The Korean Communications
in Statistics Vol. 7, No. 1, 2000
pp. 37-46

A Systematic View on Residual Plots in Linear Regression[1)]

Myung-Wook Kahng[2)], YoungIl Kim[3)], Chul H. Ahn[4)]

Abstract

We investigate some properties of commonly used residual plots in linear regression and provide some systematic insight into the relationships among the plots. We discuss three issues of linear regression in this stream of context. First of all, we introduce two graphical comparison methods to display the variance inflation factor. Secondly, we show that the role of a suppressor variable in linear regression can be checked graphically. Finally, we show that several other types of standardized regression coefficients, besides the ordinary one, can be obtained in residual plots and the correlation coefficients of one of these residual plots can be used in ranking the relative importance of variables.

1. Introduction

Consider the standard regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

where $\boldsymbol{y}$ is an $n\times 1$ response random vector, $\boldsymbol{X}$ is an $n\times p$ data matrix and $\boldsymbol{\beta}$ is a $p\times 1$ parameter vector to be estimated. We assume that the random vector $\boldsymbol{\varepsilon}$ follows $N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Scientific investigators are often confronted with the problem of explaining residuals after the removal of some particular cause of variability. Therefore, it is sometimes convenient to extend model (1.1) to include an extra carrier $\boldsymbol{z}$ into the model as follows

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{z} + \boldsymbol{\varepsilon} \tag{1.2}$$

Let's assume that the extra carrier $\boldsymbol{z}$ is coming into the model in one dimension for the time being.

When the extra carrier $\boldsymbol{z}$ is not needed to explain the variability of $\boldsymbol{y}$ not accounted for

1) This research was supported by the ChungAng University Research Grants in 1999.
2) Associate Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea.
3) Professor, Department of Information System, ChungAng University, Kyunggi-Do, 456-756, Korea.
4) Associate Professor, Department of Applied Mathematics, Sejong University, Seoul, 143-747, Korea.

X, then the reduced model (1.1) is preferred over the full model (1.2). A vice versa situation is possible. The decision between model (1.1) and model (1.2) is always difficult in practice. Three possible causes of this difficulty are first of all, the size of the magnitude of effect of z, secondly the possible association among the variables, and lastly the functional form of z entering the model. In model (1.2), we assume that z enters in linear functional form, but this is not necessarily true. Sometimes z enters the model nonlinearly. There are many statistical literatures containing this topic. We think that the book by Cook and Weisberg(1994) is the most appropriate one.

In this article, we consider the following four residual plots for obtaining a graphical evaluation of the effect of adding an explanatory variable z.

[1] The simple residual plot: Plot the vector of residuals for the regression of y on X versus z; $(y-X\hat{\beta})$ versus z, where $\hat{\beta}$ minimizes $(y-X\beta)^T(y-X\beta)$.

[2] The partial residual plot: Plot the vector of $(y-X\hat{\beta}-\hat{\gamma}z)+\hat{\gamma}z$ versus z, where $(\hat{\beta}, \hat{\gamma})$ minimizes $(y-X\beta-\gamma z)^T(y-X\beta-\gamma z)$.

[3] The added variable plot: Plot the vector of residuals of y on X versus the vector of residuals of Z on X; $(y-X\hat{\beta})$ versus $(z-X\hat{\alpha})$, where $\hat{\beta}$ minimizes $(y-X\beta)^T(y-X\beta)$ and $\hat{\alpha}$ minimizes $(z-X\alpha)^T(z-X\alpha)$.

[4] The additional R^2 plot: Plot y versus the vector of residuals of Z on X; y versus $(z-X\hat{\alpha})$, where $\hat{\alpha}$ minimizes $(z-X\alpha)^T(z-X\alpha)$.

The simple residual plot is sometimes introduced as a part of regression modelling, as can be seen in Atkinson(1985). But he mentioned this plot without any explanation on its linkage with other plots. The partial residual plot is also called a component plus residual plot. The partial residual plot has a long history, going back to Ezeikel(1924). This plot is heavily studied by many authors who are especially trying to detect the nonlinearities of z. A simple modification of this partial residual plot is developed by Mallows(1986). A more general approach is made by Cook(1993). The added variable plot is also called a partial regression plot. It is another useful plot for checking the effect of an additional regressor. It is well known that the residuals from the added variable plot are the same as the ones in full model (1.2), and that the slope takes the same value as the model (1.2). For these reasons, it is strongly favored by a majority in detecting the effects of individual observation. It is sometimes introduced as a part of explaining the fitting procedure of the least squares method. See Draper and Smith(1998) and Weisberg(1985) for examples. The additional R^2 plot is given in Guttmann(1982). This plot will be further explained as the paper progresses.

