

Substitute Paper 256-28

**Regression Diagnostic Plots**

George C. J. Fernandez, Department of Applied Economics and Statistics, UNR Reno NV 89557

**ABSTRACT**

In Multiple linear regression models, problems arise when serious multicollinearity or influential outliers are present in the data. Failure to include significant quadratic or cross-product terms result in model specification error. 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. *Mallows C(p)* plot can be used to select the optimum predictor variable subset. Regression model fit can be assessed by the explained variation plot and predicted values plot. Normality and homoscedasticity assumptions can be examined by the residual and the normal probability plots. User-friendly SAS macros for displaying these diagnostic regression plots are presented here.

**INTRODUCTION**

Multiple linear regression (MLR) 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 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 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 they 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; Montgomery 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 depends 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 to widely different coefficients and variances of the predicted values will be highly inflated. Least-squares estimates of  $\beta_i$  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 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). Other diagnostic plots used in regression model are: C(p) plot model selection; residual trend plot for detecting auto-correlations; residual plots for detecting heteroscedasticity; and normal probability residual plots for detecting deviation from normality.

I) Exploratory Plots

i) Partial Regression Plot

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

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

The partial regression plot for X<sub>1</sub> is derived as follows:

1) Fit the following two regressions:

$$Y_i = \theta_0 + \theta_2 X_{2i} + \theta_3 X_{3i} + \epsilon_{y|x_2,x_3} \text{ -----(2)}$$

$$X_{1i} = \gamma_0 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \epsilon_{x_1|x_2,x_3} \text{ -----(3)}$$

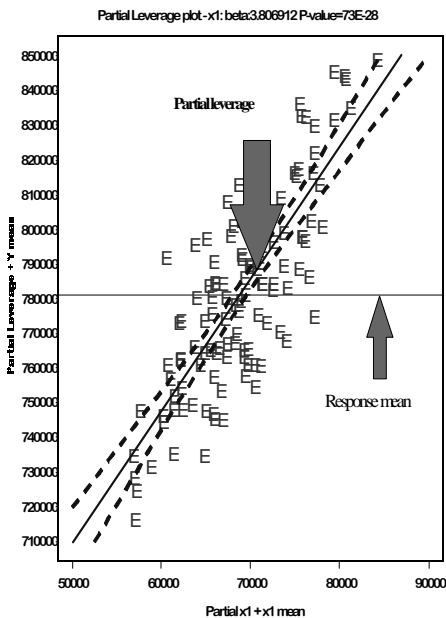


Figure 1 Partial leverage Plot

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

$$\epsilon_{y|x_2,x_3} = 0 + \beta_1 \epsilon_{x_1|x_2,x_3} + \epsilon_i$$

The partial regression plot for the X<sub>1</sub> 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<sub>1</sub>, 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

ε<sub>y|x<sub>2</sub>,x<sub>3</sub></sub> and X<sub>1</sub> mean to ε<sub>x<sub>1</sub>|x<sub>2</sub>,x<sub>3</sub></sub>. 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 = Border 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. An example of partial leverage plot showing a significant partial regression coefficient is shown in Figure 1. The partial leverage plot displays three curves: a) the vertical reference line that goes through the response variable mean; b) the partial regression line which quantifies the slope of the partial regression coefficient of the i<sup>th</sup> variable in the MLR; c) The 95% confidence band for partial regression line. The partial regression parameter estimates for the i<sup>th</sup> variable in the multiple linear regression and their significance levels are also displayed in the titles. The slope of the partial regression coefficient is considered statistically significant at the 5% level if the response mean line intersects the 95% confidence band. If the response mean line lies within the 95% confidence band without intersecting it, then the partial regression coefficient is considered not significant.

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:

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

2) Construct the Partial Residual plot:

$$(\epsilon_i + \beta_1 X_{1i}) = \beta_0 + \beta_1 X_{1i} + \epsilon_i \dots\dots\dots(4)$$

The partial residual plot for  $X_1$  is a simple linear regression between  $(\epsilon_i + \beta_1 X_{1i})$  versus  $X_1$  where  $\epsilon_i$  is the residual of the full regression model. This simple linear regression model has the same slope ( $\beta_1$ ) and residual ( $\epsilon$ ) 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:  
 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i}^2 + \epsilon_i \dots\dots\dots(5)$

2) Construct the Augmented Partial Residual plot:

$$(\epsilon_i + \beta_1 X_{1i} + \beta_4 X_{1i}^2) = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

