Journal of the Korean
Data & Information Science Society
2000, Vol. 11, No.1, pp. 73 ~ 81

Diagnostics of partial regression and partial residual plots

Jea-Young Lee [1] · **Suk-Hwa Choi** [2]

Abstract

The variance inflation factor can be expressed by the square of the ratio of t-statistics associated with slopes of partial regression and partial residual plots. Disagreement of two sides in the interpretation can be occurred, and we analyze it with some illustrations.

Key Words and Phrases: Multicollinearity, Partial residual plot, Partial regression plot, Regression diagnostics, Variance inflation factor

1. Introduction

There are various diagnostics for multicollinearity, nonlinearity, heteroscedasticity and other problems. Residual plot is widely used as a means of examining the aptness of a model and finding the true functional form of covariate. One is the partial regression plot which is discussed or illustrated by Draper and Smith(1981), Anscombe(1967), Mosteller and Tukey(1977), Belsley, Kuh and Welsch(1980), and Weisberg(1980) etc. The partial residual plot which is well-known as a method for checking direction of the nonlinearity of a regressor is discussed by Larsen and McCleary(1972). And Wood(1973) called it as a residual plus component plot. Ezekiel(1924) was the first to use such a plot to determine if a regressor should be transformed. In this paper, we show why partial residual and partial regression plots are useful in multiple regression analysis.

One way to determine the multicollinearity is looking at variance inflation factor (VIF) proposed by Marquardt(1970). We concentrate on the relationship between variance inflation factor and two alternative plots, partial regression and partial residual plots, for least squares regression to understand regression diagnostics. We

[1] Associate Professor, Department of Statistics and Institute of Natural Science, Yeungnam University, Kyongsan, 712-749, Korea

[2] Department of Statistics, Yeungnam University, Kyongsan, 712-749, Korea

have to concern something to check multicollinearity from VIF. This paper analyzes the benefit of using both partial regression plot and variance inflation factor for diagnosing the multicollinearity.

2. Partial regression and Partial residual plots

Partial regression plot(added variable plot) is refined residual plot that does show the proper relation for an independent variable while standard residual plot does not display the nature of the relation for the independent variable that should be represented in the regression model.

Suppose the full-rank regression model for independent observations as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n \tag{2.1}$$

where ϵ_i is assumed to be $N(0, \sigma^2)$. In vector form, the model is $Y = X\beta + \epsilon$ where X is the $n \times (k+1)$ matrix with columns $X_0, X_1, \cdots, X_k$ and X_0 is column vector of 1's. Here, partition X to $(X_{(j)}, X_j)$ where X_j is the new added variable and $X_{(j)}$ is $n \times k$ matrix without X_j. The new model is $Y = X_{(j)}\beta_{(j)} + X_j\beta_j + \epsilon$, $j = 1, 2, \cdots, k$.

Suppose we are interested in whether X_j is needed or not. Let U be the hat matrix about $X_{(j)}$, that is $U = X_{(j)}(X'_{(j)}X_{(j)})^{-1}X'_{(j)}$. To use more conveniently we make as orthogonal to X_j, that is $X_j{}^* = (I - U)X_j$ and T^* as the hat matrix about $X_j{}^*$.

$$\begin{aligned} T^* &= X_j^*(X_j^{*\prime}X_j^*)^{-1}X_j^{*\prime} \\ &= (I - U)X_j(X_j'(I - U)X_j)^{-1}X_j'(I - U). \end{aligned}$$

We can write $V = U + T^*$, where V is hat matrix about containing all independent variables, since $X_{(j)}$ and X_j^* are orthogonal.

