Yokohama Mathematical
Journal Vol. 30, 1982

# VARIABLE SELECTION IN REGRESSION ESTIMATION

By

Yasushi Taga and Masayuki Honda

(Received April 8, 1982)

Abstract. In estimating the population mean of the objective variable $y$ using the regression estimator with auxiliary variables $x_1, x_2, \cdots, x_p$, this paper shows that the variance of regression estimator can be minimized by suitable selection of auxiliary variables.

Key Words: Regression; variable selection.

## 1. Introduction.

Problems of variable selection in prediction using regression analysis have been studied in the last several years ([1]–[4]), but those in regression estimation have not yet dealt with thoroughly.

This article shows an optimum procedure for selecting auxiliary variables among $x_1, x_2, \cdots, x_p$ so that the variance of regression estimator of the mean of objective variable $y$ may be minimized.

The essential point of our method lies in solving the following contradiction: If a larger number of auxiliary variables are taken in, the regression model may become more accurate, but the estimation error may become larger, or vice versa.

Thus one may expect to find the optimum selection of auxiliary variables which minimizes the variance of regression estimator. Such procedures are stated in Sections 2 and 3 in both cases where the mean vector $\mu$ of auxiliary variables $x=(x_1, x_2, \cdots, x_p)^t$ is known or unknown (estimated), and two examples are shown in Section 4.

## 2. In the case where $\mu$ is known

In this section let us assume the followings:

**A1)** The $p$-variate auxiliary $x=(x_1, x_2, \cdots, x_p)^t$ and the objective $y$ have the joint density function $f(x, y)$ with mean vector $(\mu^t, \mu_0)^t$ and covariance matrix $\begin{pmatrix} \Sigma & \eta^t \\ \eta & \sigma_{00} \end{pmatrix}$ where $\sum$, $\eta$ and $\sigma_{00}$ denote the covariance matrix of $x$, $\operatorname{cov}(x, y)$ and variance of $y$ respectively.

**A2)** The mean vector $\mu$ of $x$ is known.

**A3)** The regression of $y$ on any subset of variables $(x_{i_1}, x_{i_2}, \cdots, x_{i_k})$ selected from $\{x_1, x_2, \cdots, x_p\}$ is linear, i.e.

$$y=\beta_{i(k)0}+\beta_{i(k)1}x_{i_1}+\beta_{i(k)2}x_{i_2}+\cdots+\beta_{i(k)k}x_{i_k}+\varepsilon_{i(k)}\,, \tag{2.1}$$

where subscript $i(k)$ denotes a subset of $k$ indices $(i_1, i_2, \cdots, i_k)$ selected from the integer set $\{1, 2, \cdots, p\}$.

**A4)** The conditional expectation and variance, given $x_{i(k)}=(x_{i_1}, x_{i_2}, \cdots, x_{i_k})$, of error term $\varepsilon_{i(k)}$ are given by

$$E\{\varepsilon_{i(k)}\,|\,x_{i(k)}\}=0 \quad \text{and} \quad V(\varepsilon_{i(k)}\,|\,x_{i(k)})=\sigma^2_{i(k)}=\sigma_{00}(1-\rho^2_{0\cdot i(k)}) \tag{2.2}$$

where $\rho_{0\cdot i(k)}$ denotes the multiple correlation coefficient between $y$ and $x_{i(k)}$.

Further let us consider that $n$ independent observations $\{(x_{1j}, x_{2j}, \cdots, x_{pj}, y_j);\ j=1, 2, \cdots, n\}$, distributed according to the same density $f(x, y)$ mentioned in A1, have been obtained, based on which the least squares estimator $\hat{\beta}_{i(k)}$ of $\beta_{i(k)}=(\beta_{i(k)1}, \beta_{i(k)2}, \cdots, \beta_{i(k)k})^t$ given in (2.1) may be obtained by

$$\hat{\beta}_{i(k)}=S^{-1}_{i(k)}\hat{\eta}_{i(k)}\,, \tag{2.3}$$

where $S_{i(k)}$ and $\hat{\eta}_{i(k)}$ are given by

$$S_{i(k)}=\frac{1}{n}\sum_{j=1}^{n}(x_{i(k)j}-\bar{x}_{i(k)})(x_{i(k)j}-\bar{x}_{i(k)})^t\,,$$

$$\hat{\eta}_{i(k)}=\frac{1}{n}\sum_{j=1}^{n}(x_{i(k)j}-\bar{x}_{i(k)})(y_j-\bar{y})\,,$$

$$x_{i(k)j}=(x_{i_1j}, x_{i_2j}, \cdots, x_{i_kj})^t\,, \quad (j=1, 2, \cdots, n)$$

$$\bar{x}_{i(k)}=\frac{1}{n}\sum_{j=1}^{n}x_{i(k)j}\,, \qquad \bar{y}=\frac{1}{n}\sum_{j=1}^{n}y_j\,.$$

