

VARIABLE SELECTION AND CASE ANALYSIS IN REGRESSION

Gerry Hobbs, West Virginia University

ABSTRACT

Developments in the area of regression analysis over the last fifteen years have been quite remarkable. In this paper we review some of those developments as they are reflected in PROC REG. In the first section some of the relevant literature is reviewed and notation is fixed. In Section II some examples are provided to demonstrate the inadequacy of standard summaries such as R-square. In the third and fourth sections certain influence diagnostics are reviewed. In Section V collinearity measures are defined and explained. In Section VI partial regression leverage plots are exhibited and discussed. In the seventh section some comments on model building and variable selection are given.

I. INTRODUCTION

Introductions to regression analysis are contained in many books. Some of the more popular and complete are Weisberg [1], Montgomery & Peck [2] and Draper & Smith [3]. More advanced treatments are contained in Cook & Weisberg [4] and Belsley, Kuh & Welsch [5]. An interesting review of recent developments is given by Hocking [6] and a technical report of particular interest to the current audience is [7]. A review of the issue of centering in collinearity diagnostics appears in [8] and subsequent discussion. Much of the currently available technology is based on a "leave-one-out" strategy related to jackknifing or cross validation. There are certain situations in which such diagnostics are ineffective, especially in circumstances where multiple outliers mask the effects of one another. Some additional work has been done in the area of multiple outliers by Gray & Ling in [9] and by Marasinghe in [10]. Mason & Gunat give a step by step approach to the identification of outlier induced collinearities in [11]. Users of SAS® software may find course notes [12] very useful.

Suppose we have a set of observable response variables  $y_1, y_2, \dots, y_n$  which we denote in the aggregate by the column vector  $y$ . Each of the  $y$  values is related to a set of known constants  $x_{11}, x_{12}, \dots, x_{1p}$ . For each of the  $y$ 's we assume that

$$y_i = b_0 + b_1 x_{i1} + \dots + b_p x_{ip} + e_i.$$

where the  $e_i$  are random variables with mean equal to 0. The  $b$ 's are fixed, if unknown, constants often called the regression parameters. It is convenient to represent the  $n$  equations defined above in matrix notation

$$y = Xb + e$$

where  $b = (b_1, \dots, b_p)'$ ,  $e = (e_1, \dots, e_n)'$  and the matrix

$$X = \begin{matrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{matrix}$$

The least squares solution to the matrix equation above is

$$b = (X'X)^{-1}X'y$$

The predicted values of the  $y$ 's are given by

$$y = Xb = X(X'X)^{-1}X'y$$

II. USUAL SUMMARIES

The quantities most often of concern in regression analyses are the estimated regression coefficients and a measure of goodness of fit,  $R^2$ , which is sometimes called the coefficient of determination. The problem with using these summaries, exclusively, to characterize the results of a regression problem is that very different patterns of points in a scatter plot can lead to exactly the same summary statistics. Consider, for instance, the case of simple linear regression and the data exhibited in the scatter plots of Figures 1 - 4.

