Scientia Africana, Vol. 11 (No.2), December 2012. Pp 59- 64
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

AN OBSERVATION ON THE VARIANCE OF A PREDICTED RESPONSE IN REGRESSION ANALYSIS

A. OKOLO
Department of Statistics and Operations Research
Modibbo Adama University of Technology
Yola, Nigeria

Received: 19-09-12
Accepted:19-11-12

ABSTRACT

In studying individual parameters and the predicted response in regression analysis, three important properties are usually distinguished. These are bias, variance and mean-square error. The choice of a predicted response has to be made on a balance of these properties and computational simplicity. To avoid over fitting, along with the obvious advantage of having a simpler equation, it is shown that the addition of a variable to a regression equation does not reduce the variance of a predicted response.

Key words: Linear regression; Partitioned matrix; Predicted response

INTRODUCTION

There is now a variety of estimation methods which are applicable to a single equation in a model and those which deal with the complete model in regression analysis; Schall (1991) and Fellner (1986). When planning to use a linear regression equation to predict a response, we are faced with the problem of selecting an adequate set of independent variables to include in the equation (Cantoni et al, 2005; Healy, 1990 and Bring, 1984). A reasonable objective is the selection of a set which minimizes the variance of a predicted response plus possible biases in the estimates of the parameters of the regression equation.

Draper and Smith (1998), Miller (1990) and Williams (1959) have given deserved attention to the problem of estimator bias in their texts. Gunst and Mason (1976) and Trenkler (1980) have compared several regression estimators with respect to the generalized mean squared error criterion. Peixoto (1990) has provided additional motivation for the use of variable selection algorithms that restrict search to well-formulated models. Sakallioglue *et al* (2001) have compared biased estimator constructed as alternatives to the least squares estimators when multicollinearity is present. However, it is not mentioned explicitly that the addition of a variable to a regression equation can never reduce (and in fact usually increases) the variance of a predicted response. It seems that this is worth mentioning for the sake of understanding. A convenient way to present the idea follows.

MATERIALS AND METHODS

Regression Estimates

We define y to be an (n x 1) vector of observation, x to be an (n x p) matrix of independent variables, β to be a (p x 1) vector of parameters to be estimated and ϵ to be an (n x 1) vector of errors.

Then the model under consideration can be written in the form

$$y = x\beta + \varepsilon \quad (1)$$

where $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2 I$, so that the elements of ε are uncorrelated. Consider these prediction equations:

$$\hat{y}_1 = \hat{\beta}_1 x_{1i} \quad (2)$$

and

$$\hat{y}_2 = \tilde{\beta}_1 x_{1i} + \tilde{\beta}_2 x_{2i} \; ; \quad (3)$$

where the least squares parameter estimates in these two cases are given by:

$$\hat{\beta}_1 = (x_1' x_1)^{-1} x_1' y \quad (4)$$

and

$$\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} x_1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2 \end{pmatrix}^{-1} \begin{pmatrix} x_1'y \\ x_2'y \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \begin{pmatrix} x_1'y \\ x_2'y \end{pmatrix} = \begin{pmatrix} C_{11}x_1'y + C_{12}x_2'y \\ C_{21}x_1'y + C_{22}x_2'y \end{pmatrix} \quad (5)$$

Here C_{11}, C_{12}, C_{21} and C_{22} are sub-matrices which, according to Searle (1982), are given by

$$C_{11} = (x_1'x_1)^{-1} + (x_1'x_1)^{-1} x_1'x_2 C_{22} x_2'x_1 (x_1'x_1)^{-1} \quad (5a)$$

$$C_{12} = -(x_1'x_1)^{-1} x_1'x_2 C_{22} \quad (5b)$$

$$C_{22} = (x_2'x_2)^{-1} + (x_2'x_2)^{-1} x_2'x_1 C_{11} x_1'x_2 (x_2'x_2)^{-1} \quad (5c)$$

$$C_{21} = -(x_2'x_2)^{-1} x_2'x_1 C_{11} \quad (5d)$$

Let the vector $(U_1' U_2')$ represent a point in the space of the independent variables, the two estimates of the response for equations (2) and (3) are given by:

$$\hat{y}_1 = U_1' \hat{\beta}_1 \quad (6)$$

and

$$\hat{y}_2 = U_1' \tilde{\beta}_1 + U_2' \tilde{\beta}_2 \quad (7)$$

Using the identities (5a) - (5d) for the inverse of a partitioned matrix we find:

$$\begin{aligned} \text{Cov}(y_1, y_2) &= \text{Cov}\left(U_1'\hat{\beta}_1, U_1'\tilde{\beta}_1 + U_2'\tilde{\beta}_2\right) \\ &= U_1' \text{ Cov}\left(\hat{\beta}_1, \tilde{\beta}_1\right) U_1 + U_1' \text{ Cov}\left(\hat{\beta}_1, \tilde{\beta}_2\right) U_1 \\ &= U_1' \left[(x_1'x_1)^{-1} x_1' Var(y) x_1 C_{11} + (x_1'x_1)^{-1} x_1' Var(y) x_2 C_{21}\right] U_1 \\ &\quad + U_1' \left[(x_1'x_1)^{-1} x_1' Var(y) x_1 C_{12} + (x_1'x_1)^{-1} x_1' Var(y) x_2 C_{22}\right] U_2 \\ &= \sigma^2 U_1' \left[(x_1'x_1)^{-1} x_1'x_1 C_{11} + (x_1'x_1)^{-1} x_1'x_2 C_{21}\right] U_1 \\ &\quad + \sigma^2 U_1' \left[(x_1'x_1)^{-1} x_1'x_1 C_{12} + (x_1'x_1)^{-1} x_1'x_2 C_{22}\right] U_2 \\ &= \sigma^2 U_1' \left[C_{11} - C_{12}C_{22}^{1} C_{21}\right] U_1 + \sigma^2 U_1' \left[C_{12} - C_{12}\right] U_1 \\ &= \sigma^2 U_1' (x_1'x_1)^{-1} U_1 \\ &= \text{Var}(y_1). \end{aligned} \quad (8)$$

Therefore,

$$E\left[(y_1 - E(y_1)) - (y_2 - E(y_2))\right]^2 = \text{Var}(y_1) + \text{Var}(y_2) - 2\text{Cov}(y_1, y_2)$$

Scientia Africana, Vol. 11 (No.2), December 2012. Pp 59- 64
 ISSN 1118 – 1931

= Var (y_2) – Var (y_1),

and since this expression is non-negative, we have:

$$\text{Var}\,(y_2) \geq \text{Var}\,(y_1). \qquad (9)$$

According to Searle (1982) equality holds in the above expression only if

$$\left[U_1'C_{12}C_{22}^{-1} + U_2'\right]C_{22}\left[U_1'C_{12}C_{22}^{-1} + U_2'\right]' = 0.$$

Since C_{22} is positive definite (Graybill, 1971), this implies only if

$$U_1'C_{12}C_{22}^{-1} + U_2' = 0,$$

or equivalently, according to (5b), only if

$$U_1'(x_1'x_1)^{-1}x_1'x_2 = U_2' \qquad (10)$$

will equality hold. When $x_1'x_2 \neq 0$, equality holds only if the elements of U_2 are particular linear combinations of the elements of U_1, and when $x_1'x_2 = 0$, the variances of the two estimates of the response are equal only if U_2 = 0.

These results apply to estimated regression coefficients as well as to predicted responses since the variance of a given coefficient corresponds to a particular choice of the vector (U_1', U_2').

An intuitive argument can be given by remembering that an estimator of the response having zero variance would be provided by selecting an arbitrary constant and agreeing to always predict the response to be this value. Although it could be quite inaccurate, no other estimator could provide better precision. On the other hand, estimating one or more regression coefficients would introduce variability and provide a less precise, but hopefully, more accurate estimator.

RESULTS AND DISCUSSION

Below are two simple examples to illustrate most of the points made.

Numerical Example 1

Consider the hypothetical data in Table 1.

Table 1: Hypothetical Data

X_i	Y_i
1	5
2	7
3	7
4	10
5	16
6	20

Let $\hat{y}_L$ and $\hat{y}_Q$ be the prediction equations developed from the hypothetical data, where

$$\hat{y}_{Li} = \hat{\beta}_1 x_i \ ,$$

and

$$\hat{y}_{Qi} = \tilde{\beta}_1 x_i + \tilde{\beta}_2 x_i^2$$

For this data we obtain the least squares fits:

$$\hat{y}_{Li} = (91)^{-1} (280\, x_i),$$

and

$$\hat{y}_{Qi} = (12{,}544)^{-1} (30{,}184\, x_i + 1{,}736\, x_i^2).$$

The standardized variances of the two predicted responses and their ratio at each observed x-value are given in Table 2.

Table 2: Variance of Predicted Responses for Hypothetical Data

X_i	σ^{-2} Var $\hat{y}_{Li}$	σ^{-2} Var $\hat{y}_{Qi}$	Var $\hat{y}_{Qi}$ / Var $\hat{y}_{Li}$
1	0.0110	0.1183	10.75
2	0.0440	0.2790	6.34
3	0.0989	0.3214	3.25
4	0.1758	0.2589	1.47
5	0.2747	0.2790	1.02
6	0.3956	0.7433	1.88

Numerical Example 2

The data in Table 3 were taken from Draper and Smith (1998). It contains the results of the experimental study to investigate the relation between the number of self-service coffee dispensers (X) in a cafeteria line and sales of coffee (Y) measured in hundreds of gallons of coffee.