Then it is easily proved that $\hat{\beta}_{i(k)}$ is conditionally unbiased for $\beta_{i(k)}$ given $x_{i(k)1}$, $x_{i(k)2}, \cdots, x_{i(k)n}$, and its conditional covariance matrix is given by

$$V(\hat{\beta}_{i(k)}\,|\,x_{i(k)1}, x_{i(k)2}, \cdots, x_{i(k)n})=\sigma^2_{i(k)}S^{-1}_{i(k)}/n\,, \tag{2.4}$$

under assumptions A3 and A4.

Now let us define the regression estimator $\tilde{\mu}_{0i(k)}$ of the mean $\mu_0$ of objective variable $y$:

$$\tilde{\mu}_{0i(k)}=\bar{y}-(\bar{x}_{i(k)}-\mu_{i(k)})^t\hat{\beta}_{i(k)} \tag{2.5}$$

for each selection $i(k)$.

Then it can be shown that $\tilde{\mu}_{0i(k)}$ is conditionally unbiased for $\mu_0$ given $x_{i(k)1}, x_{i(k)2}, \cdots, x_{i(k)n}$ and its conditional variance is given by

$$V(\tilde{\mu}_{0i(k)}\,|\,x_{i(k)1}, \cdots, x_{i(k)n})=\frac{\sigma^2_{i(k)}}{n}[1+(\bar{x}_{i(k)}-\mu_{i(k)})^tS^{-1}_{i(k)}(\bar{x}_{i(k)}-\mu_{i(k)})] \tag{2.6}$$

considering (2.4).

Taking the expectation of (2.6) with respect to $x_{i(k)}$'s we can get the variance of $\tilde{\mu}_{0i(k)}$ as

$$V(\tilde{\mu}_{0i(k)})=\frac{\sigma^2_{i(k)}}{n}[1+E\{(\bar{x}_{i(k)}-\mu_{i(k)})^t S^{-1}_{i(k)}(\bar{x}_{i(k)}-\mu_{i(k)})\}] . \tag{2.7}$$

Therefore the optimal selection of auxiliary variables may be defined by the set of indices $i^*(k^*)=(i_1{}^*, i_2{}^*, \cdots, i^*_{k^*})$ such that the variance $V(\tilde{\mu}_{0i(k)})$ of regression estimator $\tilde{\mu}_{0i(k)}$ takes its minimum value when $i(k)=i^*(k^*)$. Such optimal selection, however, could not be found in practical situation, because the right hand side of (2.7) does depend on some population parameters. So we will try to get estimates $\hat{i}^*(k^*)$ of $i^*(k^*)$ using unbiased estimators of $V(\tilde{\mu}_{0i(k)})$ based on the sample in two cases as follows.

(I) *General Case*

An unbiased estimator $\hat{V}_{1i(k)}$ of $V(\tilde{\mu}_{0i(k)})$ is given by

$$\hat{V}_{1i(k)}=\frac{\hat{\sigma}^2_{i(k)}}{n}\{1+(\bar{x}_{i(k)}-\mu_{i(k)})^t S^{-1}_{i(k)}(\bar{x}_{i(k)}-\mu_{i(k)})\} , \tag{2.8}$$

where $\hat{\sigma}^2_{i(k)}$ denotes an unbiased estimator of $\sigma^2_{i(k)}$ given by

$$\begin{aligned} &\hat{\sigma}^2_{i(k)}=\frac{1}{n-k-1}y'^t(I_n-Q_k)y' , \qquad y'=(y_1-\bar{y}, \cdots, y_n-\bar{y})^t , \\ &Q_k=X_{i(k)}(X^t_{i(k)}X_{i(k)})^{-1}X^t_{i(k)} , \qquad X_{i(k)}=(x_{i(k)1}-\bar{x}_{i(k)}, \cdots, x_{i(k)n}-\bar{x}_{i(k)})^t . \end{aligned} \tag{2.9}$$

An estimator $\hat{i}_1{}^*(k^*)$ of $i^*(k^*)$ may be found such that $\hat{V}_{1i(k)}$ takes its minimum value when $i(k)=\hat{i}^*(k^*)$, which is found to show fairly nice property in our simulation experiments in Section 4.

