Lecture 29

Model Selection II

In this lecture, we will discuss methods for building models; that is, for selecting the best predictor variables to include in a linear regression model. The data we will use are data about predicting baseball salaries from player statistics and are available at: http://www.amstat.org/publications/jse/jse data archive.html

Methods for Model Building

If two models are nested, they can be compared directly using a nested models F-test as discussed earlier. This is the method used for forward and backward selection procedures where variables are added (removed) from the model one at a time until the model does not become better (worse). Minitab can do something better than this - it compares all possible models using a best subset approach and spits out statistics about the models that allows you to choose your own best model. There is, however, no set way to compare non-nested models. There are many, many ways. A few of the common ways are described in the following table. In this table $\boldsymbol{k}$ is the number of predictors included in the model and $\boldsymbol{n}$ is the number of individuals (the sample size). The full model includes all possible predictors.

Criterion	Formula	Comments
$\boldsymbol{C_p}$	$(k+1)+(n-(k+1))\dfrac{\hat{\sigma}^2-\hat{\sigma}^2_{\text{Full}}}{\hat{\sigma}^2_{\text{Full}}}$	Mallow's criterion. Want the value to be closed to $\boldsymbol{k+1}$.
AIC	$n\log(\hat{\sigma}^2)+2(k+1)$	Aikake Information Criterion. Larger penalty for more predictors. Want the value as small as possible.
BIC	$n\log(\hat{\sigma}^2)+(k+1)\log n$	Bayesian Information Criterion. Even larger penalty for more predictors, especially if the sample size is large too. Tries to correct for over-fitting. Want the value as small as possible.
GCV	$\dfrac{n^2\hat{\sigma}^2}{(n-(k+1))^2}=\dfrac{n\text{SSE}}{(n-(k+1))^2}$	Generalized Cross Validation. Approximates the leaving-one-out routine for cross-validation. You want this value as small as possible too.

Many of the model selection procedures above can be described as maximizing a penalized log likelihood function. The problem is that the penalty can be arbitrary leading to an infinite number of model selection procedures. Moreover, not only is there no one way of selecting variables but there may be several nearly equally good models. You also want to use common sense and convenience when constructing a model. For example, variables that are expensive to collect should only go in the model if absolutely necessary, for instance. Also, if all else is equal, models with fewer variables are preferred to models with more variables.

We will go over how to use best subsets in Minitab with the baseball data. With that data, we should first figure out what the right transformations of the variables should be before we do variable selection procedures as described above. In the example, salaries should be log transformed.

The best subsets table from Minitab for the baseball data then looks like the following:

```
Best Subsets Regression: log_salary versus Batting_aver, On_base_perc, ...

Response is log_salary


                                                                     a
                                                                     r
                                           O                         b
                                           n                         i
                                           _                         t
                                         B b                         r
                                         a a                         a
                                         t s                         t
                                         t e                 S       i
                                         i _                 t   f   o a
                                         n p               S o   r f n r
                                         g e               t l   e r _ b
                                         _ r               r e   e e p i
                                         a c     d t       i n   _ e o t
                                         v e     o r h     k _ E a _ s r
                                         e n     u i o   w e b r g a s a
                                         r t r h b p m   a O a r e g i t
                                         a a u i l l e R l u s o n e b i
                       Mallows           g g n t e e r B k t e r c n l o
Vars  R-Sq  R-Sq(adj)       Cp        S  e e s s s s s I s s s s y t e n
   1  44.6       44.4    562.2  0.87734        X
   1  42.5       42.4    594.8  0.89317                X
   2  64.8       64.5    238.0  0.70051                          X   X
   2  62.6       62.4    272.5  0.72145        X                 X
   3  77.2       77.0     38.7  0.56395                X         X   X
   3  77.0       76.8     41.8  0.56639        X                 X   X
   4  78.3       78.1     22.6  0.55075        X       X         X   X
   4  78.3       78.0     24.2  0.55198      X         X         X   X
   5  79.0       78.7     14.2  0.54333      X         X   X     X   X
   5  78.9       78.6     16.1  0.54487        X       X         X X X
```

```
 6  79.4       79.0      9.3  0.53854        X       X   X     X X X
 6  79.4       79.0      9.4  0.53863      X         X   X     X X X
 7  79.6       79.2      7.8  0.53655        X       X X X     X X X
 7  79.6       79.2      8.3  0.53696        X       X   X X   X X X
 8  79.8       79.3      7.0  0.53506    X   X       X X X     X X X
 8  79.8       79.3      7.7  0.53566        X       X X X X   X X X
 9  79.9       79.4      6.9  0.53413    X   X       X X X X   X X X
 9  79.9       79.4      7.0  0.53421    X   X       X X X   X X X X
10  80.0       79.4      7.1  0.53351    X   X       X X X X X X X X
10  80.0       79.4      7.9  0.53418    X   X   X   X X X X   X X X
11  80.1       79.5      7.9  0.53330    X   X   X   X X X X X X X X
11  80.1       79.4      8.5  0.53380    X   X     X X X X X X X X X
12  80.1       79.4      9.5  0.53382    X   X   X X X X X X X X X X
12  80.1       79.4      9.7  0.53396  X X   X   X   X X X X X X X X
13  80.2       79.4     11.3  0.53448  X X   X   X X X X X X X X X X
13  80.2       79.4     11.4  0.53451    X   X   X X X X X X X X X X X
14  80.2       79.3     13.2  0.53518  X X   X   X X X X X X X X X X X
14  80.2       79.3     13.2  0.53521  X X   X X X X X X X X X X X X
15  80.2       79.3     15.1  0.53592  X X   X X X X X X X X X X X X X
15  80.2       79.2     15.1  0.53598  X X X X   X X X X X X X X X X X
16  80.2       79.2     17.0  0.53671  X X X X X X X X X X X X X X X X
```

Mallow's Cp criterion picks out one model as best; the one with 7 predictors: hits, runs, walks, strike-outs, free-agency, free agency possible, and arbitration possible.

We can also find the AIC and BIC criteria. However, these select different models as we will see in class.

Finally, BIC has the nice interpretation of giving posterior model distributions. That is, if all models are equally likely to begin with, then the posterior probability on a model is proportional to **exp(−**BIC**)**. Using this (and some necessary scaling) we can construct a posterior probability distribution on our sets of models. We will discuss the advantages of this approach in class.

Exercises for Lecture 29

1. –

2. –