Detection of model specification, outlier, and multicollinearity in multiple linear regression model using partial regression/residual plots.

George C. J. Fernandez, Department of Applied Economics and Statistics /204,
University of Nevada - Reno, Reno NV 89557

ABSTRACT

In multiple linear regression models problems arise when a serious multicollinearity or influential outlier present in the data. Failure to include significant quadratic or interaction terms results in model specification errors. Simple scatter plots are mostly not effective in revealing the complex relationships or detecting data problems in multiple regression models. However, partial regression plots are recommended mainly in detecting influential observations and multiple outlier and the partial residual or the added-variable or component plus-residual plots are mainly useful in detecting non-linearity and model specification errors. Neither plots in the standard format fails to detect the presence of multicollinearity. However, If these two plots are overlaid on the same plot with centered X_i values in the X-axis, the clustered data points in a partial regression plot clearly indicate the presence of multicollinearity. SAS® macros for displaying partial regression and partial residual plots using SAS/REG® and SAS/GRAPH® procedures are presented here.

INTRODUCTION

Multiple linear regression models are widely used applied statistical techniques. In regression analysis, we study the relationship between the response variable and one or more predictor variables and we utilize the relationship to predict the mean value of response variable form the known level of predictor variable or variables. Simple scatter plots are very useful in exploring the relationship between a response and a single predictor variable. However, simple scatter plots are not effective in revealing the complex relationships or detecting the trend and data problems in multiple regression models.

The use and interpretation of multiple regression depends on the estimates of individual regression coefficient. Influential outliers can bias parameter estimates and make the resulting analysis less useful. It is important to detect outliers since the outliers can provide misleading results. Several statistical estimates such as studentized residual, hat diagonal elements, Dffits, R-student, Cooks D statistics (Neter et. al, 1989; Myers 1990; Montgommery and Peck, 1992) are available to identify both outliers and influential observations. The PROC/REG procedure has an option called "INFLUENCE" to identify influential outliers. However, identifying influential outliers are not always easy in simple scatter plots.

Failure to include significant quadratic or interaction terms or omitting other important predictor variables in multiple linear regression models results in model specification errors. Significant quadratic terms and cross products can be identified by using the SAS PROC / RSREG. However, identifying significant model terms in multiple linear regression are not always easy in simple scatter plots.

The use and interpretation of multiple regression models often depend on the estimates of individual regression coefficient. The predictor variables in a regression model are considered orthogonal when they are not linearly related. But, when the regressors are nearly perfectly related, the regression coefficients tend to be unstable and the inferences based on the regression model can be misleading and erroneous. This condition is known as multicollinearity (Mason et. al, 1975).

Severe multicollinearity in OLS regression model results in large variances and covariances for the least squares estimators of the regression coefficient. This implies that different samples taken at the same X levels could lead widely different coefficients and variances of the predicted values will be highly inflated. Least-squares estimates of βi are usually too large in absolute values with wrong signs. Interpretation of the partial regression coefficient is difficult when the regressor variables are highly correlated. Multicollinearity in multiple linear regression can be detected by examining variance inflation factors (VIF) and condition indices (Neter et, al. 1989). SAS PROC REG has two options, VIF and COLINOINT to detect multicollinearity. However, identifying multicollinearity is not possible by examining simple scatter plots.

Partial plots are considered better substitutes for scatter plots in multiple linear regression. The partial regression plot for the Xi variable shows two sets of residuals, those from regressing the response variable and Xi on other predictor variables. The associated simple regression has the slope of βi , zero intercept and the same residuals (ϵ) as

the multiple linear regression.. This plot is considered useful in detecting influential observations and multiple outliers (Myers, 1990). The PARTIAL option in PROC REG produces partial regression plots (Text based plots) for all the predictor variables.

