行政院國家科學委員會專題研究計畫　成果報告

縮減估計式在迴歸與預測之研究

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-032-010-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：淡江大學統計學系(所)

計畫主持人：林志娟

計畫參與人員：崔春平

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 2 月 12 日

行政院國家科學委員會專題研究計畫成果報告

計畫編號：NSC-91-2118-M-032-010
執行期限：91 年 8 月 1 日至 92 年 7 月 30 日
主持人：林志娟　淡江大學統計系專任副教授
計畫參與人員：　崔春平淡江大學統計所碩士班

一、中文摘要

在決策理論的領域裡，Stein（1956）和 James and Stein（1961）的研究一直備受矚目，除了指出『Stein 效應』，許多研究者更將其理論延伸至其他應用上。預測的問題是日生活中非常重要的一個議題，預測的品質更是許多重大決策成敗的關鍵，因此預測模型的建立與選擇更形重要。而迴歸分析在預測的領域上一直都扮演著不可或缺的角色，在迴歸參數估計上，當參數數目大於 2 時，Stein 的理論可被延伸來建構出比最小平方估計式還好的縮減（shrinkage）估計式。只是在損失函數的選擇上卻有些微的爭議，一般而言經濟學家較偏好‘ordinary error-norm square’損失函數（$\|\hat{\beta}-\beta\|^2$）（見 Ullah and Ullah (1978)），而統計學家則較常用不受單位影響的‘normalized of weighted error-norm square’損失函數（$(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)$，X 是設計矩陣)。

本研究的主要成果為：指出給定過去資料下在迴歸模式建構下找出較佳（均方差）預測估計式和在決策理論建構下找出較佳（風險函數）預測估計式基本上是等價的。

關鍵詞：損失函數、風險函數、預測均方差、縮減（shrinkage）估計式、迴歸。

二、英文摘要 Abstract

For estimating the regression coefficients (β) in a regular regression model one can easily extend Stein's idea to construct shrinkage estimators which dominate the usual least squares estimator when the number of coefficients is greater than two. But there is a slight disagreement about the selection of a proper loss function. Some researchers, especially the econometricians have used the 'ordinary error-norm square' as a loss function ($\|\hat{\beta}-\beta\|^2$), whereas the statisticians have traditionally preferred to use the 'normalized of weighted error-norm square' ($(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)$, X being the design matrix) as a loss function. Basically, the problem of finding a better predictor of the dependent variable in the sense of smaller PMSE, in a regression set up when the past data is given, is equivalent to constructing a better estimator of β in the sense of smaller

risk (for fixed X) in a decision-theoretic set up. The connection between the above mentioned two problems has been built up in this research and show how naturally the method of 'Shrinkage Estimation' gives one a better function (ordinary or normalized error norm square) for estimating the regression coefficients.). Basically, that is the main contribution of this research.

Keywords: loss function、risk function、shrinkage estimation、predicted mean square error、regression.

三、緣由與目的

PREDICTION AND ESTIMATION OF REGRESSION COFFICIENTS

Consider a multiple linear regression model where the response (dependent) variable Y is explained by the vector $\underset{\sim}{X}_{p\times 1} = (X_1, X_2, \ldots, X_p)'$ of explanatory (independent) variables. We assume that for fixed $\underset{\sim}{X}$, Y has a normal distribution of the form

$$Y \mid \underset{\sim}{X} \sim N(\underset{\sim}{X}'\underset{\sim}{\beta}, \sigma^2) \qquad (3.1)$$

where both $\underset{\sim}{\beta} \in \Re^p$, and $\sigma^2 \in \Re^+$ are unknown and β_j (j^{th} element of $\underset{\sim}{\beta}$) is the coefficient of X_j. We further assume that $\underset{\sim}{X}$ follows a multivariate distribution $\Im$ with some mean vector $\underset{\sim}{\eta}$ and variance-covariance matrix V. Note that we are not assuming anything further regarding the distribution $\Im$ of $\underset{\sim}{X}$. The vector $\underset{\sim}{\eta}$ and the p.d. matrix V could be known or unknown and we'll discuss this later. Our above model, is the simplest form of "Errors in Variables Model" (EVM). A huge literature exists on the EVM from asymptotic point of view and one can look at Fuller (1987) for an extensive literature survey.

