Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 28 - 29, 2006, pp. 171-179

Variable Selection Theorems in General Linear Model*

Jeong Soo Park[1], *Sang Hoo Yoon*[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 : Variable selection, General linear model, Hocking's theorems, ANCOVA Model, Overfitting, Underfitting.

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 Helms(1974) and 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

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

[1]Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea. E-mail : jspark@chonnam.ac.kr

[2]Graduate student, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea. E-mail : statstar@hanmail.net

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 general linear model

$$y = x\beta + \varepsilon,\ Var(\varepsilon) = \sigma^2 V. \tag{2.1}$$

where V is a known $N\times N$ positive definite covariance matrix, $y=(y_1, y_2, \ldots, y_N)'$ is an N-dimensional vector of observed responses, $\beta=(\beta_0, \beta_1, \ldots, \beta_k)'$ 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 and $N\times N$ 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

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

Then we consider the "transformed" general linear model is the less than full rank and uncorrelated error model

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

The (generalized) least square solution is

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

where $(B'B)^{-}$ denote any g-inverse of the matrix $(B'B)$. The expectation of $\hat{\beta}$ is

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

where

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

The expectation of $\hat{\beta}$ is not unique and biased. So, we consider the expectation and variance of $c'\hat{\beta}$

$$E(c'\hat{\beta})=c'\beta, \tag{2-2}$$

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

for any estimable function of β.

The unbiased GLS estimator of σ^2 is given by

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

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

$$MSE(c'\hat{\beta}) = Var(c'\hat{\beta}) = \sigma^2 c'(X'V^{-1}X)^{-}c. \tag{2-5}$$

for any estimable function of β..

3. Model Misspecification by Underfitting

Let the models be written in matrix form as

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

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

$$E(\tilde{\beta_p}) = H_2\beta_p + A\beta_r\ . \tag{3-4}$$
$$Var(\tilde{\beta_p}) = \sigma^2(X_p'V^{-1}X_p)^{-}. \tag{3-5}$$

where,

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

Thus we know that $\widetilde{\beta_p}$ is biased. An estimator of σ^2 analogous to (2-4) is given by

$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{1}{(N-p)}(y-X\beta_p)' V^{-1}(y-X\beta_p)\\ &= \frac{1}{(N-p)} y'[\, V^{-1} - V^{-1}X_p(X_p' V^{-1}X_p)^{-}X_p' V^{-1}]y\ ,\\ E(\widetilde{\sigma}^2) &= \sigma^2 + \frac{\beta_r' X_r'(\, V^{-1} - V^{-1}X_p(X_p' V^{-1}X_p)^{-}X_p' V^{-1})X_r\beta_r}{N-p}. \end{aligned} \tag{3-7}$$

which means $\widetilde{\sigma^2}$ is also biased, The mean squared error of $c'\widetilde{\beta_p}$ is given by

$$\begin{aligned} MSE(c'\widetilde{\beta_p}) &= E(c'\widetilde{\beta_p} - c'\beta_p)(c'\widetilde{\beta_p} - c'\beta_p)'\\ &= \sigma^2 c'(X' V^{-1}X)^{-}c + c' A\beta_r\beta_r' A' c. \end{aligned} \tag{3-8}$$

for any estimable function of β_p.

Theorem 1:

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

Proof: Properties 1 and 2 are already proved above. The property 3 is shown as follows.

The variance-covariance matrix of $\widehat{\beta}$ is

$$Var(\widehat{\beta}) = \sigma^2(X' V^{-1}X)^{-} = \begin{bmatrix} Var(\widehat{\beta_p}) & Cov(\widehat{\beta_p}, \widehat{\beta_r}) \\ Cov(\widehat{\beta_r}, \widehat{\beta_p}) & Var(\widehat{\beta_r}) \end{bmatrix}^{-} = \sigma^2 \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

where

$$A_{11} = (X_p' V^{-1}X_p)^{-} + (X_p' V^{-1}X_p)^{-}X_p' V^{-1}X_r A_{22} X_r' V^{-1}X_p(X_p' V^{-1}X_p)^{-} \tag{3-9}$$

$$A_{12} = -(X_p' V^{-1}X_p)^{-}X_p' V^{-1}X_r A_{22} \tag{3-10}$$

$$A_{21} = -A_{22} X_r' V^{-1}X_p(X_p' V^{-1}X_p)^{-} \tag{3-11}$$

$$A_{22} = (X_r' V^{-1}X_r - X_r' V^{-1}X_p(X_p' V^{-1}X_p)^{-}X_p' V^{-1}X_r)^{-}. \tag{3-12}$$

using the result on the G-inverse of a partitioned matrix (Ravishanker and Dey[2001], Result 3.1.10, for example). Note that the matrix A_{22} is positive semi-definite ($A_{22} = \frac{1}{\sigma^2} Var(\widehat{\beta_r})$).