(II) *Normal Case*

Let us assume $x=(x_1, x_2, \cdots, x_p)^t$ is normally distributed with mean vector $\mu$ and covariance matrix $\sum$.

As is well known, the statistic $(n-1)(\bar{x}_{i(k)}-\mu_{i(k)})^t S^{-1}_{i(k)}(\bar{x}_{i(k)}-\mu_{i(k)})$ is distributed according to the Hotelling's $T^2$ distribution and its expectation is equal to $(n-1)k/(n-k-2)$. Therefore the right hand side of (2.7) reduces to $(n-2)\sigma^2_{i(k)}/n(n-k-2)$, and an unbiased estimator $\hat{V}_{2i(k)}$ of $V(\tilde{\mu}_{0i(k)})$ is given by

$$\hat{V}_{2i(k)}=\frac{n-2}{n(n-k-2)}\hat{\sigma}^2_{i(k)} , \tag{2.10}$$

where $\hat{\sigma}^2_{i(k)}$ is given by (2.9).

Then an estimator $\hat{i}_2{}^*(k^*)$ of $i^*(k^*)$ may be found such that $\hat{V}_{2i(k)}$ takes its minimum value when $i(k)=\hat{i}_2{}^*(k^*)$. It is noted here that $\hat{i}_2{}^*(k^*)$ may be applicable to the case where $x=(x_1, x_2, \cdots, x_p)$ is distributed according to a non-normal distribution function having the same moments as those of the normal one up to some suitable order, because of the robustness of Hotelling's $T^2$ statistic (see Appendix A).

## 3. In the case where $\mu$ is unknown (estimated)

In the case where the mean vector $\mu$ of auxiliary variables $x=(x_1, x_2, \cdots, x_p)^t$ is unknown, the first sample $\{(x_j, y_j), j=1, 2, \cdots, N\}$ of size $N$ should be taken from the population based on which an unbiased estimator $\hat{\mu}_{i(k)}$ of $\mu_{i(k)}$ may be constructed at the first stage, and then the second sample $\{(x_{(j)}, y_{(j)}); j=1, 2, \cdots, n\}$ is taken from the first sample without replacement based on which an estimator $\hat{\hat{\beta}}_{i(k)}$ of the regression coefficient vector $\beta_{i(k)}$ and a regression estimator $\tilde{\tilde{\mu}}_{0i(k)}$ of $\mu_0$, the population mean of $y$, are constructed in the second stage as follows. (It is noted here that observations on $x$ or $y$ should be made at the first or the second stage respectively.) In the following let us assume A1, A3 and A4 in Section 2.

An unbiased estimator $\hat{\mu}_{i(k)}$ of $\mu_{i(k)}$ is defined based on the first sample by

$$\hat{\mu}_{i(k)}=\frac{1}{N}\sum_{j=1}^{N} x_{i(k)j}\,, \tag{3.1}$$

where $x_{i(k)j}=(x_{i_1j}, x_{i_2j}, \cdots, x_{i_kj})^t, j=1, 2,. \cdots, N$, and $i(k)=(i_1, i_2, \cdots, i_k)$. In the same way as in (2.3), an estimator $\hat{\hat{\beta}}_{i(k)}$ of $\beta_{i(k)}$ based on the second sample is defined by

$$\hat{\hat{\beta}}_{i(k)}=S^{-1}_{i(k)}\hat{\eta}_{i(k)}\,. \tag{3.2}$$

Combining (3.1) and (3.2) the regression estimator $\tilde{\tilde{\mu}}_{0i(k)}$ of $\mu_0$ is constructed such that

$$\tilde{\tilde{\mu}}_{0i(k)}=\bar{y}-(\bar{x}_{i(k)}-\hat{\mu}_{i(k)})^t\hat{\hat{\beta}}_{i(k)}\,, \tag{3.3}$$

where

$$\bar{y}=\frac{1}{n}\sum_{j=1}^{n} y_{(j)}\,, \qquad \bar{x}_{i(k)}=\frac{1}{n}\sum_{j=1}^{n} x_{i(k)(j)}\,,$$

$$x_{i(k)(j)}=(x_{i_1(j)}, x_{i_2(j)}, \cdots, x_{i_k(j)})^t\,, \qquad j=1, 2, \cdots, n\,.$$