四、結果與討論

CONNECTION BETWEEN REGRESION SETUP AND DECISION THEORETIC SETUP

Now, we have n i.i.d. observations of $\underset{\sim}{X}$, say, $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \Lambda, \underset{\sim}{X}_n$. For given each $\underset{\sim}{X}_i$, we observe Y_i coming from the distribution $N(\underset{\sim}{X}'_i\underset{\sim}{\beta}, \sigma^2)$. The augmented independent vectors $(Y_1, \underset{\sim}{X}_1), (Y_2, \underset{\sim}{X}_2), \Lambda, (Y_n, \underset{\sim}{X}_n)$ constitute our data set (or past data set). We are interested in the future value of $Y \mid \underset{\sim}{X}$ when $\underset{\sim}{X}$ is given. Suppose the future value of $\underset{\sim}{X}$ is $\underset{\sim}{X}_{new}$ (which again comes from the distribution $\Im$ and is independent of the past observations). Our main goal is to predict the value of Y, say Y_{new}, when $\underset{\sim}{X}_{new}$ is observed.

A predictor $\hat{Y}_{new}$ of Y_{new} (the future observation of Y) is evaluated by its Prediction Mean Squared Error (PMSE) defined as

$$PMSE(\hat{Y}_{new}) = E(Y_{new} - \hat{Y}_{new})^2, \qquad (4.1)$$

where the expectation is taken with respect to the joint distribution of $\underset{\sim}{Y}$, X, X_{new} and Y_{new}.

Let the matrix X be the usual design matrix where i^{th} row of X is $\underset{\sim}{X}'_i$. Then for the past data,

$$Y \mid X \sim N(X\underset{\sim}{\beta}, \sigma^2 I_n) \qquad (4.2)$$

where $\underset{\sim}{Y} = (Y_1, ..., Y_n)'$. Using the least squares (LS) method our LS estimator of $\underset{\sim}{\beta}$

is $\hat{\underset{\sim}{\beta}}^0 = (X'X)^{-1}X\underset{\sim}{Y}$. As we mentioned before a future value of $\underset{\sim}{X}$ also arises at random according to a multivariate probability distribution with mean $\underset{\sim}{\eta}$ and variance covariance matrix V.

Let $\hat{Y} = \underset{\sim}{X}'\hat{\underset{\sim}{\beta}}^0$ be the LS predictor of Y and the fitted value of Y_i is

$$\hat{Y}_i = X'_i\,\hat{\beta}^0, \quad i = 1, 2, \Lambda\ , n.$$

Consider a future observation (Y_{new}, $\underset{\sim}{X}_{new}$), where Y_{new} is unknown and $\underset{\sim}{X}_{new}$ is given, and we are interested in Y_{new}. Then

$$E(Y_{new} \mid X, \underset{\sim}{Y}) = E[\ E(Y_{new} \mid \underset{\sim}{X}_{new}, X, \underset{\sim}{Y})]$$
$$= \underset{\sim}{\eta}'\underset{\sim}{\beta}$$

$$E(\hat{Y}_{new} \mid X, \underset{\sim}{Y}) = E\ E(\hat{Y}_{new} \mid \underset{\sim}{X}_{new}, X, \underset{\sim}{Y})$$
$$= \underset{\sim}{\eta}'\hat{\underset{\sim}{\beta}}^0$$

$$Var(\hat{Y}_{new} \mid X, \underset{\sim}{Y}) = Var(E(\hat{Y}_{new} \mid \underset{\sim}{X}_{new}, X, \underset{\sim}{Y}))$$
$$+ E(Var(\hat{Y}_{new} \mid \underset{\sim}{X}_{new}, X, \underset{\sim}{Y}))$$
$$= \hat{\underset{\sim}{\beta}}^{0\prime} V\,\hat{\underset{\sim}{\beta}}^0$$

$$Var(Y_{new} \mid X, \underset{\sim}{Y}) = Var(E(Y_{new} \mid \underset{\sim}{X}_{new}, X, \underset{\sim}{Y}))$$
$$+ E(Var(Y_{new} \mid \underset{\sim}{X}_{new}, X, \underset{\sim}{Y}))$$
$$= \underset{\sim}{\beta}'V\,\underset{\sim}{\beta} + \sigma^2$$

$$Cov[(Y_{new}, \hat{Y}_{new}) \mid X, \underset{\sim}{Y}]$$
$$= E(Y_{new}\hat{Y}_{new} \mid X, \underset{\sim}{Y}) - E(Y_{new} \mid X, \underset{\sim}{Y})\,E(\hat{Y}_{new} \mid X, \underset{\sim}{Y})$$
$$= \underset{\sim}{\beta}'V\,\hat{\underset{\sim}{\beta}}^0, \text{ since } V = E(\underset{\sim}{X}_{new}\underset{\sim}{X}'_{new}) - \underset{\sim}{\eta}\,\underset{\sim}{\eta}'$$

