iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4199

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Effect of Additional Variables on Regression Diagnostics

Disha K. Rank[1]
[1]*Department of Statistics, Saurashtra University, Rajkot*

Abstract: ***We obtain the effect of adding concomitant variables in a linear regression model on estimators of parametric functions of interest and on some commonly used regression diagnostics.***
Keywords: Regression model, Regression diagnostics, linear function, DFFIT, DFBETA.

I. INTRODUCTION

Regression diagnostics and detection of influential observations play an important role in statistical modelling (see for example Belsley, Kuh, Welsch (1980), Chatterjee and Haldi (1988)). In this paper, we consider the usual linear regression model $(Y, X\beta, \sigma^2 I)$ where $X'X$ could be singular and obtain the effect of adding more independent variables to this model on some of the most commonly used regression diagnostics DFBETA and DFFIT. This is important in usual regression set up as one might like to look into this aspect also while doing a step-wise regression. Also the stability of the regression diagnostics as the regressors increase would help in identifying influential observations in a more meaningful manner. The results of this paper can be interpreted in the design set up also. Consider the original model as the ANOVA model and the latter model as the analysis of covariance model. We are usually intersted in inferences on linear functions of a subset of β parameters (for example treatment contrasts). The present study enables us to assess the influence of an observation on the inferences concerning the desired parametric functions in the light of concomitant variables. A detailed study in this regard is underway and will be reported elsewhere.The following notations will be used in the paper. For a matrix A : ρ(A) , C (A), R(A) and A' denote the rank, column space, row space, and transpose, respectively. A denotes a g-inverse of A. For a matrix B, b_{i*} and b_{*j} denote the i^{th} row and the j^{th} column of B, respectively, and $B_{(i)}$ denotes the matrix obtained from B by dropping the ith row. P_B denotes the orthogonal projector projecting into C *(B).* (The inner product considered is the Euclidean inner product.)

II. MAIN RESULTS

We consider four models

(*a)* $(Y, x\beta, \sigma^2 I)$ *(2.1)*
b) $(Y_{(i)}, X_{(i)}\beta, \sigma^2 I)$ (2.2)
c) $(Y, X\beta + C\gamma, \sigma^2 I)$ *(2.3)*
d) $(Y_{(i)}, X_{(i)}\beta + C_{(i)}\gamma, \sigma^2 I)$. (2.4)

where *X* and *C* are fixed matrices of orders *n* x *m* and n x *k*, respectively. In all the above models *X' X* could be singular. Let $\hat{\beta}^{(i)}$, j = 1, 2 and $\begin{pmatrix}\hat{\beta}^{(i)} \\ \hat{\beta}^{(i)}\end{pmatrix}$, j = 3, 4. satisfy the normal equations and $R_0^{2(j)}$, j = 1, . , 4, denote the residual sum of squares for the models (2.1), (2.2), (2.3) and (2.4) respectively.

We start with the definitions of a few regression diagnostics for the model (2.1). $(\hat{\beta}^{(1)} - \hat{\beta}^{(2)})$ is called $DFBETA_i$. Let λ'β be estimable under models (2.1) and (2.2) then $DFBETA_{i,\lambda}$ is the quantity $\lambda'(\hat{\beta}^{(1)} - \hat{\beta}^{(2)})$, namely the difference in BLUEs of λ'β under the models (2.1) and (2.2). $DFBETA_{i}, x_j'$ is called $DFFIT_{ij}$. The matrix $H = D(X\beta^{(1)})/\sigma^2 = X(X'X)^{-}X' = P_x$ is called the Hat matrix and its diagonal elements are called the hat matrix diagonals (or leverage values).

A. Lemma 2.1. Let A be a symmetric matrix and A^{-} be a g-inverse of A. Let $u \epsilon C(A)$. Then for any nonzero real number k,

$$(A + kuu') = \begin{cases} A^- & if\ u'A^-u = \frac{1}{k} \\ A^- - \frac{A^-uu'A^-}{\left(\frac{1}{k}\right) + u'A^-u} & otherwise. \end{cases}$$

This lemma is well known. (See, for example, Rao and Mitra (1971) p. 40).