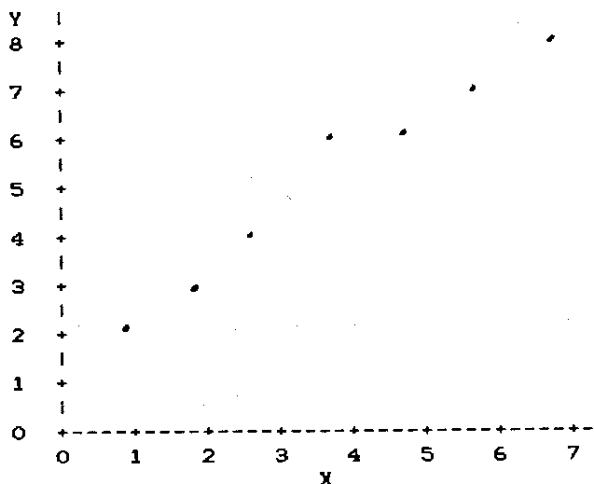
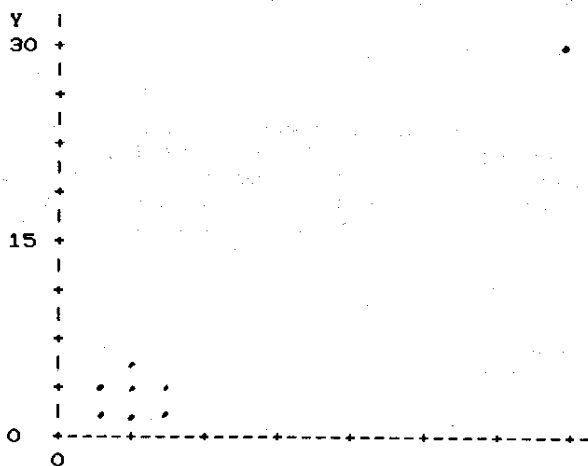
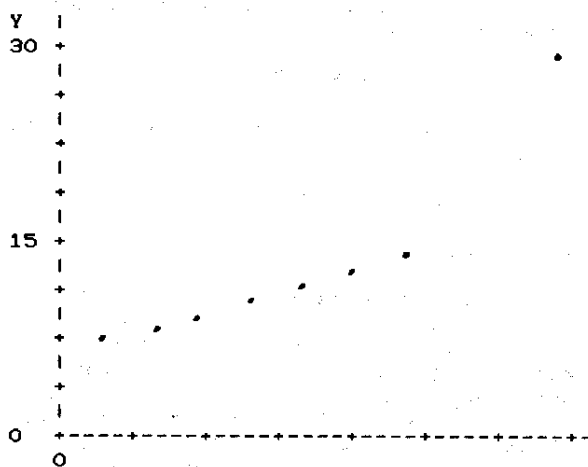


Figure 1

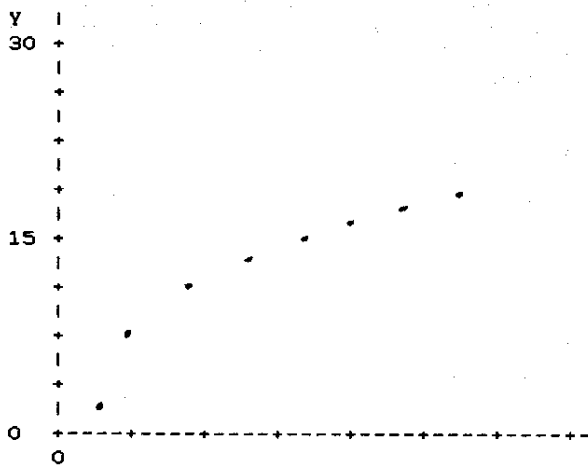
SAS is a registered trademark of the SAS Institute, Cary, N.C.



X  
Figure 2



X  
Figure 3



X  
Figure 4

In all four cases the intercept is about 1 the slope is about 1 and the value of  $R^2$  is about .95. In fact, similar examples can be constructed wherein all three summaries are exactly the same. It is quite obvious that there are different lessons to be learned from the different situations. In the case of simple linear regression we can always resort to the use of plots in order to clarify any ambiguities of this type. In circumstances where there are several predictor variables it is less clear how we can use plots to actually "see what is going on".

### III. INFLUENCE STATISTICS

One of the problems inherent in least squares regression solutions highlighted in Section II is the influence that a small subset of the observations may exert on the overall fit and on the values of individual coefficients in the estimated regression model. In order to assess the influence a particular observation exerts on a particular coefficient we can proceed as follows. For any case, say the  $i$ -th, the influence exerted on the  $j$ -th component of  $b$  is the difference between the estimated parameter with and without the  $i$ -th case deleted from the calculations

$$b_j(i) - b_j$$

where  $b_j(i)$  is the value of the estimate of the  $j$ -th coefficient using all of the data except for the  $i$ -th case and  $b_j$  is the same estimate except that all of the data is used. This difference is called  $DFBETA_{ij}$  in [5]. A scaled form of this quantity is available under PROC REG's INFLUENCE option and is exhibited in Table 1 for a data set which may be found in [1]. The data set consists of fuel consumption,  $Y$ , and four explanatory variables for each of the 48 contiguous states. In that table, also used in [13], the influence of each of the 48 cases for each of the five coefficients is given. One common criterion for "significance" is to judge values greater than two times the square root of  $n$  as indicative of significant influence. In a similar vein the influence a given observation exerts on model "fit" can be measured as the difference between the predicted values given by the model when the  $i$ -th observation is left out and when the full data set is used. Of course that difference depends on which point in factor space we consider. Most of our interest focuses on the difference

