HANDOUT - Stat 230 - Variable Selection (c4.2)

P.B.Matheson adadpted from S.M. Liao and A.S. Wagaman

At the top of this file, note we loaded a new library: *leaps*, which is needed for these functions. You should check that it is installed/loaded if trying to work with these commands. There are also options using the *HH* library (optional) where you'd need to install this package once to use it.

The HELPrct data set contains 453 observations on 27 variables. The data is the data set on Health Evaluation and Linkage to Primary Care study results. The HELP study was a clinical trial for adult inpatients recruited from a detoxification unit. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking them to primary medical care.

```
data(HELPrct)
#names(HELPrct) # shows the variable names in the dataset
#help(HELPrct) #will tell you more about the dataset
dim(HELPrct) #displays the number of rows and columns in the dataset
```

```
## [1] 453  30
```

```
#str(HELPrct) #quickly shows you the data
```

We want to eliminate those subjects with missing data (for our methods to run - dealing with missing data may be a topic that needs further investigating for your projects).

```
HELPrct <- with(HELPrct, HELPrct[ !is.na(drugrisk), ])
dim(HELPrct)
```

```
## [1] 452  30
```

This eliminates the ONE data point with missing values. Notice the dimensions of the data file has gone from 453 rows(observations)with 30 columns(variables) to 452 rows.

Let's predict a baseline measure of depression (cesd) using other baseline variables including age, number of previous hospitalizations (d1), drug risk (drugrisk), average (i1) and maximum (i2) number of drinks in a day (last 30 days), inventory of drug use score (indtot), mental (mcs) and physical (pcs) component score, perceived social support (pss_fr), and sex risk score (sexrisk).

Using what we know now we will just throw in all the predictors.

```
modall <- lm(cesd ~ age + d1 + drugrisk + i1 + i2 + indtot +
               mcs + pcs + pss_fr + sexrisk, data = HELPrct)
msummary(modall)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.9575     4.3436   14.95   <2e-16 ***
## age          -0.0411     0.0565   -0.73    0.467
## d1           -0.0820     0.0702   -1.17    0.244
## drugrisk      0.0467     0.0993    0.47    0.638
## i1            0.0408     0.0397    1.03    0.304
## i2            0.0128     0.0282    0.45    0.649
## indtot        0.0697     0.0648    1.08    0.283
## mcs          -0.6171     0.0352  -17.55   <2e-16 ***
## pcs          -0.2396     0.0410   -5.84    1e-08 ***
## pss_fr       -0.2275     0.1057   -2.15    0.032 *
## sexrisk      -0.3085     0.1492   -2.07    0.039 *
##
## Residual standard error: 8.66 on 441 degrees of freedom
## Multiple R-squared:  0.532,  Adjusted R-squared:  0.521
## F-statistic: 50.1 on 10 and 441 DF,  p-value: <2e-16
```

```
car::vif(modall)
```

```
##      age       d1 drugrisk       i1       i2   indtot      mcs      pcs
##  1.13368  1.13800  1.11399  3.79328  3.76596  1.28994  1.22316  1.17986
##   pss_fr  sexrisk
##  1.06699  1.05145
```

QUESTIONS:

1. What variables are significant in predicting cesd?
2. How well does the model fit?
3. Do we observe any issues with multicollinearity that might make interpreting coefficients difficult?

Let's see if we can simplify the model building process. There are 4 variable selection techniques (methods) we will review (best subsets, backward, forward, and stepwise) to come up with a good set of predictors for predicting cesd.

We will start by considering all 10 predictors and use **best subsets**. This technique helps determine the best model(s) for each set of models of a given size, e.g., 1 predictor, 2 predictors, ..., all way through the full-term (10 predictor) model.

To run this procedure (regsubsets), you must specify what criteria will be used to define the *best*. One commonly used measure is "Mallows' Cp". R defaults to Akaike Information Criteria (AIC) unless you specify otherwise. Note: the lower the Cp or AIC, the better. The Bayesian Information Criteria (BIC) is also available.

DEFINITIONS:

Mallow's C_p is defined as:

$$C_p = \frac{SSE_m}{MSE_k} + 2(m+1) - n$$

where there is a subset of m predictors from a full set of k predictors, with a sample size n.