The regression estimator $\tilde{\tilde{\mu}}_{0i(k)}$ just defined above is easily shown to be unbiased for $\mu_0$, and its variance $V(\tilde{\tilde{\mu}}_{0i(k)})$ is given, after some calculations, by

$$\begin{aligned} V(\tilde{\tilde{\mu}}_{0i(k)})&=\frac{1}{n}\,\sigma^2_{i(k)}[1+E\{(\bar{x}_{i(k)}-\hat{\mu}_{i(k)})^tS^{-1}_{i(k)}(\bar{x}_{i(k)}-\hat{\mu}_{i(k)})\}]+\frac{1}{N}\,\beta^t_{i(k)}\Sigma_{i(k)}\beta_{i(k)} \\ &=\frac{1}{n}\,\sigma^2_{i(k)}\Bigg[1+\left(\frac{N-n}{N}\right)^2E\{(\bar{x}_{i(k)}-\mu_{i(k)})^tS^{-1}_{i(k)}(\bar{x}_{i(k)}-\mu_{i(k)}) \\ &\qquad +(\bar{z}_{i(k)}-\mu_{i(k)})^tS^{-1}_{i(k)}(\bar{z}_{i(k)}-\mu_{i(k)})\}\Bigg]+\frac{1}{N}\,\beta_{i(k)}\Sigma_{i(k)}\beta_{i(k)} \end{aligned} \tag{3.4}$$

where $\sigma^2_{i(k)}=\sigma_{00}(1-\rho^2_{0\cdot i(k)})$, $\sigma_{00}=V(y)$, $\rho_{0\cdot i(k)}$ is the multiple correlation coefficient between $y$ and $x_{i(k)}$, $S_{i(k)}$ is the sample covariance matrix of $x_{i(k)}$, and $\bar{z}_{i(k)}=(N\hat{\mu}_{i(k)}-n\bar{x}_{i(k)})/(N-n)$ (Appendix B). The last term in the right-hand side of (3.4) may be negligible if the first sample size $N$ is sufficiently large.

In the case where $x_{i(k)}$ is normally distributed with mean vector $\mu_{i(k)}$ and covarience matrix $\sum_{i(k)}$, (3.4) reduces to

$$(3.5) \qquad V(\tilde{\bar{\mu}}_{0i(k)})=\frac{1}{n}\,\sigma^2_{i(k)}\left[1+\frac{k}{n-k-2}\left(\frac{N-n}{N}\right)^2\left(1+\frac{n}{N-n}\right)\right]+\frac{1}{N}\beta^t_{i(k)}\Sigma_{i(k)}\beta_{i(k)}\,.$$

From (3.4) and (3.5) unbiased estimators of $V(\tilde{\bar{\mu}}_{0i(k)})$ are given by

$$(3.6) \qquad \hat{V}_{3i(k)}=\frac{1}{n}\,\hat{\sigma}^2_{i(k)}[1+(\bar{x}_{i(k)}-\hat{\mu}_{i(k)})^t S^{-1}_{i(k)}(\bar{x}_{i(k)}-\hat{\mu}_{i(k)})]+\frac{1}{N}\,(\hat{\sigma}_{00}-\hat{\sigma}^2_{i(k)})$$

in the general case, and by

$$(3.7) \qquad \hat{V}_{4i(k)}=\frac{1}{n}\,\hat{\sigma}^2_{i(k)}\left[1+\frac{k}{n-k-2}\left(1-\frac{n}{N}\right)\right]+\frac{1}{N}(\hat{\sigma}_{00}-\hat{\sigma}^2_{i(k)})$$

in the normal case, respectively, where

$$\hat{\sigma}_{00}=\frac{1}{n}\sum_{j=1}^{n}(y_{(j)}-\bar{y})^2\,, \qquad \bar{y}=\frac{1}{n}\sum_{j=1}^{n}y_{(j)}$$

and $\hat{\sigma}^2_{i(k)}$ is an unbiased estimator of $\sigma^2_{i(k)}$ based on the second sample as defined in (2.9).

Then estimators $\hat{i}_3{}^*(k^*)$ or $\hat{i}_4{}^*(k^*)$ of $i^*(k^*)$ may be found such that $\hat{V}_{3i(k)}$ or $\hat{V}_{4i(k)}$ are minimized respectively. As noted in Section 2, $\hat{i}_4{}^*(k)$ may be applicable to the non-normal case where $x_{i(k)}$'s have the same moments as those of the normal (statures) of 82 male students in some junior high school.

