Communications of the Korean Statistical Society
2009, Vol. 16, No. 3, 495–500

Two Diagnostic Plots in Constrained Regression

Myung Geun Kim[1,a]

[a]Department of Mathematics Education, Seowon University

Abstract

Two diagnostic plots, added variable plot and partial residual plot, are proposed when a new explanatory variable is linearly added to constrained regressions. They are useful for investigating the effect of adding an explanatory variable to the constrained regression. They visually give an overall impression of the strength of linear relationship between response variable and added variable. A numerical example is provided for illustration.

Keywords: Added variable plot, constrained regression, partial residual plot.

1. Introduction

Mosteller and Tukey (1977) introduced the added variable plot, also called the partial regression plot. The added variable plot is a graphical method of visualizing the contribution of each explanatory variable to the fit of an adequate target model, as noted in Cook and Weisberg (1994). This added variable plot is useful for understanding the role of the added variable and for identifying outliers. Also, the added variable plot indicates which observations are contributing to the linear relationship between response variable and added variable and which observations are not.

A diagnostic plot was introduced for diagnosing the need to transform an explanatory variable by Ezekiel (1924). It was called the partial residual plot by Larsen and McCleary (1972), and also called the component-plus-residual plot by Wood (1973). It is a computationally convenient plot substituting for the added variable plot.

Constrained regressions are used widely in the field of econometrics, for example in the estimation of Cobb-Douglas production functions as noted in Chipman and Rao (1964). However, no diagnostic plots are available for constrained regressions. In this work, added variable plot and partial residual plot are suggested when an explanatory variable is linearly added to constrained regressions. An illustrative example is provided.

2. Diagnostic Plots

2.1. Preliminaries for constrained regression

In this subsection some results for the multiple linear regression with linear constraints on regression coefficients are reviewed.

The constrained regression can be written as

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \tag{2.1}$$

with linear constraints on β

$$A\beta = c,$$

[1] Professor, Department of Mathematics Education, Seowon University, 231 Mochung-Dong, Cheongju, Chungbuk 361-742, Korea. E-mail: mgkim@seowon.ac.kr

where $\boldsymbol{y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors which are independent and identically distributed as a normal distribution with mean zero and unknown variance, $\boldsymbol{A}$ is a specified $q \times p$ $(q \leq p)$ matrix of rank q, and $\boldsymbol{c}$ is a specified $q \times 1$ vector.

The least squares estimator of $\boldsymbol{\beta}$ is

$$\breve{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} - \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{A}^T\left[\boldsymbol{A}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{A}^T\right]^{-1}\left(\boldsymbol{A}\tilde{\boldsymbol{\beta}} - \boldsymbol{c}\right),$$

where $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. The residual vectors are denoted by $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\breve{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$. Denoting the hat matrix by $\tilde{\boldsymbol{H}} = (\tilde{h}_{ij}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ for unconstrained regression, we define

$$\boldsymbol{H} = (h_{ij}) = \tilde{\boldsymbol{H}} - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{A}^T\left[\boldsymbol{A}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{A}^T\right]^{-1}\boldsymbol{A}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T.$$

We note that $\boldsymbol{H}$ is symmetric and idempotent. More details can be found in Chapter 4 of Seber (1977).

2.2. Addition of a new variable

Let $\boldsymbol{z}$ be an $n \times 1$ vector of measurements on a new explanatory variable. Then the effect of adding the new variable $\boldsymbol{z}$ to the original model (2.1) can be measured by considering the following regression model including $\boldsymbol{z}$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{z} + \boldsymbol{\varepsilon}. \tag{2.2}$$

Let

$$\boldsymbol{A}_* = [\boldsymbol{A}, 0] \quad \text{and} \quad \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\beta} \\ \gamma \end{bmatrix}.$$

Since $\boldsymbol{A}$ is assumed to be of full column rank, it is easy to show that the rows of $\boldsymbol{A}_*$ are linearly independent, so that the rank of $\boldsymbol{A}_*$ is also q. The linear constraints $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$ for the model (2.1) should be incorporated into the model (2.2) and thus the following matrix form

$$\boldsymbol{A}_*\boldsymbol{\eta} = \boldsymbol{c}$$

becomes new linear constraints on $\boldsymbol{\eta}$ for the model (2.2).

