A COMPARATIVE STUDY ON COUNTER-MEASURES FOR MULTICOLLINEARITY IN REGRESSION ANALYSIS

Morito TSUTSUMI
Research Associate
Department of Civil Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113 Japan
Fax: +81-3-5689-7290
E-mail:tsutsumi@planner.t.u-tokyo.ac.jp

Eihan SHIMIZU
Associate Professor
Department of Civil Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113 Japan
Fax: +81-3-5689-7290
E-mail : shimizu@planner.t.u-tokyo.ac.jp

Yasutaka MATSUBA
Taisei Corporation
Nishi-Shinjuku, Shinjuku-ku, Tokyo,
163 Japan

abstract: For transport and regional analysis with regression model, multicollinearity is one of the most difficult problems. Many conventional counter-measures have been provided. Among them are ridge regression and principal component regression which are the most well-known ones. There has not been, however, a systematic comparative study in both their theoretical and practical aspects.
This paper discusses the theoretical relationship between ridge and principal component regression in the common framework of inverse analysis and gives a practical comparison through the parameter estimation of land price function.

1. INTRODUCTION

For economic analysis of transport projects, regression models are widely used. It is rare, however, that the data we have for estimating a regression model conform exactly to the theory underlying the model. Multicollinearity is one of the most difficult problems and many conventional counter-measures have been provided. Among them are ridge regression and principal component regression which are the most well-known ones. There has not been, however, a systematic comparative study in both their theoretical and practical aspects.

This paper focuses on the multicollinearity problem and attempts to compare the theoretical and practical characteristics between ridge regression and principal component regression.

The next chapter gives the explanations of these two counter-measures. Chapter 3, which is the distinguished part of the paper, discusses their theoretical relationship in the common framework of the inverse analysis. In the following chapter, ridge regression and principal components regression are applied to a common parameter estimation problem of land price function and their practical characteristics are demonstrated.

2. CONVENTIONAL COUNTER-MEASURES FOR MULTICOLLINEARITY

2.1 Multicollinearity

Consider the standard model for multiple linear regression,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of explained variable; $\mathbf{X}$ is an $n \times m$ matrix of non-stochastic variables of rank m ($<n$); $\boldsymbol{\beta}$ is a $m \times 1$ parameter vector; and $\mathbf{u}$ is an $n \times 1$ vector of residuals with

$$E(\mathbf{u}) = \mathbf{0} \tag{2}$$

$$Var(\mathbf{u}) = \sigma^2\mathbf{I} \tag{3}$$

Without essential loss of generality, let the X-variables be standardized, so that the diagonal elements of $\mathbf{X}'\mathbf{X}$ are all 1. The Ordinary Least Squares (OLS) is to minimize the sum of squares of residual

$$\min_{\boldsymbol{\beta}} \quad \Phi(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \tag{4}$$

OLS estimator $\boldsymbol{\beta}_O$ is obtained by solving the normal equation

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \tag{5}$$

yielding

$$\boldsymbol{\beta}_O = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{6}$$

$$E(\boldsymbol{\beta}_O) = \boldsymbol{\beta}^* \tag{7}$$

$$Var(\boldsymbol{\beta}_O) = E[(\boldsymbol{\beta}_O - \boldsymbol{\beta}^*)(\boldsymbol{\beta}_O - \boldsymbol{\beta}^*)'] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{8}$$

where $\boldsymbol{\beta}^*$ is the true value of $\boldsymbol{\beta}$.
The mean squared length of $\boldsymbol{\beta}_O$ is

$$E(\boldsymbol{\beta}_O^t\boldsymbol{\beta}_O) = \boldsymbol{\beta}^{*\prime}\boldsymbol{\beta}^* + \sigma^2 tr\,(\mathbf{X}'\mathbf{X})^{-1} \tag{9}$$

$$= \boldsymbol{\beta}^{*\prime}\boldsymbol{\beta}^* + \sigma^2\sum_{j=1}^{m}\left(\frac{1}{\lambda_j}\right) > \boldsymbol{\beta}^{*\prime}\boldsymbol{\beta}^* + \frac{\sigma^2}{\lambda_{\min}} \tag{10}$$

where λ_j is eigenvalue of $\mathbf{X}'\mathbf{X}$ (Maddala (1988)).

Define $\mathbf{P}$ to be the orthogonal matrix which puts ($\mathbf{X}'\mathbf{X}$) in diagonal form, that is,

$$\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P} = \boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)\ ;\ \mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I} \tag{11}$$

Thus,

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}' \qquad (\boldsymbol{\Lambda}^{-1} = diag(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_m})) \tag{12}$$

The situation where the explanatory variables are highly intercorrelated is referred to as multicollinearity.