B. Lemma 2.2. (Rohde (1965), Bhimasankaram (1971)). Let $\sum$ be a nonnegative definite (nnd) matrix. Partition $\sum$ as

$$\sum = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$$

Where $\sum_{11}$ and $\sum_{22}$ are square matrices. let $\sum_{11}^-$ be a g - inverse of $\sum_{11}$. Denote $F = \sum_{22} - \sum_{21}\sum_{11}^-\sum_{12}$. Then one choice of a g-inverse of $\sum$ is

$$\begin{pmatrix} \Sigma_{11}^- & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \Sigma_{11}^- \\ -I \end{pmatrix} \Sigma_{12} F^- (\Sigma_{21}\Sigma_{11}^- \quad -I)$$

where F^- is any g-inverse of F.

C. Lemma 2.3. Consider the model (2.1). Let $x_{i*} \notin R(X(i))$. Then the following hold :

1) The error space is the same for the models (2.1) and (2.2).

2) For *all* $p \in C(X'_{(i)})$ the BLUES of $p'\beta$ *are* the same in the models (2.1) and (2.2) and $x_i'\hat{\beta} = Yi$

3) The usual estimator of σ^2 is the same under the models (2.1) and (2.2).

Proof : (a) l' Y belongs to the error space $\Rightarrow$ ' $l^1X = 0$

$l^1{}_{(i)} X_{(i)} + l_i x_{i*} = 0 \Rightarrow l_i = 0$ since, $x_i \notin ((X'(i))$. $\Rightarrow l'Y = l'_{(i)}\ Y_{(i)})$. Thus error space of the model (2.1) is contained in the error space of the model (2.2). Thus other inclusion is trivial. Proofs of (b) and (c) are easy in view of (a) and are omitted. (These are also available in Bhimasankaram and Jammalamadaka (1990), henceforth denoted as B & J 1990)).

D. Lemma 2.4 (B&J (1990)). Consider the model (2.1). Let $x_{i*} \in R(X_{(i)})$ and let

$e_i = y_i - x_{i*}\hat{\beta}^{(1)}$. Then the following hold :

(a) $\hat{\beta}^{(2)} = \hat{\beta}^{(1)} - \frac{(X'X)^-x_{i*}'\ e_i}{1-h_i}$

(b) $(R_0^2)^2 = \left(R_0^{(1)}\right)^2 - \frac{e_i^2}{1-h_i}$

where $h_i = x_i *(X'X)^-x'_{i*}$.

E. Lemma 2.5. $P_{(x:c)} = P_x + P_{(I-PX)}C$.

The proof when (*X : C)* is of full column rank is given in Chatterjee and Hadi (1988). The proof when (*X : C)* is not of full rank follows from lemma (2.2) along the same lines.

F. Lemma 2.6. Consider the models (2.1) and (2.3). Let θ_r denote a solution of the normal equations for the Model $(c_{*p}, X\theta, \sigma^2 I)$. Let $s_{pq} = (c_{*p} - X\hat{\theta}_p)'\ (c_{*q} - X\hat{\theta}_q)$, p, q = 1, . . . , k and $s_{po} = (c_{*p} - X\hat{\theta}_p)'\ (Y - X\hat{\beta}^{(1)})$. Write $T = (\hat{\theta}1 \ldots\ldots\ldots \hat{\theta}k), S_{k\ x\ k} =$ ***((Spq))*** and $S_o = (S_{10}, \ldots\ldots\ldots, S_{k}o)'$. *Then* $(\hat{\beta}^{(3)}, \hat{\gamma}^{(3)'})'$ *can be obtained as follows :* $\hat{\gamma}^{(3)}$ *is a solution of* $s_\gamma = s_0$ *and* $\hat{\beta}^{(3)} = \hat{\beta}^{(1)} - T\hat{\gamma}^{(3)}$. $\left(R_0^{(3)}\right)^2 = \left(R_0^{(1)}\right)^2 - S_0'\hat{\gamma}^{(3)}$.

Proof. If *S* is singular then the system of equations $s_\gamma = S_0$ will be consistent because $S = [C'\ (1 - P_x)]\ [C'\ (I - Px)]'$ and $S_0 = C'\ (I - Px)Y$ and the normal equations for the model (2.3) are

$X'X\beta + X'C\gamma = X'Y$ *(1)*

$C'X\beta + C'C\gamma = C'Y$ (2)

and the reduced normal equations for γ is C' *(1— P_x)* Cy = C' (1— P_x) *Y*. Therefore β and y will be estimated along the same lines of Rao's method.