Table 3: Data for Cafeteria Coffee Sales

Cafeteria I	Number of Dispensers X_i	Coffee Sales Y_i
1	0	508.1
2	0	498.4
3	1	568.2
4	1	577.3
5	2	651.7
6	2	657.0
7	4	755.3
8	4	758.9
9	5	787.6
10	5	792.1
11	6	841.4
12	6	831.8
13	7	854.7

Scientia Africana, Vol. 11 (No.2), December 2012. Pp 59- 64
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

14	7	871.4

Consider the prediction equations

$$\hat{y}_{Li} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and

$$\hat{y}_{Qi} = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \tilde{\beta}_2 x_i^2 .$$

For this cafeteria coffee sales data, we obtain the least squares fits:

$$\hat{y}_{Li} = 527.326 + 51.427\, x_i$$

and

$$\hat{y}_{Qi} = 503.346 + 78.941\, x_i - 3.969\, x_i^2 .$$

The variances of the predicted responses and their ratio at each observed x-value are given in Table 4.

Table 4: Variances of Predicted Responses for Cafeteria Coffee Sales Example

X_i	σ^{-2} Var $\hat{y}_{Li}$	σ^{-2} Var $\hat{y}_{Qi}$	Var $\hat{y}_{Qi}$ / Var $\hat{y}_{Li}$
1	0.1507	0.1507	1.00
2	0.101	0.1535	1.52
3	0.0753	0.1942	2.58
4	0.0736	0.1896	2.58
5	0.0959	0.1420	1.48
6	0.1421	0.1426	1.00
7	0.2123	0.3647	1.72

In example 1, we observe from Table 2 that, over the set of points at which data was taken, use of the second degree term can increase the variance of the predicted response to more than ten times its value when the simpler equation is used. The increase in the variance of the predicted response, when the second degree term is used, is also observed from Table 4 of example 2 to be more than two times its value when the simpler equation is used. However, this is quite apart from any consideration of which, if either, of the two equations is correct. That is, estimates from both equations may be biased in amounts depending on the nature of the true model.

It follows that adding a variable to the equation can never improve the precision but only remove possible biases from the various estimates obtained from the regression analysis. Simultaneous reduction of both variance and bias may be achieved only by the substitution of a new variable for one already in the equation.

CONCLUSION

In many regression and scientific studies, there is an ambition to compare the relative importance of different variables. We are faced with situations where we fit one model (e.g., a straight line) but we fear that this model may be somewhat inadequate (e.g., there may in fact be a little quadratic curvature).

We can talk in terms of the fitted model and the true model but it is better to think in terms of the fitted model and the feared model alternative. Interest should be on what might be wrong with the model fitted if some specified alternative were true. It has been shown in this

paper that the addition of a variable to a regression equation increases the variance of a predicted response.

REFERENCES

Bring, J. (1994). How to Standardize Regression Coefficients. *The American Statistician*, **48**, 209 – 213.

Cantoni, E., Flamming, J. M. and Ronchetti, E. (2005). Variable Selection for Marginal Longitudinal Generalized Linear Models. *Biometrics*, Vol. **61**, 507 – 514.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd Edition. John Wiley and Sons, Inc., New York, 235pp.

Fellner, W. H. (1986). Robust Estimation of Variance Components. *Technometrics*, Vol. **28**, No.1, 51 - 60.

Grayhill, F. A. (1971). *Introduction to Matrices with Applications in Statistics*. Wadsworth Publishing Company, Belmont, CA, 163pp.

Gunst, R. F. and Mason, R. L (1976). Generalized Mean Squared Error Properties of Regression Estimators. *Communications in Statistics – Theory and Methods*, A5 (**15**), 1501 – 1508.

Healy, M. J. R. (1990). Measuring Importance. *Statistics in Medicine*, Vol. **9**, 633 – 637.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall, London, 110pp.

Peixoto, J. L. (1990). A Property of Well Formulated Polynomial Regression Models. *The American Statistician*, Vol. **44**, No. 1, 26 – 30.

Sakallioglu, S., Kaciranlar, S. and Akdeniz, F. (2001). Mean Squared Error Comparisons of Some Biased Regression Estimators. *Communications in Statistics – Theory and Methods*, Vol. **30**. No. 2, 347 – 361.

Schall, R. (1991). Estimation in Generalized Linear Models with Random Effects. *Biometrika*, Vol. **78**, No. 4, 719 – 727.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* John Wiley and Sons, Inc., New York, 257pp.

Trenkler, G. (1980). Generalized Mean Square-Error Comparisons of Biased Regression Estimators. *Communications in Statistics-Theory and Methods*, A9 (**12**), 1247 – 1259.

Williams, E J. (1959) *Regression Analysis*. John Wiley and Sons, Inc., New York, 70pp.