In most textbooks about the regression, the residual analysis immediately follows after mathematical treatment of the least squares method for estimating the parameters in the linear

regression. Furthermore, the residual plots themselves are often treated in dealing with the overall check of the model only. As a result, the residual analysis sometimes put the students into difficulties in understanding how it contributes to the process of regression model build-up. Each of the four plots mentioned above is related to each other. Therefore, much more information can be obtained from the residual plots with careful comparison of them. These plots are discussed further by Berk and Booth(1995).

In this article, we mentioned three issues; 1) variance inflation factor, 2) suppressor variable and 3) its related issue, ranking of variables in the model. We explain the mechanism of residual plots to show that these three issues can be analyzed graphically using four plots.

The three issues mentioned above will be discussed in the sequence in sections 2, 3 and 4. Conclusion will be made in the 5th section.

2. Variance Inflation Factor

It can be shown that the estimated slope $\hat{\gamma}^*$ of the simple residual plot is related to $\hat{\gamma}$ under the full model (1.2) as

$$\hat{\gamma}^* = (1-R^2_{zX})\hat{\gamma} \tag{2.1}$$

where R^2_{zX} is the coefficient of determination when z is regressed on $\boldsymbol{X}$. Note that the slope in simple residual plot is always smaller than one in model (1.2) in absolute value.

It is immediate (although not explicit in the literature) from equation (2.1) that

$$\hat{\gamma}/\hat{\gamma}^* = 1/(1-R^2_{zX}) \tag{2.2}$$

which is just the variance inflation factor VIF for z, VIF_z. Thus, the comparison of $\hat{\gamma}$ and $\hat{\gamma}^*$ will give us an idea of how large VIF_z will be. For their graphical comparison, we need a plot in addition to simple residual plot. Since the x-axis of simple residual plot is z, the plot that has the same x-axis is the most appropriate one. We propose that the partial residual plot is the one to be compared with. It is well known that the slope of the partial residual plot is the same as the one in the full model (1.2).

VIF_z can also be presented as the ratio of R^2 associated with the simple residual and added variable plot as follows:

$$VIF_z = r^2_{add}/r^2_{sim} \tag{2.3}$$

where r^2_{sim} and r^2_{add} are the correlation coefficients associated with the simple residual and added variable plot, respectively. Note that the r^2 in the simple residual plot is always smaller than one in the added variable plot. Both of equations (2.2) and (2.3) will be useful in

presenting the variance inflation factors of variables. But, in some cases when large VIF cannot be detected by the comparison of two slopes, which can occur when both of their fits are not good, this comparison of two correlation coefficients may work well.

Detecting VIF_z is further aided by another comparison. Stine(1995) has noted that the ratio of two variances of estimated slopes of the added variable and partial residual plot is just VIF_z. As Cook and Weisberg(1982, p. 51) mentioned, if R^2_{zX} is large, then the variability around the slope in the residual plot can be much smaller than the one in the added variable plot, and the partial residual plot will present an incorrect image of the strength of the relationship between y and z. Comparing two variabilities along the slope will give us some intuition on how significant the VIF_z would be. This seems to be comparable to Atkinson's comment(1985, p. 75) that the ratio of horizontal scatter in the partial residual plot to that in the added variable plot quantifies the extent to which the partial residual plot over-emphasizes the importance of the relationship between y and z.

In summary, when all three plots are displayed as a part of routine diagnosis checking, we get three methods to obtain additional information about VIF_z. When the x and z are orthogonal, then three methods should be exactly the same.

3. Suppressor Variable

We teach in class that as we add a variable to the model the value of R^2 increases monotonically. But we usually do have mis-conception about the delicate mechanism of this R^2. We take the usual notations

$$SSR(\boldsymbol{X}, z) = SSR(\boldsymbol{X}) + SSR(z|X)$$

where SSR denotes the regression sum of squares and $SSR(z|X)$ is the extra sum of squares obtained after entering z.