The augmented partial residual plot for  $X_1$  is a simple linear regression between  $(\epsilon_i + \beta_1 X_{1i} + \beta_4 X_{1i}^2)$  versus  $X_1$  where  $\epsilon_i$  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

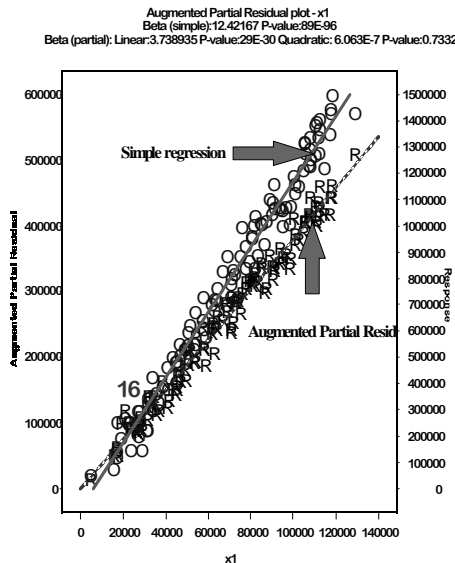


Figure 2 Augmented partial residual plot

transformation for  $X_i$ .

Currently, no option is available in SAS to readily produce partial residual plots. An example of augmented partial residual plot showing a significant partial regression coefficient and the regression

relationship from a simple regression model are shown in Figure 2.

The linear/quadratic regression parameter estimates for the simple and multiple linear regressions and their significance levels are also displayed in the titles. The simple linear regression line describes the relationship between the response and the predictor variable in a simple linear regression. The APR line shows the quadratic regression effect of the  $i^{th}$  predictor on the response variable after accounting for the linear effects of other predictors on the response. The APR plot is very effective in detecting significant outliers and non-linear relationships. Significant outliers and/or influential observations are identified and marked on the APR plot if the absolute 'STUDENT' value exceeds 2.5 or the 'DFFITS' statistic exceeds 1.5. These influential statistics are derived from the MLR model involving all predictor variables. If the correlations among the predictor variables are negligible, the simple and the partial regression lines should have similar slopes.

iii). VIF PLOT

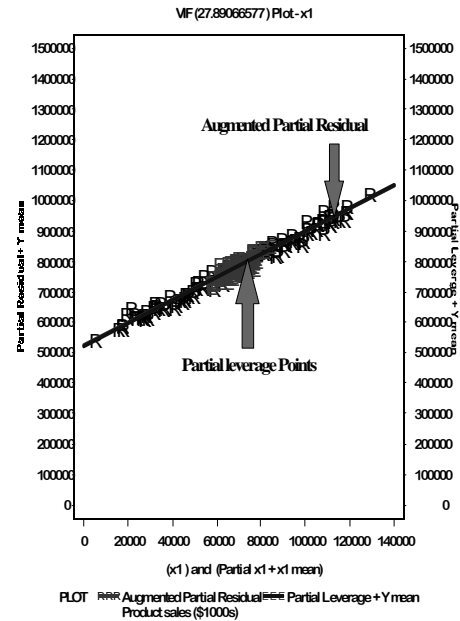


Figure 3 VIF- Plot

Augmented partial residual and partial regression plots in the standard format generally fail to detect the presence of multicollinearity. However, the leverage plot, the partial regression plot expressed in the scale of the original  $X_i$  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  $X_i$  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. An example of VIF plot showing a significant partial regression coefficient and moderate level of multicollinearity is shown in Figure 3

The VIF plot displays two overlaid curves: a) The first curve shows the relationship between partial residual + response mean and the  $i^{th}$  predictor variable b) the second curve displays the relationship between the partial leverage + response mean and the partial  $i^{th}$  predictor value + mean of  $i^{th}$  predictor value. The slope of the both regression lines should be equal to the partial regression coefficient estimate for the  $i^{th}$  predictor. When there is no high degree multicollinearity, both the partial residual (Symbol 'R') and the partial leverage (Symbol 'E') values should be evenly distributed around the regression line. But, in the presence of severe multicollinearity the partial leverage values, 'E' shrinks and distributed around the mean of the  $i^{th}$  predictor variable. Also, the partial regression for the  $i^{th}$  variable shows a non-significant relationship in the partial leverage plots whereas the partial residual plot shows a significant trend for  $i^{th}$  variable. Furthermore, the degree of multicollinearity can be measured by the 'VIF' statistic in a MLR model and the 'VIF' statistic for each predictor variable is displayed on the title statement of the 'VIF' plot.

