Variable Selection Theorems in General Linear Model

Sang Hoo Yoon[1] , Jeong Soo Park[2]

Abstract

For the problem of variable selection in linear models, we consider the errors are correlated with V covariance matrix. Hocking's theorems on the effects of the overfitting and the underfitting in linear model are extended to the less than full rank and correlated error model, and to the ANCOVA model

Keywords : Linear model, General Linear Model, Hocking's Theorems, Variable Selection,

1. INTRODUCTION

The primary purpose of this paper is to provide a review of the concepts associated with variable selection in general linear models, the errors are correlated with V covariance matrix. Also, we discuss general results for the situation where the matrix of predictors need not have full rank.

The problem of determining the "best" subset of variables has long been of interest to applied statisticians and, primarily because of the current availability of high-speed computations, this problem has received considerable attention in the recent statistical literature(Seber and Lee, 2003).

The problem of overfitting(i.e, putting too many predictors in a linear model) has been addressed by Hocking(1976). It supports that deleting independent variables corresponding to small coefficients(relative to their standard errors) will lead to high precision in the estimates of coefficients corresponding to the retained variables.

Hocking's theorems have been described in many textbook on linear model, for example in Park(2001) and Ravishanker and Dey(2001). These theorems are extended to the less than full rank and correlated errors model, and to the analysis of covariance model.

2. NOTATION AND BASIC CONCEPTS

Consider the generalized linear model

$$\boldsymbol{y} = X\beta + \varepsilon, \quad Var(\varepsilon) = \sigma^2 V . \tag{2-1}$$

where V is a known $N \times N$ positive definite covariance matrix, $\boldsymbol{y} = (Y_1, Y_2, \cdots, Y_N)'$ is an

† This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (KRF-2005-202-C00072)

1) Graduate, Department of Statistics, Chonnam National University, Gwangju, Korea, 500-757
2) Professor, Department of Statistics, Chonnam National University, Gwangju, Korea, 500-757

N-dimensional vector of observed responses, $\beta=(\beta_0,\beta_1,\cdots,\beta_k)'$ is a $(k+1)$-dimensional vector of unknown parameters, and X is an $N\times(k+1)$ matrix of rank q(less than full rank) of known predictors. Since V is positive definite, there exists an $N\times N$ matrix K with $r(K)=N$, such that $V=KK'$.

Let, $Z=K^{-1}y$, $B=K^{-1}X$ and $\eta=K^{-1}\varepsilon$, then

$$E(\eta)=0,\quad Var(\eta)=K^{-1}(\sigma^2V)K^{-1'}=\sigma^2K^{-1}KK'K^{-1'}=\sigma^2I_N.$$

Consider the "transformed" general linear model

$$Z=B\beta+\eta\ ,\quad Var(\eta)=\sigma^2I_N.$$

The (generalized) least square solution is

$$\hat{\beta}=(B'B)^-B'=(X'K^{-1'}K^{-1}X)^-X'K^{-1'}K^{-1}y$$
$$=(X'V^{-1}X)^-X'V^{-1}y\ ,$$

where $(B'B)^-$ denote any g-inverse of the matrix $(B'B)$, and the expectation and variance of $\hat{\beta}$ are

$$E(\hat{\beta})=E[(X'V^{-1}X)^-X'V^{-1}y]=H_1\beta\ . \tag{2-2}$$

where

$$H_1=(X'V^{-1}X)^-X'V^{-1}X\ ,$$

$$Var(\beta)=Var[(X'V^{-1}X)^-X'V^{-1}y]=\sigma^2(X'V^{-1}X)^-. \tag{2-3}$$

The unbiased GLS estimator of σ^2 is given by

$$\hat{\sigma}^2=\frac{1}{(N-r)}(z-B\hat{\beta})'(z-B\hat{\beta})$$
$$=\frac{1}{(N-r)}y'[V^{-1}-V^{-1}X(X'V^{-1}X)^-X'V^{-1}]y\ , \tag{2-4}$$

$$E(\hat{\sigma}^2)=\sigma^2. \tag{2-5}$$

The mean squared error of $\hat{\beta}$ is given by

$$MSE(\hat{\beta})=Var(\hat{\beta})=\sigma^2(X'V^{-1}X)^-. \tag{2-6}$$

3. UNDERFITTING

Let the models be written in matrix form as

(full model) $\quad y=X_p\beta_p+X_r\beta_r+\varepsilon,\ Var(\varepsilon)=\sigma^2V\ ,$ (3-1)