OLS is a good estimation procedure if $\mathbf{X}'\mathbf{X}$ is nearly a unit matrix. However, when multicollinearity occurs, OLS estimator is too sensitive to errors in data. $Var(\boldsymbol{\beta}_O)$ and mean squared length of $\boldsymbol{\beta}_O$ given by (8) and (9) are extremely large.
The obvious remedy for this problem is to drop variables suspected of causing the problem from the regressions, which might be frequently used. Other approaches are suggested, which we consider further.

2.2 Ridge Regression

Ridge Regression (RR) (Hoerl and Kennard (1959) (1962)) is an estimation procedure based upon

$$(\mathbf{X}^t\mathbf{X} + k\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y} \tag{13}$$

yielding

$$\begin{aligned}\boldsymbol{\beta}_R &= (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{y} \\ &= (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}^*\end{aligned} \tag{14}$$

where $\mathbf{I}$ is a unit matrix, and the scalar k (> 0) is called ridge parameter which is chosen arbitarily.

$$\begin{aligned}E(\boldsymbol{\beta}_R) &= (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}\, E(\boldsymbol{\beta}_O) \\ &= (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}^*\end{aligned} \tag{15}$$

$$\begin{aligned}Var(\boldsymbol{\beta}_R) &= (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}\, Var(\boldsymbol{\beta}_O)\, \{ (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{X}\}^t \\ &= \sigma^2\, (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\, (\mathbf{X}^t\mathbf{X})\, (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\end{aligned} \tag{16}$$

Ridge estimator is biased,

$$\begin{aligned}B(\boldsymbol{\beta}_R) &\equiv E(\boldsymbol{\beta}_R) - \boldsymbol{\beta}^* \\ &= -k\, (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\boldsymbol{\beta}^* \\ &\neq 0\end{aligned} \tag{17}$$

However, each diagonal component of $Var(\boldsymbol{\beta}_R)$ is always less than that of $Var(\boldsymbol{\beta}_O)$.

Ridge regression is to be explained in some ways such as constrained least squares, and Bayesian interpretation (Maddala (1988)).

Ridge estimator $\boldsymbol{\beta}_R$ is the solution to the problem:

$$\min_{\boldsymbol{\beta}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \tag{18}$$

$$s.t. \quad \|\boldsymbol{\beta}\|^2 = c^2 \quad (const.)$$

Introducing the Lagrangian we find that the above problem is equivalent to

$$\min_{\boldsymbol{\beta},k} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + k(\|\boldsymbol{\beta}\|^2 - c^2) \tag{19}$$

where k is a Lagrange multiplier.

Differentiating this expression (19) with respect to $\boldsymbol{\beta}$ equating the derivation zero, we get the equation (13) and $\boldsymbol{\beta}_R$ as a constrained least squares estimator. $\boldsymbol{\beta}_R$ is also called damped squares solution (Menke(1989)). However, it is rare that we have a prior information of norm of $\boldsymbol{\beta}$. Constrained least squares shall be discussed in chapter 3 again.

If we assume that the prior information is of the form that $\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_B^2\mathbf{I})$, the Bayesian approach also leads ridge estimator (Maddala (1988)).

Generalization of RR estimator is able to be considered, which is called shrinkage estimator (Chatterjee and Price (1977)) .But we don't mention it in this paper. Shrinkage estimator is discussed in the paper of Goldstein and Smith (1973) in detail.

The most essential and difficult problem of ridge regression is how to determine ridge parameter k, which shall be discussed in section 3.5.

2.3 Principal Component Regression

Another remedy suggested and used for the multicollinearity problem is the Principal Component Regression (PCR). We consider linear functions of variables,

$$\mathbf{Z} = \mathbf{X}\mathbf{P} \tag{20}$$

where $\mathbf{Z}$ is a principal component $n \times m$ matrix.
Suppose that we use l ($<m$) principal components of the m columns of $\mathbf{Z}$,. Thus, we regress $\mathbf{y}$ on $\mathbf{X}\,\mathbf{P}_L$, where $\mathbf{P}_L$ is a $n \times l$ matrix containing l characteristic vectors of $\mathbf{X}'\mathbf{X}$.

$$\mathbf{y} = \mathbf{Z}_L\,\boldsymbol{\alpha} \tag{21}$$

The estimator is

$$\boldsymbol{\alpha} = (\mathbf{Z}_L{}^t\,\mathbf{Z}_L)^{-1}\,\mathbf{Z}_L{}^t\,\mathbf{y} \tag{22}$$

Obviously the estimator given by principal components regression is

$$\begin{aligned}\boldsymbol{\beta}_s &= \mathbf{P}_L\,\boldsymbol{\alpha}\\ &= \mathbf{P}_L\,(\mathbf{P}_L{}^t\,\mathbf{X}^t\mathbf{X}\mathbf{P}_L)^{-1}\,\mathbf{P}_L{}^t\,\mathbf{X}^t\mathbf{y}\end{aligned} \tag{23}$$