II. Model selection diagnostic plots

The  $C(p)$  plot (Figure 4) shows the  $C(p)$  statistic against the number of predictor variables for the full model and the best two models for each subset. The Mallows  $C(p)$  measures the total squared error for a subset that equals to total error variance plus the bias introduced by not including the important variables in the subset. Additionally, the root mean squared ( $RMSE$ ) statistic for the full model and best two regression models in each subset is also shown in the  $C(p)$  plot. Furthermore, the diameter of the bubbles in the  $C(p)$  plot is proportional to the magnitude of  $RMSE$ . Thus, the  $C(p)$  plot produced by the 'REGDIAG' macro can be used effectively in selecting the best subset in regression models with many (5 to 25) predictor variables.

III. Plots illustrating regression model fit

The overall model fit is illustrated in Figure 5 by displaying the relationship between the observed response variable and predicted values. The  $N$ ,  $R^2$ ,  $R^2$  (adjusted), and  $RMSE$  statistics that were useful in comparing regression models and the regression

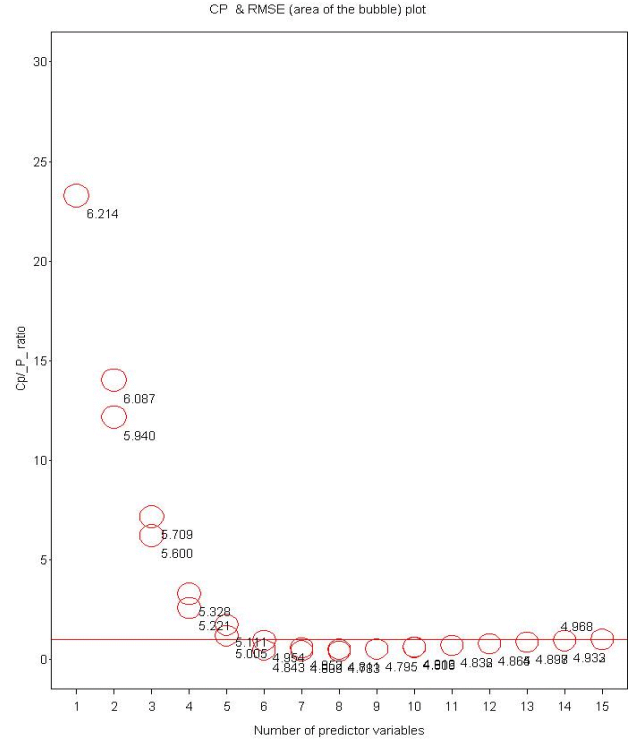


Figure 4 Model selection using Cp model are also included on the plot.

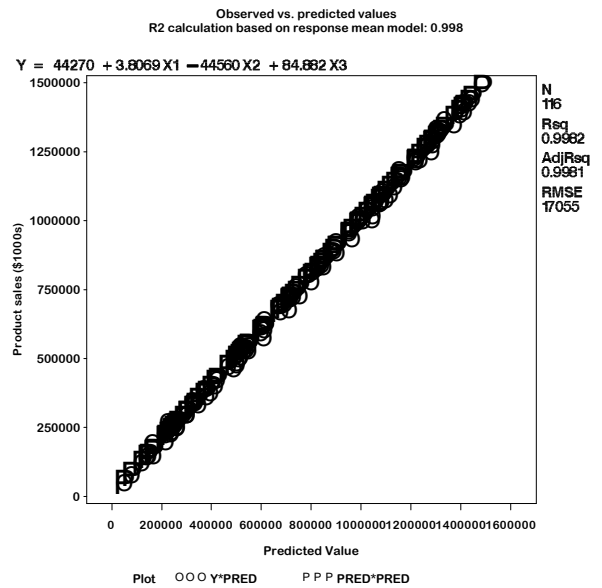


Figure 5 Plot illustrating model fit

If the data contained replicated observations, the deviation from the model includes both 'pure error' and 'deviation from the regression'. The  $R^2$  estimates can be computed from a regression model using the means of the replicated observations as the response. Consequently, the  $R^2$  computed based on the means ( $R^2_{(mean)}$ ) is also displayed in the title statement. If there

is no replicated data,  $R^2_{(mean)}$  and the  $R^2$  estimate reported by the PROC REG will be identical