Let $e_{Y|X_j}$ be residuals from the regression of Y on the independent variables without X_j.

$$\begin{aligned} e_{Y|X_j} &= (I - U)Y = (I - (V - T^*))Y \\ &= e + (I - U)X_j\hat{\beta}_j \end{aligned} \tag{2.2}$$

where e are residuals from regression of Y on the all independent variables. Taking expectations over e in (2.2) gives

$$E[e_{Y|X_j}] = (I - U)X_j\hat{\beta}_j$$

which suggests that a plot of $e_{Y|X_j}$ versus $(I - U)X_j$ will be linear through the origin. Because it is a simple linear regression of $e_{Y|X_j}$ on $(I - U)X_j$, the slope of partial regression plot is $\hat{\beta}$ same as the full model (2.1).

Since the horizontal axis of partial regression plot is $(I-U)X_j$, it is not sufficient to show the direct relation between X_j and Y. Considering this aspect we define the new residuals e^* as the partial residuals replacing U by 0 in (2.2), that is

$$e^* = e + X_j\hat{\beta}_j. \tag{2.3}$$

We call the plot of e^* versus X_j as partial residual plot(component plus residual plot). It should be noted that the e^* are actually pseudo-residuals in that they are not residuals obtained from either using X_j or not using X_j. The slope of the partial residual model (2.3) is the same as the slope of full model, $\hat{\beta}$, since the model (2.3) is a simple linear model. Partial residual plot is very useful to detect nonlinear term about new-added variable in multiple regression. But this plot ignores the effect of multicollinearity and conveys a misleading impression of the significance of the fit, as noted by various authors like Larsen and Mccleary(1973) and Cook and Weisberg(1982).

3. Relationship between Variance inflation factor(VIF) and Two residual plots

In multiple regression analysis, one is often concerned with the nature and significance of the relation between the independent and dependent variables. When the independent variables are correlated among themselves, it is said to exist multicollinearity among them. Since multicollinearity inflates var($\hat{\beta}_j$) and causes problems with the sign of the $\hat{\beta}_j$ and confidence interval for $\hat{\beta}_j$, it is important to be able to detect multicollinearity when it exists. One way to detect the presence of multicollinearity is looking at VIF proposed by Marquardt(1970). Let's consider about the model (2.1) that is discussed in previous section. In vector form, $Y = X\beta + \epsilon$, the name of the diagnostic arises from writing the variance of the least squares estimator $\hat{\beta}_j (j = 1, \cdots, k)$ as

$$\begin{aligned} var(\hat{\beta}_j) &= \sigma^2 (X'X)^{-}1_{jj} \\ &= \frac{\sigma^2}{\sum_i (x_{ij}-\bar{x}_j)^2} \frac{1}{1-R_j{}^2} \end{aligned}$$

where R_j^2 is the R^2 statistic from the regression of X_j on the other independent variables. Then variance inflation factor for $\hat{\beta}$ is

$$VIF_j = \frac{1}{1-R_j^2}$$

The variance inflation factor is equal to 1 when R_j^2 is zero, i.e., when X_j is not linearly related to the other covariates. When R_j^2 is not zero, VIF_j is greater than

1 indicating an inflated variance for $\hat{\beta}_j$. Unfortunately, there is no perfect critical value for what is needed to have a "large" VIF. But multicollinearity is declared to exist if the VIF value is in excess of 10 by Chatterjee and Price(1991).

Although both the partial regression and partial residual plots have the same slope $\hat{\beta}_j$ as discussed in previous section and the same residuals, $\hat{\epsilon} = Y - X\hat{\beta}$, their appearance can be remarkably different since the X-axis' are not equal. Consider the variances of slope in two alternative residual plots. First, the estimated variance of the slope for $\hat{\beta}_j$ in the partial regression plot is

$$\begin{aligned} Var_j^{reg} &= \frac{n-(k+1)}{n-2}\hat{\sigma}^2(X'X)^{-1}_{jj} \\ &= \frac{n-(k+1)}{n-2}\frac{\hat{\sigma}^2}{\sum_i (x_{ij}-\bar{x}_j)^2(1-R_j^2)} \end{aligned} \tag{3.1}$$

where $\hat{\sigma}^2 = \sum_i \hat{\epsilon}_i^2/(n-(k+1))$. With adjustment for degrees of freedom the apparent estimated variance of the slope based on the partial regression plot is the same as the estimated variance of $\hat{\beta}_j$ from the full regression.

