PARTIAL PLOTS IN REGRESSION ANALYSIS

George C. J. Fernandez, Department of Applied Economics and Statistics /204,
University of Nevada, USA

In Multiple linear regression models, problems arise when serious multicollinearity or influential outliers are present in the data. Simple scatter-plots are most of the time not effective in revealing the complex relationships of predictor variables or data problems in multiple linear regression. However, partial regression plots are considered useful in detecting influential observations and multiple outliers; partial residual plots or the added-variable or component-plus-residual plots are useful in detecting non-linearity and model specification errors. The leverage plots available in SAS/JMP software are considered effective in detecting multicollinearity and outliers. The VIF-plot, which is very effective in detecting multicollinearity, can be obtained by overlaying both partial regression and partial residual plots with a common centered X-axis. A SAS macro, PARTIAL, for displaying these partial plots is 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 the response variable from a 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 depend 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 they can provide misleading results. Several statistical estimates such as *student residual, hat diagonal elements, DFFITS, R-student, Cooks D* statistics (Neter et. al, 1989; Myers 1990; Montgomery and Peck, 1992) are available to identify both outliers and influential observations. The SAS/REG procedure has an option, “INFLUENCE” to identify influential outliers. However, identifying influential outliers 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 independent 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 covariance for the least squares estimators of the regression coefficient. This implies that different samples taken at the same X levels could lead to 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/REG has two options, VIF and COLINOINT to detect multicollinearity. However, identifying multicollinearity is not realistic by examining simple scatter plots.

Partial plots are considered better substitutes for scatter plots in multiple linear regression. These partial plots illustrate the partial effects or the effects of a given predictor variable after adjusting for all other predictor variables in the regression model. Two kinds of partial plots, partial regression and partial residual or added variable plot are documented in the literature (Belsley et al 1980; Cook and Weisberg 1982).

PARTIAL REGRESSION PLOT

A multiple regression model with 3 (X1-X3) predictor variables and a response variable Y is defined as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad \text{----------(1)}$$

The partial regression plot for X1 is derived as follows:

1) Fit the following two regressions:

$$Y_i = \theta_0 + \theta_2 X_2 + \theta_3 X_3 + \varepsilon_{|x2,x3} \quad \text{----------(2)}$$

$$X_{1i} = \gamma_0 + \gamma_2 X_2 + \gamma_3 X_3 + \varepsilon_{x1|x2,x3} \quad \text{----------(3)}$$

2) Fit the following simple linear regression using the residuals of models 2 and 3.

$$\varepsilon_{y|x2,x3} = 0 + \beta 1 \ \varepsilon_{x1|x2,x3} + \varepsilon_i$$

The partial regression plot for the X_1 variable shows two sets of residuals, those from regressing the response variable (Y) and Xi on other predictor variables. The

associated simple regression has the slope of β_1, 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.

Sall (1990) proposed an improved version of the partial regression plot and called it leverage plot. He modified both X and Y axis scale by adding the response mean to $\varepsilon_{y|x2,x3}$ and X1 mean to $\varepsilon_{x1|x2,x3}$. In his leverage plots, Sall also included a horizontal line through the response mean value and a 95% confidence curves to the regression line. This modification helps us to view the contribution of other regressor variables in explaining the variability of the response variable by the degree of response shrinkage in the leverage plot. This is very useful in detecting severe multicollinearity. Also based on the position of the horizontal line through response mean and the confidence curves, the following conclusions can be made regarding the significance of the slope.

Confidence curve crosses the horizontal line = Significant slope
Confidence curve asymptotic to horizontal line = Boarder line significance
Confidence curve does not cross the horizontal line = Non Significant slope

Thus, the leverage plots are considered useful in detecting outliers, multicollinearity, non-linearity, and the significance of the slope. Currently, SAS has no option to generate these leverage plots. However, SAS/JMP has option to generate these leverage plots.

II) Partial residual (added-variable or component plus-residual) plot (Larson and McCleary, 1972).

The Partial residual plot is derived as follows:

1) Fit the full regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad \text{----------(4)}$$

2) Construct the Partial Residual plot:

$$(\varepsilon_i + \beta_1 X_1) = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad \text{----------(5)}$$

The partial residual plot for X_1 is a simple linear regression between $(\varepsilon_i + \beta_1 X_1)$ versus X_1 where ε_i is the residual of the full regression model. This simple linear regression model has the same slope (β_1) and residual (ε) of the multiple linear regression. The partial residual plot display allows to easily evaluate the extent of departures from linearity. These plots are also considered useful in detecting influential outliers and

inequality of variance. Currently, no option is available in SAS to readily produce partial residual plots.

Mallows (1986) introduced a variation of partial residual plot in which a quadratic term is used both in the fitted model and the plot. This modified partial residual plot is called an augmented partial residual plot.