PCR estimator is also a biased estimator

$$\begin{aligned}E(\boldsymbol{\beta}_s) &= \mathbf{P}_L(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\,\mathbf{P}_L{}^t\,\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}^*\\ &\neq \boldsymbol{\beta}^*\end{aligned} \tag{24}$$

However, each diagonal component of $Var(\boldsymbol{\beta}_s)$ is also always less than that of $Var(\boldsymbol{\beta}_O)$.

$$Var(\boldsymbol{\beta}_s) = \sigma^2\mathbf{P}_L\boldsymbol{\Lambda}_L{}^{-1}\mathbf{P}_L{}^t \tag{25}$$

There are some problems with PCR as well (Maddala (1988)).

3. THEORETICAL COMPARISON

3.1 Properties of Diagonal Matrices

Let $\mathbf{X}^t\mathbf{X}$ have eingevalues that is grouped qualitatively into two types－substantially greater than zero (λ_j ($1 \leqq j \leqq r$)), slightly greater than zero(λ_j ($r+1 \leqq j \leqq m$)).
In ridge regression, the matrix $(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}$ is to be diagonalized below (Marquart (1970)).

$$(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1} = \mathbf{P}(\boldsymbol{\Lambda} + kI)^{-1}\mathbf{P}^t = \mathbf{P}(\boldsymbol{\Lambda}_{(R)})^{-1}\mathbf{P}^t = \sum_{j=1}^{m}\frac{1}{\lambda_j + k}\mathbf{p}_j\mathbf{p}_j{}^t \tag{26}$$

where $\mathbf{p}_j$ is characteristic vector of $\mathbf{X}^t\mathbf{X}$ corresponding to λ_j.
And in principal component regression

$$(\mathbf{X}^t{}_{(S)}\mathbf{X}_{(S)})^{-1} = \mathbf{P}(\boldsymbol{\Lambda}_{(S)})^{-1}\mathbf{P}^t = \sum_{j=1}^{r}\frac{1}{\lambda_j}\mathbf{p}_j\mathbf{p}_j{}^t \tag{27}$$

$\mathbf{P}$ and $\mathbf{P}^t$ work as linear operators that do not change the scale or Euclidian norm of vectors. The sensitivity of estimator depends upon the diagonal matrices $\boldsymbol{\Lambda}^{-1}, \boldsymbol{\Lambda}_{(R)}{}^{-1}, \boldsymbol{\Lambda}_{(S)}{}^{-1}$. Fig.1 helps us to understand it intuitively.

Even when λ_j($r+1 \leqq j \leqq m$) is precisely zero, generalized inverse of $\mathbf{X}^t\mathbf{X}$ is to be defined

$$(\mathbf{X}^t\mathbf{X})^-_{\mathrm{MP}} = \mathbf{P}\boldsymbol{\Lambda}^-\mathbf{P}^t = \sum_{j=1}^{r}\frac{1}{\lambda_j}\mathbf{p}_j\mathbf{p}_j{}^t \tag{28}$$

where $\boldsymbol{\Lambda}^-$ is generalized inverse matrix of $\boldsymbol{\Lambda}$

$$\boldsymbol{\Lambda}^- = diag(\lambda_1, \lambda_2, \cdots, \lambda_r, 0, \cdots, 0) \tag{29}$$