On the other hand, the estimated variance of the slope for based on the partial residual plot is

$$Var_j^{res} = \frac{n-(k+1)}{n-2}\frac{\hat{\sigma}^2}{\sum_i (x_{ij}-\bar{x}_j)^2} \tag{3.2}$$

which ignores any effect due to fitting the other variables. We can easily see that the variance in (3.2) can be much smaller than the variance in (3.1) if R_j^2 is large. Also, the partial residual plot will present an incorrect image of the strength of the relationship between Y and X_j (conditional on the other Xs).

Comparing (3.1) with (3.2) the difference between two variances is $1/(1-R_j^2)$. As Robert A. Stine(1995) suggested, VIF can be written as the ratio of variances of the slopes in two residual plots. That is

$$\begin{aligned} VIF_j &= \frac{Var^{reg}{}_j}{Var^{res}{}_j} \\ &= \left[\frac{\hat{\beta}_j/\sqrt{Var^{res}}}{\hat{\beta}_j/\sqrt{Var^{reg}}}\right]^2 \end{aligned} \tag{3.3}$$

Since the slopes of two alternative plots are equal, VIF_j also can be expressed by square of the ratio of the t-statistics dividing both numerator and denominator by $\hat{\beta}_j^2$ in (3.3). The value of VIF is more than 10 means the square of the ratio of the t-statistics exceed 10 and there exists severe multicollinearity. The large value of VIF_j tells us X_j variable is highly correlated with the other covariates and don't need to add to the model. But, although the square of ratio is very large, if the slope of the partial regression is significant, we are in something of a dilemma. Here, we should consider both VIF_j and significance of $\hat{\beta}_j$ in partial regression plot to check multicollinearity.

4. Illustrations

We analyze 2 examples to understand the two plots and VIF we have discussed so far. The wildcat example in exercise 7.21 of Gujarati(1995) considers a time-series regression with substantial collinearity. The data consists of the number of wildcats drilling activity as dependent variable, price at the wellhead in the previous period(X_1), domestic production(X_2), GNP constant(X_3) and time trend(X_4).

In this example, we can see that the slopes of two residual plots from Table 2 are equal to the estimated regression coefficients in full model from Table 1. And all of the slopes in partial residual plots are more significant than those of partial regression plots.