Therefore, the conditional bivariate distribution of $(Y_{new}, \hat{Y}_{new})$ given the past data $(X, \underset{\sim}{Y})$ will have mean $(\underset{\sim}{\eta}'\underset{\sim}{\beta}, \underset{\sim}{\eta}'\hat{\underset{\sim}{\beta}}^0)$ and variance-covariance matrix

$$\begin{bmatrix} \underset{\sim}{\beta}'V\,\underset{\sim}{\beta} + \sigma^2 & \underset{\sim}{\beta}'V\,\hat{\underset{\sim}{\beta}}^0 \\ \underset{\sim}{\beta}'V\,\hat{\underset{\sim}{\beta}}^0 & \hat{\underset{\sim}{\beta}}^{0\prime}V\,\hat{\underset{\sim}{\beta}}^0 \end{bmatrix}$$

Given the past data $(X, \underset{\sim}{Y})$, our usual guess for Y_{new} is $Y_{new} = \hat{Y}_{new}$, i.e., we follow the line with slope=1 if regress Y_{new} on $\hat{Y}_{new}$. But it can be seen, from a simple linear regression point of view, that for a better prediction of Y_{new}, we need to have slope < 1. Recall that in the simple linear regression set up, if (Z_1, Z_2) has mean $\underset{\sim}{\mu} = (\mu_1, \mu_2)$ and variance-covariance matrix $\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$ and if we use the simple linear regression model

$Z_1 \mid Z_2 = a + bZ_2 + \varepsilon,\ where \quad \varepsilon \sim (0, \sigma^2{}_*)$, then the optimal values of a and b which minimize $L = (Z_1 - (a + bZ_2))^2$ are $a = \mu_1 - \frac{\sigma_{12}}{\sigma_2^2}\mu_2$, which is the intercept, and $b = \frac{\sigma_{12}}{\sigma_2^2}$, which is the slope of the line. Therefore, instead of using the line $\hat{Y}_{new} = Y_{new}$, we need to use $\hat{Y}_{new} = KY_{new}$, where $K = \frac{Cov(Y_{new}, \hat{Y}_{new})}{Var(\hat{Y}_{new})}$. It is expected (we will show) that K should be strictly between 0 and 1. Note that K is unknown since it depends on the unknown parameters σ^2 and $\underset{\sim}{\beta}$, and so it has to be estimated by $(X, \underset{\sim}{Y})$. This motivates us to look at the preshrunk predictor

$$\tilde{Y}_{new} = \hat{K}\,\underset{\sim}{X}'_{new}\,\hat{\underset{\sim}{\beta}}^{0}, \qquad (4.3)$$

where $\hat{K}$ is a suitable estimator of K depending on $(X, \underset{\sim}{Y})$ only.

The overall PMSE of $\tilde{Y}_{new}$ (given in (4.3)), defined as

$E(Y_{new} - \tilde{Y}_{new})^2$, can be simplified as below.

$$E(Y_{new} - \tilde{Y}_{new})^2 = E(Y_{new} - \hat{K}\,\underset{\sim}{X}'_{new}\,\hat{\underset{\sim}{\beta}}^{0})^2$$
$$= B_1 + B_2 - 2B_3(say) \qquad (4.4)$$

The above terms B_1, B_2 and B_3.can be simplified (Lin,(2002)) as following：

$$B_1 = E\left[Y_{new} - \underset{\sim}{X}'_{new}\,\underset{\sim}{\beta}\right]^2$$
$$= \sigma^2$$
$$B_2 = E[(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'\underset{\sim}{X}_{new}\underset{\sim}{X}'_{new}(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})]$$
$$= E[(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'(V + \underset{\sim}{\eta}\,\underset{\sim}{\eta}')(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})],$$

since $Var(\underset{\sim}{X}_{new}) = V$.

$$B_3 = E[(Y_{new} - \underset{\sim}{X}'_{new}\,\underset{\sim}{\beta})\,\underset{\sim}{X}'_{new}(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})]$$
$$= 0$$

Hence, if we write $V_0 = V + \underset{\sim}{\eta}\,\underset{\sim}{\eta}'$, then

The expression (4.4)
$= \sigma^2 + \mathrm{E}[(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'V_0(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})]$

=(constant free from $\underset{\sim}{\beta}$) + [risk of estimating $\underset{\sim}{\beta}$ by $\hat{K}\,\hat{\underset{\sim}{\beta}}^{0}$ under the loss

$$L(\hat{\underset{\sim}{\beta}}, \underset{\sim}{\beta} \mid V_0) = (\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})'V_0(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})]. \qquad (4.5)$$

This was also shown by Sclove (1968), but his main goal was to estimate partitioned coefficient vectors. Note that if we predict Y_{new} by the LS predictor $\hat{Y}_{new} = \underset{\sim}{X}'_{new}\,\hat{\underset{\sim}{\beta}}^{0}$, then

$$E(Y_{new} - \hat{Y}_{new})^2 = \sigma^2 + \mathrm{E}[(\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'V_0(\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})] \qquad (4.6)$$