Hamilton(1987) mentioned that sometimes $SSR(z|X) > SSR(z)$ is caused by the entering variable z. Sharpe and Roberts(1997) named z a suppressor variable, a variable that increases the importance of another variable when it is added to the regression. Although this phenomenon is rare in real data analysis, it happens under some certain conditions. Hamilton (1987) showed that the following was the necessary and sufficient condition for his claim

$$r^2_{yz \cdot X} > r^2_{yz}(1 - R^2_{yX}) \tag{3.1}$$

The lefthand side of (3.1) is the squared partial correlation coefficient. Weisberg(1985, p. 40) has mentioned that when the association reflected by r^2 in added variable plot, which is just

the squared correlation coefficient between y and z given X is greater than r^2_{yz}, then X and z interact to explain more than the sum of R^2_{yX} and r^2_{yz}. Obviously he did not take into consideration the effect of R^2_{yX}. Therefore, we had some thoughts to devise a proper and simple graphical comparison. Note that $r^2_{yz\cdot X}(1-R^2_{yX})$ is just the additional increase of R^2 when the variable z enters the model. This is just the squared correlation coefficient between y and $(z-X\hat{\alpha})$. It can be shown that although the y-axis is different from the one in the added variable plot, the estimated slope of the additional R^2 plot is the same as the one in full model (1.2). See Guttmann(1982) for details. But since the residuals are different from the ones in the full model, it is not frequently used in practice, unlike the added variable plot. Still it would be helpful to explain the concept of the additional increase of R^2 graphically. Furthermore it is nice to get additional information about the peculiar issue raised by Hamilton(1987).

In conclusion, when the additional R^2 plot shows much stronger association than the simple plot of y vs z, then we say the sum of *SSRs* due to individual X and z is less than the overall *SSR* due to both X and z. Routinely we do have information about $(z-X\hat{\alpha})$ when constructing the added variable plot, so this comparison does not require additional computational work. When X and z are orthogonal to each other, then obviously $SSR(z|X)=SSR(z)$.

4. Standardized Regression Coefficient

In most social science research work there are some interests concerning the rank of relative importance of different variables in the model. Statistical packages such as *SPSS* provide the printouts on the standardized coefficient denoting the following relationship between usual $\hat{\gamma}$ and the standardized coefficient B_z for z

$$B_z = \hat{\gamma} \times S_z / S_y$$

where S_z and S_y are the standard deviation of the variables z and y, respectively. But many people make cautionary remarks that the rankings of the standardized coefficients in terms of absolute magnitude does not necessarily reflect the importance of variables in explaining the variability of y. Many textbooks give warnings against the misuse of this automatic computer generated output. But none of the textbooks had explained the relationship between this standardized coefficient and the correlation coefficients of various residual plots. The correlation coefficients of the plots from [1], [2], [3], and [4] are computed algebraically

as follows:

$$
\begin{aligned}
&[1]\ \hat{\gamma}^* \cdot S_z / (S_y \cdot \sqrt{1-R^2_{yX}}) \\
&[2]\ \hat{\gamma} \cdot S_z / \sqrt{S^2_y(1-R^2_{y\cdot Xz}) + \hat{\gamma}^2 S^2_z} \\
&[3]\ \hat{\gamma} \cdot (S_z\sqrt{1-R^2_{zX}}) / (S_y \cdot \sqrt{1-R^2_{yX}}) \\
&[4]\ \hat{\gamma} \cdot (S_z\sqrt{1-R^2_{zX}}) / S_y
\end{aligned}
\tag{4.1}
$$

From (4.1), we immediately see that all of these correlation coefficients are other measures of standardized regression coefficient except those appropriate adjustments taking place in each formula in (4.1). And this suggests that the appropriate correlation coefficients may be used in ranking the relative importance of variables. The appropriate correlation coefficients are those associated with [3] and [4].

The correlation coefficient derived from added variable plot is a good measure in determining the ranks of variables according to the size of partial t-values for regression coefficient. In other words, the ranks according to the partial t-values for coefficients are the same as those according to these partial correlation coefficients. This is based on the result of Stapleton(1995), $r^2_{yz\cdot X} = d/(1+d)$, where $d = t^2_z/(n-p-1)$. But all except the last one still lack interpretations on determining the exact size of contribution of each variable in the model.

The correlation coefficient from the additional R^2 plot compares each variable's contribution in terms of additional increase of R^2 over the variability explained by other variables in the model. Bring(1994) had derived a very similar measure of standardized regression coefficient like this, with a different adjustment. His measure seemed to follow a hard path for interpretation. He took the degrees of freedom into consideration. However, ours has a more clear interpretation. We admit that the exact contribution of different variables we defined here may not necessarily be accepted by the social researchers. Neverthless, when we are interested in the relative importance of different variables we think that this is the right choice. The best thing is that we visually compare the relative importance of different variables in terms of increase(reduction) of R^2 if a variable entered(omitted). Futhermore, the ratio of two correlation coefficients from the additional R^2 plot is exactly the same as the ratio of corresponding partial t-values. Therefore, comparing partial t-values is equivalent to considering the reduction in R^2, obtained by eliminating each of the variables. See Bring(1994) for similar results. In our opinion, it seems very appealing to define the exact contribution of variables in terms of increase of R^2.