The partial residual (added-variable or component plus-residual) plot (Myers, 1990) corresponding to Xi shows the relationship between ($\epsilon + \beta i\ Xi$) versus Xi where ϵ is the residual of the full model. This simple linear regression model has the same slope (βi) and residual (ϵ) of the multiple linear regression. The partial residual plot display allows to easily evaluate the extent of departures from linearity. Currently, no option is available in SAS to readily produce partial residual plots.

Neither plots in the standard format fails to detect the presence of multicollinearity. However, If these two plots are overlaid on the same plot with the X-axis uses centered Xi values, the clustered in a partial regression plot clearly indicate the presence of multicollinearity (Stine, 1995). Since the overlaid plot is mainly useful in detecting multicollinearity. I named this plot as VIF plot.

The objective of this study is to develop SAS macros to produce high quality partial regression, partial residual, and VIF plots for all predictor variables in a multiple linear regression. The effectiveness of these plots in detecting influential outliers, model specification errors, and multicolliniarity are evaluated in this paper. Three separate data sets namely, DATA1 (contain highly influential outlier), DATA2 (need significant quadratic term for X2 variable and significant cross-product between X1 and X2), and DATA3 (three predictor variables involved in multicollinearity) are used in this investigations.

EXAMPLE

DATA1: Data with influential outlier (Neter et. al., , 1989)

```
data score;
input x1 x2 x3 y @@;
label y ='Patient satisfaction score'
    x1='patients age in years'
    x2='severity index'
    x3='anxiety index';
cards;
50  51  2.3  48  36  46  2.3  57
40  48  2.2  66  41  44  1.8  70
28  43  1.8  89  49  54  2.9  36
42  50  2.2  46  45  48  2.4  54
52  62  2.9  26  29  50  2.1  77
29  48  2.4  89  43  53  2.4  67
38  55  2.2  47  34  51  2.3  51
53  54  2.2  57  36  56  2.5  79
29  46  1.9  88  89  70  4.0  90
33  49  2.1  60  55  51  2.4  49
29  52  2.3  77  44  58  2.9  52
43  50  2.3  60
;
proc print  label noobs;
run;
* The influential outlier is highlighted;
```

DATA2: Response surface data - (contain significant quadratic and cross product)

```
DATA regx1x2;
    INPUT X1 X2  Y @@;
    LABEL  Y = 'UNITS OF ALGAE'    X1= 'MG COPPER'
          X2= 'DAYS';
CARDS;
2 5 .3    2 12 .40   2 18 .38   2 25  .32
2 5 .34   2 12 .36   2 18 .30   2 25  .22
1 5 .38   1 12 .46   1 18 .38   1 25  .34
1 5 .36   1 12 .44   1 18 .39   1 25  .32
1 5 .34   1 12 .38   1 18 .29   1 25  .23
3 5 .28   3 12 .32   3 18 .28   3 25  .16
3 5 .2    3 12 .28   3 18 .22   3 25  .18
3 5 .24   3 12 .32   3 18 .30   3 25  .20
4 5 .04   4 12 .10   4 18 .08   4 25  .06
4 5 .00   4 12 .18   4 18 .13   4 25  .04
4 5 .06   4 12 .16   4 18 .10   4 25  .08
5 5 .01   5 12 .14   5 18 .04   5 25  .03
5 5 .02   5 12 .10   5 18 .07   5 25  .02
5 5  0    5 12 .11   5 18 .05   5 25  .01
;
proc print label noobs; run;
```

Data3: Data with severe multicollinearity (Neter et. al 1989).

```
Data fat;
input x1-x3 y;
label x1='Triceps skin fold thickness'    x2='Thigh circumference'
      x3='mid arm circumference'          y='body fat';
cards;
        19.5        43.1        29.1        11.9
        24.7        49.8        28.2        22.8
        30.7        51.9        37.0        18.7
        29.8        54.3        31.1        20.1
        19.1        42.2        30.9        12.9
        25.6        53.9        23.7        21.7
        31.4        58.5        27.6        27.1
        27.9        52.1        30.6        25.4
        22.1        49.9        23.2        21.3
        25.5        53.5        24.8        19.3
        31.1        56.6        30.0        25.4
        30.4        56.7        28.3        27.2
        18.7        46.5        23.0        11.7
        19.7        44.2        28.6        17.8
        14.6        42.7        21.3        12.8
        29.5        54.4        30.1        23.9
        27.7        55.3        25.7        22.6
        30.2        58.6        24.6        25.4
        22.7        48.2        27.1        14.8
        25.2        51.0        27.3        21.1
;
proc print label noobs;  run;
```

ANALYSIS:

The partial regression, partial residual, and overlaid partial regression/residual plots of given predictor variables in a multiple linear regression can be obtained easily by running the SAS macro VIFPLOT. The macro-call file with the descriptions of macro parameters for running this SAS macro is given below:

```
%inc 'a:\macro\vifplot.mac';
%vifplot(data= fat       ,  /*RQ : SAS data file */
        resp= y          ,  /* RQ: Name of the response */
        pred= x1 x2 x3   ,  /* RQ: Model terms */
          Term =x1 x2 X3 ,/* RQ: Identify the terms for
                             Partial plots */
       size = 1          , /* Text size in SAS graphics */
        dir = a:\plot    ,  /* Dir to save the graphics */
    dev = win) /* Change to CGM to save the graphics*/
  *RQ: Required ;
```

Results and Discussion

The partial residual, partial regression, and the overlaid VIF plots for the DATA1 that contain a highly influential outlying observation are presented in Fig.1. The impact of the influential outlier is clearly evident in all three partial residual plots (Fig.1 A, D, G). The linear regresion line is pulled toward the influential point. Also, because of the single influential observation, a significant quadratic effect is evident for predictor variable X1 and X3.

Surprisingly, the partial regression plots (Fig.1 B, E, H) are not very effective in detecting the influential outlier in this data set. Partial regression plots are specially recommended to detect influential and multiple outliers (Myers, 1990). However, in this example, partial regression plot failed to detect the single influential point. Multicollinearity is not a problem in this data set. Therefore, any unusual clustering of partial regression points is not evident in these plots (Fig.1 C, F, I).

The partial residual, partial regression, and the overlaid VIF plots for the DATA2 with the following model terms (X1, X2, X1*X2) are presented in Fig.2. The need for a significant quadratic term for X2 is clearly evident in Fig.2 D. The partial residual plot for the cross product term (X1*X2) also shows a quadratic trend. This might be due to the fact that X2 - quadratic term is not included in the model..

Once again, the partial regression plots (Fig. 2 B, E, H) are not very effective as the partial residual plots in detecting the non linearity in this data set. Because, the X-axis in partial regression plot is based on residual, rather than the actual Xi, it complicates the usefulness of these plots.

Small amount of multicollinearity is expected in this data because we have included both the main effects of X1 and X2 and their cross products. This is clearly shown in the VIF plot (Fig. 2-I) for cross product term and some degree of clustering of partial regression points is evident here.

The partial residual, partial regression, and the overlaid VIF plots for the DATA3 that contain a high degree of multicollinearity among the predictor variables are presented in Fig.3. The partial residual plots (Fig 3. A, D, G) or partial regression plots (Fig. 3 B, E, H) alone failed to detect the problem of multicollinearity in the data set . The VIF plots (Fig. 3 C, F, I) very clearly indicate the impact of multicollinearity in all three predictor variables. In the presence of other two variables, the influences of the third predictor variable become unimportant when multicollinearity is present.

Summary

The features in SAS systems for detecting influential outliers, model specification errors, and multicollinearity using partial regression, partial residual, and overlaid partial regression/residual plots are presented her by using SAS macro called VIFPLOT. This macro can be obtained from the author by sending e-mail.

References

Mason, R. L. , Gunst, R.F. and Webster, J.T. 1975. Regression analysis and problem of multicollinearity. Commun. Statistics. 4(3): 277-292.

Montgomery D.C. And Peck E.A. 1992. Introduction to Linear regression analysis 2nd edition. John Wiley. New York.

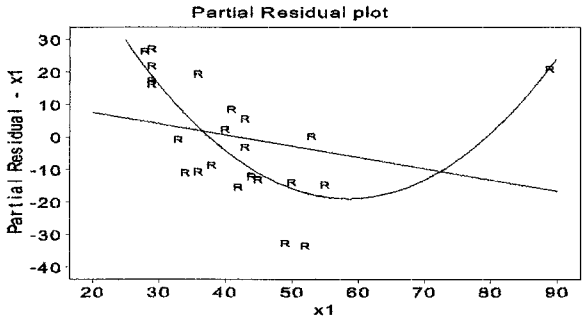
Myers, R.H. 1990. Classical and modern regression application. 2nd edition. Duxbury press. CA.

Neter, J. Wasserman, W., and Kutner, M.H. 1989. Applied Linear regression Models. 2nd Edition. Irwin Homewood IL.

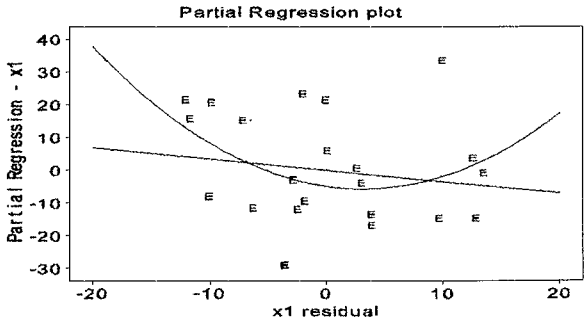
Stine Robert A. 1995. Graphical Interpretation of Variance Inflation Factors. The American Statistician vol 49: 53-56.

Author's Contact address:

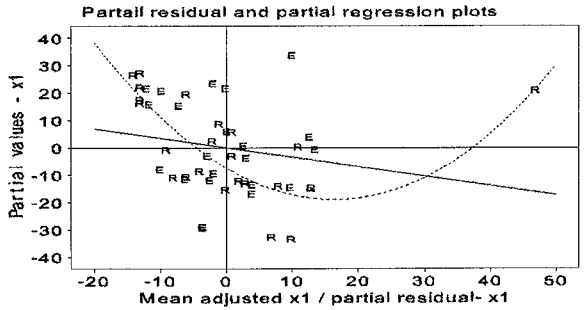
Dr. George C.J. Fernandez
Associate Professor in Applied Statistics
Department of Applied Economics/Statistics/204
University of Nevada- Reno Reno NV 89557.
(702) 784-4206 E-mail: GCJF@scs.unr.edu