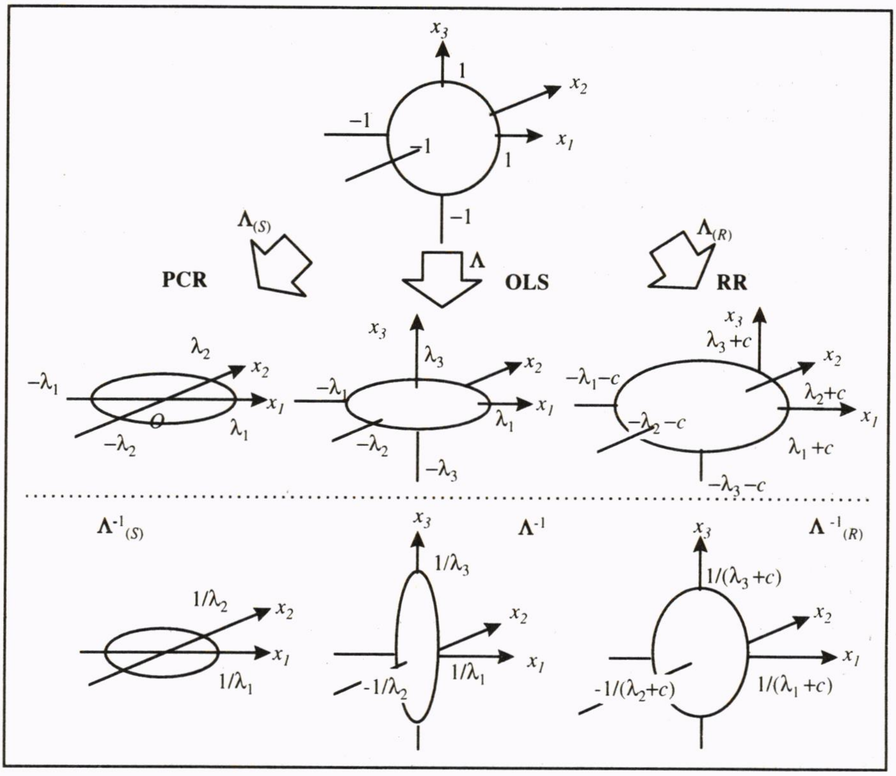

Fig.2 Comparison among OLS, RR, and PCR.

3.2 *r-k* estimator

Ridge regression and principal component regression can be combined. Baye and Parker (1984) proposed *r-k* class estimator for $\boldsymbol{\beta}$, which includes the OLS estimator, RR estimator and PCR estimator as special cases. This estimator is given as

$$\boldsymbol{\beta}_r(k) = \mathbf{P}_r(\mathbf{P}_r\mathbf{X}'\mathbf{X}\mathbf{P}_r + k\mathbf{I})^{-1}\mathbf{P}_r'\mathbf{X}'\mathbf{y} \qquad (30)$$

Special cases of the *r-k* class estimators are as follows (Sarkar (1989)):

(i) $\boldsymbol{\beta}_m(0) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta}_O$ (31)

(ii) $\boldsymbol{\beta}_m(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta}_R$ (32)

(iii) $\boldsymbol{\beta}_l(0) = \mathbf{P}_L(\mathbf{P}_L'\mathbf{X}'\mathbf{X}\mathbf{P}_L)^{-1}\mathbf{P}_L'\mathbf{X}'\mathbf{y} = \boldsymbol{\beta}_s$ (33)

3.3 Inverse and Ill-posed Problem

We have until now discussed from the chiefly statistical point of view. It is interesting and helpful to discuss from another viewpoint.

Again consider the standard model below.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} \qquad (34)$$

When we treat this model as a direct problem, estimating parameter $\boldsymbol{\beta}$ is an inverse problem. The problem (34) is said to be *correct, correctly posed* or *well-posed* if the following two conditions hold:

(a) for each $\mathbf{y}$ the equation has a unique solution
(b) the solution of (34) is stable, i.e. the operator $\mathbf{X}^{-1}$ is defined on all of the space which $\mathbf{y}$ belongs and is continuous.

Otherwise, the problem (34) is said to be *incorrectly posed* or *ill-posed*. As the number of data n is greater than that of parameter m, we are not able to solve this equation and determine $\boldsymbol{\beta}$ generally. Thus, parameter estimation is a typical example of an ill-posed problem.

One way to get an approximate solution is least squares method which minimize the sum of squared residuals

$$\min_{\boldsymbol{\beta}} \quad \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 \tag{35}$$

Equating its first derivatives with respect to $\boldsymbol{\beta}$ to zero yields normal equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta}=\mathbf{X}'\,\mathbf{Y} \tag{36}$$

If $\mathbf{X}'\mathbf{X}$ is not singular, the solution is

$$\boldsymbol{\beta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{Y} \tag{37}$$

Thus $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is an inverse operator, which is called least squares generalized inverse of $\mathbf{X}$. However, if $\mathbf{X}'\mathbf{X}$ is singular or near singular, inverse problem (34) is again ill-posed.

There are two typical ways to improve the ill-posedness:

(i) to change the solution space introducing some constrained conditions
(ii) to change the operator and solve the related problem

Constrained least squares method discussed in section 2.2 is interpreted as type (i).

There are also some methods in type (ii). Truncated Singular Value Decomposition (TSVD) is one of traditional methods in numerical analysis (Groetsch (1993)). As singular value of square matrix is called eigenvalue, principal component regression is equivalent to TSVD.
Another method in type (ii) is to solve the problem of minimizing a smoothing function, which is called Tikhonov regularization.