Putting $\boldsymbol{X}_* = [\boldsymbol{X}, \boldsymbol{z}]$, the least squares estimator of $\boldsymbol{\eta}$ for the model (2.2) with the linear constraints $\boldsymbol{A}_*\boldsymbol{\eta} = \boldsymbol{c}$ is denoted by

$$\hat{\boldsymbol{\eta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\gamma} \end{bmatrix},$$

and it is given by

$$\hat{\boldsymbol{\eta}} = \tilde{\boldsymbol{\beta}}_* - \left(\boldsymbol{X}_*^T\boldsymbol{X}_*\right)^{-1}\boldsymbol{A}_*^T\left[\boldsymbol{A}_*\left(\boldsymbol{X}_*^T\boldsymbol{X}_*\right)^{-1}\boldsymbol{A}_*^T\right]^{-1}\left(\boldsymbol{A}_*\tilde{\boldsymbol{\beta}}_* - \boldsymbol{c}\right),$$

where $\tilde{\boldsymbol{\beta}}_* = (\boldsymbol{X}_*^T\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\boldsymbol{y}$. In order to get $\hat{\boldsymbol{\eta}}$, we need some notations as follows

$$\begin{aligned}
\boldsymbol{u}_1 &= \left[\boldsymbol{A}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{A}^T\right]^{-1}\boldsymbol{A}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{z} = \left[\boldsymbol{A}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{A}^T\right]^{-1}\boldsymbol{A}\boldsymbol{u}_2 \\
\boldsymbol{u}_2 &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{z} \\
s_1 &= \boldsymbol{z}^T\left(\boldsymbol{I} - \tilde{\boldsymbol{H}}\right)\boldsymbol{z} \\
s_2 &= \boldsymbol{u}_1^T\boldsymbol{A}\boldsymbol{u}_2.
\end{aligned}$$

By computing 2×2 partitioned matrix $(X_*^T X_*)^{-1}$, we get

$$\tilde{\beta}_* = \begin{bmatrix} \tilde{\beta} - \dfrac{u_2(z^T\tilde{e})}{s_1} \\ \dfrac{z^T\tilde{e}}{s_1} \end{bmatrix}$$

which enables us to obtain

$$\left[A_*\left(X_*^T X_*\right)^{-1} A_*^T\right]^{-1} = \left[A\left(X^T X\right)^{-1} A^T\right]^{-1} - \frac{u_1 u_1^T}{s_1 + s_2}.$$

Thus a little more complicated computation shows that an appropriate partition of the resulting $\hat{\eta}$ gives

$$\hat{\beta} = \breve{\beta} + \frac{\left(X^T X\right)^{-1} A^T u_1 u_1^T \left(A\tilde{\beta} - c\right)}{z^T(I - H)z} - \frac{u_2 u_1^T \left(A\tilde{\beta} - c\right)}{z^T (I - H)\, z} + \frac{\left(X^T X\right)^{-1} A^T u_1 z^T \tilde{e}}{z^T(I - H)z} - \frac{u_2 z^T \tilde{e}}{z^T (I - H)\, z}$$

$$\hat{\gamma} = \frac{z^T e}{z^T(I - H)z}$$

since $s_1 + s_2 = z^T(I - H)z$. Hence the residual vector from the model (2.2) with linear constraints $A_*\eta = c$ can be obtained as

$$\begin{aligned} e_* &= y - X_*\hat{\eta} \\ &= e - (I - H)z\left[\frac{u_1^T(A\tilde{\beta} - c)}{z^T(I - H)z} + \frac{z^T\tilde{e}}{z^T(I - H)z}\right] \\ &= e - \hat{\gamma}(I - H)z. \end{aligned} \tag{2.3}$$