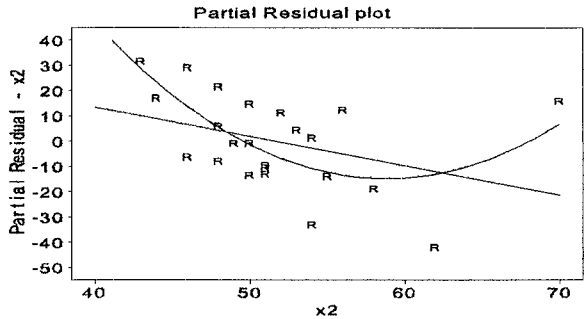
1-A



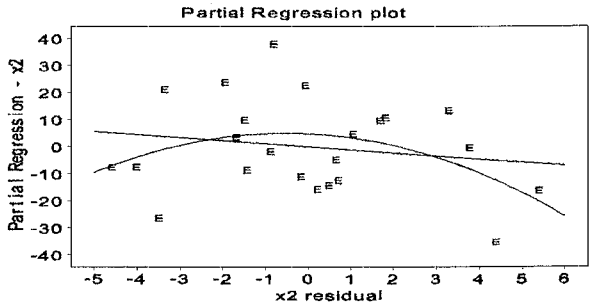
1-B



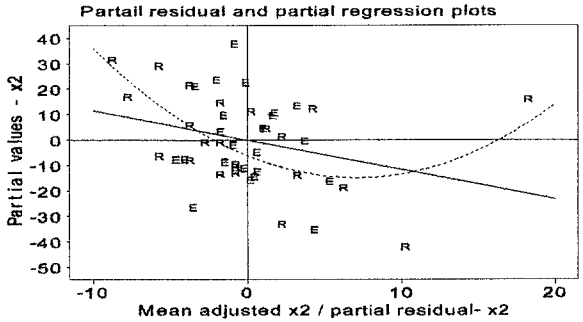
1-C



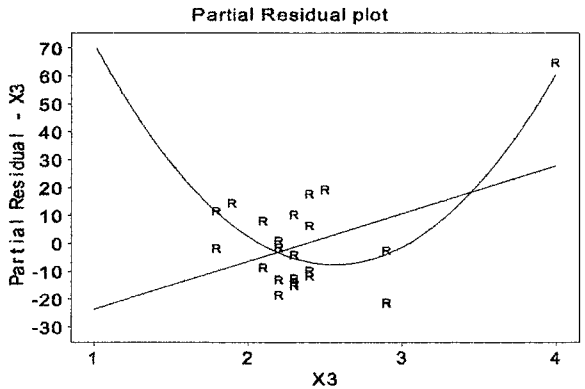
1-D



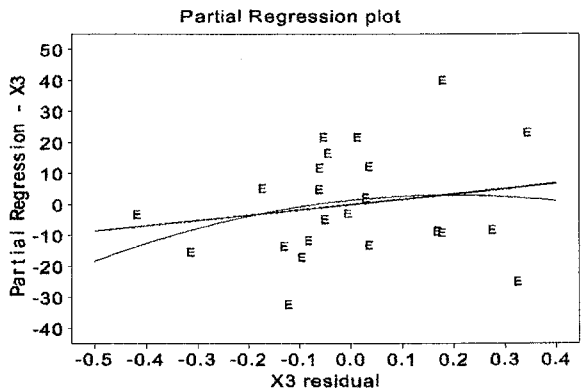
1-E



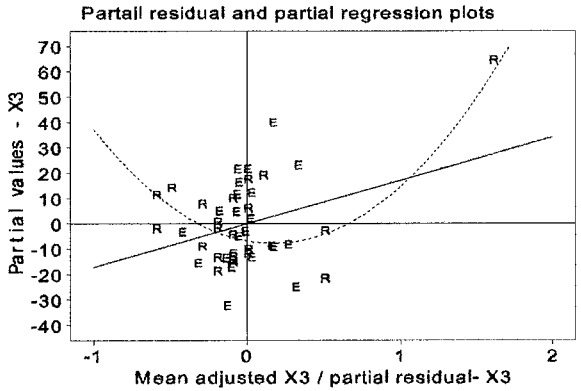
1-F



1-G

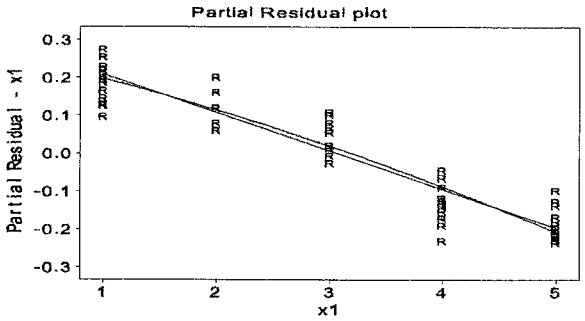


1-H

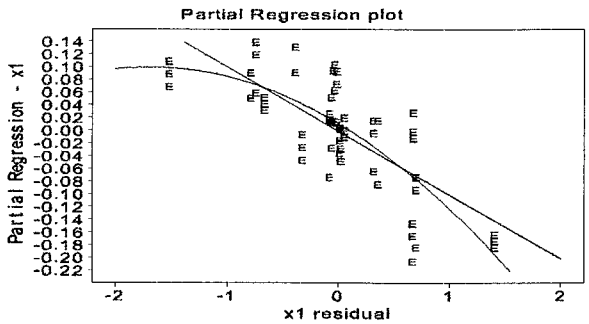


1-I

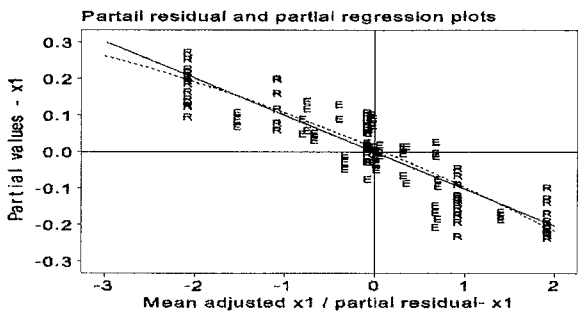
FIG. 1. DATA1: Patient satisfaction score data containing three predictor variables and one highly influential outlier.