3.4 Tikhonov Regularization

The idea of Tikhonov regularization is given by

$$\min_{\boldsymbol{\beta}} \quad \Phi^{\gamma}(\boldsymbol{\beta})=\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\gamma\cdot f(\boldsymbol{\beta}) \tag{38}$$

instead of the problem (35), where $f(\boldsymbol{\beta})$ a smoothing function and γ is a regularization parameter. Tikhonov regularization is applicable to many situations, for it is an abstract theory generalized to the discussion in Hilbert space. For example, Levenberg-Marquardt method in optimization theory is interpreted as Tikhonov regularization (Groetsch (1993)). If we assume that

$$f(\boldsymbol{\beta})=\|\boldsymbol{\beta}\|^2, r=k \tag{39}$$

the problem (39) is equivalent to ridge regression as we consider section 2.3.

The form $(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'$ in RR estimator has a property that it leads an approximate Moore-Penrose generalized inverse $\mathbf{X}^{-}_{MP}$

$$\mathbf{X}^{-}_{MP} = \lim_{\delta\to 0}(\mathbf{X}'\mathbf{X}+\delta\mathbf{I})^{-1}\mathbf{X}' \tag{40}$$

M-P g-inverse is a solution of the problem

$$\begin{aligned} &\mathbf{X}^{-}_{MP} = \arg\{\min_{\mathbf{X}^{-}\in C}\|\boldsymbol{\beta}'\|^2 = \|\mathbf{X}^{-}\mathbf{y}\|^2\} \\ &C = \{X^{-}|\boldsymbol{\beta}' = \mathbf{X}^{-}\mathbf{y}\}, \boldsymbol{\beta}' \in B = \{\boldsymbol{\beta}|\boldsymbol{\beta} = \arg\{\min_{\beta}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2\}\end{aligned} \tag{41}$$

Using normal equation, the estimator with M-P g-inverse is given by

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \ \|\boldsymbol{\beta}\|^2 \\ &s.t. \ \ \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}\end{aligned} \tag{42}$$

equivalently

$$\min_{\boldsymbol{\beta},\mathbf{k}} \ \|\boldsymbol{\beta}\|^2 + \mathbf{k}'\left(\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\mathbf{y}\right) \tag{43}$$

$\boldsymbol{k}$: Lagrange multiplier vector

This equation (40) is also interpreted as Tikhonov regularization(Groetsch C. W. (1993)).

If $\mathbf{X}'\mathbf{X}$ is singular and k is small enough, the ridge estimator is an approximate solution with M-P g-inverse (Marquart (1970)).

3.5 Choosing a Good Regularization Parameter

Regularization for ill-posed problem is a kind of compromise, it is necessary to introduce some criteria for the choice of better alternative. As the estimation error is generally unavoidable, our hope is that, by accepting some bias, we can achieve a larger reduction in variance and an overall reduction in mean square error (MSE)

$$\begin{aligned} \mathrm{MSE}(\boldsymbol{\beta}) &\equiv E[(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'(\boldsymbol{\beta}-\boldsymbol{\beta}^*)] \\ &= \mathrm{tr}\,[Var(\boldsymbol{\beta})] + \boldsymbol{B}'(\boldsymbol{\beta})\,\boldsymbol{B}(\boldsymbol{\beta}) \\ &= \text{variance} + (\text{bias})^2 \end{aligned} \tag{44}$$

The variance term in $\mathrm{MSE}(\boldsymbol{\beta})$ is a monotonic decreasing function of k. The bias term in $\mathrm{MSE}(\boldsymbol{\beta})$ is a monotonic increasing function of k. Thus, as k increases away from zero, the variance decreased, and the bias increases. Hoerl and Kennard (1970 a),(1970 b) show that there always exists a constant $k>0$ such that

$$\begin{aligned} \mathrm{MSE}(\boldsymbol{\beta}_R) &= tr[\sigma^2\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\,(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}] + k^2\boldsymbol{\beta}^{*\prime}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta}^* \\ &< \mathrm{MSE}(\boldsymbol{\beta}_O) \end{aligned} \tag{45}$$

However, the parameter k which minimizes MSE or any other nontrivial quadratic loss function depends on σ^2 and unknown $\boldsymbol{\beta}$.

Among several proposed methods for determination of a value of k that gives a better $\boldsymbol{\beta}$, ridge trace (Hoerl and Kennard (1970a)) has been commonly used. Ridge trace is a plot of the $\beta_j(k)$ shown in Fig.2. Hoerl and Kennard suggested that the best strategy is the choice of a value of k that stabilizes ridge trace. As Minotani(1992) points out, however, this method is extremely subjective and the choice of k that stabilizes ridge trace may bear large bias.

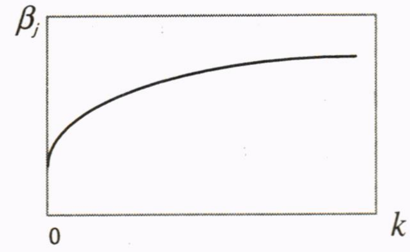

Fig.2 Ridge Trace.