|    | ORS      | RESIDUAL | RSTUDENT | HAT    | DIAG    | COV     | DFBETAS  | DFRETAS | DFBETAS | DFBETAS | DFBETAS |
|----|----------|----------|----------|--------|---------|---------|----------|---------|---------|---------|---------|
|    |          |          |          | H      | RATIO   | DFIT    | INTERCEP | TAX     | X1      | X2      | X3      |
| 1  | 17.7386  | 0.2785   | 0.0963   | 1.2333 | 0.0909  | 0.0105  | 0.0365   | -0.0462 | -0.0149 | -0.0139 |         |
| 2  | -29.1207 | -0.4323  | 0.0740   | 1.1856 | -0.1279 | 0.0352  | -0.0662  | 0.0209  | 0.0309  | -0.0247 |         |
| 3  | -17.1248 | -0.2671  | 0.0847   | 1.2186 | -0.0813 | 0.0250  | -0.0465  | 0.0307  | 0.0099  | -0.0257 |         |
| 4  | -79.2873 | -1.2887  | 0.1252   | 1.0591 | -0.4875 | -0.2431 | 0.2655   | -0.2883 | 0.3245  | 0.2892  |         |
| 5  | -121.991 | -1.9980  | 0.0923   | 0.7871 | -0.6372 | -0.3026 | 0.2511   | -0.1648 | 0.5272  | 0.2767  |         |
| 6  | 23.4763  | 0.3993   | 0.2287   | 1.4310 | 0.2175  | -0.1337 | 0.1364   | 0.1170  | 0.0053  | 0.0295  |         |
| 7  | 25.3291  | 0.4471   | 0.2832   | 1.5324 | 0.2810  | -0.0051 | 0.0429   | 0.1591  | 0.1216  | -0.1504 |         |
| 8  | -24.4389 | -0.3871  | 0.1107   | 1.2425 | -0.1366 | -0.0164 | 0.0335   | -0.1035 | 0.0734  | 0.0484  |         |
| 9  | -25.002  | -0.3854  | 0.0608   | 1.1767 | -0.0980 | 0.0179  | -0.0397  | -0.0263 | -0.0602 | 0.0246  |         |
| 10 | -52.3458 | -0.8061  | 0.0479   | 1.0942 | -0.1809 | -0.0460 | 0.0609   | -0.0646 | -0.0435 | 0.0627  |         |
| 11 | 79.5364  | 1.2293   | 0.0358   | 0.9776 | 0.2370  | 0.0337  | 0.0320   | 0.0645  | 0.0239  | -0.1174 |         |
| 12 | 28.8008  | 0.4910   | 0.2305   | 1.4205 | 0.2687  | -0.0824 | 0.0788   | 0.1161  | 0.2006  | -0.0369 |         |
| 13 | -38.2476 | -0.5893  | 0.0557   | 1.1431 | -0.1431 | -0.0295 | 0.0632   | -0.0937 | 0.0118  | 0.0302  |         |
| 14 | -57.9841 | -0.8923  | 0.0433   | 1.0704 | -0.1898 | -0.1270 | 0.1175   | -0.0132 | 0.0386  | 0.0989  |         |
| 15 | -72.029  | -1.1146  | 0.0441   | 1.0171 | -0.2394 | 0.0543  | 0.0171   | -0.0027 | -0.0975 | -0.1032 |         |
| 16 | 30.7275  | 0.4749   | 0.0641   | 1.1702 | 0.1243  | -0.0283 | 0.0119   | -0.0010 | 0.0882  | 0.0287  |         |
| 17 | 5.53573  | 0.0842   | 0.0381   | 1.1685 | 0.0168  | 0.0024  | -0.0036  | -0.0011 | 0.0067  | -0.0002 |         |
| 18 | 117.618  | 1.8985   | 0.0734   | 0.8043 | 0.5343  | 0.4460  | -0.3428  | -0.1947 | -0.2492 | -0.2462 |         |
| 19 | 92.0837  | 1.5701   | 0.1904   | 1.0446 | 0.7614  | -0.4553 | 0.0684   | 0.0884  | 0.0923  | 0.6559  |         |
| 20 | -42.7102 | -0.7107  | 0.1874   | 1.3040 | -0.3414 | 0.2914  | -0.2314  | 0.0436  | -0.1568 | -0.2951 |         |
| 21 | -45.9339 | -0.7246  | 0.0954   | 1.1686 | -0.2353 | 0.1317  | -0.0228  | -0.0315 | -0.0803 | -0.1770 |         |
| 22 | -30.1802 | -0.4782  | 0.1094   | 1.2292 | -0.1676 | 0.0692  | 0.0338   | -0.0997 | 0.1082  | -0.0048 |         |
| 23 | 48.9547  | 0.7805   | 0.1127   | 1.1796 | 0.2781  | -0.0141 | 0.0682   | 0.1586  | -0.0647 | -0.1221 |         |