In view of lemma 2.3, it is clear that the case $x_{i*} \notin R(X_{(i)})$ is of no practical interest in connection with the regression diagnostics. We shall consider the case $x_{i*} \in R(X_{(i)})$ and prove

G. Theorem 2.7. Consider the models (2.1) and (2.3). Let $x_{i*} \in R(X_{(i)})$. Then the following hold :

Let λ'β be estimable under both the models; then the change in $DFBETA_{i,\lambda}$

(i.e. $\lambda'((\hat{\beta}^{(3)} - \hat{\beta}^{(4)}) - (\hat{\beta}^{(1)} - \hat{\beta}^{(2)}))$ is given by $X'(W\hat{\gamma}^{(4)} - T\hat{\gamma}^{(3)})$,

the change in the residual sum of squares is given by $V_o'\hat{\gamma}^{(4)} - s_0'\hat{\gamma}^{(3)}$ where W and V_o correspond to *T* and S_o respectively when the models (2.1) and (2.3) are replaced by (2.2) and (2.4), respectively, in lemma 2.6.

The proof follows by applying lemma 2.6 twice for models (2.1) and (2.3) and then for models (2.2) and (2.4).

H. Theorem 2.8. Consider the models (2.1) and (2.3). Let $x_{i*} \in R(X_{(i)})$. Then the following hold :

(a) $x_{i*}\beta$ is estimable under all the models (2.1), (2.2), (2.3) and (2.4). Then the change in

DFFI Tij i.e. $(\hat{y}_j^{(3)} - \hat{y}_j^{(4)}) - (\hat{y}_j^{(1)} - \hat{y}_j^{(2)})$ is given by

$$x_{j*}(\hat{W}\gamma^{(4)} - \hat{T}\gamma^{(3)} + \frac{c_{*j}' s^{-} \eta (e_i - \eta' \hat{\gamma}^{(3)})}{1 - h_i - \eta' s^{-} \eta}$$

Where η' = *($x_{i*}T$ - c_{i*})* and W is the same as above.

The proof follows by straight forward computations using lemmas 2.6 and 2.4. (b)The effect on hat matrix for *(X : C)* $H_{(x : c)}$ = *(X : C) [(X : C)' (X : C)]⁻(X:C)'*

Now

$$[(X : C)'(X : C)]^{-} = \begin{pmatrix} (X'X)^{-} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} (X'X)^{-}X'C \\ -I \end{pmatrix} F^{-}(C'X(X'X)^{-} \quad -I)$$

where *F = C'(I — Px) C.*

Hence H(x:c) = Hx + *(I — Hx)C[C' (I — Hx)C]⁻ C' (I — Hx).* For single additional variable discussion is given in Chatterjee and Haldi (1988).

The variance of an estimable parametric function $\lambda'\hat{\beta}^{(3)}$, assuming *X* and C tobe non-stochastic, is $\sigma^2\lambda'[(X^IX)^{-} + T\bar{s}T']\lambda$, which will be needed to compute *DFBETAS and DFFITS etc*.