## 4. Examples

Two examples in case of simple random sampling are shown below under the assumption that $x=(x_1, x_2, \cdots, x_p)^t$ is a $p$-variate normal random vector (exactly or approximately).

The 1st example is quite artificial one in which interesting results are obtained on multicollinearity. The 2nd example is based on actual measurement data on heights (statures) of 82 male students in some junior high school.

*Example* 1.

Let us assume $(y, x_1, \cdots, x_p)^t$ is distributed exactly according to the $(p+1)$-variate normal distribution with mean vector $(\mu_0, \mu_1, \cdots, \mu_p)$ and covariance matrix $\sum$ where $V(y)=V(x_i)=\sigma_{00}$, $\mathrm{Cov}(y, x_i)=\rho_0\sigma_{00}$, $\mathrm{Cov}(x_i, x_j)=\rho\sigma_{00}$, $(i, j=1, 2, \cdots, n;$ $i\neq j)$. Then it is easily shown that the variance $V(\tilde{\mu}_{0i(k)})$ of the regression estimator $\tilde{\mu}_{0i(k)}$ given in (2.7) is obtained as

$$(4.1) \qquad V(\tilde{\mu}_{0i(k)})=\frac{(n-2)\sigma_{00}}{n(n-k-2)}\left(1-\frac{k\rho_0{}^2}{1+(k-1)\rho}\right), \qquad k=0, 1, \cdots, p.$$

The optimum number $k^*$ of variable selection are obtained by minimizing $V(\tilde{\mu}_{0i(k)})$, i.e. calculating

$$g_n(k)=\frac{1}{n-k-2}\left(1+\frac{k\rho_0{}^2}{1+(k-1)\rho}\right),\qquad k=0,1,\cdots,p \tag{4.2}$$

for $n=10$, 25, 50, 75, 100, 200, 1000; $\rho_0=-0.9(0.1)0.9$; $\rho=0.0(0.1)0.9$. Table 1 shows the optimum number $k^*$'s for $\rho_0=0.3$ and $\rho_0=0.5$. $(p=5)$

Looking at Table 1, we can easily see that the effect of multicollinearity appears remarkably for small $n$ even though $\rho$ is fairly low, and that it appears quite a little for larger $n$ even though $\rho$ is fairly high.

Table 1. Optimum number $k^*$ $(p=5)$

| $\rho_0$ | $\rho$ | $n$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 10 | 25 | 50 | 75 | 100 | 200 | 1000 |
| 0.3 | 0.1 | 0 | 5 | 5 | 5 | 5 | 5 | 5 |
| | 0.3 | 0 | 2 | 4 | 5 | 5 | 5 | 5 |
| | 0.5 | 0 | 1 | 2 | 3 | 3 | 5 | 5 |
| | 0.7 | 0 | 1 | 1 | 2 | 2 | 3 | 5 |
| | 0.9 | 0 | 1 | 1 | 1 | 1 | 2 | 3 |
| 0.5 | 0.1 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| | 0.3 | 2 | 5 | 5 | 5 | 5 | 5 | 5 |
| | 0.5 | 1 | 3 | 5 | 5 | 5 | 5 | 5 |
| | 0.7 | 1 | 2 | 3 | 4 | 4 | 5 | 5 |
| | 0.9 | 1 | 1 | 1 | 2 | 2 | 3 | 5 |

*Example* 2.

Let us apply our method to the regression estimation of the mean $\mu_0$ of height $y$, in the population of 82 male students of the first grade in 1977 at some junior high school, with auxiliary variables $x_1, x_2, \cdots, x_5$, where $x_i$ denotes the height when they were in the $i$th grade at their elementary schools.

We made simulation experiments in the following way:

1) Take a sample of size $n$ with replacement from the population repeatedly 100 times, so that 100 sets of samples are obtained for each $n=10$, 20, 30, 40, 50.
2) Compute the regression estimate $\tilde{\mu}_{0i(k)}$ and its variance estimates $\hat{V}_{i(k)1}$ and $\hat{V}_{2i(k)}$ for all possible selection sets $i(k)$'s of indices.
3) Count frequencies $f_1, f_2$, $\hat{i}^*(k^*)$ minimizing $\hat{V}_{1i(k)}$ or $\hat{V}_{2i(k)}$, and calculate the mean values of $\hat{V}_{1i(k)}$ and $\hat{V}_{2i(k)}$ with respect to 100 sets of samples which give approximate values to $V(\tilde{\mu}_{0i(k)})$ in general case and in normal case respectively.
4) Calculate the mean square errors of regression estimator $\tilde{\mu}_{0\hat{i}^*(k^*)}$ in both cases where $\hat{i}^*(k^*)$ is obtained by minimizing $\hat{V}_{1i(k)}$ or $\hat{V}_{2i(k)}$.

The results are summarized as follows.

a) Optimally selected sets of variables: $\{x_5\}$ for $n=10$, 20, 30, using both $\hat{V}_{1i(k)}$ and $\hat{V}_{2i(k)}$, $\{x_4, x_5\}$ for $n=40$, 50, using both $\hat{V}_{1i(k)}$ and $\hat{V}_{2i(k)}$.

b) Mean square errors.

Table 2. Results by simulation

| $n$ | MSE by $\hat{V}_{1i(k)}$ (cm$^2$) | MSE by $\hat{V}_{2i(k)}$ (cm$^2$) | $\hat{V}(\bar{y})$ (cm$^2$) |
|---|---|---|---|
| 10 | 1.5767 | 2.0438 | 7.7867 |
| 20 | 0.5835 | 0.6064 | 3.7224 |
| 30 | 0.3049 | 0.3311 | 2.4228 |
| 40 | 0.2272 | 0.2467 | 1.8191 |
| 50 | 0.1630 | 0.1917 | 1.4692 |

(Note): $\hat{V}(\bar{y})$ denotes the estimated variance of the sample mean $\bar{y}$ (in case no auxiliary variable is used).

## 5. Discussion

In this article the method for optimum selection of auxiliary variables in regression estimation is stated from the standpoint of multivariate analysis in the infinite population.

However it must be interesting and important to treat the same problem from the other standpoint of sampling theory in the finite population with "super-population model" as found in [7], which we are going to study in the near future.

*Acknowledgement.* The authors express their hearty thanks to the referee for his comments, and to Misses A. Inako and K. Kaneda for preparing typescript.

## References

[1] Akaike, H.: *Information theory and extension of the maximum likelihood principle*, Proc. 2nd Inter. Symp. on Information Theory (B. N. Petrov and F. Csaki eds.), Akademia Kaido, (1973), 268–281.

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

[3] Mallows, C. L.: *Some comments on $C_p$*, Technometrics, **15** (1973), 661–675.

[4] Oliker, V. I.: *On the relationship between the sample size and the number of variables in a linear regression model*, Commun. Statist.—Theory and Method, **A7(6)** (1978), 509–516.

[5] Raj, D.: *On a method of using multi-auxiliary information in sample surveys*, JASA, **60** (1965), 270–277.

[6] Rao, P. S. R. S.: *On two-phase regression estimator*, Sankhyā (A), **34** (1972), 473–476.

[7] Royall, R. M.: *Linear regression models in finite population sampling theory*, Foundations of Statistical Inference (V. P. Godambe and D. A. Sprott eds.), Holt, Rinehart and Winston of Canada, (1971), 259–279.

Department of Mathematics
Yokohama City University
Yokohama 236, Japan

and

Division of Medical Information Science
Chiba University Hospital
Chiba 281, Japan

**Appendix A.** Hotelling's $T_p{}^2$ statistic has the Taylor expansion up to the $r$th degree as follows:

$$\begin{aligned}T_p{}^2 &= n(\bar{x}-\mu)^t\left(\frac{n}{n-1}S\right)^{-1}(\bar{x}-\mu)\\ &= n(\bar{x}-\mu)^t\left\{I_p+\left(I_p-\frac{n}{n-1}S\right)+\cdots+\left(I_p-\frac{n}{n-1}S\right)^r\right.\\ &\qquad\left.+\frac{n-1}{n}S^{-1}\left(I_p-\frac{n}{n-1}S\right)^{r+1}\right\}(\bar{x}-\mu)\end{aligned}$$

where $I_p$ denotes the unit matrix of order $p$. It is noted here that $x=(x_1, \cdots, x_p)^t$ may be supposed to have the zero mean and the unit covariance matrix without loss of generality by considering the transformation $z=\sum^{-1/2}(x-\mu)$.