The Augmented Partial residual plot is derived as follows:

1) Fit the full regression model with a quadratic term:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_1^2{}_i + \varepsilon_i \quad \text{----(6)}$$

2) Construct the Augmented Partial Residual plot:

$$(\varepsilon_i + \beta_1 X_1 + \beta_4 X_1^2{}_i) = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad \text{----------(7)}$$

The augmented partial residual plot for X_1 is a simple linear regression between $(\varepsilon_i + \beta_1 X_1 + \beta_4 X_1^2{}_i)$ versus X_1 where ε_i is the residual of the full regression model. The augmented partial residual plot effectively detects the need for a quadratic term or the need for a transformation for Xi. Currently, no option is available in SAS to readily produce partial residual plots.

Partial residual and partial regression plots in the standard format fail to detect the presence of multicollinearity. However, the leverage plot, the partial regression plot expressed in the scale of the original Xi variable , clearly shows the degree of multicollinearity. Stine (1995) proposed overlaying the partial residual and partial regression plots on the same plot to detect the multicollinearity. Thus by overlaying the partial residual and regression plots with the centered Xi values on the X-axis, the degree of multicollinearity can be detected by amount of shrinkage of partial regression residuals. Since the overlaid plot is mainly useful in detecting multicollinearity, I named this plot as VIF plot.

Even though these diagnostic plots are very useful tools in regression analyses, currently SAS options are not readily available to generate these plots. To generate partial residual, partial regression, and VIF plot, a SAS macro called VIFPLOT (Fernandez, 1997) can be used very effectively. More information about this macro can be obtained by the author. However, SAS codes are not available to produce augmented partial residual and leverage plots. Therefore, the objective of this study is to develop a SAS macro to produce high quality augmented partial residual, leverage, and VIF plots for all predictor variables in a multiple linear regression.

ANALYSIS

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

```
**************************************************************************
%inc 'a:\macro\partial.mac';
%partial  (
  DATA      = score     ,        /* RQ : SAS data file name*/
   RESP     = y         ,        /* RQ: Name of the response */
   PRED     = x1 x2 x3  ,        /* RQ: Predictor variables and any cross-products */
   TERM     = x1        ,        /* RQ: Select the variables/terms for plot */
    DIR     = c:        ,        /* Folder to save the graphics */
    Z       = 1         ,  /* Counter value */
   DEV   = win        ) * Graphic device options: WIN  CGMMWWC  CGMWPCA  */
**************************************************************************
```

By running the SAS macro, Partial will produce three plots for each predictor variable specified in the TERM option. The first plot is an overlay plot of a simple scatter plot between the response (Y) (Y-axis) and the predictor variable Xi (X-axis) and the augmented partial residual plot for Xi. The X-axis of the augmented partial residual plot displays Xi value. The augmented partial residual values are scaled to the response variable by adding the response mean value and display on the Y-axis. Influential (DFFITS>1) or outlier (|STUDENT| > 2.5) observations based on the original multiple linear regression model are identified on this plot by the observation number.

The second plot is the partial regression leverage plot. The Y-axis displays the partial residual of Y, scaled back to the original response by adding the response mean. The X-axis displays the partial residual of Xi, scaled back to the original Xi by adding the Xi mean. A simple regression line, regression equation, and the 95% confidence curves are also included in this plot. A horizontal line representing the response mean is also displayed. Influential (DFFITS>1) or outlier (|STUDENT| > 2.5) observations based on the original multiple linear regression model are also marked on this plot by the observation number.

The third plot is the overlay plot between the augmented partial residual plot and the leverage plot. A second order quadratic regression line is fitted to the augmented partial regression plot. This quadratic regression line will help to evaluate the need for a quadratic term in the model. A simple linear regression line is fitted to the leverage plot. Both regression equations are also displayed on this plot.

Examples of these three plots will be presented at the ICOTS-5 symposium by running the SAS macro PARTIAL. A copy of the SAS macro can be obtained by sending an E-mail to the author.

REFERENCES

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression diagnostics. N.Y. John Wiley.

Cook, R. D. and Weisberg, S. (1982) Residuals and Influence in Regression. N.Y. Chapman and Hall.

Fernandez, G. C. J. (1997). Detection of model specification , outlier, and multicollinearity in multiple linear regression models using partial regression/residual plots. SAS institute inc., Proceedings of the 22nd annual SAS users group international conference. 1246–1251.

Larsen W. A. and McCleary S. J. (1972). The use of partial residual plots in Regression analysis. *Technometrics 14*, 781-790.

Mallows, C. L. (1986). Augmented partial residual Plots. *Technometrics 28*, 313–319

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.

Sall, J. (1990). Leverage plots for general linear hypothesis. *The American Statistician, 44*, 308-315

Stine R. A. (1995). Graphical Interpretation of Variance Inflation Factors. *The American Statistician, 49,* 53-56.