Misspecification in Multivariate Regression

Min-Woong Shin* & Chi-Hoon Choi**

ABSTRACT

Hocking (5), Rao (7), Rosenberg and Levy (8) and Walls and Weeks (10) have given some results for the misspecification in univariate regression model. We give similar results for a multivariate regression model.

1. Introduction

Various authors have considered the consequences of eliminating variables in univariate regression analysis. Hocking (6) reviewed the consequences of incorrect model specification which provide a theoretical motivation for variable selection.

We would like to offer similar results for a multivariate regression analysis. Especially, we show that multivariate regression coefficients may be estimated with smaller variances using the subset model.

2. Notation and Basic Concepts

The identifying notion of multivariate regression is that of the presence of several correlated dependent variables (p of them). In the standard multivariate regression model there is a separate univariate regression

* Department of Mathematics, Hankuk University of Foreign Studies.
** Department of Computer Science and Statistics, Seoul National University. The authors are grateful to professor Sung H. Park for several helpful suggestion.

equation corresponding to each of the p-dependent variables, and N observations are taken jointly on the dependent and independent variables.

2.1. The Standard Multivariate Regression Model

Let $y_1,\ldots,y_p$ be $N\times 1$ vectors representing N independent observations on each of p correlated dependent random variables.

Assume the linear model (linear in the parameters β_j)

(2.1) $\quad y_j = X\cdot\beta_j + u_j,\ j=1,\ldots,p,$

where X is an $(N\times q)$ matrix of known form and be thought of as arising either as a "functional" or a "conditional" regressor matrix; is a $(q\times 1)$ vector of parameters; u is an $(N\times 1)$ vector of errors and $E(u)=0$ var(u) $=I\sigma^2$, so the elements of u are uncorrelated. For a given j, (2.1) is a univariate regression.

The basic model equation may be written in a more compact form in the following way. Define

$$Y\equiv(y_1,\ldots,y_p),\quad U\equiv(u_1,\ldots,u_p),\quad B\equiv(\beta_1,\ldots,\beta_p)$$

Then (2.1) becomes

(2.2) $$\underset{(N\times p)}{Y} = \underset{(N\times q)}{X}\cdot\underset{(q\times p)}{B} + \underset{(N\times p)}{U}.$$

2.2. Assumptions

To define the multivraiate regression model completely, impose the following assumptions and constraints on the quantities in (2.2)

(2.3) $$p+q\leq N.$$

This assumption is used so that there will be a sufficient number of observations available to estimate all of the parameters. Let N_0 denote the total number of observations, and let N_p denote the total number of parameters. Then, since there are N observations per equation, and p equations, $N_0=Np$. Since a row of U has p correlated errors [and all rows will have the same covariance matrix, by assumption (2.5)], there will be a corresponding $p\times p$ covariance matrix with $p(p+1)/2$ distinct elements in general, and fewer distinct elements if the covariance matrix is structured. Moreover, unless there are some a priori constraints on the coefficient matrix, B will have pq distinct parameters. Thus,

$$N_p \leq \frac{p(p+1)}{2} + pq.$$

(2.4) $\underset{(N\times q)}{\text{rank}(X)} = q.$

This condition will be required in order to obtain a unique solution to the normal equations. If (2.4) is not satisfied, use a generalized inverse.

Define the rows of U by

$$U \equiv (v_1 \ldots, v_N)'$$

where v_j is a $p\times 1$vector, $j=1,\ldots,N$. $\Sigma > 0$ will be used to denote that Σ is positive definite. Assume for all j that

(2.5) $E(v_j)=0$, $\text{var}(v_j)=\Sigma \equiv (\sigma_{ij})$, and $\Sigma > 0$.

[Remark: An alternative form of this assumption, which is often useful, is obtained by stringing out the columns of U into a long $N_p \times 1$ vector u, where $u' \equiv (u_1', u_2', \ldots, u_p')$. Then (2.5) becomes

(2.6) $E(u)=0$, $\text{var}(u)=\Sigma \otimes I_N$

where $\otimes$ is the direct product.]