Figure 6 shows graphically the total and the unexplained variation in the response variable after accounting for the regression model. The ordered and the centered response variable versus the ordered sequence display the total variability in the response. If the ordered response shows a linear trend without any sharp edges at the both ends then response variable has a normal distribution. The unexplained variability in the response variable is given by the residual distribution. The residual variation shows a random distribution without any sudden peaks, trends or patterns if the regression model assumptions are not violated. The differences between the total and residual variability show the amount of variation in the response accounted for by the regression model and are estimated by the  $R^2$  statistic. The predictive potential of the fitted model can be determined by estimating the  $R^2_{(prediction)}$  by substituting 'PRESS ( $i^{th}$  deleted residual)' for SSE in the formula for the  $R^2$  estimation. The predictive power of the estimated regression model is considered high if the  $R^2_{(prediction)}$  estimate is large and closer to the model  $R^2$ . The estimates of  $R^2_{(mean)}$  and the  $R^2_{(prediction)}$  described previously are also displayed in the 'title' statement. These estimates and the graphical display of explained and unexplained variation help to judge

the quality of the model fit.

IV. Model violation diagnostic plots

Figure 7 shows the trend plot of the residual over the observation sequence. If the data is a time series data, we can examine the residual plot for a cyclic pattern when there is a sequence of positive residuals following negative residuals. This cyclical pattern might be due to the presence of first order auto-correlation where the  $i^{th}$  residual is correlated with the  $lag1$  residual. The Durbin-Watson (DW)  $d$  statistic measures the degree of first order auto-correlation. An estimate of the DW statistic and the significance of  $1^{st}$  order auto-correlation are estimated using the PROC REG and displayed on the title statement. Observations used in the modeling are identified as

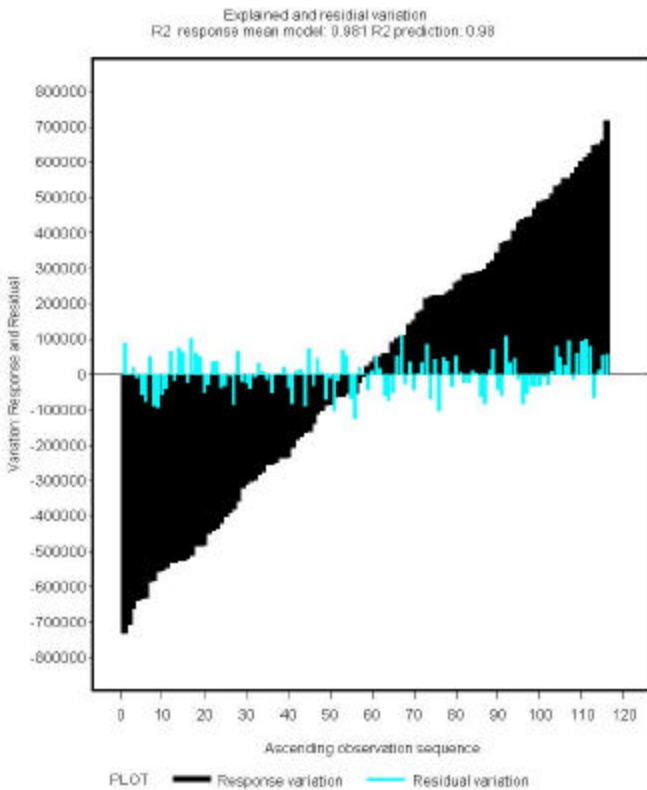


Figure 6 Explained variation plot

Residual plot for examining autocorrelation by obs number  
Durbin-Watson D= 1.763; 1st Order Autocorrelation= 0.117 NS

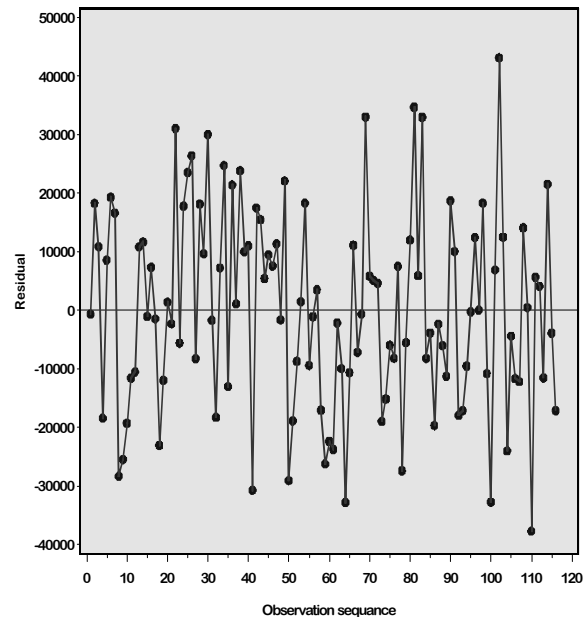


Figure 7 First order auto-correlation detection plot

outliers if the absolute STUDENT value exceeds 2.5. Also, observations are identified as influential if the 'DFFITS' statistic value exceeds 1.5. An outlier detection bubble plot between 'student' and 'hat value' identifies the outliers if they falls outside the 2.5 boundary line and detects influential points if the diameter of the bubble plot, which is proportional to DFFITS is relatively big (Figure 8). A fan pattern like the profile of a megaphone, with a noticeable flare either to the right or to the left in the residual plot against predicted value is the indication of significant heteroscedasticity.

The Breusch-Pagan test based on the significance of linear model using the squared absolute residual as the response and all combination of variables as

predictors is recommended for detecting heteroscedasticity. However, the presence of significant outliers and non-normality may confound with heteroscedasticity and may interfere with the detection. The results of the 'Breusch-Pagan' test and the random pattern of the residuals in the residual plot (Figure 8) both can confirm if the residuals have equal variance. Multiple linear regression models are fairly robust against violation of non-normality especially in large samples. Signs of non-normality are significant skewness (lack of symmetry) and/or kurtosis light-tailedness or heavy-tailedness. The normal probability plot (Figure 8 normal Q-Q plot), along with the normality test statistics, can provide information on the normality of the residual distribution.

### 'REGDIAG' SAS macro

The regression diagnostic plots described above can be obtained easily by running the SAS macro application 'REGDIAG' (Fernandez 2002). The user-friendly SAS macro application 'REGDIAG' integrates the statistical and graphical analysis tools available in SAS systems and provides complete regression diagnostic solutions without writing SAS program codes or using the point-and-click approach. Step-by-step instructions for using the SAS macro 'REGDIAG' and interpreting the results are emphasized. Thus, by following the step-by-step instructions and downloading the user-friendly SAS macros described in the book, data analysts can perform regression diagnostics quickly and effectively.

### Summary

The features in SAS systems for detecting influential outliers, non-linearity, and multicollinearity using augmented partial residual, partial regression leverage and overlaid augmented partial residual and leverage (VIF PLOT) plots, model selection plot using Cp statistic, plots showing model fit, residual plots for detecting first-order auto correlation, heteroscedasticity, influential outliers, and departure from normality are presented here by using a SAS macro called user-friendly SAS macro application 'REGDIAG'. The complete details of using and instructions for obtaining the macro are reported elsewhere (Fernandez, 2002).

### 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 22<sup>nd</sup> annual SAS users group international conference. 1246–1251.
- Fernandez, G.C.J. 2002 Data mining using SAS applications CRC/Chapman-Hall Publications FL <http://www.ag.unr.edu/gf/dm.html>
- 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 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.
- Neter, J. Wasserman, W., and Kutner, M.H. 1989. Applied Linear regression Models. 2nd Edition. Irwin Homewood IL.
- Sall, John 1990. Leverage plots for general linear hypothesis. The Amer. Statistician. Vol.44. 308-315
- Stine Robert A. 1995. Graphical Interpretation of Variance Inflation Factors. The American Statistician vol 49: 53-56.
- SAS, SAS/GRAPH, and SAS/STAT are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