2.3. Added variable plot

Premultiplying both sides of (2.2) by $I - H$, we have

$$(I - H)y = (I - H)X\beta + \gamma(I - H)z + (I - H)\varepsilon$$

which becomes

$$e = \gamma(I - H)z + (I - H)\varepsilon \tag{2.4}$$

since $A\beta = c$ and

$$(I - H)y = e + X\left(X^T X\right)^{-1} A^T \left[A\left(X^T X\right)^{-1} A^T\right]^{-1} c$$

$$(I - H)X = X\left(X^T X\right)^{-1} A^T \left[A\left(X^T X\right)^{-1} A^T\right]^{-1} A.$$

Taking expectations of both sides of (2.4) yields

$$E(e) = \gamma(I - H)z$$

which suggests a plot of

$$e = e_* + \hat{\gamma}(I - H)z \quad \text{versus} \quad (I - H)z. \tag{2.5}$$

A justification for the use of the added variable plot in (2.5) can be made as follows. The regression of $\boldsymbol{e}$ on $(\boldsymbol{I}-\boldsymbol{H})z$ without the intercept can be written as

$$\boldsymbol{e} = \zeta(\boldsymbol{I}-\boldsymbol{H})z + \text{error}. \tag{2.6}$$

The least squares estimator of the slope ζ for the model (2.6) is easily computed as

$$\hat{\zeta} = \frac{z^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{e}}{z^T(\boldsymbol{I}-\boldsymbol{H})z}.$$

Since $\boldsymbol{H}\boldsymbol{e} = 0$, we have $z^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{e} = z^T\boldsymbol{e}$. Hence $\hat{\zeta}$ is equivalent to $\hat{\gamma}$. Furthermore, the residuals from the model (2.6) are equivalent to those from the model (2.2) by (2.3). Hence, as in the unconstrained regression models, the added variable plot in (2.5) should indicate a linear relationship through the origin with the slope $\hat{\gamma}$. When the variable z is added linearly to the original model (2.1) as in the model (2.2), a visual inspection of the scatter of points in the added variable plot (2.5) enables us to get an overall impression of the strength of relationship between response variable and added variable. The added variable plot may show points separated from the main body of data that are highly influential in estimating the regression coefficient γ.

2.4. Partial residual plot

Replacing $\boldsymbol{H}$ by $\boldsymbol{0}$ in (2.5) yields a plot

$$\boldsymbol{e}_* + \hat{\gamma}z \quad \text{versus} \quad z \tag{2.7}$$

which is a companion with a partial residual plot suggested in unconstrained regression models.

The use of the partial residual plot in (2.7) can be justified as follows. For the regression of $\boldsymbol{e}_* + \hat{\gamma}z$ on z without the intercept

$$\boldsymbol{e}_* + \hat{\gamma}z = \psi z + \text{error} \tag{2.8}$$

the least squares estimator of the slope ψ is computed as

$$\hat{\psi} = \frac{z^T(\boldsymbol{e}_* + \hat{\gamma}z)}{z^Tz}.$$

Since $\boldsymbol{e}_* = \boldsymbol{e} - \hat{\gamma}(\boldsymbol{I}-\boldsymbol{H})z$, we have

$$z^T(\boldsymbol{e}_* + \hat{\gamma}z) = \hat{\gamma}z^Tz.$$

Hence $\hat{\psi}$ is equivalent to $\hat{\gamma}$ and therefore the residuals from the model (2.8) are equivalent to those from the model (2.2).

3. A Numerical Example

We will fit a constrained regression to the body fat data set (Neter *et al.*, 1996, p.261), which contains twenty measurements on three explanatory variables X_1, X_2, X_3 and a response variable y.

