The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 473-478

# Test for an Outlier in Multivariate Regression with Linear Constraints

**Myung Geun Kim**[1)]

## Abstract

A test for a single outlier in multivariate regression with linear constraints on regression coefficients using a mean shift model is derived. It is shown that influential observations based on case-deletions in testing linear hypotheses are determined by two types of outliers that are mean shift outliers with or without linear constraints. An illustrative example is given.

*Keywords* : Linear constraints, mean shift, multivariate regression, outlier, test.

## 1. Introduction

Outliers are everywhere and they can distort relevant statistical analyses. Hence methods of detecting outliers are necessary for statistical models of our interest. Outlier detections in multivariate regression have been pursued by some authors, for example Barrett and Ling (1992), Kim (1995) and Srivastava and Rosen (1998). Tests of linear hypotheses about regression coefficients are a fundamental step in analyzing regression data. When there are linear relationships among regression coefficients, we need an appropriate method of detecting outliers for this situation. However, no method is available for multivariate regression with linear constraints.

In this work we will derive a test for a single outlier in multivariate regression with linear constraints using a mean shift model. In Section 2 some basic results in multivariate regression are reviewed. In Section 3 a mean shift model for detecting a single outlier with linear constraints is defined and a test for detecting an outlier is derived. In Section 4 we will show that the change in Wilks lambda statistic for testing linear hypotheses due to single case-deletions is determined by two types of statistics that are used for detecting mean shift outliers with or without linear constraints. An illustrative example is given in Section 5.

---

1) Professor, Department of Applied Statistics, Seowon University, Cheongju, 361-742.
E-mail: mgkim@seowon.ac.kr. This work was supported by grant No. 2001-1-10400-006-1 from the Basic Research Program of the Korea Science & Engineering Foundation.

## 2. Preliminaries

Consider the multivariate regression model

$$Y = XB + U,$$

where $Y = (y_1, \ldots, y_n)^T$ is an n by p matrix of response variables, $X = (x_1, \ldots, x_n)^T$ is an n by q matrix with rank q of independent variables, and $B$ is a q by p matrix of unknown parameters. We assume that the rows of $U = (u_1, \ldots, u_n)^T$ are independent and identically distributed as a p-variate normal distribution with mean $0$ and covariance matrix $\Sigma$. Further assume that we have linear constraints on $B$

$$AB = C, \tag{1}$$

where $A$ is a specified r by q matrix of rank r ( $r \leq q$ ) and $C$ is a specified r by p matrix. The least squares estimator of $B$ under the linear constraints (1) is given by

$$\hat{B}_0 = \hat{B}_F - (X^T X)^{-1} A^T [A(X^T X)^{-1} A^T]^{-1} (A\hat{B}_F - C), \tag{2}$$

where $\hat{B}_F = (X^T X)^{-1} X^T Y$. For more details, see Chap. 8 of Seber (1984).

## 3. Test For An Outlier

When there is a shift in the mean of the i-th observation $y_i$ among n observations, a mean shift model for detecting a single outlier can be expressed as

$$Y = XB + d_i \phi^T + U = X_* B_* + U, \tag{3}$$

where $\phi$ is a p by 1 vector of unknown shift parameters, $d_i$ is the i-th column of the identity matrix $I_n$ of size n, $X_* = [X \; d_i]$ and $B_* = \begin{bmatrix} B \\ \phi^T \end{bmatrix}$. Let $A_* = [A \; 0]$. For the mean shift model (3) the linear constraints (1) imposed on $B$ is converted into

$$A_* B_* = AB = C. \tag{4}$$

Hence the least squares estimator of $B_*$ for the mean shift model (3) under the linear constraints (4) has the same form as that of $\hat{B}_0$ in (2) with $X$ and $A$ replaced by $X_*$ and $A_*$, respectively. An appropriate partitioning of the least squares estimator of $B_*$ and a little complicated algebra provide the least squares estimator of $\phi$ as

$$\hat{\phi}_0 = \frac{e_{0,i}}{1-h_{0,ii}}, \tag{5}$$

where $e_{0,i} = y_i - \hat{B}_0{}^T x_i$, $h_{0,ii}$ is the i-th diagonal element of

$$H_0 = H - X(X^TX)^{-1}A^T[A(X^TX)^{-1}A^T]^{-1}A(X^TX)^{-1}X^T$$

and $H = X(X^TX)^{-1}X^T$.