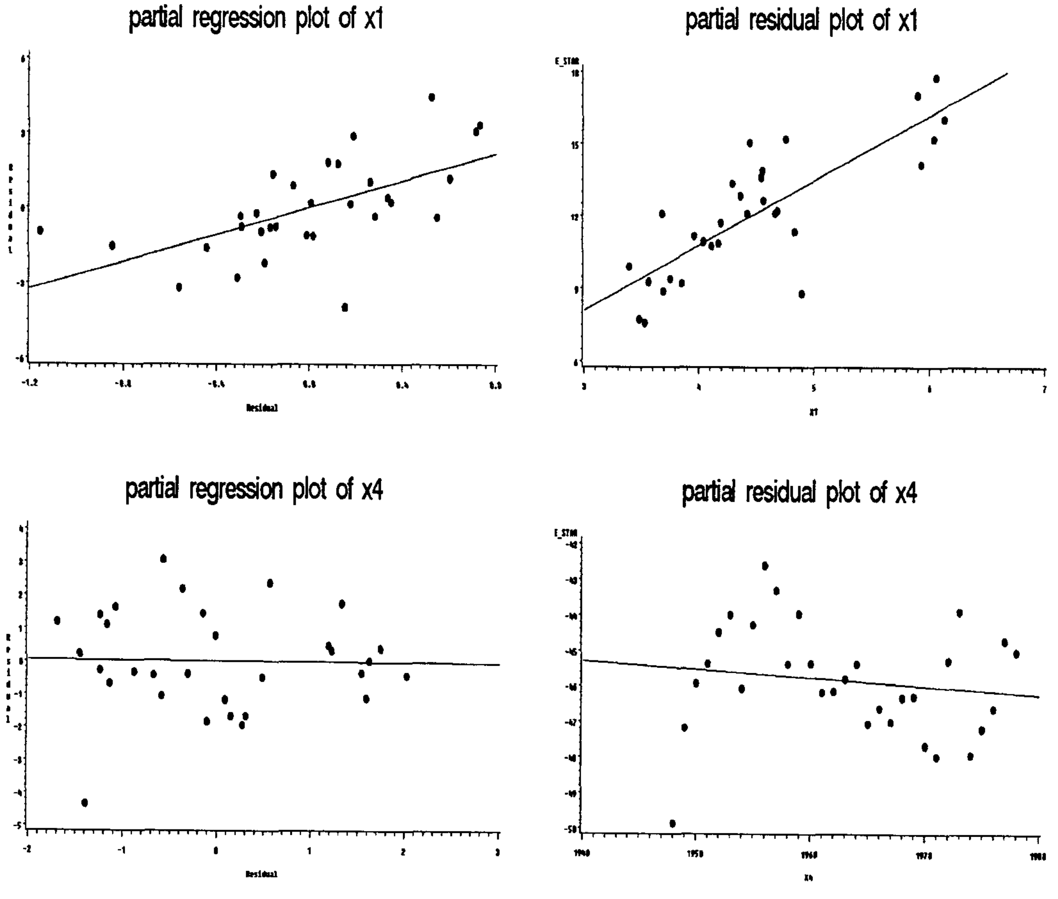

Figure 1. Two residual plots with X_1 and X_4 in Wildcat data

<Table 1> Summary statistic of the Least Squares Regression Model fitted using Wildcat data for 31 observations. Goodness of Fit statistic is $f = 8.917 (p = 0.0001)$ and the square of the multiple correlation is $R^2 = 0.587$.

Variable	Estimate	Standard Error	t-statistic	VIF
constant	35.659	525.94	0.068	n.a.
X_1	2.700	0.700	3.865	3.59
X_2	3.045	0.941	3.236	12.50
X_3	-0.016	0.008	-1.948	55.19
X_4	-0.023	0.273	-0.085	68.69

It is because of multicollinearity as we already discussed. If the variable of interest is X_1 or X_4, then the interpretation of VIF is same as the partial regression plot. That is, X_4 is highly correlated with the other covariates(VIF=68.69) and we don't need to add it to the model because the slope of partial regression plot is not

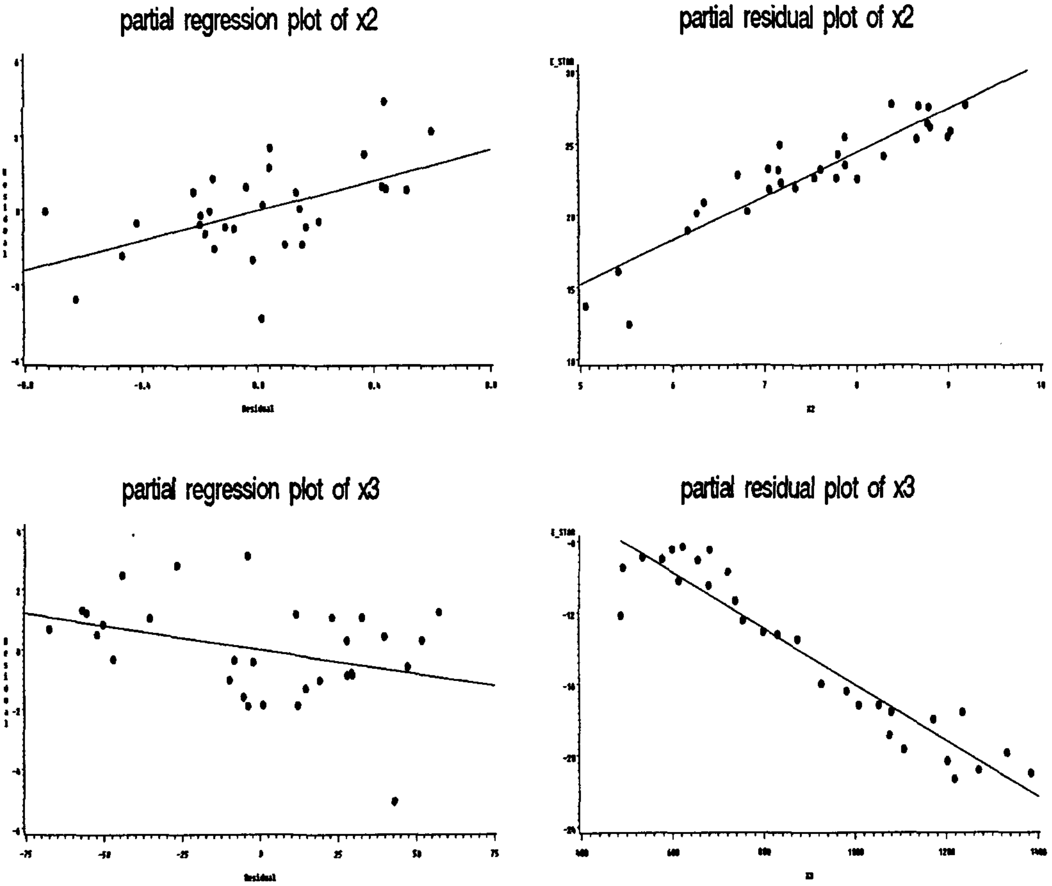

Figure 1. Two residual plots with X_2 and X_3 in Wildcat data

significant(p=0.4608). In the other case, the slope of partial regression plot for X_2 is very significant (=0.0019). It means that the independent variable X_2 has much information to predict dependent variable and is needed to the model. But the value of VIF(=12.50) which is more than 10 tell us that X_2 is highly correlated. At this point the analyst is in something of a dilemma. It's very similar with X_3. If we only see VIF_2, we will conclude X_2 is not an important variable. But both slopes of two residual plots are significant in figure 2.

<Table 2> Summary statistic on partial regression plots(left) and partial residual plots(right) when X_2 is the added variable.

New Added variable	slope $(=\hat{\beta}_j)$	S.E.	t-stat.	p-value	slope $(=\hat{\beta}_j)$	S.E.	t-stat.	p-value
X_1	2.700	0.6625	4.082	0.0003	2.700	0.349	7.735	.0001
X_2	3.045	0.891	3.417	.0019	3.045	0.252	12.082	0.0001
X_3	-0.016	0.008	-2.057	0.0488	-0.016	0.001	-15.281	0.0001
X_4	-0.023	0.259	-0.090	0.9288	-0.023	0.0312	-0.747	0.4608

Second example given by Wood(1973) is for comparing the behavior of the partial regression plot and the partial residual plot when nonlinear term is needed. 40 observations were generated as $Y = 20 + 20X_1 - 3X_2 - X_1^2 + \epsilon$, where $\epsilon sim N(0,1)$. The values of X_1 and X_2 used to generate observations 1-20 were repeated for observations 21-40.

Using the first 20 data points like Mansfield and Conerly(1987), the plots are presented in figure 3. The second graph which is the partial residual plot shows the nature of the curvilinear relation between Y and X_1, when X_2 is already in the regression model. Vertical deviations around the line with slope $\hat{beta}_1$ are negative

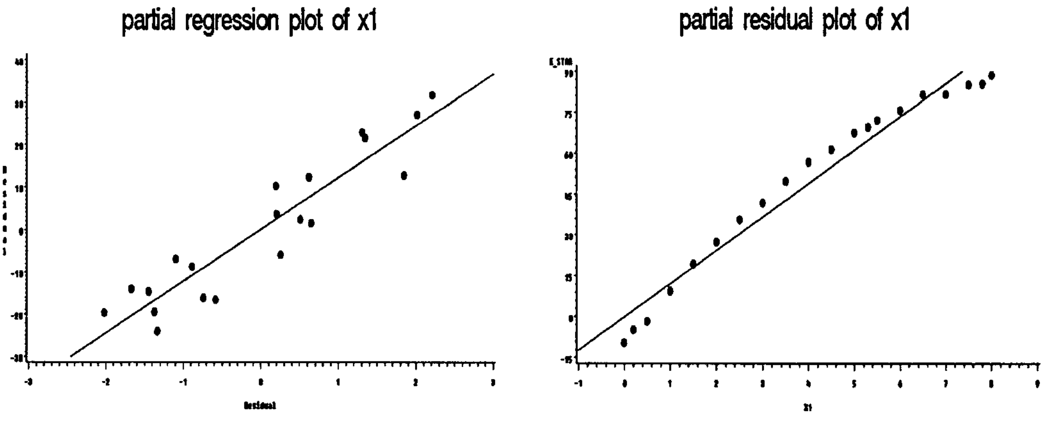

Figure 3. Two residual plots of X_1 Wood data

at the left, positive in the middle, and negative again at the right. But the partial regression plot fails to indicate the need of quadratic term.

<Table 3> Summary statistic using Wood data for 20 observations. Goodness of Fit statistics is $F = 80.918(p = 0.0001)$ *and the square of the multiple correlation is* $R^2 = 0.905$.

Variable	Estimate	S.E	t-statistic	p-value	VIF
constant	29.612	2.836	10.441	.0001	n.a
X_1	12.222	1.165	10.491	.0001	4.07
X_2	-3.046	0.240	-12.679	.0001	4.07

5. Conclusion

The difference between partial residual and partial regression plots comes from $\hat{R}^2$. Although partial residual plot ignores any effect due to fitting the other variables, partial residual plot is more sensitive to display the correct form of the regressor variable. We analyzed the relationship through VIF and two alternative residual plots. Since the slopes of two alternative plots are equal, VIF can be expressed by square of the ratio of the t-statistics for slopes as we showed. The large value of VIF means the regressor is highly correlated with the others. But if the slope of the partial regression is significant, although the square of ratio is very large($VIF > 10$), we are in something of a dilemma to interpret. We concluded that it should be considered both VIF and significance of $\hat{\beta}_j$ in partial regression plot to check multicollinearity. Of course, we also consider a variety of remidial measures for dealing with these multicollinearity problems, including ridge regression, factor and principal component analysis, and bayesian regression.

ACKNOWLEDGEMENTS

The authors thank the referees for their useful comments.

References

1. Anscombe, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares(with discussion). *Journal of Royal Statistical Society*, Series, B, 29, 1-52

2. Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: Wiley

3. Chatterjee, S., and Price, B. (1991). *Regression Diagnostics*, New York: John Wiley

4. Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, New York: Chapman and Hall

5. Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd edition. New York: Wiley

6. Mansfield, E. R. and Conerly, M. D. (1987). Diagnostic Value of Residual and Partial Residual Plots, *American Statistical Association*, 41, No.2, 107-116

7. Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variables, *Journal of the American Statistical Association*, 19, 431-453

8. Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis, *Technometrics*, 14, 781-791

9. Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, *Technometrics*, 12, 591-612

10. Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*, reading, MA: Addison-Wesley

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

12. Weisberg, s. (1980a). *Applied Linear Regression.* New York: Wiley.

13. Wood, F. S. (1973). The Use of Individual Effects and Residuals in Fitting Equations to Data, *Technometrics*, 15, 677-695