Script $\mathcal{L}$ will be used to denote probability law, so that, for example, $\mathcal{L}(\mathrm{x})=N(\theta, \Sigma)$ should be interpreted as: The probability law of the random x is multivariate normal with mean vector θ and covariance matrix Σ.

(2.7) $\mathcal{L}(v_j)=N(0,\Sigma)$, $j=1,\ldots, N$.

2.3. Estimation

The unknown parameters which must be estimated are B and Σ. They will be estimated by the methods of least squares and the maximum likelihood.

Least Squares Estimation: Since each equation in (2.1) corresponds to a univariate regression, minimize $(u_j'u_j)$ with respect to β_j, for each fixed j and one find

(2.8) $\hat{\beta}_j = (X'X)^{-1}X'y_j$, $j=1,\ldots,p$.

where $\hat{\beta}_j$ is a $(q\times 1)$ vecter of parameters.

Then

$$\hat{B}=(\hat{\beta}_1,\ldots,\hat{\beta}_p)=[(X'X)^{-1}X'y_1,\ldots,(X'X)^{-1}X'y_p]$$

$$=(X'X)^{-1}X'Y. \tag{2.9}$$

The residuals may be used to estimate Σ. Thus, define the matrix of residuals:

$$\hat{U}=Y-X\hat{B}.$$

Then the residual sums of squares is $\hat{U}'\hat{U}$. Since each residual variance and covariance involves an estimate of a q-vector $\hat{\beta}$, an unbiased estimation of Σ is given by

$$\hat{\Sigma}=\frac{1}{N-q}\hat{U}'\hat{U}=\frac{1}{N-q}(Y-X\hat{B})'(Y-X\hat{B}). \tag{2.10}$$

2.4. Properties of Estimators.

$\hat{B}$ is consistent, unbiased, and efficient estimator of B under the assumption of the model defined above.

Note that the space of observationis is orthrgonal to the space of residuals; that is

$$X'U=0,\quad Y'U=(XB)'U=0. \tag{2.11}$$

From, (2.8)

$$\text{cov}(\hat{\beta}_i,\hat{\beta}_j)=\sigma_{ij}(X'X)^{-1}.$$

Hence, for the strung out vectors

$$\underset{(1\times pq)}{\hat{\beta}'}\equiv(\underset{(1\times q)}{\hat{\beta}_1'},\ldots,\underset{(1\times q)}{\hat{\beta}_p'}),\qquad \beta'\equiv(\beta_1',\ldots,\beta_p'),$$

$$\text{var}(\hat{\beta})=\Sigma\otimes(X'X)^{-1}\equiv\Phi$$

which is a $pq\times pq$ matrix. Thus,

$$\mathcal{L}(\hat{\beta})=N(\beta,\Phi).$$

Moreover, $\hat{\beta}$ and $\hat{\Sigma}$ are independent.

3. Consequences of Incorrect Model Specification

There are a variety of practical and economical reasons for reducing the number of independent variables in the final equation. In addition, variable deletion may be desirable in terms of the statistical properties of the parameter estimates and the estimate of the final equation. This section provides the consequences of incorrectly specifying the model either in terms of retaining extraneous variables or deleting relevant variables.

Let the model (2. 2) be written in matrix form as

(3. 1) $$Y=X_pB_q+X_rB_r+U$$

where the X matrix has been partitioned into X_p of dimension $N\times p$ and X_r of dimension $N\times r$, $p+r=q$. The B vector is partitioned conformably. Let $\hat{B}$, with components $\hat{B}_p$ and $\hat{B}_r$, denote the least squares estimate of B and let $\tilde{B}_p$ denote the subset least squares estimator of B_p if the variables in X_r are deleted from the model. That is,