From (4.5) and (4.6), we know that

$$\mathrm{E}(Y_{new} - \tilde{Y}_{new})^2 \le \mathrm{E}(Y_{new} - \hat{Y}_{new})^2$$

if and only if

$$E(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'V_0(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})$$
$$\le E(\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'V_0(\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta}),$$

where $V_0 = V + \underset{\sim}{\eta}\,\underset{\sim}{\eta}'$

i.e. , $E\,E[(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'V_0(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta}) \mid X]$
$\le E\,E[(\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'V_0(\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta}) \mid X]$. (4.7)

Therefore the improvement in terms of PMSE can be achieved by the improvement in the risk (under $L(\hat{\underset{\sim}{\beta}}, \underset{\sim}{\beta} \mid V_0)$ in (4.5)) of estimating $\underset{\sim}{\beta}$ by $\hat{K}\,\hat{\underset{\sim}{\beta}}^{0}$ instead of $\hat{\underset{\sim}{\beta}}^{0}$ for the given design matrix X. Therefore, the problem of finding a better predictor of Y_{new} in the sense of smaller PMSE, in a regression set up when the past data is given, is equivalent to constructing a better estimator of $\underset{\sim}{\beta}$ in the sense of smaller risk (for fixed X) in a decision-theoretic set up. So, to have a smaller PMSE, it is enough to show that

$$E[(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'V_0(\hat{K}\,\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta}) \mid X]$$
$$\le E[(\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta})'V_0(\hat{\underset{\sim}{\beta}}^{0} - \underset{\sim}{\beta}) \mid X]. \qquad (4.8)$$

where $\hat{\underset{\sim}{\beta}}^{0}$ is the LS estimator of $\underset{\sim}{\beta}$ and $0 \le \hat{K} \le 1$, and this is how the method of shrinkage estimator is applied to the prediction of the future in a regression setup. Then we can apply the well-established results in shrinkage normal mean estimation to get the improved results in regression set up later.

Acknowledgement: The author's research has been supported by a research grant from the National Science Council (nsc-91-2218-M-032-010).

五、參考文獻

1. Baranchik, A. J. (1964). Multiple regression and estimation of the mean of

a multivariate normal distribution. *Technical Report No.51*, Department of Statistics, Stanford University, Stanford, California.

2. Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution . *Annals of Mathematical Statistics* , 41, 642 -645.
3. Chang, C., Lin, J. and Pal, N. (1993). Improvements over the James-Stein estimator: A risk Analysis. *Journal of Statistical Computation Simulation*, 48, 117-126.
4. Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of Royal Statistical Society,* Series B, 45, 311-354.
5. Efron, B. and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *The Annals of Statistics*, 4, 22-32.
6. Fuller, W. A. (1987). Measurement Errors Models, Wiley, New York.
7. George, E. I. (1986a). Minimax multiple shrinkage estimator. *Annals of Statistics*, vol 14, 188-205.
8. George, E. I. (1986b). A formal Bayes multiple shrinkage estimator. *Communications in Statistics, Theory & methods*, 7, 2099-2114.
9. Guo, Y. Y. and Pal, N. (1992). A sequence of improvements over the James-Stein estimator. *Journal of Multivariate Analysis*, 42, 302-317.
10. Harewood, S. (1992). A Stein rule estimator which shrinks towards the ridge regression estimator. *Economics Letter*. 40, 127-133.
11. James, W. and Stein, C. (1961). Estimation with quadratic loss. In Proceedings of the *Fourth Berkeley Symposium on Mathematical Statistics & Probability,* Vol. 1, 361-380, University of California Press, Berkeley.
12. Kubokawa, T. (1991). An approach to improving the James-Stein estimator, *Journal of Multivariate Analysis,* 36, 121-126.
13. Sclove, S. L. (1968). Improved estimators for coefficients in linear regression, *JASA*, 63,599-606.
14. Sengupta, D. (1991). On shrinkage towards an arbitrary estimator. *Statistics and Decisions*, 9, 81-105.
15. Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. In Proceedings of the *Third Berkeley Symposium on Mathematical Statistics & Probability,* Vol.1, 197-206, University of California Press, Berkeley.
16. Stein, C. (1981). Estimation of the Mean of a Multivariate Normal distribution. *Annals of Statistics*, 9, 1135-1151.
17. Ullah, A. and Ullah, S. (1978). Double k-class estimators of coefficients in linear regression. *Econometrica*. 46, No. 3, 705-722.