### Author's Bio

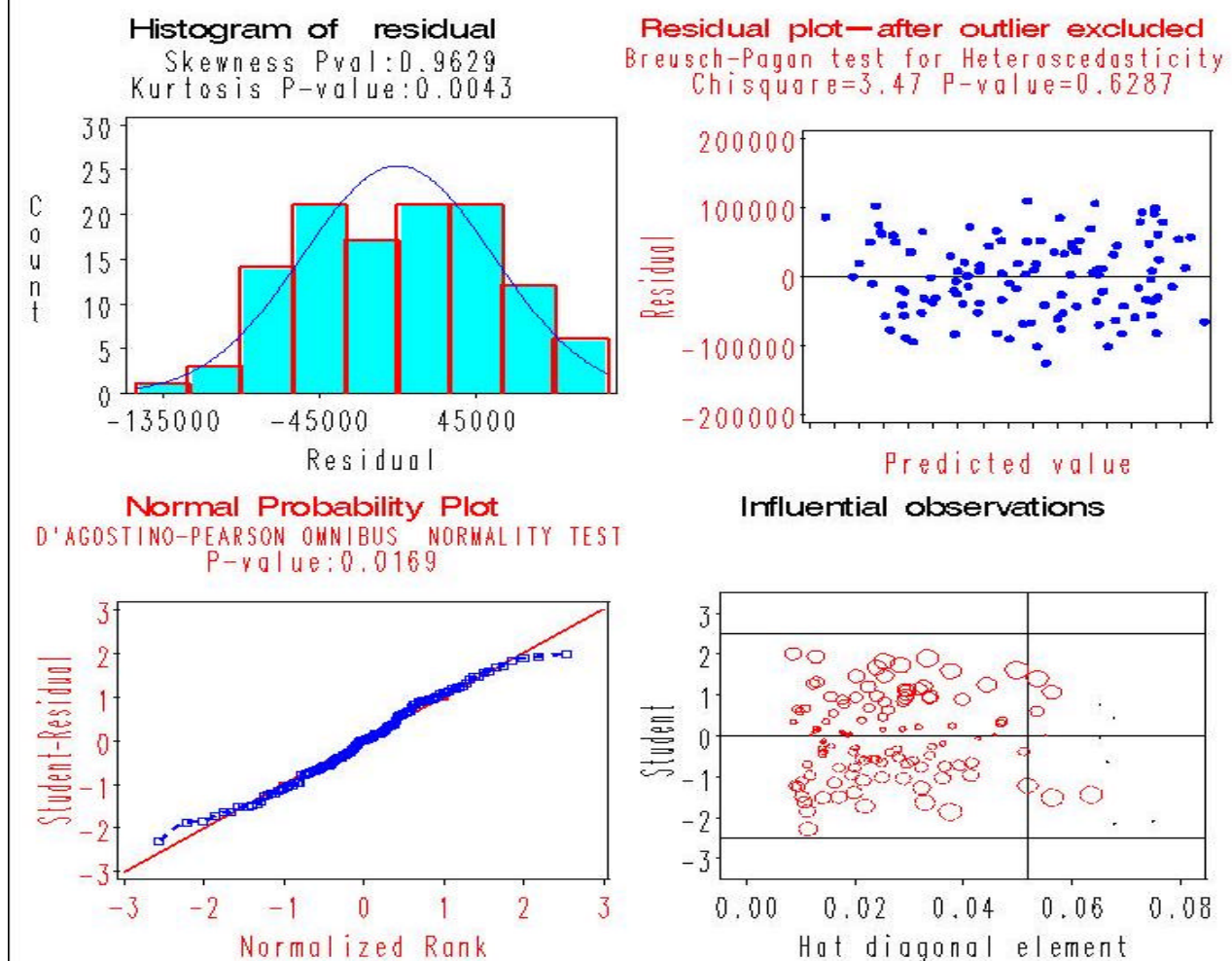
Dr. George C.J. Fernandez serves as the statistician for the Nevada Experimental station and Cooperative Extension and an associate professor in the department of applied economics and statistics at UNR. He has over 21 years experience in SAS/BASE, SAS/IML, SAS/STAT, SAS/QC SAS/ETS, SAS/INSIGHT, SAS/ANALYST, SAS/LAB, SAS/ASSIST and SAS/GRAPH. He has won best paper and poster presentation awards at the regional and international SAS conferences. He has presented invited full-day workshops on the "**Applications of user-friendly statistical methods in datamining**" at the American Statistical Association Joint meeting in Atlanta and at the –WUSS conferences in Arizona and in San Diego. Both international and national SAS users are currently using his user-friendly SAS macros for data analysis via on-line. He has organized 7<sup>th</sup> WUSS conference at Los Angeles in 1999 and served as the WUSS executive committee member. He has published more than 75 research papers including refereed journal papers, invited and contributed articles in proceedings, and book chapters. His new book

entitled "**Datamining using SAS applications**" published by the CRC press/Chapman&Hall contains 13 userfriendly SAS macro applications for performing data mining.

**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.  
(775) 784-4206 E-mail: GCJF@unr.edu  
Data mining book web page:  
<http://www.ag.unr.edu/gf/dm.html>

Checking for violations of assumptions Data= sales Response=y



**Figure 8** Diagnostic plots for checking violations of assumptions