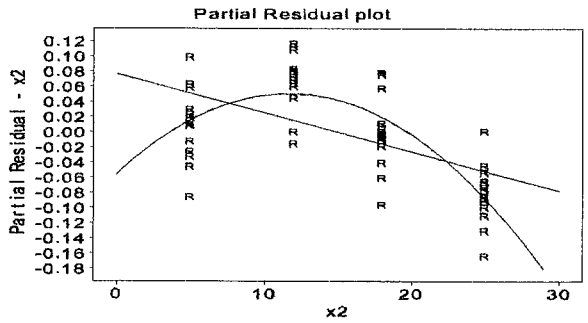
2-A



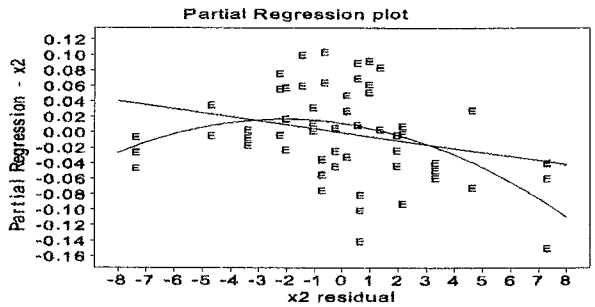
2-B



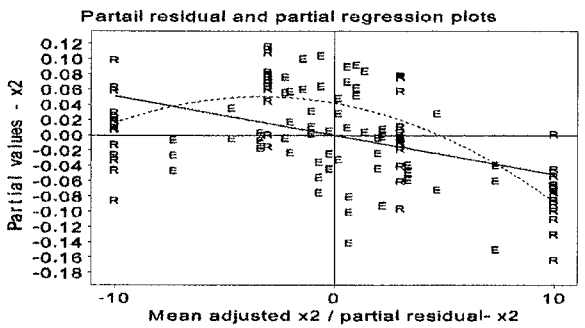
2-C



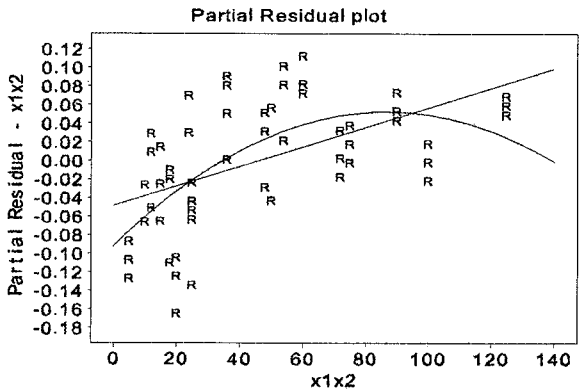
2-D



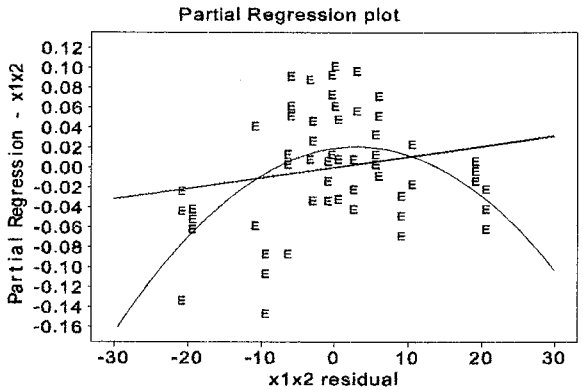
2-E



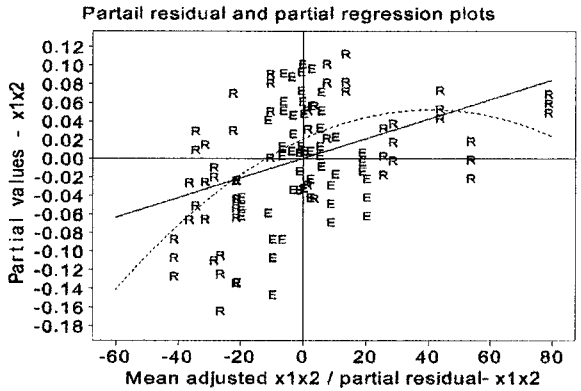
2-F



2-G

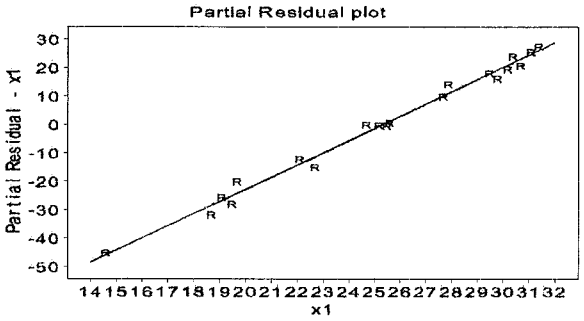


2-H

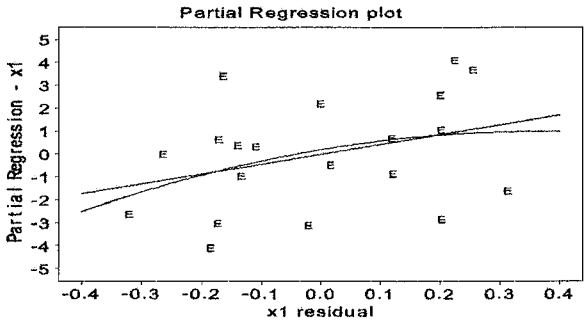


2-I

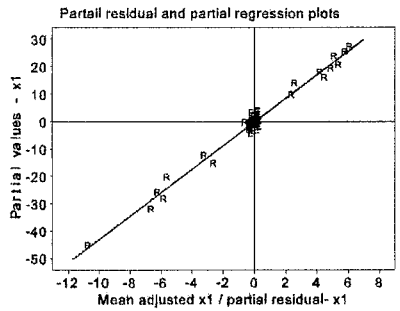
FIG. 2. DATA2 Response surface data with 2 predictor variables and X1X2 cross product: The x2 variable has significant quadratic effect.