| 24 | 86.7772  | 1.3743   | 0.0737   | 0.9746 | 0.3877  | -0.0803 | 0.2483   | 0.0127  | 0.0862  | -0.0946 |         |
| 25 | -47.0001 | -0.7242  | 0.0518   | 1.1149 | -0.1692 | 0.0115  | -0.0332  | -0.0711 | 0.0619  | 0.0380  |         |
| 26 | 34.0532  | 0.5345   | 0.0912   | 1.1965 | 0.1693  | -0.0494 | 0.1247   | -0.0818 | 0.0569  | 0.0223  |         |
| 27 | -11.6649 | -0.1800  | 0.0657   | 1.1993 | -0.0478 | -0.0098 | -0.0109  | 0.0373  | -0.0064 | 0.0004  |         |
| 28 | 18.9142  | 0.2919   | 0.0646   | 1.1904 | 0.0767  | -0.0178 | 0.0275   | -0.0376 | 0.0540  | 0.0247  |         |
| 29 | 15.9913  | 0.2418   | 0.0265   | 1.1475 | 0.0399  | -0.0086 | 0.0172   | -0.0049 | 0.0132  | 0.0031  |         |
| 30 | 61.9872  | 0.9906   | 0.1090   | 1.1248 | 0.3466  | 0.0338  | 0.1737   | -0.1554 | 0.0619  | -0.1087 |         |
| 31 | 4.1076   | 0.0638   | 0.0781   | 1.2196 | 0.0186  | 0.0146  | -0.0088  | -0.0080 | -0.0020 | -0.0102 |         |
| 32 | -27.4205 | -0.4341  | 0.1088   | 1.2343 | -0.1517 | -0.1174 | 0.0635   | 0.0895  | 0.0188  | 0.0723  |         |
| 33 | -74.6908 | -1.2204  | 0.1377   | 1.0958 | -0.4876 | 0.0300  | -0.1804  | 0.4252  | -0.1681 | -0.1578 |         |
| 34 | 14.1258  | 0.2198   | 0.0803   | 1.2160 | 0.0649  | 0.0390  | -0.0171  | -0.0468 | -0.0194 | -0.0130 |         |
| 35 | -19.4224 | -0.3063  | 0.1041   | 1.2416 | -0.1044 | -0.0753 | 0.0221   | 0.0434  | 0.0370  | 0.0680  |         |
| 36 | -72.8019 | -1.1500  | 0.0909   | 1.0481 | -0.3412 | -0.0090 | 0.0802   | 0.1625  | -0.0839 | -0.1694 |         |
| 37 | -7.22739 | -0.1302  | 0.3151   | 1.6390 | -0.0883 | -0.0165 | 0.0258   | 0.0095  | -0.0517 | 0.0052  |         |
| 38 | 60.8893  | 0.9365   | 0.0404   | 1.0572 | 0.1923  | 0.0692  | -0.0746  | -0.0884 | -0.0088 | 0.0231  |         |
| 39 | -69.6743 | -1.1590  | 0.1709   | 1.1590 | -0.5261 | 0.2898  | -0.2789  | 0.2731  | -0.0878 | -0.4200 |         |
| 40 | 234.989  | 4.4938   | 0.0997   | 0.1755 | 1.4956  | -0.3483 | -0.3183  | -0.0223 | -0.3820 | 1.0181  |         |
| 41 | -75.8213 | -1.1818  | 0.0544   | 1.0100 | -0.2833 | -0.0020 | 0.1228   | -0.0530 | 0.1014  | -0.1055 |         |
| 42 | 65.9531  | 1.0352   | 0.0745   | 1.0715 | 0.2936  | 0.2143  | -0.1802  | -0.1326 | -0.1547 | -0.0706 |         |
| 43 | -12.4431 | -0.1911  | 0.0570   | 1.1877 | -0.0470 | -0.0171 | 0.0299   | -0.0047 | 0.0293  | -0.0026 |         |
| 44 | 34.0627  | 0.5525   | 0.1488   | 1.2745 | 0.2310  | 0.2105  | -0.1633  | -0.0367 | -0.1563 | -0.1492 |         |
| 45 | 68.3119  | 1.2095   | 0.2661   | 1.2915 | 0.7282  | 0.1068  | -0.4744  | 0.3580  | -0.4245 | 0.1021  |         |
| 46 | -9.67192 | -0.1498  | 0.0723   | 1.2093 | -0.0418 | 0.0258  | -0.0321  | -0.0062 | -0.0092 | -0.0107 |         |
| 47 | -60.3544 | -0.9360  | 0.0563   | 1.0751 | -0.2287 | -0.0302 | 0.1118   | -0.0065 | 0.1056  | -0.0750 |         |
| 48 | -45.6729 | -0.7127  | 0.0879   | 1.1614 | -0.2243 | 0.0706  | 0.0034   | -0.1305 | -0.1020 | -0.0275 |         |