Assume that $x$ has the same moments as those of the $p$-variate normal distribution with zero mean and unit covariance matrix up to the sixth order, then the expectation of $T_p{}^2$ is approximated by using the above expansion as in the following (in case $r=2$):

$$E\{T_p{}^2\} \fallingdotseq p\left\{1+\frac{p+1}{n}+\frac{1}{n^2}(4p^2-5p+3)-\frac{1}{n^2(n-1)}(2p^2-13p+9)\right\}.$$

This result is identical to the exact expectation of $T_p{}^2$ up to the order of $O(n^{-1})$. Indeed, $E\{T_p{}^2\}$ is represented as

$$E\{T_p{}^2\}=\frac{(n-1)p}{n-p-2}=p\left(1-\frac{p+1}{n-1}\right)^{-1}=p\left\{1+\frac{p+1}{n}+\frac{1}{n^2}(p+1)(p+2)+\cdots\right\}.$$

In case $r=4$, the same approximation of $E\{T_p{}^2\}$ as above is expected to be identical to the exact $E\{T_p{}^2\}$ up to the order of $O(n^{-2})$, as examined correctly as such when $p=1$.

**Appendix B.** Let us denote the first and second samples by $\{(\xi_j, \eta_j); j=1, 2, \cdots, N\}$ and $\{(x_\alpha, y_\alpha); \alpha=1, 2, \cdots, n\}$ respectively, where the latter is drawn from the former equally without replacement. Further denote the sample means of the 1st and 2nd sample by $(\bar{\xi}, \bar{\eta})$ and $(\bar{x}, \bar{y})$ respectively. (In this appendix the suffix $i(k)$ expressing variable selection will be omitted for simplification.)

Taking the conditional expectation of the regression estimator $\tilde{\tilde{\mu}}_0$ defined by (3.3), given $\xi=(\xi_1, \xi_2, \cdots, \xi_N)$, we can get easily

$$E\{\tilde{\tilde{\mu}}_0 \mid \xi\}=\binom{N}{n}^{-1}\sum_{\nu=1}^{\binom{N}{n}} E\{\bar{y}_{(\nu)}-(\bar{x}_{(\nu)}-\bar{\xi})^t\hat{\hat{\beta}}_{(\nu)} \mid \xi\}=\mu_0+(\bar{\xi}-\mu)^t\beta\,,$$

and then

$$E\{\tilde{\tilde{\mu}}_0\}=\mu_0\,,$$

noting that

$$\binom{N}{n}^{-1}\sum_{\nu}\bar{y}_{(\nu)}=\bar{\eta}\,,\qquad \binom{N}{n}^{-1}\sum_{\nu}\bar{x}_{(\nu)}=\bar{\xi}\,,$$

and $E\{\hat{\hat{\beta}}_{(\nu)}\,|\,\xi\}=\beta$, where the suffix ($\nu$) denotes the 2nd sample in the $\nu$th order taken from the 1st sample.

Then the variance of $\tilde{\tilde{\mu}}_0$ is obtained as

$$\text{(B.1)}\qquad V(\tilde{\tilde{\mu}}_0)=V(E\{\tilde{\mu}_0\,|\,\xi\})+E\{V(\tilde{\tilde{\mu}}_0\,|\,\xi)\}\,.$$

The 1st and 2nd terms of the right-hand side of (B.1) is easily

$$\text{(B.2)}\qquad V(\mu_0+(\bar{\xi}-\mu)^t\beta)=\frac{1}{N}\,\beta^t\textstyle\sum\beta\,,$$

and

$$\text{(B.3)}\qquad \begin{aligned} E\{V(\tilde{\tilde{\mu}}_0\,|\,\xi)\}&=\frac{\sigma^2}{n}\left[1+\binom{N}{n}^{-1}\sum_{\nu}E\{(\bar{x}_{(\nu)}-\bar{\xi})^t S_{(\nu)}^{-1}(\bar{x}_{(\nu)}-\bar{\xi})\}\right]\\ &=\frac{\sigma^2}{n}\left[1+\left(1-\frac{n}{N}\right)^2 E\{(\bar{x}-\mu)^t S^{-1}(\bar{x}-\mu)+(\bar{z}-\mu)^t S^{-1}(\bar{z}-\mu)\}\right], \end{aligned}$$

where $\bar{z}=(N\bar{\xi}-n\bar{x})/(N-n)$. Substituting (B.2) and (B.3) into (B.1) we can get (3.4).