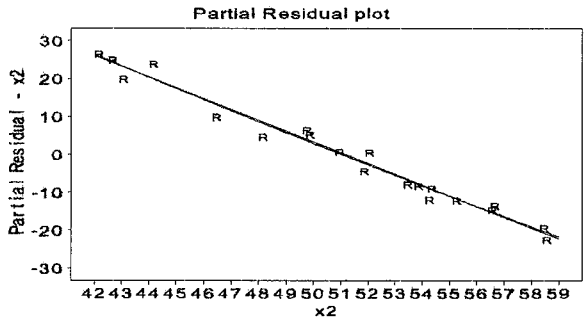
3-A



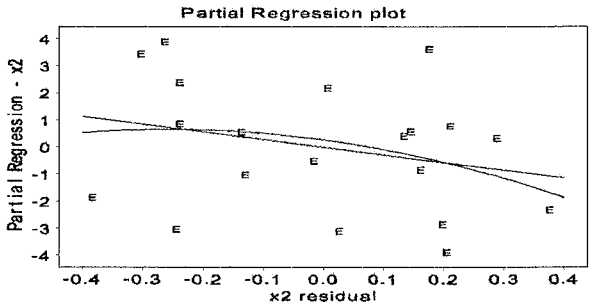
3-B



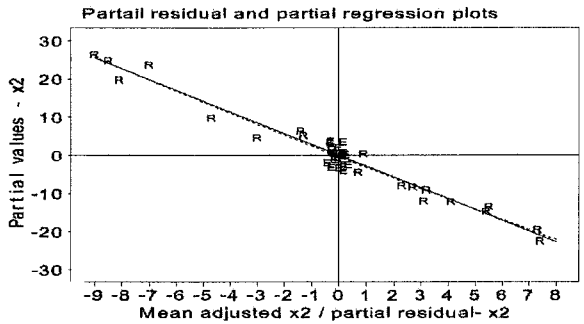
3-C



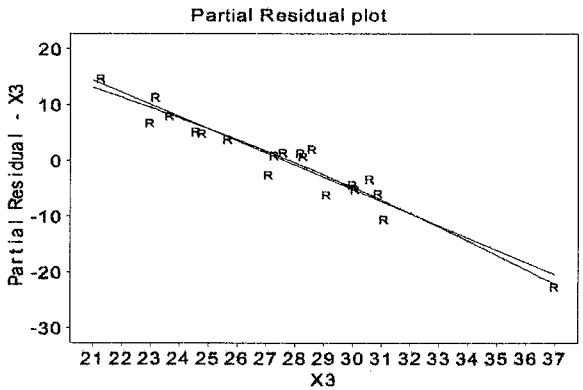
3-D



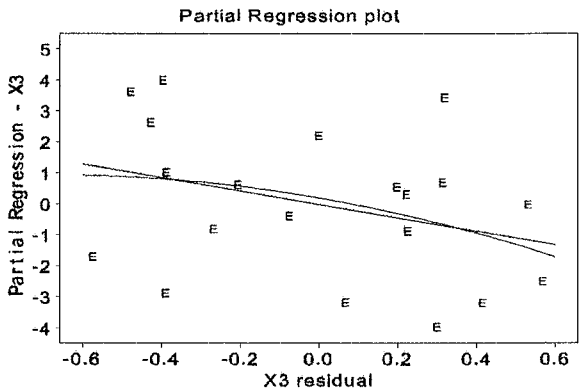
3-E



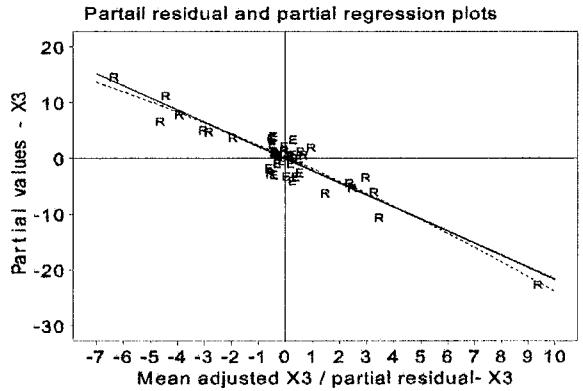
3-F



3-G



3-H



3-I

FIG. 3. DATA3: Body fat data containing three predictor variables involved in multicollinearity.