A significant deviation of $\phi$ from zero vector implies that the i-th observation is an outlier, and an outlier may occur in $y_i$, $x_i$ or both under the mean shift model (3) with linear constraints (4). In order to check the outlyingness of the i-th observation, we need to perform a test of the following hypothesis

$$H: \phi = 0 \tag{6}$$

which can be done as in what follows. To this end, we first need the sampling distribution of $\hat{\phi}_0$ in (5) that is derived as follows. Under the linear constraints it is easily shown that $E(e_{0,i}) = \phi$. We can write $e_{0,i}$ as a linear combination of the $y_i$ plus a constant term and then easily get $cov(e_{0,i}) = (1-h_{0,ii})\Sigma$, since $H_0^2 = H_0$. Hence $\hat{\phi}_0$ is distributed as

$$\hat{\phi}_0 \sim N(\phi/(1-h_{0,ii}), \Sigma/(1-h_{0,ii})).$$

Next, let $S_0 = (Y - X\hat{B}_0)^T(Y - X\hat{B}_0)$. For an arbitrary matrix $W$, $W_{(i)}$ is interpreted as $W$ computed from the sample without the i-th observation. Then we have

$$S_{0(i)} = S_0 - \frac{e_{0,i}\, e_{0,i}^T}{1-h_{0,ii}} \tag{7}$$

(see Tang & Fung, 1997). The sampling distribution of $S_{0(i)}$ can be derived as follows. Let $E_0 = Y - X\hat{B}_0$. Then it is easily shown that $E_0 = (I_n - H_0)U$ under the linear constraints. Hence we have $e_{0,i} = U^T(I_n - H_0)d_i$. By using (7) we obtain $S_{0(i)} = U^TQU$, where $Q = I_n - H_0 - \frac{1}{1-h_{0,ii}}(I_n - H_0)d_i d_i^T(I_n - H_0)$. It can be shown that $Q$ is idempotent, and we have rank($Q$) = n-q+r-1. By Theorem 3.4.4(2) of Mardia et al. (1979), $S_{0(i)}$ is distributed as a Wishart distribution $W_p(\Sigma, n-q+r-1)$. Since $Q$ is idempotent and $Q(I_n - H_0)d_i = 0$, Corollary 3 of Seber (1984, p.25) shows that $S_{0(i)}$ and $e_{0,i}$ are independent. Hence we can obtain a test statistic for performing a test of the hypothesis (6) given by

$$T_i = \frac{n-p-q+r}{p} \frac{e_{0,i}^T S_0^{-1} e_{0,i}/(1-h_{0,ii})}{1- e_{0,i}^T S_0^{-1} e_{0,i}/(1-h_{0,ii})} \quad (8)$$

which is distributed under $\phi = 0$ as a F-distribution $F_{p,n-p-q+r}$ with degrees of freedom p and n-p-q+r. A significantly large value of $T_i$ indicates the outlyingness of the i-th observation.

The test statistic in (8) is used only when the i-th observation is known as an outlier. When we do not know which observation is an outlier, the test is usually based on the maximum of the $T_i$ over all $i$. However, it is not easy to derive the sampling distribution of $\max_{1\le i\le n} T_i$ in most cases and we often use the following Bonferroni upper bound

$$\Pr(\max_{1\le i\le n} T_i \ge t) \le \sum_{i=1}^{n} \Pr(T_i \ge t) = nPr(T_1 \ge t).$$

For a significance level $\alpha$, the test based on the Bonferroni upper bound rejects the hypothesis (6) if $\max_{1\le i\le n} T_i > F_{p,n-p-q+r}(1-\alpha/n)$, where $F_{p,n-p-q+r}(\gamma)$ is the $100\times\gamma$th percentile of the $F_{p,n-p-q+r}$ distribution and the Bonferroni upper bound for the p-value of this test is

$p-value \le n\Pr(F_{p,n-p-q+r}$ *random variable* $>$ *the observaed value of* $\max_{1\le i\le n} T_i)$.

## 4. Types of Influential Observations

A test of the linear hypothesis defined by (1) can be performed by using some test statistics in which Wilks lambda statistic is often adopted and it is given by $\Lambda = |S|/|S_0|$, where $S = (Y - X\hat{B}_F)^T(Y - X\hat{B}_F)$.