Table 1

observed at the set of  $x$ 's corresponding to the observation left out. This difference measures the influence the observation has near its own region in the  $x$ -space and is called DFFIT. A scaled form of this quantity is shown in the fifth column, labeled DFFITS, of Table 1.

The most commonly used diagnostic tool is the set of ordinary residuals. One problem with ordinary residuals is that they are not independent of one another and while they all have mean 0 (if the model is right) the variances are not all the same. In fact, the variance of the  $i$ -th ordinary residual is  $(1-v_{ii}) \text{Var}(e_i)$ . The  $v_{ii}$  can be shown

to be related to the distance from the  $i$ -th set of  $x$ 's to the centroid of the set of all  $x$ 's. The  $v_{ii}$  are the diagonals of the "hat" matrix  $X(X'X)^{-1}X'$ . Cases where  $v_{ii}$  is large correspond to cases which are remote in factor space. Thus, cases with large  $v_{ii}$  must necessarily have small residuals regardless of the value of  $y_i$ . It is clear that such points exert a very strong potential influence in their own region and that they can influence the fit in other regions as well. The HAT DIAG column of Table 1 lists the values of  $v_{ii}$ . Large values of  $v_{ii}$  indicate points which have unusual combinations of  $x$  values and are called high leverage

points. Of particular note in Table 1 are cases 37 (Texas) and 7 (New York) which may be seen in the original data set to have rather extreme values for some of the  $x$ 's. The ordinary residuals can be "adjusted" to unit variance using the square root of  $MSE \cdot (1-v_{ii})$  as a divisor. The results of that division are given in the RSTUDENT column of Table 1. Those residuals are more nearly comparable than are the ordinary residuals since they all have unit variance.

#### IV. USING RESIDUALS

The R option in PROC REG produces output which is in some respects similar to that produced by the INFLUENCE option. The most notable statistic produced is Cook's D which is quite similar to DFFITS. This statistic is a function of the Studentized residual

$$r_i = e_i / \text{MSE} (1-v_{ii}).$$

The Studentized residual differs from RSTUDENT only in that for the latter MSE is calculated after the  $i$ -th case has been deleted while the former utilizes the entire data set. Cook's D may be expressed in many ways, one of which is

$$D_i = (r_i^2/p) / (v_{ii}/(1-v_{ii}))$$

The first factor is a measure of "lack of fit" while the second is a measure of potential influence. DFFITS can be expressed in the same way except that  $r_i$  is replaced by RSTUDENT and a proportionality constant is used. There is some controversy as to which measure is the most useful. Some edited sample output is given in Table 2 for the data of Figure 4. Notice that while the ordinary residual is quite unremarkable both the Studentized residual and Cook's D "flag" the discrepant point very clearly (and so, of course, would the DFSTATISTICS discussed earlier).

Table 2

| OBS | ACTUAL | PREDICT |          | STUDENT COOKS |        |
|-----|--------|---------|----------|---------------|--------|
|     |        | VALUE   | RESIDUAL | RES           | D      |
| 1   | 4.000  | 2.299   | 1.701    | -1.002        | 0.173  |
| 2   | 5.000  | 3.955   | 1.045    | 0.595         | 0.046  |
| 3   | 7.000  | 5.610   | 1.390    | 0.773         | 0.059  |
| 4   | 7.000  | 7.266   | -0.266   | -0.146        | 0.002  |
| 5   | 8.000  | 8.922   | -0.922   | -0.501        | 0.018  |
| 6   | 9.000  | 10.578  | -1.578   | -0.858        | 0.053  |
| 7   | 9.000  | 12.234  | -3.234   | -1.771        | 0.254  |
| 8   | 29.000 | 27.136  | 1.864    | 2.373         | 14.883 |

#### V. COLLINEARITY DIAGNOSTICS

One collinearity diagnostic is produced under the INFLUENCE option. It is the statistic exhibited in the COV RATIO column of Table 2. COV RATIO is the ratio of the determinants of MSE times  $(X'X)^{-1}$  with and without the  $i$ -th

case deleted. Since collinearity is often characterized by a condition in which the determinant of  $X'X$  is small this ratio flags observations which affect the determinant of  $X'X$  greatly. It is values of this ratio which are far from 1 which concern us. In our example case 40 (Wyoming) is the most notable. As it happens Wyoming has at once the highest per capita fuel consumption of the 48 observations and a very high (not the highest) percentage of licenced drivers. It is an observation which we should pay particular attention to in this problem.

The  $p$ -th condition index is the ratio of the largest eigenvalue to the  $p$ -th largest eigenvalue. The largest condition index is the ratio of the largest eigenvalue to the smallest eigenvalue and is called the condition number. Large condition numbers are indicative of collinearity problems in that they suggest that the data in some problem involving say  $p$  predictor variables actually falls essentially in some space of dimension lower than  $p$ . One of the uses of condition indices involves regression coefficient variance decompositions. The idea is that high condition indices are related to small eigenvalues and that if an unusually high proportion of the variance in two or more coefficients is associated with the same large condition index (small eigenvalue) then the collinearity is being caused by the variables associated with the high decomposition proportions. Both COLLIN and COLLINDINT can be used as options to PROC REG to get these diagnostics. COLLINDINT centers the data before doing the calculations while COLLIN operates on the original data. Though it is controversial, see [8], it would seem that most analyses are better served by the statistics produced using COLLINDINT than by the other option. The issue was thoroughly discussed by Myers in another invited paper given at this conference.

#### VII. PARTIAL REGRESSION LEVERAGE PLOTS

Partial  $t$ -tests are certainly the most common way of assessing the importance of each variable used in a given regression model. A graphical way of assessing the usefulness of a particular term involves the partial regression leverage plot. These plots may be thought of as plots of the residuals of  $y$  generated when a variable is left out of the model against the residuals generated when the left out variable is regressed against the other variables. Such plots give a good sense of the importance of each variable in the full model. They also can be used to determine if the

coefficient associated with that term is determined by the data generally or by only one (or a few) observation(s). These plots are special cases of added variable plots which are more general and can be used to examine data for curvilinear relationships also. Partial regression leverage plots can be gotten using the PARTIAL option in PROC REG. One such plot is shown below in Figure 5 for the fuel consumption data. The variable involved is the proportion of licenced drivers which happens to be the most important of the explanatory variables.

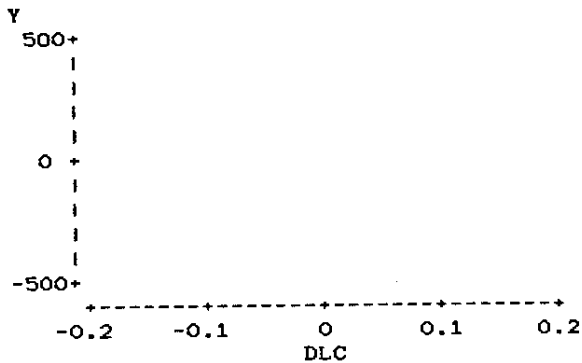


Figure 5

Such plots are produced for each of the independent variables in the model statement. The slopes of the least squares lines through the point clouds can be shown to be the same as the coefficient of the corresponding term in the full model.