Before we conclude this section, we'd like to say that it is a common practice to have plots in original scale of measurement in order to preserve the symptoms such as heterogeneity of

variance or non-linearity of entering variable. Standardized versions of residual plots, i.e. plots obtained after standardizing all the variables before analysis, are not recommended for these reasons.

5. Examples

To illustrate the concepts explained in Sections 2, 3 and 4, we use the data as described by Neter et al. (1996, p.335). The data consist of four explanatory variables - blood clotting score(X_1), prognostic index(X_2), enzyme function test score(X_3), liver function test score (X_4). The response is survival time(Y). The original response variable is common logarithm ($\log_{10}$) transformed to build a regression model according to their suggestion.

1. *VIF*: Suppose all other explanatory variables but X_3 are in the model. Figures 1 to 3 represent simple residual plot, partial residual plot, and added variable plot, repectively. The *VIF* value for X_3 is 1.678. The impact of *VIF* for X_3 is graphically displayed as the ratio of two slopes, 0.009475 for Figure 2 and 0.005646 Figure 1. Both the ratio of two r^2, 0.9211 for Figures 3 and 0.5489 for Figures 1, and the ratio of two estimated variance of slope regression coefficients, 0.0003847^2 for Figures 3 and 0.0002969^2 for Figures 2, show the same value equal to VIF for X_3.

2. Suppressor Variable: After a preliminary examination of the full model, the variable X_4 is dropped. Unlike the full model, the reduced model shows that there exists two suppressor variables, X_1 and X_3. Figure 4 is the additional R^2 plot for X_3, of which r^2 is shown to be greater than that of Figure 5, which is just the plot of $\log_{10} y$ versus X_3. The two actual r^2 values are 0.5343 and 0.4424 for Figures 4 and 5, respectively.

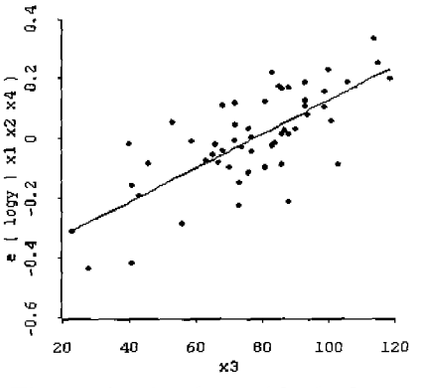

Figure 1: simple residual plot

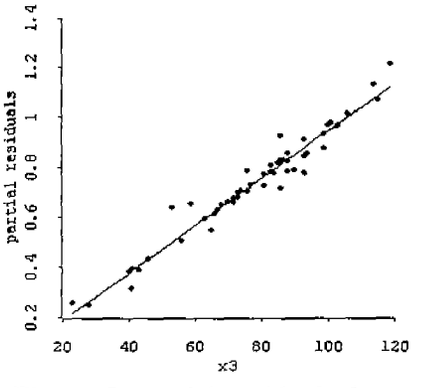

Figure 2: partial residual plot

3. Ranking of Variables: Two more additional R^2 plots are displayed for X_1, X_2 together with X_3. New versions of standardized regression coefficients are obtained from Figure 6, Figure 7 and Figure 4. The values of correlation coefficient are 0.399, 0.571 and 0.731 for X_1, X_2, and X_3, respectively. According to the size of these values it seems that X_3 is the most important and X_1 the least important from the additional R^2 viewpoint. For reference the conventional standardized regression coefficients are computed for X_1, X_2, and X_3. They are 0.405, 0.574, and 0.739, respectively. Additional R^2 for X_1, X_2, and X_3 are 0.159, 0.327, and 0.534, respectively.

Traditionally in most regression textbooks explanations are given only to the individual residual plot neglecting the systematic view of various plots. We believe that a comparative study of various residual plots not only reveals much more information on the three issues raised here but also helps students understand the basic mechanism of the underlying fitting process of regression.