Observations that have a large influence in testing the linear hypothesis are usually called influential observations. Case-deletion method is one way to investigate the influence of observations in testing the linear hypothesis. Tang and Fung (1997) obtained the change in the value of Wilks lambda statistic due to removal of the i-th observation from which we can get the ratio of $\Lambda_{(i)}$ to $\Lambda$ as follows

$$\Lambda_{(i)}/\Lambda = \frac{1- e_i^T S^{-1} e_i/(1-h_{ii})}{1- e_{0,i}^T S_0^{-1} e_{0,i}/(1-h_{0,ii})}, \quad (9)$$

where $e_i = y_i - \hat{B}_F^T x_i$ and $h_{ii}$ is the i-th diagonal element of $H$. In view of (8), the quantity in the denominator of (9) determines the outlyingness of observations for the mean shift model with linear constraints because a significantly large value of $e_{0,i}^T S_0^{-1} e_{0,i}/(1-h_{0,ii})$ indicates that the i-th observation is a mean shift outlier. Similar roles of $e_i^T S^{-1} e_i/(1-h_{ii})$ of the numerator of (9) in the mean shift model

without any constraint can be seen by putting $A = C = 0$ in Section 3. Hence influential observations based on case-deletions are determined by two types of outliers that are mean shift outliers with or without linear constraints (1).

## 5. Example

For illustration we consider the adaptive score data (Cook and Weisberg, 1982, p.22). The adaptive score data includes 21 observations for children with one independent variable $x$ (the age of a child in months) and one response variable $y_1$ (Gesell adaptive score). In order to prepare bivariate data we drew 21 observations for the error term by generating random numbers from a normal distrbution with mean 0 and variance 1 and then computed the values of the second response variable $y_2$ according to the equation $y_2 = 1 - x + error$. The full data set is included in Table 1 for easy reference.

| No | $x$ | $y_1$ | $y_2$ | $T_i$ | No | $x$ | $y_1$ | $y_2$ | $T_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 15 | 95 | -13.5 | 0.41 | 11 | 7 | 113 | -5.2 | 2.09 |
| 2 | 26 | 71 | -26.1 | 2.55 | 12 | 9 | 96 | -8.1 | 0.14 |
| 3 | 10 | 83 | -8.7 | 1.13 | 13 | 10 | 83 | -8.7 | 1.16 |
| 4 | 9 | 91 | -8.8 | 1.90 | 14 | 11 | 84 | -10.0 | 1.01 |
| 5 | 15 | 102 | -14.5 | 0.40 | 15 | 11 | 102 | -9.9 | 0.16 |
| 6 | 20 | 87 | -18.6 | 0.24 | 16 | 10 | 100 | -8.9 | 0.02 |
| 7 | 18 | 93 | -17.3 | 0.06 | 17 | 12 | 105 | -11.8 | 0.71 |
| 8 | 11 | 100 | -8.9 | 1.68 | 18 | 42 | 57 | -40.5 | 0.61 |
| 9 | 8 | 104 | -7.6 | 0.37 | 19 | 17 | 121 | -17.7 | 8.91 |
| 10 | 20 | 94 | -19.3 | 0.18 | 20 | 11 | 86 | -8.7 | 1.79 |
| | | | | | 21 | 10 | 100 | -8.6 | 0.19 |

**Table 1.** Adaptive score data and the values of the $T_i$

For the bivariate regression model $y_1 = \beta_{01} + \beta_{11}x + error$ and $y_2 = \beta_{02} + \beta_{12}x + error$, we consider linear constraints

$$\beta_{01} + 100\beta_{11} = -2 \text{ and } \beta_{02} + 100\beta_{12} = -100 \tag{10}$$

under which we will find a mean shift outlier using the test in Section 3. First we will check whether these linear relationships hold for the adaptive score data. The corresponding Wilks lambda test yields the p-value = 0.972. Hence it is reasonable to assume the linear

relationships (10) for the adaptive score data. Next the values of the $T_i$ in (8) are included in Table 1. The maximum of the $T_i$ is $T_{19}=8.91$ and the Bonferroni upper bound for the p-value for the test based on $\max_{1\le i\le n} T_i$ is 0.042. Hence we can conclude that observation 19 is a mean shift outlier under linear constraints (10) at any significance level greater than or equal to 0.042.

## References

[1] Barrett, B.E. and Ling, R.F. (1992) General classes of influence measures for multivariate regression, *Journal of the American Statistical Association*, **87**, 184-191.

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

[3] Kim, M.G. (1995) Local influence in multivariate regression, *Communications in Statistics: Theory and Methods*, **24**, 1271-1278.

[4] Mardia K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*, Academic Press.

[5] Seber, G.A.F. (1984). *Multivariate Observations*, Wiley, New York.

[6] Srivastava, M.S. and von Rosen, D. (1998) Outliers in multivariate regression models, *Journal of Multivariate Analysis*, **65**, 323-337.

[7] Tang, M.K. and Fung, W.K. (1997) Case-deletion diagnostics for test statistics in multivariate regression, *Australian Journal of Statistics*, **39**, 345-353.

[ 2002년 5월 접수, 2002년 7월 채택 ]