Second Special Case: Two Continuous Variables

The model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$ can be rewritten as follows:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ Y &= \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon \end{aligned}$$

The coefficient of one explanatory variable depends on the value of the other explanatory variable.

Variable Selection and Model Building

We usually want to choose a model that includes a subset of the available explanatory variables.

Two separate but related questions:

- How many explanatory variables should we use (i.e., subset size)? Smaller sets are more convenient, but larger sets may explain more of the variation (SS) in the response.
- Given the subset size, which variables should we choose?

Criteria for Model Selection

To determine an appropriate subset of the predictor variables, there are several different criteria available. We will go through them one at a time, noting their benefits and drawbacks. They include R^2, adjusted R^2, Mallow's C_p, MSE, $PRESS$, AIC, SBC. SAS will provide these statistics, so you should pay more attention to what they are good for than how they are computed. To obtain them from SAS, place after the `model` statement `/selection = MAXR ADJRSQ CP`. Note that the different criterion may not lead to the same model in every case.

R^2 and Adjusted R^2 (or MSE) Criterion

- The text uses $R^2_p = R^2 = 1 - \frac{SSE}{SSTO}$ (see page 354). Their subscript is just the number of variables in the associated model.
- The goal in model selection is to maximize this criterion. One MAJOR drawback to R^2 is that the addition of any variable to the model (significant or not) will increase R^2 (perhaps not enough to notice depending on the variable). At some point, added variables just get in the way!
- The *Adjusted R^2 criterion* penalizes the R^2 value based on the number of variables in the model. Hence it eventually starts decreasing as unnecessary variables are added.

$$R^2_a = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} \text{ (we end up subtracting off more as } p \text{ is increased)}$$

- Maximizing the Adjusted R^2 criterion is one way to select a model. As the text points out this is equivalent to minimizing the MSE since
$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} = 1 - \frac{MSE}{SSTO/(n-1)} = 1 - \frac{MSE}{\text{constant}}$$

Mallow's C_p Criterion

- The basic idea is to compare subset models with the full model.
- The full model is good at prediction, but if there is multicollinearity our interpretations of the parameter estimates may not makes sense. A subset model is good if there is not substantial "bias" in the predicted values (relative to the full model).
- The C_p criterion looks at the ratio of error SS for the model with p variables to the MSE of the full model, then adds a penalty for the number of variables.
$$C_p = \frac{SSE_p}{MSE(Full)} - (n - 2p)$$
- SSE is based on a specific choice of $p-1$ variables (p is the number of regression coefficients *including the intercept*); while MSE is based on the full set of variables.
- A model is good according to this criterion if $C_p \leq p$. We may choose the smallest model for which $C_p \leq p$, so a benefit of this criterion is that it can achieve for us a "good" model containing as few variables as possible.
- One might also choose to pick the model that minimizes C_p.
- See page 357-359 for details.

PRESS Statistic

Stands for PREdiction Sums of Squares
Obtained by the following algorithm: For each observation i, delete the observation and predict Y for that observation using a model based on the $n-1$ cases. Then look at SS for the observed minus predicted.
$$PRESS_p = \sum (Y_i - \hat{Y}_{i(i)})^2$$
Models with small PRESS statistic are considered good candidates.

SBC* and *AIC

Criterion based log(likelihood) plus a penalty for more complexity.
$$AIC \quad -- \quad \text{minimize } n\log\left(\frac{SSE_p}{n}\right) + 2p$$
$$SBC \quad -- \quad \text{minimize } n\log\left(\frac{SSE_p}{n}\right) + p\log(n)$$
Note that different criteria will not give the identical answer.

Model Selection Methods

There are three commonly available. To apply them using SAS, you use the option `selection = *****` after the `model` statement. The methods include

- *Forward Selection* (`FORWARD`) - starts with the null model and adds variables one at a time.
- *Backward Elimination* (`BACKWARD`) - starts with the full model and deletes variables one at a time.
- *Forward Stepwise Regression* (`STEPWISE`) - starts with the null model and checks for adds/deletes at each step. This is probably the preferred method (see section on multicollinearity!). It is forward selection, but with a backward glance at each step.

These methods all add/delete variables based on Partial F-tests. (See page 364)

