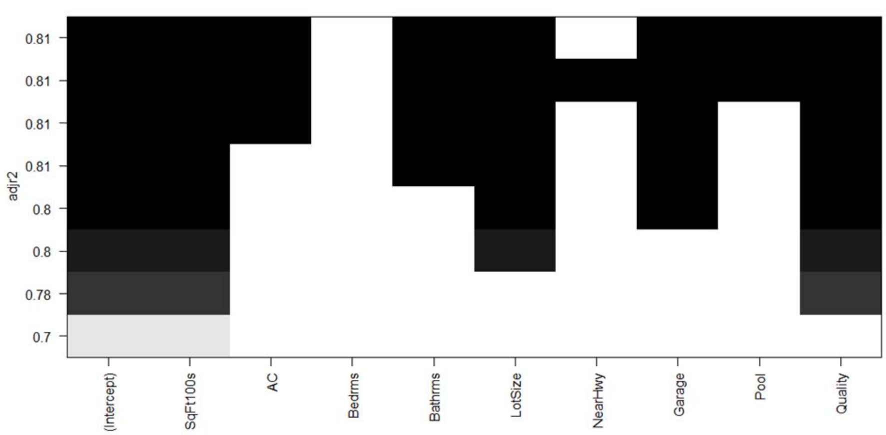
- Black box in a row means variable is included, where variable names are show at the bottom.
- Example: Top row *omits* Bedrms, NearHwy, Pool



Graphical display showing Adjusted R-squared

```
> plot(Best,scale="adjr2")
```

Best 4 models all have Adjusted R-Sq = 0.81.



One more option: HH, for nicer output

In R or R Studio: Tools -> Install packages, then type HH in the box or go to a CRAN site and select HH from the list shown.

```
#Load the HH package and the leaps package
> library(leaps)
> library(HH)


#Ask for "nice" output from the regsubsets,
  after you've run the model and called it
  "Best"
> summaryHH(Best)
```

```
> summaryHH(Best)
                model p   rsq  rss adjr2     cp  bic stderr
1                   S 2 0.705 28.6 0.704 285.50 -625  0.235
2                 S-Q 3 0.785 20.9 0.784  69.54 -783  0.201
3               S-L-Q 4 0.797 19.7 0.796  39.31 -807  0.195
4             S-L-G-Q 5 0.805 19.0 0.803  19.69 -821  0.191
5          S-Bt-L-G-Q 6 0.809 18.5 0.807   9.29 -827  0.189
6        S-A-Bt-L-G-Q 7 0.811 18.3 0.809   6.41 -826  0.189
7      S-A-Bt-L-G-P-Q 8 0.812 18.3 0.809   6.82 -821  0.189
8 S-A-Bt-L-N-G-P-Q 9 0.812 18.3 0.809   8.29 -816  0.189

Model variables with abbreviations
                                                                       model
S                                                                   SqFt100s
S-Q                                                         SqFt100s-Quality
S-L-Q                                               SqFt100s-LotSize-Quality
S-L-G-Q                                      SqFt100s-LotSize-Garage-Quality
S-Bt-L-G-Q                           SqFt100s-Bathrms-LotSize-Garage-Quality
S-A-Bt-L-G-Q                      SqFt100s-AC-Bathrms-LotSize-Garage-Quality
S-A-Bt-L-G-P-Q               SqFt100s-AC-Bathrms-LotSize-Garage-Pool-Quality
S-A-Bt-L-N-G-P-Q SqFt100s-AC-Bathrms-LotSize-NearHwy-Garage-Pool-Quality

model with largest adjr2
7

Number of observations
522
```

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 test statistic t).
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

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 *SSR*, 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 but never gets deleted

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.

Implementing in R

First fit the full model with all variables:

```
Full<-lm(LogSales~SqFt100s+AC+Bedrms+Bathrms+LotSize
+ NearHwy + Garage+Pool+Quality,data=AppendixC7)
```

Fit a “base” model with just the intercept:

```
>Base<-lm(LogSales~1,data=AppendixC7)
```

Then you can use forward or backwards:

```
> step(Base, scope = list(upper=Full, lower=~1),
direction = "forward", trace=FALSE)
```

The above tells it to start with just the intercept, and possibly go up to the Full model.

```
> step(Full, direction = "backward", trace=FALSE)
```

Final results are the same in this example, but the order is different. In "forward" it shows them in the order entered. In "backward" it leaves the "good" variables in the same order they were given in the Full model.

```
> step(Base, scope = list(upper=Full, lower=~1), direction = "forward",
trace=FALSE)

Call:
lm(formula = LogSales ~ SqFt100s + Quality + LotSize + Garage +
    Bathrms + AC, data = AppendixC7)

Coefficients:
(Intercept)     SqFt100s      Quality      LotSize       Garage      Bathrms
  11.856608     0.028320    -0.206487     0.000004     0.064316     0.041774
         AC
   0.055222

>
> step(Full, direction = "backward", trace=F)

Call:
lm(formula = LogSales ~ SqFt100s + AC + Bathrms + LotSize + Garage +
    Quality, data = AppendixC7)

Coefficients:
(Intercept)     SqFt100s           AC      Bathrms      LotSize       Garage
  11.856608     0.028320     0.055222     0.041774     0.000004     0.064316
    Quality
  -0.206487
```

```
> step(Base, scope = list(upper=Full, lower=~1), direction = "forward",
trace=T)
Start:  AIC=-876
LogSales ~ 1

           Df Sum of Sq  RSS   AIC
+ SqFt100s  1      68.4 28.6 -1511
+ Quality   1      60.2 36.9 -1379
+ Bathrms   1      53.7 43.4 -1295
+ Garage    1      34.5 62.5 -1104
+ Bedrms    1      22.8 74.3 -1014
+ AC        1      12.1 85.0  -944
+ LotSize   1       5.7 91.4  -906
+ Pool      1       2.8 94.3  -889
<none>                  97.1  -876
+ NearHwy   1       0.2 96.9  -875

Step:  AIC=-1511
LogSales ~ SqFt100s

          Df Sum of Sq  RSS   AIC
+ Quality  1      7.77 20.9 -1674
+ Garage   1      2.99 25.7 -1567
+ Bathrms  1      2.72 25.9 -1561
+ AC       1      1.72 26.9 -1541
+ LotSize  1      1.20 27.4 -1532
<none>                 28.6 -1511
+ Pool     1      0.10 28.6 -1511
+ Bedrms   1      0.04 28.6 -1510
+ NearHwy  1      0.00 28.6 -1509

Step:  AIC=-1674
LogSales ~ SqFt100s + Quality
```

First two steps of "forward," with "trace = TRUE"

In Step 1, "SqFt100s" is added (best Sum of Sq)

In Step 2, "Quality" is added (best Sum of Sq from what's left)

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.