대한산업공학회/한국경영과학회 2000 춘계공동학술대회 논문집. 2000년 4월 21-22일. 경남대학교
Session OC6.2

선형회귀모형에서 잔차분석에 대한 시스템적 관점
Systematic View on Residual Plots in Linear Regression

강명욱[*], 김영일[**], 안철환[***]
*숙명여자대학교 통계학과 **중앙대학교 정보시스템학과***세종대학교 통계학과

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

$$y = X\beta + \varepsilon \qquad (1.1)$$

where y is an $n\times 1$ response random vector, X is an $n\times p$ data matrix and β is a $p\times 1$ parameter vector to be estimated. We assume that the random vector ε follows $N(0, \sigma^2 I)$.

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 z into the model as follows

$$y = X\beta + \gamma z + \varepsilon \ (1.2)$$

Let's assume that the extra carrier z is coming into the model in one dimension for the time being. When the extra carrier z is not needed to explain the variability of y not accounted for X, then the reduced (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.

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
[2] The partial residual plot.
[3] The added variable plot
[4] The additional R^2 plot·

The simple residual plot is sometimes

introduced as a part of regression modelling as can be seen in Atkinson(1985). The partial residual plot has a long history, going back to Ezeikel(1924). The added variable plot is also called a partial regression plot. It is another useful plot for checking the effect of an additional regressor. 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.

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

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

where R^2_{zX} is the coefficient of determination when z is regressed on X.

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

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

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:

$$VIF_z = r^2_{add} / r^2_{sim} \qquad (2.3)$$

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.

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

$$SSR(X, z) = SSR(X) + SSR(z|X)$$

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

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

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. 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). 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).

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.

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:

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

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].

5. 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.

References

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

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

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

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

[5] 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.

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