First, for the following regression model of y regressed on X_1 and X_2,

$$y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \text{error}, \tag{3.1}$$

we consider a linear constraint

$$3\beta_1 - \beta_2 = 0. \tag{3.2}$$

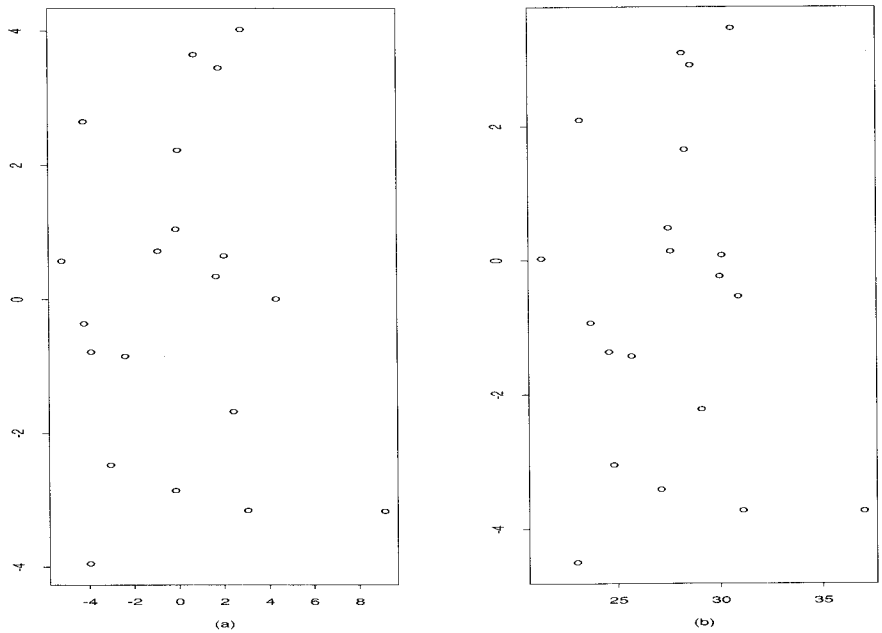

Figure 1: *Added variable plot (a) and partial residual plot (b)*

Using the usual F-test for linear hypotheses (see Chapter 4 of Seber, 1977), we can check whether the linear constraint (3.2) holds or not for the body fat data set. The value of the F-test statistic is 4.2×10^{-5} and the corresponding p-value is 0.995. Thus we may assume the constrained regression model (3.1) with linear constraint (3.2). For this constrained regression, the least squares estimates of the regression coefficients are

$$\hat{\beta}_0 = -19.217, \qquad \hat{\beta}_1 = 0.220, \qquad \hat{\beta}_2 = 0.661.$$

Next, the effect of adding the third explanatory variable X_3 to the constrained regression (3.1) with linear constraint (3.2) is investigated. When the following regression model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \text{error} \tag{3.3}$$

with linear constraint (3.2) is fitted to the body fat data set, the least squares estimates of the regression coefficients are

$$\hat{\beta}_0 = -18.766, \qquad \hat{\beta}_1 = 0.221, \qquad \hat{\beta}_2 = 0.663, \qquad \hat{\beta}_3 = -0.021.$$

Hence the regression coefficients for X_1 and X_2 hardly change when the third explanatory variable X_3 is added to the constrained regression model (3.1) with linear constraint (3.2). For the regression model (3.3) with linear constraint (3.2), the value of the general linear test statistic (Neter *et al.*, 1996, p.279) for testing $H_0 : \beta_3 = 0$ is 0.02, and its associated p-value is 0.90. Thus the regression coefficient for X_3 is insignificant at reasonable significance levels. This investigation is in parallel with two diagnostic plots provided in Figure 1. The added variable plot in Figure 1(a) does not show

any pattern and therefore the third explanatory variable X_3 hardly contributes to the formation of constrained regression equation. The partial residual plot in Figure 1(b) is almost similar to the added variable plot

References

Chipman, J. S. and Rao, M. M. (1964). The treatment of linear restrictions in regression analysis, *Econometrica*, **32**, 198–209.

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

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

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

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

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Regression Models*, 3rd ed., Irwin.

Seber, G. A. F. (1977). *Linear Regression Analysis*, John Wiley & Sons, New York.

Wood, F. S. (1973), The use of individual effects and residuals in fitting equations to data, *Technometrics*, **15**, 677–695.

Received January 2009; Accepted March 2009