Another criterion is the Prediction Sum of Squares (PRESS). Allen's PRESS estimate of k (Allen (1974)) is an ordinary cross-validation estimate, which is the minimizer of

$$PRESS(k) = \frac{1}{n}\sum_{i=1}^{m}\left(\left[\mathbf{X}\boldsymbol{\beta}^{(i)}(k)\right]_i - y_i\right)^2 \tag{46}$$

where $\boldsymbol{\beta}^{(i)}(k)$ is a ridge estimate of $\boldsymbol{\beta}$ with the ith data y_i omitted, $[\mathbf{X}\,\boldsymbol{\beta}^{(i)}(k)]_i$ is the ith component of $\mathbf{X}\,\boldsymbol{\beta}^{(i)}(k)$. Although the idea of PRESS is intuitively appealing, it would not do very well in the case where $\mathbf{X}\,(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t$ is near diagonal (Golub et al.(1979)).

Golub et al (1979) suggested that General Cross Validation (GCV) for obtaining a good estimate of regularization parameter. The GCV estimate of k is the minimizer of

$$V(k) = \frac{\frac{1}{n}\left\|(\mathrm{I} - \mathbf{A}(k)\mathbf{y})\right\|^2}{\left[\frac{1}{n}Trace(\mathrm{I} - \mathbf{A}(k))\right]^2} \tag{47}$$

where

$$\mathbf{A}(k)=\mathbf{X}\,(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t \tag{48}$$

The GCV is a extension of Allen's PRESS. It can be shown that the GCV estimate of k in the ridge regression is given by

$$GCV = \frac{SSE_R}{\left\{n - \left[\frac{n}{n+k} + \sum_{j=1}^{m}\left(\frac{\lambda_j}{\lambda_j + k}\right)\right]\right\}^2} \tag{49}$$

where SSE_R is sum of squared error in ridge regression (Minotani(1992)).

Although it is needless to say that even GCV is not an absolute criterion for choosing regularization parameter, this estimate does not require an estimate of σ^2 ,and thus may be used for ill-posed problem such as multicollinearity.

4. APPLICATION

4.1 Data Set for Comparison between Counter-Measures

We simulate and compare the counter-measures for multicollinearity with a simple land price function below.

$$Y_i = \beta_0 + \sum_{j=1}^{6} \beta_j x_j \tag{50}$$

where suffix i shows point number from 1 to 130, j shows category number of explanatory variables from 1 to 6, x_j is explanatory variables shown in Table.2, Y_i is land price [thousand yen / m^2], and β_0, β_j (i=1,⋯,6) are unknown parameters.

Although the result is shown with no standardized form, practical application was done under the condition that x_j was standardized so that $\mathbf{X}'\mathbf{X}$ has the form of a correlation matrix.

Study area is a part of Chiba Prefecture of Tokyo Metropolitan Area, locating from about 25 to 45 minutes from Ueno, near Tokyo Station, by Joban Line. For estimation of the function we use officially assessed land price data set (1995) of National Land Agency.

4.2. Detection of Multicollinearity

Results of Regression Diagnostics (Minotani (1992)) is below.

(i) condition index K_j

$$K_j = \frac{\lambda_1}{\lambda_j} \quad (j=1,\cdots,m\,;\ \lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_m) \tag{51}$$

Condition index K_m is what is called condition number in numerical analysis. The result in our example is shown in Table 1.

Table 1. Condition Index K_j.

Eigenvalue λ_j	K_j
λ_1=2.24	K_1=1.00
λ_2=1.22	K_2=1.84
λ_3=1.01	K_3=2.22
λ_4=0.80	K_4=2.81
λ_5=0.68	K_5=3.27
λ_6=0.05	**K_6=43.7**

Table 2. *Variance Inflation Factor.*

x_j	VIF_j
x_1 : Width of Front Road (m)	VIF_1 = 1.12
x_2 : Sewer System (dummy)	VIF_2 = 1.11
x_3 : Use Zoning (dummy)	VIF_3 = 1.11
x_4 : Time Distance to Ueno (min)	VIF_4 = 1.10
x_5 : Distance to Nearest Station (km)	**VIF_5 = 10.1**
x_6 : Distance to Nearest Shopping Center (km)	**VIF_6 = 9.90**

(ii) Variance Inflation Factor (VIF)

$$VIF_j = \frac{1}{1 - R_j^2} \tag{52}$$

where R_j^2 is the squared multiple correlation coefficient between x_j and the other explanatory variables $x_1, x_2, \cdots, x_{j-1}, x_{j+1}, \cdots$. The result is shown in Table 2.
Table 1. shows K_6 (=λ_1/λ_6) is very high and there is a possibility that multicollinearity occurs and the result of parameter estimation is unstable. Table 2. shows that VIF_5 , VIF_6 is

high and x_5 , x_6 are correlated.