(3. 2) $$\tilde{B}_p=(X_p'X_p)^{-1}X_p'Y.$$

Further

(3. 3) $$\tilde{\Sigma}=\frac{1}{N-p}(Y-X_p\tilde{B}_p)'(Y-X_p\tilde{B}_p).$$

If model (3. 1) is correct, the properties of $\hat{B}$ and $\hat{\Sigma}$ as estimates of B and Σ are well known from general linear model theory. In particular, $\hat{\beta}$ and $\hat{\Sigma}$ are unbiased estimators for β and Σ with

$$\mathcal{L}(\hat{\beta})=N(\beta,\ \Sigma\otimes(X'X)^{-1})$$

and

$$\mathcal{L}((N-q)\hat{\Sigma})=W(\Sigma, p,\ N-q).$$

Lemma 3. 1 If we let

(3. 4) $$A=(X_p'X_p)^{-1}X_p'X_r$$

then

$$E(\tilde{B}_p)=B_p+AB_r.$$

Proof. Suppose we postulate the model

$$E(Y)=X_pB_p.$$

This leads to the least squares estimators

$$\tilde{B}_p=(X_p'X_p)^{-1}X_p'Y.$$

Suppose, now the true response relationship is in fact

$$E(Y)=X_pB_p+X_rB_r$$
$$E(\tilde{B}_p)=(X_p'X_p)^{-1}X_p'E(Y)$$
$$=(X_p'X_p)^{-1}X_p'(X_pB_p+X_rB_r)$$
$$=B_p+ABr.$$

We know that

(3.5) $$VAR(\tilde{\beta}_p)=\Sigma\otimes(X_p'X_p)^{-1}$$

The mean squared error is given by

(3.6) $$MSE(\tilde{\beta}_p)=E(\tilde{\beta}_p-\beta_p)(\tilde{\beta}_p-\beta_p)'$$
$$=E[(\tilde{\beta}_p-E(\tilde{\beta}_p))(\tilde{\beta}_p-E(\tilde{\beta}_p))']+(E(\tilde{\beta}_p)-\beta_p)(E(\tilde{\beta}_p)-\beta_p)'$$
$$=\Sigma\otimes(X_p'X_p)^{-1}+(E(\tilde{\beta}_p)-\beta_p)(E(\tilde{\beta}_p)-\beta_p)'$$

Lemma 3.2 If A is an $n\times n$ positive semi-definite (p.s.d.)matrix and C is $p\times n$ matrix of rank p, then CAC' is p.s.d..

Proof. see Graybill (6).

Lemma 3.3 The matrix cov $(\hat{\beta}_{ip}, \hat{\beta}_{jp})-\text{cov}(\tilde{\beta}_{ip}, \tilde{\beta}_{jp})$ is $\sigma_{ij}(A_{pp}^{-1}A_{pr})\ B_{rr}\ (A_{pp}^{-1}A_{pr})'$ and especially positive semi-definite, when $\sigma_{ij}>0$, where $y_j=X_p\beta_{jp}+X_r\beta_{jr}+u_j$.

Proof. Let the natrix $A=[a_{ij}]$ of order q and its inverse $B=[b_{ij}]$ be partitioned into submatrices of indicated orders, where $A=X'X$;

$$\begin{pmatrix} \underset{(p\times p)}{A} & \underset{(p\times r)}{A} \\ \underset{(r\times p)}{A} & \underset{(r\times r)}{A} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \underset{(p\times p)}{B} & \underset{(p\times r)}{B} \\ \underset{(r\times p)}{B} & \underset{(r\times r)}{B} \end{pmatrix}$$

Then, since A_{pp} is non-singular,

$$B_{pp}=A_{pp}^{-1}+(A_{pp}^{-1}A_{pr})\xi^{-1}(A_{pp}^{-1}A_{pr})', \quad B_{rp}=B_{pr}'$$
$$B_{pr}=-(A_{pp}^{-1}A_{pr})\xi^{-1}, \quad B_{rr}=\xi^{-1}$$

where $\xi=A_{rr}-A_{pr}(A_{rr}^{-1}A_{pr})$.

Hence

$$\text{cov}(\hat{\beta}_{ip}, \hat{\beta}_{jp})-\text{cov}(\tilde{\beta}_{ip}, \tilde{\beta}_{jp})=\sigma_{ij}B_{pp}-\sigma_{ij}A_{pp}^{-1}$$
$$=\sigma_{ij}A_{pp}^{-1}+\sigma_{ij}(A_{pp}^{-1}A_{rr})B_{rr}(A_{pp}^{-1}A_{pr})'-\sigma_{ij}A_{pp}^{-1}$$
$$=\sigma_{ij}(A_{pp}^{-1}A_{pr})B_{rr}(A_{pp}^{-1}A_{pr})'$$

which is positive semi-definite for $\sigma_{ij}>0$ by Lemma 3.2.