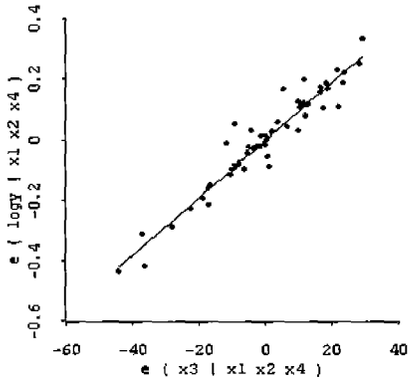

Figure 3: added variable plot

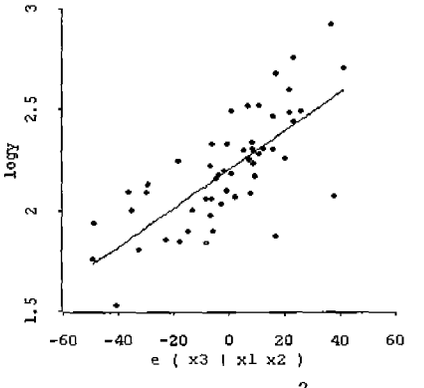

Figure 4: additional R^2 Plot of X_3

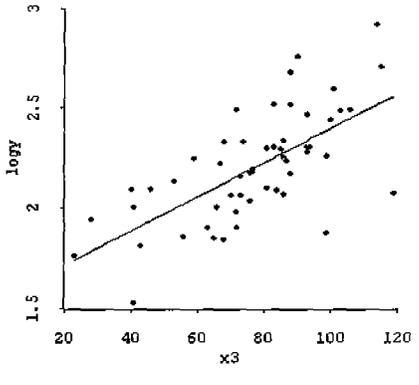

Figure 5: plot of $\log_{10} y$ versus X_3

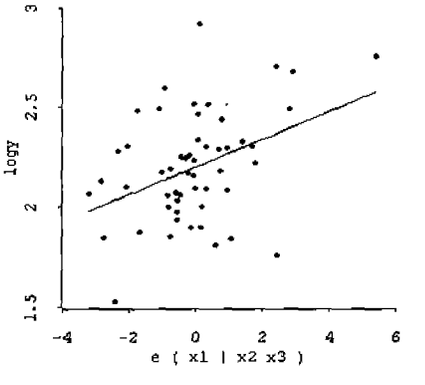

Figure 6: additional R^2 Plot of X_1

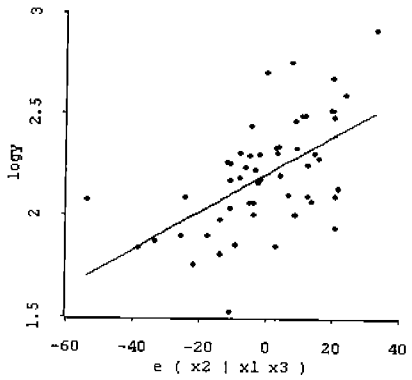

Figure 7: additional R^2 Plot of X_2

6. Conclusions

We have made some useful remarks on the commonly cited residual plots in linear regression to get additional information about 1) the *VIF*, 2) the suppressor variable, and 3) the interpretations of correlation coefficients. All of these can be readily obtained from direct interpretation of the residual plots. Some discussions on the relationship among the plots will help teachers in convincing students the usefulness of residual plots. We hope that these materials will be helpful in teaching regression to students inside and outside the field of statistics.

References

[1] Atkinson, A. C. (1985). *Plots, Transformations, and Regression,* Oxford University Press: Oxford.

[2] Berk, K. N. and Booth, D. E. (1995). Seeing a curve in multiple regression, *Technometrics,* Vol. 37, 385-398.

[3] Bring, J. (1994). How to standardize regression coefficients, *The American Statistician,* Vol. 48, 209-213.

[4] Cook, R. D. (1993). Exploring partial residual plots, *Technometrics,* Vol. 35, 351-362.

[5] Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression,* Chapman and Hall: London.

[6] Cook, R.D. and Weisberg, S. (1994). *Introduction to Regression Graphics,* John Wiley & Sons, New York.

[7] Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis, 3rd Edition,* John Wiley & Sons, New York.

[8] Ezekiel, M. (1924). A method for handling curvilinear correlation for any numbers of variables, *Journal of the American Statistical Association,* Vol. 19, 431-453.

[9] Guttmann, I. (1982). *Linear Models: An Introduction,* John Wiley & Sons, New York.

[10] Hamilton, D. (1987). Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$, *The American Statistician,* Vol. 41, 129-132.

[11] Mallows, C. L. (1986). Augmented partial residual plots, *Technometrics,* Vol. 28, 313-320.

[12] Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Model, 4th Edition,* Irwin.

[13] Sharpe, N. R., and Roberts, R. A. (1997). The relationship among sums of squares, correlation coefficients, and suppression, *The American Statistician,* Vol. 51, 46-48.

[14] Stapleton, J. H. (1995). *Linear Statistical Models,* John Wiley & Sons, New York.

[15] Stine, R. A. (1995). Graphical interpretation of variance inflation factor, *The American Statistician,* Vol. 49, 53-56.

[16] Weisberg, S. (1985). *Applied Linear Regression, 2nd Edition,* John Wiley & Sons, New York.