BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 18 DATE: May 5, 1989

SUBJECT: Multiple Regression: Selecting the best subset

Researchers with a multiple regression at hand, frequently wonder if all the independent variables are necessary. Suppose a multiple regression with k variables fits the data quite well. The question is: "Would a regression with, say, p variables (p < k) fit just as well?" Stepwise methods have traditionally been used to answer this question, primarily because of computing limitations. Weisberg (1980, page 195) states that:

> "Unfortunately, there are important drawbacks to the use of stepwise procedures. Firstly, ... the model chosen by stepwise regression need not be the best of any criterion of interest; indeed, because of the nature of the one-at-a-time philosophy of stepwise methods, there is no guarantee that the model chosen will in fact include any of the variables that would be in the best subset. Stepwise methods are best when the independent variables are nearly uncorrelated, the condition under which finding a subset model is least likely to be relevant. Also, it is possible to construct examples in which the best subset of size p = 2 is completely disjoint from the best subset for p = 3, and so on.
>
> Probably the worst indictment of stepwise techniques, at least for the user who is not statistically sophisticated, is that they produce a single result that appears to be the model. Similarly, many users pay undue attention to the order in which the variables are entered or deleted from a model ...The ordering of the variables that we get from stepwise regression is an artifact of the algorithm used and need not reflect relationships of substantive interest."

Why not fit all possible subsets? Many computing packages do this quite readily now. Computing limitations are not the problem anymore.

SAS, for instance, will fit all subsets, or, selected "best" subsets if there are many independent variables. Suppose that Y is the dependent variable with X_1, X_2, ..., X_k, as independent variables.

i) The older versions of SAS (i.e. those before Version 6) would use the following code:

```
PROC RSQUARE CORR;
  MODEL Y = X1 X2 -- Xk/CP ADJRSQ MSE;
```

ii) Version 6 uses:

```
PROC REG CORR;
  MODEL Y = X1 X2 -- Xk/SELECTION = CP ADJRSQ MSE;
```

DEFINITIONS:

1. SS = sums of squares
2. MS = mean square
3. CP = Mallow's C(p) statistic:

$$C(p) = \frac{SS(\text{k variable model}) - SS(\text{p variable model})}{MS(\text{k variable model})} + 2p - (k + 1)$$

If the p variable model is as good as the k variable model then $C(p) \leq p+1$

4. RSQUARE = R^2 = squared correlation coefficient:

$$R^2 = SS(\text{Model}) / SS(\text{total}) = 1 - SS(\text{residuals}) / SS(\text{total})$$

R^2 is often interpreted as the proportion of the variability in the data which is explained by the model.

5. ADJRSQ = Adjusted R^2:

$$R^2_{adj} = 1 - (n-1)(1-R^2)/(n-p)$$

= R^2 adjusted for the influence of the sample size, n, relative to the number of variables, p. Thus, if n >> p then $R^2_{adj} \simeq R^2$.

An example of the output (SAS/PC version) is shown below:

```
                         EXAMPLE FOR SAS/PC PRACTICE
                           ---- PROGRAM SEC41 ----
                       REGRESSION OF Y ON X1, X2, and X3

N = 192           Regression Models for Dependent Variable: Y

     C(p)        R-square           Adjusted       MSE        Variables in
                             In     R-Square                  Model

      0.39001    0.39522833     1    0.39204532    32.979444     X2
      2.10964    0.39612839     2    0.38973821    33.104596     X1 X2
      2.29141    0.39554487     2    0.38914852    33.136585     X2 X3
      4.00000    0.39648035     3    0.38684971    33.261287     X1 X2 X3
      4.11890    0.38325781     1    0.38001179    33.632221     X1
      6.02869    0.38354739     2    0.37702408    33.794294     X1 X3
    120.73759    0.00888723     1    0.00367085    54.047419     X3
```

Note that the subsets are sorted by C(p) since it was listed first in the `MODEL` statement. R^2 is also output although it was not specified in the `MODEL` statement. The column titled `In` is the number of variables (p) in the model. Thus, look for models with C(p) ≤ `In` + 1.

The output can be used:

i) to test an individual model or regression by:

$$F = \frac{R^2/df_n}{(1 - R^2)/df_d} \ , \ df_n = p, \ df_d = n - (p + 1)$$

ii) to compare two models when one model (with p_1 variables and R_1^2) contains more variables than the other (with p_2 variables and R_2^2). The comparison tests whether adding the extra variables explains a significant amount of variation in the data (using the additional sums of squares principle). The test is calculated by:

$$F = \frac{(R_1^2 - R_2^2)/df_n}{(1 - R_1^2)/df_d} \ , \ df_n = p_1 - p_2, \ df_d = n - (p_1 + 1)$$

Note: In this case, you **might** choose the R^2 for the model with ALL the variables (instead of R_1^2) for the denominator (see Example 3 below), since an assumption of Mallow's C(p) statistic is that the model with all variables in it is a "correct" model.

EXAMPLES:

1. Test model: Y = X2

$$F = \frac{(0.3952/1)}{(1\text{-}0.3952)/(192\text{-}2)} = 124.2, \ df = 1, 190$$

2. Test model: Y = X1, X2

$$F = \frac{(0.3961)/2}{(1\text{-}0.3961)/(192\text{-}3)} = 62.0, \ df = 2, 189$$

3. Test contribution of X1 to model: Y = X1, X2

$$F = \frac{(0.3961\text{-}0.3952)/(2\text{-}1)}{(1\text{-}0.03965)/(192\text{-}4)} = 0.28, \ df = 1, 188$$

Tests 1 and 2 indicate that both models provide a statistically "significant" fit. Nevertheless the third test indicates that X2 does as good a job explaining the variation in the data as does X1 and X2 together. Thus a model with only X2 in it may be sufficient.

The next step in the data analysis is to examine the model with X2 for lack of fit, heterogeneity of variance etc. In other words, you still don't know if X2 is the best model. For instance, the model with X1 and X2 may have better looking residuals, etc.

Reference:

Weisberg, S., 1980. *Applied linear regression.* John Wiley and Sons, New York, New York

CONTACT: Wendy Bergerud
387-5676

PROBLEMS

Using the example output, test the 3 variable model and the contribution made by X1 and X3. In other words, does X2 do as good a job as all three variables?

ANSWER: In the next pamphlet.