4.3 Application of Ridge Regression

Ridge trace and GCV are shown in Fig.3 and Fig.4. As is shown in Fig.4, GCV is least at k=0.0662, where the ridge traces of β_5,β_6 is not stable (Fig. 3). Fig.1 and Fig.2 show that it is not a good way to choose such k that makes the ridge trace stable. Because the larger k is , the more biased estimator is. So if we use ridge trace for determination of k, mean square error of ridge regression might be larger than that of k=0.

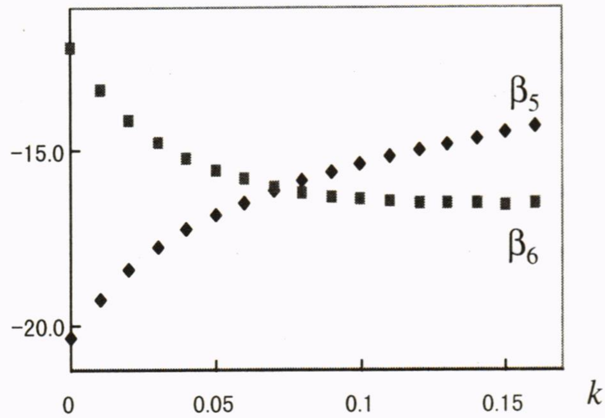

Fig.3 Ridge Trace for β_5 and β_6.

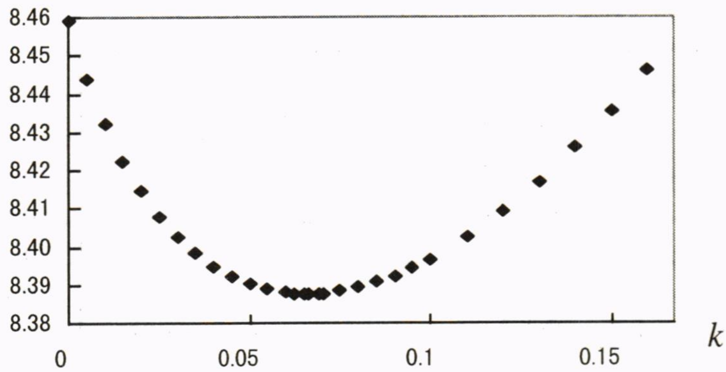

Fig.4 GCV in Ridge Regression.

4.4 Application of Principal Component Regression

The smallest eigenvalue corresponded with the 6th principal components is 0.051 and its contribution rate is only 0.9%.

Table 3. Eigenvalue and Contribution Rate.

eigenvalue	$\lambda_1 = 2.24$	$\lambda_2 = 1.22$	$\lambda_3 = 1.01$	$\lambda_4 = 0.80$	$\lambda_5 = 0.68$	$\lambda_6 = 0.05$
contribution rate	37.3%	20.3%	16.9%	13.3%	11.4.%	**0.9%**

Table 4. Principal Components Analysis of Explanatory Variables

Variables	1st	2nd	3rd	4th	5th	6th
x_1	0.46	0.21	-0.49	0.70	-0.082	-0.0080
x_2	0.37	-0.50	-0.57	-0.24	0.48	0.0033
x_3	-0.43	-0.45	0.44	0.50	-0.41	0.0012
x_4	-0.058	0.85	0.065	-0.052	0.52	0.0081
x_5	0.93	-0.042	0.33	-0.0048	-0.013	0.16
x_6	0.92	0.024	0.36	0.034	0.057	-0.16

We drop the 6th components for principal regression, because

(i) the eigenvalue of 6th principal components is extremely small

and

(ii) the share of x_5 , x_6 is relatively large.

4.5 Comparison among Ordinary Least Squares, Ridge Regression, and Principal Component Regression

Table 5. shows the comparison among ordinary least squares, ridge regression, and principal regression.

Although the correlation coefficient is almost the same, the estimates to the parameter are different and the standard errors of the parameters in RR and RPC are improved compared to OLS. It is important to note that the difference among the estimated β_5 ,β_6 affects transport project evaluation.

Table 5. The Results of OLS, RR. PCR.