Lemma 3.4. Let A and B be positive (semi) definite matrices then $A\otimes B$ is a positive (semi) definite matrix.

Proof. See Craybill (6).

Theorem 3.1. Let Σ be a positive definite matrix.
Matrix $\mathrm{VAR}(\hat{\beta}_p)-\mathrm{VAR}(\tilde{\beta}_p)$ is potitive semi-definite.

Proof. $\mathrm{VAR}(\hat{\beta}_p)-\mathrm{VAR}(\tilde{\beta}_p)=\Sigma\otimes(A_{pp}^{-1}A_{pr}B_{rr}A_{rp}A_{pp}^{-1})$ which establishes Theorem 3.1 by the Lemma 3.3. and Lemma 3.4.

The regression equation is frequently used to predict the response to a particular input, say $x'=(x_p', x_r')$. If we use the full model then the predicted value of the response is $\hat{y}=x'\hat{B}$ which has mean $x'B$ and prediction variance, that is,

$$\text{(3.7)}\quad \mathrm{VARP}(\hat{y}')=E(\hat{y}'-y')(\hat{y}'-y')'$$
$$=E(\hat{y}'-E(y'))(\hat{y}'-E(y'))+E[(E(y')-y')(E(y')-y')']$$
$$=\mathrm{VAR}(\hat{y}')+\Sigma.$$

On the other hand, if the subset model with x_r deleted is used, the predicted response is $x_p=x_p'\tilde{B}_p$ with mean

$$E(\tilde{y}_p)=E(x_p'\tilde{B}_p)$$
$$=x_p'B_p+x_p'AB_r$$

and

$$\mathrm{VARP}(\tilde{y}_p')=\mathrm{VAR}(\tilde{y}_p')+\tilde{\Sigma}$$

The prediction mean squared error $\mathrm{MSEP}(\tilde{y}_p')$ is given by

$$\mathrm{MSEP}(\tilde{y}_p')=E(\tilde{y}'-y')(\tilde{y}'-y')'$$
$$=E[(\tilde{y}-E(\tilde{y}'))(\tilde{y}'-E(\tilde{y}'))']+[E(\tilde{y}')-E(y')][E(\tilde{y}')-E(y')]'+\Sigma$$
$$=\mathrm{VAR}(\hat{y}_p)+\Sigma+(x_p'AB_r-x_r'B_r)(x_p'AB_r-x_r'B_r)'$$

Lemma 3.5. $\mathrm{VARP}(\hat{y}_{jp})\geq\mathrm{VARP}(\hat{y}_{ip})$ $j=1,2,\ldots,p$.
Proof. see Hocking(10).

The motivation for variable elimination is provided by Theorem 3. 1 and Lemma 3. 5. That is, even if $\beta_r \neq 0$, β_p may be estimated or future responses may be predicted with smaller variance using the subset model.

REFERENCES

[1] Anderson, T.W., (1976). An introduction to multivariate statistical analysis.

[2] Donald F. Morrison, (1976). Multivariate statistical methods.

[3] Graybill, F.A., (1969) Introduction to matrices with applications in statistics.

[4] Drapher, N.R. and Smith, H., (1966). Applied regression analysis.

[5] Hocking, R.R., (1974), Misspecification in regression. The American Statistician 28, 39-40.

[6] Hocking, R.R., (1976). The analysis and selection of variables in linear regression, Biometrics 32, 1-49.

[7] Rao, P., (1971) Some Notes on Misspecification in Multiple Regression. The American Statistician 25, 37-39.

[8] Rosenberg, S.H. and Levy, P.S. (1972). A characterization on Miss pecification in the General Linear Regression Model. Biometrics 28, 1129-1133.

[9] S. James Press, (1977). Applied multivariate analysis.

[10] Walls, R.E. and Weeks, D.L., (1969). A note on the variance of a predicted response in regression. The American Statistician 23, 24-26. (1977).