(reduced model) $\quad y=X_p\beta_p+\varepsilon,\ Var(\varepsilon)=\sigma^2V\ ,$ (3-2)

where the X matrix has been partitioned into X_p of dimension $N\times(p+1)$ and X_r of dimension $N\times r$. Suppose the true model is given by the full model. The β vector is partitioned conformable. Let $\hat{\beta}$, with components $\hat{\beta}_p$ and $\hat{\beta}_r$, denote the least squares solution of β and let $\tilde{\beta}_p$ denote the least squares solution of β_p in the reduced model. That is, when we underfitting the true model by the reduced model, we have

$$\tilde{\beta}_p=(X_p'V^{-1}X_p)^-X_p'V^{-1}y\ . \tag{3-3}$$

Now the expectation and variance of $\tilde{\beta}_p$ are

$$E(\tilde{\beta}_p) = H_2\beta_p + A\beta_r . \tag{3-4}$$
$$Var(\tilde{\beta}_p) = \sigma^2(X_1'V^{-1}X_1)^-$$

where,

$$H_2 = (X_p'V^{-1}X_p)^- X_p'V^{-1}X_p , \tag{3-5}$$
$$A = (X_p'V^{-1}X_p)^- X_p'V^{-1}X_r . \tag{3-6}$$

Thus we know that $\tilde{\beta}_p$ is biased. An estimator of σ_p^2 analogous to (2-5) is given by

$$\tilde{\sigma}_p^2 = \frac{1}{(N-p)}(\boldsymbol{y}-X\beta_p)'V^{-1}(\boldsymbol{y}-X\beta_p)$$
$$= \frac{1}{(N-p)}\boldsymbol{y}'[V^{-1}-V^{-1}X_p(X_p'V^{-1}X_p)^-X_p'V^{-1}]\boldsymbol{y} , \tag{3-7}$$
$$E(\tilde{\sigma}_p^2) = \sigma^2 + \frac{\beta_r'X_r'(V^{-1}-V^{-1}X_p(X_p'V^{-1}X_p)^-X_p'V^{-1})X_r\beta_r}{N-p} . \tag{3-8}$$

which means $\tilde{\sigma}_p^2$ is also biased, The mean squared error of $\tilde{\beta}_p$ is given by

$$MSE(\tilde{\beta}_p) = E(\tilde{\beta}_p-\beta_p)(\tilde{\beta}_p-\beta_p)'$$
$$= \sigma^2(X'V^{-1}X)^- + A\beta_r\beta_r'A' . \tag{3-9}$$

Theorem 3.1:

1. β_p is generally biased, interesting exceptional cases being (a) $\beta_r=0$ or (b) $X_p'X_r=0$.
2. The matrix $Var(\hat{\beta}_p) - Var(\tilde{\beta}_p)$ is positive semi-definite.
3. If the matrix $Var(\tilde{\beta}_r)-\beta_r\beta_r'$ is positive semi-definite, then the matrix $Var(\hat{\beta}_p) - MSE(\tilde{\beta}_p)$ is positive semi-definite.
4. $\tilde{\sigma}^2$ is generally biased.