	Estimated of Prameters			Standard Errors		
Explanatory Variables	OLS	RR	PCR.	OLS	RR	PCR
x_1:Width of Front Road (m)	11.6	10.4	11.2	2.86	2.64	2.83
x_2:Sewer System (0,1)	22.1	20.7	22.4	6.40	5.95	6.39
x_3:Use Zoning (0,1)	22.0	20.9	22.2	6.57	6.10	6.56
x_4:Time to Ueno Station (min)	-1.89	-1.75	-1.81	0.60	0.55	0.58
x_5:Distance to Nearest Station (km)	**-20.3**	**-16.3**	**-14.1**	**7.13**	**3.28**	**1.19**
x_6:Distance to Nearest Shopping Area (km)	**-12.1**	**-16.0**	**-20.5**	**9.69**	**4.48**	**1.69**
Constant	295.7	294.7	296.0	27.4		

Table 5. (Continued)

	t-value			Correlation Coefficient		
Explanatory Variables	OLS	RR	PCR.	OLS	RR	PCR
x_1	4.06	3.92	3.97	0.783	0.781	0.781
x_2	3.45	3.48	3.51			
x_3	3.36	3.38	3.43			
x_4	-3.17	-3.07	-3.19			
x_5	-2.85	-11.9	-4.96			
x_6	-1.25	-12.2	-3.57			
Constant	10.7					

It is reasonable that a slightly biased predictor with a very small prediction error variance can be better in terms of MSE that an unbiased prediction with a large prediction error variance. However, as we have no way to determine ridge parameter k directly as discussed in section 3.5, both of the figures are expected to be shown for the analysis of the project.

5. CONCLUSION

We discussed the counter-measures for multicollinearity in regression anaylysis, especially ridge regression and principal component regression among many ones.

In chapter 3, we explained the theoretical comparison between ridge regression and principal component regression in the framework of the inverse and ill-posed problem.

Two typical ways are known to improve the ill-posedness: (i) to change the solution space introducing some constrained conditions, (ii) to change the operator and solve the related problem. By regarding the parameter estimation procedure as an inverse problem, we find that principal component regression is equivalent to Truncated Singular Value Decomposition and classified as type (ii). By using singular value decomposition, ridge regression is also classified as type (ii). On the other hand, ridge regression can be classified as type (i), if we interpreter ridge regression as a Tikhonov regularization. We confirmed that the concept of inverse analysis is very useful for us to understand the relationships among some counter-measures in statistical problems.

We also compared ridge regression and principal component regression practically by applying them to a common parameter estimation problem of land price function and demonstrated the performance of ridge regression and principal component regression in chapter 4.

For the choice of ridge parameter, we used the General Cross Validation (GCV), which has been scarcely used in econometrics. It is noted that although Ridge Trace has been commonly used as criterion for determination of ridge parameter, it does not work well in term of mean square error of the parameters, compared with GCV. Needless to say, as the researchers in econometrics have little experience in application of GCV, more experiments of GCV are required for ridge regression.

REFERENCES

Allen, D. M. (1974) The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction, **Technometrics,** 16, No.2, 215-223

Baye, M. R. and Parker, D. F. (1984) Combining Ridge and Principal Component Regression : A Money Demand Illustration, **Communications in Statistics *A* - Theory and Methods**, Vol.13, 197-205

Chatterjee, S. and Price, B. (1977) **Regression Analysis by Example**, John Wiley & Sons, New York

Goldestein, M. And Smith A. F. M.(1973) Ridge-type Estimators for Regression Analysis, **Journarl of Royal Statistics *B***, Vol.38, 248-250.

Golub, G. H., Heath, M. and Wahba, G. Golub(1979) Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, **Technometrics,** 21, No.2, 215-223

Groetsch, G. W. (1993) **Inverse Problems in the Mathematical Sciences**, Vieweg, Wiesbade

Hoerl, A. E.(1962) Application of rideg analysis to regression problems, **Chemical Engeering Progress**, 58, No.3, 54-59

Hoerl, A. E. (1959) Optimum solution of many variables equations, **Chemical Engineering Progress**, 55, No.11, 69-78

Hoerl, A. E. and Kennard, R. W. (1970 a) Ridge Regression : Biased Estimation for Nonorthogonal Problems, **Technometrics**, 12, No.1, 55-67

Hoerl, A. E. and Kennard, R. W. (1970 b) Ridge Regression : Applications to Nonrothogonal Problems, **Technometrics**, 12, No.1, 69-82

Maddala, G.S. (1988) **Introduction to Econometrics**, Macmillan, New York.

Marquardt, D. W. (1970) Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation, **Technometrics**, 12, No.3, 591-612

Marquardt, D. W. (1963) An Algorithm for Least-Squares Estimation of Nonlinear Parameters, **Journal of Society of Industrial Applied Mathematics**, 11, No.2, 431-441

Menke, W. (1989) **Geophysical Data Analysis : Discrete Inverse Theory**, Academic Press

Minotani, C. (1992) **New Development of Economterics** , Taga Shuppan (in Japanese)

Sarkar, N. (1989) **Annals of the Institute of Statistical Mathematics**, Vol.41, No.4, 717-724

Tikhonov, A. N. et. al. (1990) **Numerical Methods for the Solution of Ill-Posed Problems** , Kluwer Academic Press