Some additional options in the `model` statement

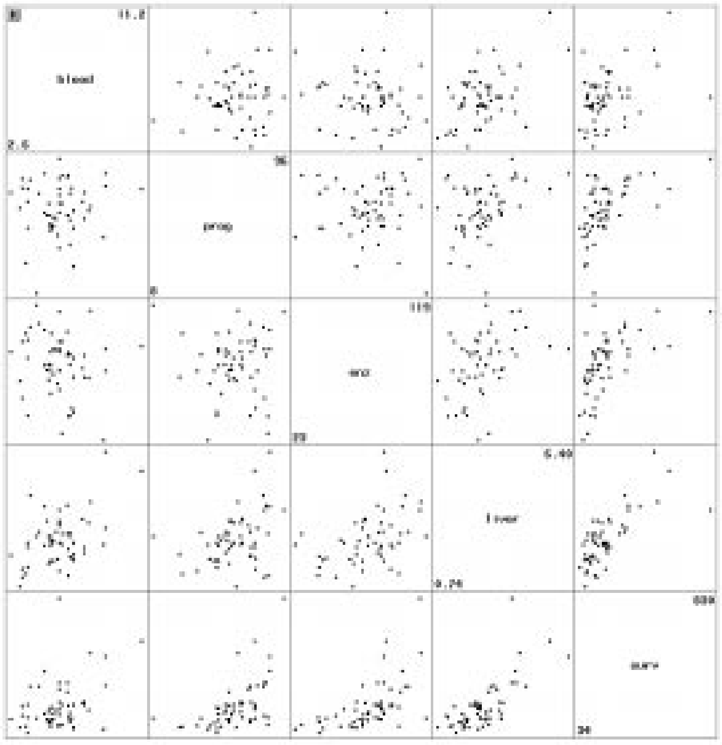
`INCLUDE=n` forces the first n explanatory variables into all models
`BEST=n` limits the output to the best n models of each subset size
`START=n` limits output to models that include at least n explanatory variables

Ordering Models of the Same Subset Size

Use R^2 or SSE.
This approach can lead us to consider several models that give us approximately the same predicted values.
May need to apply knowledge of the subject matter to make a final selection.
If prediction is the key goal, then the choice of variables is not as important as if interpretation is the key.

Surgical Unit Example

- References: KNNL Section 9.2 (p350ff), `knnl350.sas`
- Y is the survival time
- Potential X's include Blood clotting score (X_1), Prognostic Index (X_2), Enzyme Function Test (X_3), and Liver Function Test (X_4).
- $n = 54$ patients were observed.
- Initial diagnostics note curved lines and non-constant variance, suggesting that Y should be transformed with a log. Take a look at the plots in the SAS file and play with some analyses on your own.



```
data surgical;
   infile 'H:\System\Desktop\Ch08ta01.dat';
   input blood prog enz liver surv;
```

Take the log of survival

```
data surgical;
   set surgical;
   lsurv=log(surv);
proc reg data=surgical;
   model lsurv=blood prog enz liver/
   selection=rsquare cp aic sbc b best=3;
```

```
Number in
  Model      R-Square        C(p)           AIC           SBC
       1       0.5274    787.9471      -87.3085     -83.33048
       1       0.4424    938.6707      -78.3765     -74.39854
       1       0.3515    1099.691      -70.2286     -66.25061
--------------------------------------------------------------
       2       0.8129    283.6276     -135.3633    -129.39638
       2       0.6865    507.8069     -107.4773    -101.51034
       2       0.6496    573.2766     -101.4641     -95.49714
--------------------------------------------------------------
       3       0.9723      3.0390     -236.5787    -228.62281
       3       0.8829    161.6520     -158.6434    -150.68745
       3       0.7192    451.8957     -111.4189    -103.46299
--------------------------------------------------------------
       4       0.9724      5.0000     -234.6217    -224.67680
```

One model stands out: the first one with 3 variables ($C_p = 3.04 < p = 4$). The full model has $C_p = 5 = p$. The parameter estimates indicate that the desired model is the one with `blood`, `prog` and `enz`, but not `liver`.

```
Number in                 --------------------------Parameter Estimates--------------------------
  Model      R-Square      Intercept        blood           prog             enz            liver

       1       0.5274        3.90609             .               .               .        0.42771
       1       0.4424        3.55863             .               .         0.01973              .
       1       0.3515        3.68138             .         0.02211               .              .
-----------------------------------------------------------------------------------------------
       2       0.8129        2.08947             .         0.02271         0.02015              .
       2       0.6865        3.19784             .               .         0.01301        0.32010
       2       0.6496        3.24325             .         0.01403               .        0.34596
-----------------------------------------------------------------------------------------------
       3       0.9723        1.11358       0.15940         0.02140         0.02193              .
       3       0.8829        2.16970             .         0.01819         0.01612        0.18846
       3       0.7192        2.68966       0.09239               .         0.01604        0.22556
-----------------------------------------------------------------------------------------------
       4       0.9724        1.12536       0.15779         0.02131         0.02182        0.00442
```

In this particular example you would probably come to the same conclusion based on the Type II *SS*, but not on the individual correlations: *liver* has the *highest* individual correlation with *lsurv* (but also is correlated with the other three).

Below we see that this is the same model chosen by forward stepwise regression:

```
proc reg data=surgical;
   model lsurv=blood prog enz liver / selection=stepwise;
```

```
                                      Summary of Stepwise Selection
        Variable      Variable      Number      Partial      Model
Step    Entered       Removed       Vars In     R-Square     R-Square      C(p)       F Value     Pr > F
  1     liver                          1         0.5274       0.5274     787.947       58.02     <.0001
  2     enz                            2         0.1591       0.6865     507.807       25.89     <.0001
  3     prog                           3         0.1964       0.8829     161.652       83.83     <.0001
  4     blood                          4         0.0895       0.9724      5.0000      158.65     <.0001
  5                   liver            3         0.0000       0.9723      3.0390        0.04     0.8442
```