Proof: Properties 1 and 4 are already proved above. The property 2 is shown as follows.

$$(X'V^{-1}X)^- = \begin{bmatrix} X_1'V^{-1}X_1 & X_1'V^{-1}X_2 \\ X_2'V^{-1}X_1 & X_2'V^{-1}X_2 \end{bmatrix}^-$$
$$= \begin{bmatrix} (X_p'V^{-1}X_p)^- + (X_p'V^{-1}X_p)^-X_p'V^{-1}X_rEX_r'V^{-1}X_p(X_p'V^{-1}X_p)^- & -(X_p'V^{-1}X_p)^-X_p'V^{-1}X_rE \\ -EX_r'V^{-1}X_p(X_p'V^{-1}X_p)^- & E \end{bmatrix} . \tag{3-10}$$

using the result on the G-inverse of a partitioned matrix (Ravishanker and Dey[2001], Result 3.1.10, for example), where

$$E = (X_2'V^{-1}X_2 - X_2'V^{-1}X_1(X_1'V^{-1}X_1)^-X_1'V^{-1}X_2)^- , \tag{3-11}$$

and denote the above matrix as D. Note that the matrix E is $\frac{1}{\sigma^2}Var(\hat{\beta}_r)$, from (2-3).

$$Var(\hat{\beta}_p) - Var(\tilde{\beta}_p)$$
$$= \sigma^2[(X_p'V^{-1}X_p)^- + (X_p'V^{-1}X_p)^-X_p'V^{-1}X_r\,E\,X_r'V^{-1}X_p(X_p'V^{-1}X_p)^-]$$

$$-\sigma^2(X_p'V^{-1}X_p)^-$$
$$=\sigma^2 A E A' : \text{p.s.d}$$

because E is p.s.d.

The property 3 is shown as follows.

$Var(\tilde{\beta}_r)=\sigma^2(X_p'V^{-1}X_p)^-+\sigma^2AEA'$ and $MSE(\tilde{\beta}_p)=(X_p'V^{-1}X_p)^-\sigma^2+A\beta_r\beta_r'A'$. Thus $Var(\hat{\beta}_p)-MSE(\tilde{\beta}_1)=A[E\sigma^2-\beta_2\beta_2']A'$ is p.s.d if the matrix $Var(\tilde{\beta}_r)-\beta_r\beta_r'$ is p.s.d. ■

Consider predicted value of the response to a particular input, say $x'=(x_p'x_r')$.

$$E(y)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k=x'\beta=x_p'\beta_p+x_r'\beta_r,$$

If we use the full model then the predicted value of the response is

$$\text{(full model)} \qquad \hat{y}=x'\hat{\beta}=x_p'\beta_p+x_r'\beta_r, \tag{3-12}$$

which has the expectation and prediction variance are given by

$$E(\hat{y})=x'\beta,$$
$$Var(\hat{y})=x'(X'V^{-1}X)^-x\sigma^2. \tag{3-13}$$

On the other hand, if the reduced model with x_r deleted is used, the predicted response is

$$\text{(reduced model)} \quad \tilde{y}_p=x_p'\tilde{\beta}_p,$$

which has the expectation and prediction variance are given by

$$E(\tilde{y}_p)=x_p'\beta_p+x_p'A\beta_r,$$
$$Var(\tilde{y}_p)=x_p'(X_p'V^{-1}X_p)^-x_p\sigma^2. \tag{3-14}$$

Thus we know that $\tilde{y}_p$ is biased. The prediction mean squared error is given by

$$MSE(\tilde{y}_p)=E(\tilde{y}_p-x'\beta)^2$$
$$=x_p'(X_p'V^{-1}X_p)^-x_p\sigma^2+(x_p'A\beta_r-x_r'\beta_r)^2. \tag{3-15}$$

Theorem 3.2:

1. $\tilde{y}_p$ is biased unless (a) $\beta_r=0$ or (b) $X_p'X_r=0$.
2. $Var(\hat{y})\geq Var(\tilde{y}_p)$.
3. If the matrix $Var(\tilde{\beta}_r^0)-\beta_r\beta_r'$ is positive semi-definite, then $Var(\hat{y})\geq MSE(\tilde{y}_p)$.

Proof: Properties 1 is already proved above. The property 2 is shown as follows.

$$Var(\hat{y})=x'(X'V^{-1}X)^-x\sigma^2$$
$$=(x_p', x_r')\,D\begin{bmatrix}x_p\\x_r\end{bmatrix}\sigma^2$$
$$=x_p'(X_p'V^{-1}X_p)^-x_p\sigma^2+x_p'AEA'x_p\sigma^2-x_r'E^-A'x_p\sigma^2-x_p'AEx_r\sigma^2+x_r'E^-x_r\sigma^2.$$

where A is same as (3-6), D is same as (3-10), and E is same as (3-11).

$$Var(\hat{y})-Var(\tilde{y}_p)=[A'x_p-x_p]'E[A'x_p-x_r]\geq 0.$$

because E is p.s.d.

The property 3 is shown as follows.

$Var(\hat{y}) - MSE(\tilde{y}_p) = [A'x_p - x_r]'(E\sigma^2 - \beta_r\beta_r')[A'x_p - x_r]$ is p.s.d , if the matrix $Var(\tilde{\beta}_r) - \beta_r\beta_r'$ is p.s.d. ■

4. OVERFITTING

We consider the general linear model be partitioned as (3-1). If we include $X_r\beta_r$ when it should be excluded(that is, when $\beta_r = 0$), we are *overfitting*. When overfitting, the expectation $\hat{\beta}_p$ is

$$E(\hat{\beta}_p) = H_2\beta_p ,$$

where H_2 is same as (3-5). And $MSE(\hat{\beta}_p)$ is an unbiased estimate of σ^2

$$MSE(\beta_p) = \sigma^2(X_p'V^{-1}X_p)^- .$$

5. ANCOVA MODEL

A general formulation of the ANCOVA model is

$$y = X\tau + Z\beta + \varepsilon \tag{5-1}$$

where y is and N-dimensional vector, X is an $N\times p$ design matrix with rank $(X) = r < p$, τ is a p-dimensional vector of fixed-effects parameters, Z is an $N\times q$ regresion matrix with rank $(Z) = q$, β is a q-dimensional vector of regression parameters, the columns of X are linearly independent of the columns of Z, and ε has an N-variate normal distribution with mean vector 0 and covariance matrix $\sigma^2 I_N$. We can rewrite the model in (5-1) as

$$y = W\gamma + \varepsilon, \text{ where } W = (X\ Z) \text{ and } \gamma = \begin{pmatrix}\tau\\ \beta\end{pmatrix}.$$

We develop results similar to the theorems 3.1 and 3.2 for the ANCOVA model.

6. CONCLUSION

The motivation for variable elimination is provided by theorems 3.1 and 3.2. That is, if the variance become the criteria to decision for parameter $\tilde{\beta}_p$ and prediction response $\tilde{y}_p$, reduced model predicted with smaller variance(by property 2 of theorem 3.1 and 3.2). The penalty is in the bias. In the mean squared error, property 3 of theorem 3.1 and 3.2 describe a special condition(the matrix $Var(\tilde{\beta}_r^0) - \beta_r\beta_r'$ is positive semi-definite) under which the gain in precision is not offset by the bias.

References

Helms, R. (1974). The average estimated variance criterion for the selection of variable problem in general linear models, *Technometrics*, Vol 16, pp. 261~274

Hocking, R.R (1976). The analysis and selection of variables in linear regression, *Biometrics*, Vol. 32, pp. 1~49

Littel, R.C., Freund, R.J. and Spector, P.C. (1991). *SAS System for Linear Models*, SAS Institute INC., Cary, NC, USA

McCullagh, P. and Nelder, J.A. (1991). *Generalized Linear Models*, Chapman & Hall, London

Ravishanker, and Dey, D.K. (2001). *A First Course in Linear Model Theory*, Chapman & Hall, London

Park, S.H. (2001), *The Regression Analysis , 3/e*, Minyoungsa , Korea

Seber, G.A.F. and Lee, A.J. (2003). *Linear Regression Analysis, 2/e*, John Wiley & Sons, New York