OTHER SELECTION CRITERIA that can be used to evaluate best model (FYI only):

$$AIC = n \cdot log(MSE_m) + 2(m+1)$$
$$BIC = n \cdot log(MSE_m) + log(n) \cdot (m+1)$$

Best Subsets using Mallow's Cp to determine best option for variable inclusion

First we set the list of explanatory variables for the leaps function to work on. Then we tell it to run using the Cp method and plot the results.

```
best <- regsubsets(cesd ~ age + d1 + drugrisk + i1 + i2 + indtot +
              mcs + pcs + pss_fr + sexrisk, data = HELPrct, nbest = 1)
with(summary(best), data.frame(rsq, adjr2, cp, rss, outmat))
```

```
##              rsq    adjr2       cp     rss age d1 drugrisk i1 i2 indtot mcs
## 1  ( 1 ) 0.464077 0.462886 56.79474 37870.4                               *
## 2  ( 1 ) 0.512002 0.509828 13.65368 34483.8                               *
## 3  ( 1 ) 0.518003 0.514775 10.00083 34059.8                 *             *
## 4  ( 1 ) 0.523356 0.519091  6.95863 33681.5                 *             *
## 5  ( 1 ) 0.527413 0.522115  5.13689 33394.8                 *             *
## 6  ( 1 ) 0.529268 0.522921  5.38979 33263.7      *          *             *
## 7  ( 1 ) 0.530761 0.523363  5.98401 33158.2      *          *         *   *
## 8  ( 1 ) 0.531377 0.522914  7.40393 33114.7   *  *          *         *   *
##          pcs pss_fr sexrisk
## 1  ( 1 )
## 2  ( 1 )   *
## 3  ( 1 )   *
## 4  ( 1 )   *      *
## 5  ( 1 )   *      *       *
## 6  ( 1 )   *      *       *
## 7  ( 1 )   *      *       *
## 8  ( 1 )   *      *       *
```

The HH package has a nicer function for displaying the regsubsets results.

```
summaryHH(best)
```

```
##                  model p   rsq   rss adjr2    cp  bic stderr
## 1                    m 2 0.464 37870 0.463 56.79 -270   9.17
## 2                 m-pc 3 0.512 34484 0.510 13.65 -306   8.76
## 3              i1-m-pc 4 0.518 34060 0.515 10.00 -305   8.72
## 4           i1-m-pc-p_ 5 0.523 33681 0.519  6.96 -304   8.68
## 5         i1-m-pc-p_-s 6 0.527 33395 0.522  5.14 -302   8.65
## 6      d1-i1-m-pc-p_-s 7 0.529 33264 0.523  5.39 -298   8.65
## 7   d1-i1-in-m-pc-p_-s 8 0.531 33158 0.523  5.98 -293   8.64
## 8 a-d1-i1-in-m-pc-p_-s 9 0.531 33115 0.523  7.40 -288   8.65
##
## Model variables with abbreviations
##                                                          model
## m                                                          mcs
## m-pc                                                     mcs-pcs
## i1-m-pc                                                i1-mcs-pcs
## i1-m-pc-p_                                          i1-mcs-pcs-pss_fr
## i1-m-pc-p_-s                                     i1-mcs-pcs-pss_fr-sexrisk
## d1-i1-m-pc-p_-s                                d1-i1-mcs-pcs-pss_fr-sexrisk
## d1-i1-in-m-pc-p_-s                       d1-i1-indtot-mcs-pcs-pss_fr-sexrisk
## a-d1-i1-in-m-pc-p_-s age-d1-i1-indtot-mcs-pcs-pss_fr-sexrisk
##
```

```
## model with largest adjr2
## 7
##
## Number of observations
## 452
```

We can examine lots of properties about the best subsets solutions in the output above. You can also adjust the summary to give you only certain statistics if you want to focus on just one or two. Here, we want to examine the model with the minimum Cp value, which we can see is the model with 5 predictors.

Remember, the procedure regsubsets does not actually run the regressions and show them to us. We have to actually fit the model chosen by best subsets, using only those variables marked by the best subsets process (a model with the 5 predictors of i1, mcs, pcs, pss_fr and sexrisk).

```
Cpmod <- lm(cesd ~ i1 + mcs + pcs + pss_fr + sexrisk, data = HELPrct)
msummary(Cpmod)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.0192     2.3428   28.18  < 2e-16 ***
## i1            0.0505     0.0209    2.42    0.016 *
## mcs          -0.6324     0.0324  -19.51  < 2e-16 ***
## pcs          -0.2291     0.0388   -5.91  6.9e-09 ***
## pss_fr       -0.2526     0.1040   -2.43    0.016 *
## sexrisk      -0.2887     0.1476   -1.96    0.051 .
##
## Residual standard error: 8.65 on 446 degrees of freedom
## Multiple R-squared:  0.527,  Adjusted R-squared:  0.522
## F-statistic: 99.5 on 5 and 446 DF,  p-value: <2e-16
```