By subtracting (3-5) from $A_{11}\sigma^2$ is

$$Var(\widehat{\beta_p}) - Var(\widetilde{\beta_p}) = \sigma^2(X_p'V^{-1}X_p)^- - A_{11}\sigma^2$$
$$= (X_p'V^{-1}X_p)^- X_p'V^{-1}X_r A_{22} X_r'V^{-1}X_p(X_p'V^{-1}X_p)^-\sigma^2.$$

positive semi-definite.

The property 4 is shown as follows.

The mean squared error of $c'\widehat{\beta_p}$ is given by

$$MSE(c'\widehat{\beta_p}) = Var(c'\widehat{\beta_p}) = c'A_{11}c\sigma^2. \tag{3-13}$$

By subtracting (3-8) from (3-13) is

$$MSE(c'\widehat{\beta_p}) - MSE(c'\widetilde{\beta_p}) = c'A_{11}c\sigma^2 - [(\sigma^2 c'(X_p'V^{-1}X_p)^- c + c'A\beta_r\beta_r'A'c]$$
$$= A[A_{22}\sigma^2 - \beta_r\beta_r']A' = A[Var(\widehat{\beta_r}) - \beta_r\beta_r']A'$$

positive semi-definite, if the matrix $Var(\widetilde{\beta_r}) - \beta_r\beta_r'$ is p.s.d. ■

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

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = x'\beta = x_p'\beta_p + x_r'\beta_r,$$

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

$$\text{(full model)} \quad \widehat{y} = x'\widehat{\beta} = x_p'\beta_p + x_r'\beta_r,$$

which has the expectation and prediction variance are given by

$$E(\widehat{y}) = x'\beta, \tag{3-14}$$
$$Var(\widehat{y}) = x'(X'V^{-1}X)^- x\sigma^2. \tag{3-15}$$

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

$$\text{(reduced model)} \quad \widetilde{y_p} = x_p'\widetilde{\beta_p},$$

which has the expectation and prediction variance are given by

$$E(\widetilde{y_p}) = x_p'\beta_p + x_p'A\beta_r, \tag{3-16}$$
$$Var(\widetilde{y_p}) = x_p'(X_p'V^{-1}X_p)^- x_p\sigma^2. \tag{3-17}$$

where A is same as in (3-6). Thus we know that $\widetilde{y_p}$ is biased. The prediction mean squared error is given by

$$MSE(\widetilde{y_p}) = E(\widetilde{y_p} - x'\beta)^2 = Var(\widetilde{y_p}) + (x_p'A\beta_r - x_r'\beta_r)^2. \tag{3-18}$$

Theorem 2:

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

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

The variance of $\hat{y}$ is

$$\begin{aligned} Var(\hat{y}) &= x'(X'V^{-1}X)^{-}x\sigma^2 = (x_p', x_r')\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{pmatrix} x_p \\ x_r \end{pmatrix}\sigma^2 \\ &= x_p'(X_p'V^{-1}X_p)^{-}x_p\sigma^2 + x_p'AA_{22}A'x_p\sigma^2 - x_r'A_{22}A'x_p\sigma^2 \\ &\quad - x_p'AA_{22}x_r\sigma^2 + x_r'A_{22}x_r\sigma^2. \end{aligned} \tag{3-19}$$

where A is same as (3-6), from A_{11} to A_{22} are same as (3-9), (3-10), (3-11) and (3-12). By subtracting (3-17) from (3-19) is

$$Var(\hat{y}) - Var(\widetilde{y_p}) = [A'x_p - x_r]'A_{22}\sigma^2[A'x_p - x_r] \geq 0. \tag{3-20}$$

The property 3 is shown as follows.

The mean squared error of $\hat{y}$ is

$$MSE(\hat{y}) = Var(\hat{y}). \tag{3-21}$$

then we subtract the mean squared error of $\widetilde{y_p}$ from (3-21)

$$MSE(\hat{y}) - MSE(\widetilde{y_p}) = [A'x_p - x_r]'(Var(\widehat{\beta_r}) - \beta_r\beta_r')[A'x_p - x_r]$$

is positive semi-definite, if the matrix $Var(\widetilde{\beta_r}) - \beta_r\beta_r'$ is positive semi-definite. ■

4. Model Misspecification by Overfitting

We consider the general linear model be partitioned as (3-1). If the model include $X_r\beta_r$ when it should be excluded(that is, when $\beta_r = 0$), we say overfitting. When overfitting, the

(generalized) least square solution of $\widehat{\beta_p}$ is

$$\widehat{\beta_p} = (X_p' V^{-1} X_p)^{-} X_p' V^{-1}(y - X_r \widehat{\beta_r}) \tag{4-1}$$

The expectation of $\widehat{\beta_p}$ is

$$E(\widehat{\beta_p}) = E[(X_p' V^{-1} X_p)^{-} X_p' V^{-1}(y - X_r \widehat{\beta_r})] = H_2 \beta_p.$$

where H_2 is same as (3-6)($\because E(\widehat{\beta_r}) = 0$). The expectation of $\widehat{\beta_p}$ is not unique and biased. So, we consider the expectation and variance of $c'\widehat{\beta_p}$

$$\begin{aligned} E(c'\widehat{\beta_p}) &= E[c'(X_p' V^{-1} X_p)^{-} X_p' V^{-1}(y - X_r \widehat{\beta_r})] \\ &= c'(X_p' V^{-1} X_p)^{-} X_p' V^{-1} X_p \beta_p = c'\beta_p \end{aligned} \tag{4-2}$$

$$Var(c'\widehat{\beta_p}) = c'(X_p' V^{-1} X_p)^{-} c\sigma^2 + c' A X_r A_{22} X_r' A' c\sigma^2. \tag{4-3}$$

for any estimable function of β_p.