## VII. VARIABLE SELECTION

Many of the diagnostics already discussed can be used in the context of variable selection in a regression problem. For instance a partial regression leverage plot might make it clear that a particular term in one model is dominated by a single outlying data point. In such cases the analyst may choose to delete the point or to remedy the situation in such a way that the term is no longer needed. This author's preference is to let the form of the model be dictated by substantive considerations. Model departures can then be discussed and improvements in the model made with the advice of experts in the field of interest. Occasionally the analyst will be on something of a "fishing expedition" and be asked to retrieve whatever is important from a large number of possibly important explanatory variables. Such situations are dangerous and we should be careful of whatever apparent "truths" we glean from the data. Any of several SAS procedures can be used in these situations. Two of the most directly

applicable are PROC RSQUARE and PROC STEPWISE.

PROC RSQUARE prints the value of  $r^2$  for every possible multiple linear regression model using the variables at hand. The output is organized into all of the one variable models, all of the two variable models, etc.. Within each grouping the candidate models are sorted by the value of  $r^2$ . Users may then select models for which the value of  $r^2$  is large and the set of explanatory variables is both reasonable and parsimonious. Since a regression problem with  $p$  candidate variables gives rise to  $2^p$  different models (ignoring transformations, powers, cross products and other adjustments we might try) it is clear that for large values of  $p$  computational difficulties are likely to arise. Our experience is that at about  $p=15$  those concerns become paramount. You can sometimes alleviate the difficulty by forcing certain variables into the model so that fewer possibilities need to be explored.

PROC STEPWISE can be used to build a multifactor model also. Options exist to build FORWARD. That is, add one variable at a time so that the best one variable model is first selected, then the next variable selected is the one which uses the first variable chosen and the other variable which along with the first variable gives the largest  $r^2$  value. Note that this does not necessarily result in the best two variable model. This process continues until all variables are used or until a given stopping rule is satisfied. In a similar vein the BACKWARD option causes us to begin with the full model and delete the least useful variable then the next least useful, etc.. Note that this approach only guarantees that the best  $p-1$  variable model is selected. Again the process continues until all variables are deleted from the model or a given stopping rule is satisfied. PROC STEPWISE can also be made to run both forwards and backwards. In that case the program checks after each variable is added to the model to see if one of the previously added variables should now be deleted and each time a variable is deleted from the model it checks to see if any of the remaining variables should now be added. Again, certain stopping rules define termination of the process. No general guarantees exist as to the optimality of the solutions thus generated. A good discussion of stepwise procedures can be found in [14].

### VIII. LITERATURE CITED

- [1] Weisberg, S., Applied Linear Regression, John Wiley & Sons, New York, 1985
- [2] Montgomery, D.C. & Peck, E.A., Linear Regression Analysis, John Wiley & Sons, New York, 1982
- [3] Draper, N.R. & Smith, H., Applied Regression Analysis, John Wiley & Sons, New York, 1981
- [4] Cook, R.D., & Weisberg, S., Residuals and Influence in Regression, Chapman & Hall, New York, 1982
- [5] Belsley, D.A., Kuh, E. & Welsch, R.E., Regression Diagnostics, John Wiley & Sons, 1980
- [6] Hocking, R.R., Developments in Linear Regression Methodology: 1959-1982, Technometrics, Vol 25 No 3 August, 1983
- [7] SAS Regression Applications, Technical Report A-102, SAS Institute, Cary, N.C., 1981
- [8] Belsley, D.A., Demeaning Conditioning Diagnostics Through Centering, The American Statistician, Vol 38 No 2, May, 1984. See also discussion by Cook, Gunst, Marquardt & Snee and Wood
- [9] Gray, J.B., & Ling, R.F., K-Clustering as a Tool for Influential Subsets in Regression, Technometrics Vol 26 No 4, November, 1984
- [10] Marasinghe, M.G., A Multistage Procedure for Detecting Several Outliers in Linear Regression, Technometrics, Vol 27 No 4, November, 1985
- [11] Mason, R.L. & Gunst, R.F., Outlier Induced Collinearities, Technometrics, Vol 27 No 4, November, 1985
- [12] Bolling, J., et al, Principles of Regression Analysis, Course notes, SAS Institute, Cary, N.C., 1984
- [13] Hobbs, G.R., Regression and Diagnostics, Proc. of the SAS Users Group, International, SAS Institute, Cary, N.C., 1985
- [14] Daniel, C. & Wood, F.S., Fitting Equations to Data, John Wiley & Sons, New York, 1980