4. How does this solution compare to the model with all 10 predictors?

```
anova(Cpmod,modall) # conducts nested F test to answer question 4
```

```
## Analysis of Variance Table
##
## Model 1: cesd ~ i1 + mcs + pcs + pss_fr + sexrisk
## Model 2: cesd ~ age + d1 + drugrisk + i1 + i2 + indtot + mcs + pcs + pss_fr +
##     sexrisk
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1    446 33395
## 2    441 33084  5     310.4 0.827  0.531
```

```
AIC(Cpmod,modall) # gives us the AIC values for the 2 models
```

```
##        df     AIC
## Cpmod   7 3241.44
## modall 12 3247.22
```

Backward elimination

Recall that backward elimination starts from the full model. The logic asks which variable can I remove? and repeat! We can use regsubsets with the backward method (see below). Since your book focuses on Cp as the criterion, it is selected here. The default maximum number of predictors it will test is 8. If you have a very large set of predictors, you may need to adjust that value as an option (nvmax).

```
backward <- regsubsets(cesd ~ age + d1 + drugrisk + i1 + i2 + indtot +
               mcs + pcs + pss_fr + sexrisk, data = HELPrct, method = "backward", nbest = 1)
with(summary(backward), data.frame(cp, outmat))
```

```
##               cp age d1 drugrisk i1 i2 indtot mcs pcs pss_fr sexrisk
## 1  ( 1 ) 56.79474                              *
## 2  ( 1 ) 13.65368                              *   *
## 3  ( 1 ) 10.00083                  *           *   *
## 4  ( 1 )  6.95863                  *           *   *      *
## 5  ( 1 )  5.13689                  *           *   *      *       *
## 6  ( 1 )  5.38979      *           *           *   *      *       *
## 7  ( 1 )  5.98401      *           *         *   *   *      *       *
## 8  ( 1 )  7.40393   *  *           *         *   *   *      *       *
```

In this particular case, the final model is the same one (line 5) achieved via minimizing the Cp from the best subset model. This does NOT always occur. Again, you still have to fit the model (run lm) selected to get its summary output.

Forward selection

For forward selection, we start from a model with just an intercept and build up a model one predictor at a time. The logic asks what variable can I add now? repeat!

```
forward <- regsubsets(cesd ~ age + d1 + drugrisk + i1 + i2 + indtot +
               mcs + pcs + pss_fr + sexrisk, data = HELPrct, method = "forward", nbest = 1)
with(summary(forward), data.frame(cp, outmat))
```

```
##               cp age d1 drugrisk i1 i2 indtot mcs pcs pss_fr sexrisk
## 1  ( 1 ) 56.79474                              *
## 2  ( 1 ) 13.65368                              *   *
## 3  ( 1 ) 10.00083                  *           *   *
## 4  ( 1 )  6.95863                  *           *   *      *
## 5  ( 1 )  5.13689                  *           *   *      *       *
## 6  ( 1 )  5.38979      *           *           *   *      *       *
## 7  ( 1 )  5.98401      *           *         *   *   *      *       *
## 8  ( 1 )  7.40393   *  *           *         *   *   *      *       *
```

This method also obtains the same model as our bestsubset option (line 5), and as backward selection. Again, the methods do not always agree on final models!

You still have to fit the final model yourself to get the model summary.

Stepwise Regression

Starts off looking like forward selection but allows for predictors to be kicked out (like backwards) as the process goes. It takes the unidirectional blinders off.

```
stepwise <- regsubsets(cesd ~ age + d1 + drugrisk + i1 + i2 + indtot +
              mcs + pcs + pss_fr + sexrisk, data = HELPrct, method = "seqrep", nbest = 1)
with(summary(stepwise), data.frame(cp, outmat))
```

```
##               cp age d1 drugrisk i1 i2 indtot mcs pcs pss_fr sexrisk
## 1  ( 1 ) 56.79474                                *
## 2  ( 1 ) 13.65368                                *   *
## 3  ( 1 ) 10.00083                 *              *   *
## 4  ( 1 )  6.95863                 *              *   *      *
## 5  ( 1 )  5.13689                 *              *   *      *       *
## 6  ( 1 )  5.38979      *          *              *   *      *       *
## 7  ( 1 )  5.98401      *          *          *   *   *      *       *
## 8  ( 1 ) 15.28611   *  *        *  *  *      *   *   *
```

The final decision about which model is best (based on Cp) remains the same, but you can lso see the size 8 model was different. You still have to fit the final model yourself to get the model summary. AND you have to check conditions, multicollinearity and outliers.

ANOTHER WAY to do this in R

A quick way to use `AIC` as the selection criteria can be with the command `stepAIC` and it is shown below. If you are too eager to watch all the steps, you can add the option after backward of trace=FALSE.

```
#stepAIC(modall,direction="backward")$anova
#stepAIC(modall,direction="backward",trace=0)$anova
```

You still have to fit the final model yourself to get the model summary.