The unbiased GLS estimator of σ^2 is given by

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{(N-r)}(y - X\widehat{\beta})'(y - X\widehat{\beta}) \\ &= \frac{1}{(N-r)} y'[V^{-1} - V^{-1} X(X' V^{-1} X)^{-} X' V^{-1}] y, \\ E(\widehat{\sigma}^2) &= \sigma^2. \end{aligned}$$

which means $\widehat{\sigma^2}$ is unbiased.

Theorem 3:

1. $\widehat{\beta_p}$ and $\widehat{\sigma^2}$ are unbiased.
2. The matrix $Var(c'\widehat{\beta_p}) - Var(c'\widetilde{\beta_p})$ is positive semi-definite.

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

The variance of $c'\widetilde{\beta_p}$ is

$$Var(c'\widetilde{\beta_p}) = c'(X_p' V^{-1} X_p)^{-} c\sigma^2. \tag{4-4}$$

By subtracting (4-4) from (4-3) is

$$Var(c'\widehat{\beta_p}) - Var(c'\widetilde{\beta_p}) = c' A X_r A_{22} X_r' A' c\sigma^2.$$

Note that the matrix A_{22} is positive semi-definite ($A_{22} = \frac{1}{\sigma^2} Var(\widehat{\beta_r})$) by (3-12). ■

5. Misspecification in 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

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

The least squares solutions for β and τ are

$$\hat{\beta} = [Z'QZ]^{-1}Z'Qy,$$
$$\hat{\tau} = (X'X)^{-}X'y - (X'X)^{-}X'Z\hat{\beta},$$

where $Q = I - X(X'X)^{-}X'$. The expectation of $\hat{\beta}$ and $\hat{\tau}$ are

$$E(\hat{\beta}) = \beta,\ E(\hat{\tau}) = (X'X)X'X\tau.$$

The expectation of $\hat{\beta}$ is unbiased. But the expectation of $\hat{\tau}$ is not unique and biased. So, we consider the expectation of $c'\hat{\tau}$

$$E(c'\hat{\tau}) = c'\tau,$$

for any estimable function of τ. The mean squared error of $\hat{\beta}$ and $c'\hat{\tau}$ are

$$MSE(\hat{\beta}) = Var(\hat{\beta}) = [Z'QZ]^{-1}\sigma^2,$$
$$MSE(c'\hat{\tau}) = Var(c'\hat{\tau}) = c'(X'X)^{-}X'c\sigma^2 + c'(X'X)^{-}X'WX(X'X)^{-}c,$$

where $W = Z(Z'QZ)^{-1}Z'$, for any estimable function of τ.

We have the following theorem as above theorems (with proofs omitted here),

Theorem 4:

1. $\widetilde{\beta_p}$ and $\tilde{\tau}$ are generally biased.
2. The matrix $Var(\widehat{\beta_p}) - Var(\widetilde{\beta_r})$ is positive semi-definite.

3. If the matrix $Var(\widehat{\beta_r}) - \beta_r\beta_r'$ is positive semi-definite, then $MSE(c'\widehat{\beta_p}) - MSE(c'\widetilde{\beta_p})$ is positive semi-definite.
4. If the matrix $W - W_p$ is positive semi-definite, then $Var(c'\widehat{\tau}) - Var(c'\widetilde{\tau})$ is positive semi-definite, for any estimable function of τ..
5. If the matrix $(W - W_p) - (Z_r\beta_r - Z_pD\beta_r)(Z_r\beta_r - Z_pD\beta_r)'$ is positive semi-definite, then $MSE(c'\widehat{\tau}) - MSE(c'\widetilde{\tau})$ is positive semi-definite, for any estimable function of τ., where $W_p = Z_p(Z_p'QZ_p)^{-1}Z_p'$ and $D = (Z_p'QZ_p)^{-1}Z_p'QZ_r$.

6. Conclusion

The motivation for variable elimination is provided by theorems 1, 2, 3 and 4. That is, if only the variances of parameter estimates and predictions are concerned, the reduced model may be preferable. But, since some estimates are biased, the mean squared errors should be considered. Property 4 of the theorem 1, property 3 of the theorem 2, property 3 and 5 of the theorem 4 describe that the reduced model is better than the full model in the mean squared error sense under some conditions. That is, the gain in precision (reduction of variance) is not offset by the (increased) bias, under some conditions. We will develop the similar result for the error-in-variable linear model as a future work.

References

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

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

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

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

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