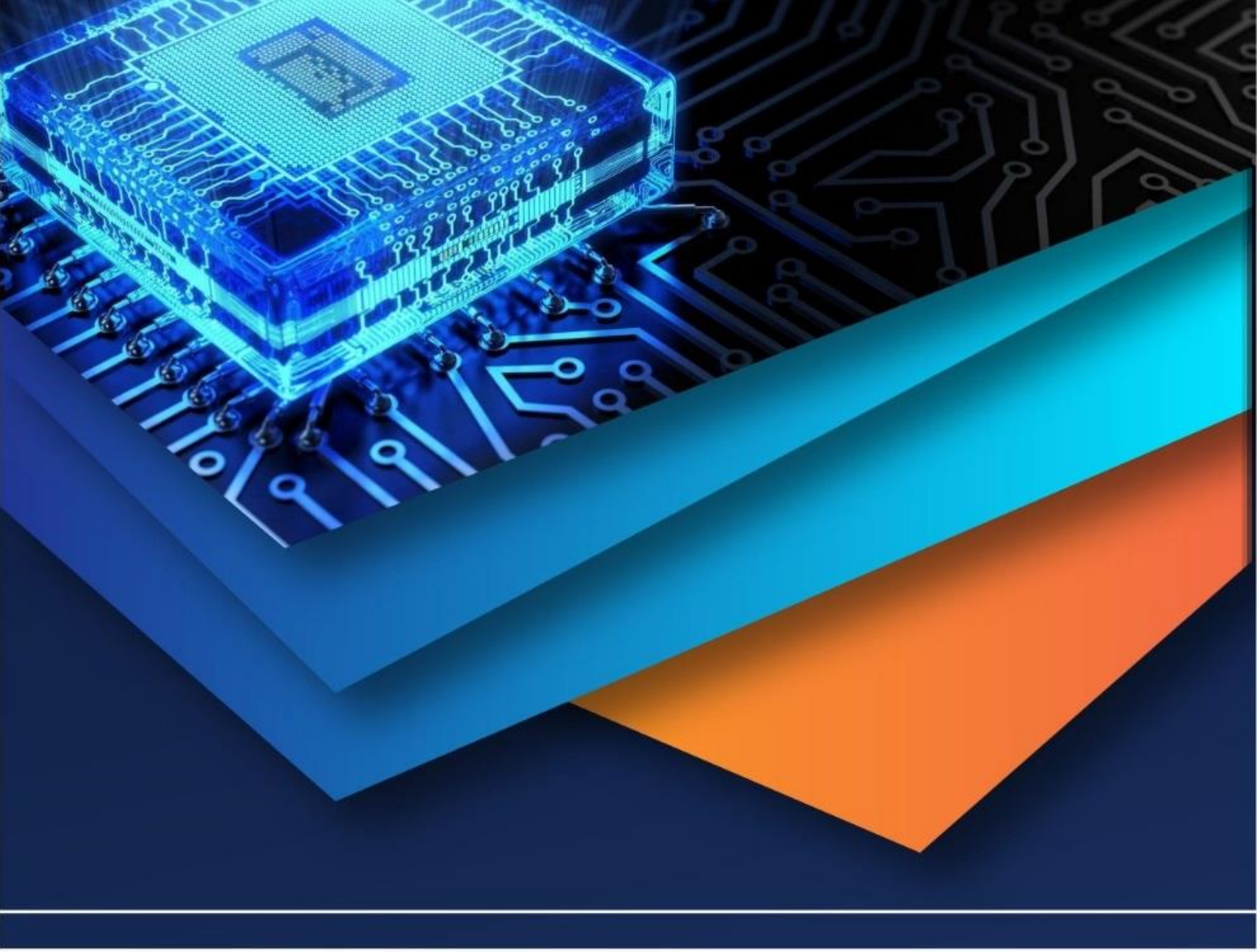
Now let us consider one additional variable *c* case. Then the change in *D F B E T Ai* using Rao's method of estimating β and γ is

$$\frac{\delta i\ [ei\ ((1 — hi)\hat{\theta}^{(3)} — \delta i\ V)\ +\ \hat{\gamma}^{(3)}V]}{(1 - h_i)(\delta'\delta - \delta_i^2)}$$

Where V= *(X'* X)⁻ x'$_{i*}$ and δ = *c —$X\hat{\theta}$*, and the other regression diagnostics change correspondingly.

REFERENCES

[1] Belsley, D.A., Kuh, E. and Welsch, R.E., (1980), "Regression Diagnostics", Wiley, New Y
[2] P. Bhimasankaram, D. Sengupta and S. Ramanathan (1995). Recursive inference in the general linear model, Sankhya Series A, 57(2), 227-25
[3] Chatterjee, S. and Hadi, A.S. (1988), "Sensitivity Analysis in Linear Regression", Wiley, New York.
[4] Rao, C.R. (1973), "Linear Statistical Inference and its Applications", John Wiley, New York.
[5] Rao, C.R. & Mitra, S. (1971) : Generalized inverse of matrices and its applications, John Wiley and Sons.
[6] Rohde, C.A. (1965) : Generalized inverses of partitioned matrices. J. SIAM., 13, 1033-1035.
[7] Bhimasankaram, P. (1971) : On generalized inverses of partitioned matrices. Sankhya, Series , 33, 311-314.
[8] D. Sengupta and S.R. Jammalamadaka (2003). Linear Models: An Integrated Approach, World Scientific Publishing Company, River Edge, NJ,



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY