Large Sample Theory for Some Ridge-Type Regression Estimators

Yu Jin and David J. Olive*
Southern Illinois University

September 11, 2022

Abstract

This paper gives large sample theory for some ridge-type multiple linear regression estimators, including Liu-type regression estimators, when the number of predictors is fixed.

KEY WORDS: Ridge Regression, Liu-Type Regression Estimator

1 INTRODUCTION

This section reviews the multiple linear regression model, some ridge-type regression estimators, and the large sample theory for the ordinary least squares estimator. Suppose that the response variable Y_i and at least one predictor variable $x_{i,j}$ are quantitative with $x_{i,1} \equiv 1$. Let $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p})$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ where β_1 corresponds to the intercept. Then the multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{1}$$

for $i = 1, ..., n$. Here n is the sample size, and assume that the random variables e_i are independent and identically distributed (iid) with mean $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. The ith fitted value $\hat{Y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ and the ith residual $r_i = Y_i - \hat{Y}_i$ where $\hat{\boldsymbol{\beta}}$ is any $p \times 1$ estimator of $\boldsymbol{\beta}$. Ordinary least squares (OLS) is often used for inference if n/p is large.

*David J. Olive is Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA.

Liu (2003) defined the Liu-type estimator

$$\hat{\boldsymbol{\beta}}_{k,d} = (\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{Y} - d\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}_{R,k} - \frac{d}{n}n(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\hat{\boldsymbol{\beta}} \tag{3}$$

where $k = k_n \geq 0$, $d = d_n$ is a real number, and the Hoerl and Kennard (1970) ridge regression estimator $\hat{\boldsymbol{\beta}}_{R,k}$ corresponds to $d = 0$. The Liu (1993) estimator

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{Y} + c\hat{\boldsymbol{\beta}})$$

is another special case with $k = 1$ and $d = -c$ where $0 < c < 1$.

Kurnaz and Akay (2015) showed that several ridge-type regression estimators in the literature can be written as $\hat{\boldsymbol{\beta}}_f = (\boldsymbol{X}^T\boldsymbol{X}+k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{Y}+f(k)\hat{\boldsymbol{\beta}})$ where $k \geq 0$ and $f(\cdot)$ is a continuous function of k, including ridge-type estimators given by Özkale and Kaçiranlar (2007), Sakallioğlu and Kaçiranlar (2008), and Yang and Chang (2010). Note that $\hat{\boldsymbol{\beta}}_f = \hat{\boldsymbol{\beta}}_{k,d}$ with $d = -f(k)$. If $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{R,k}$, then $\hat{\boldsymbol{\beta}}_f = (\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X} + [k + f(k)]\boldsymbol{I})\hat{\boldsymbol{\beta}}_{R,k}$.

Kibria and Lukman (2020) defined the estimator

$$\hat{\boldsymbol{\beta}}_{KL} = (\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X} - k\boldsymbol{I})\hat{\boldsymbol{\beta}}_{OLS}.$$

Since $(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X} - k\boldsymbol{I}) = (\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I} - 2k\boldsymbol{I}) =$
$\boldsymbol{I} - 2k(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}$,

$$\hat{\boldsymbol{\beta}}_{KL} = [\boldsymbol{I} - 2k(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}]\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}}_{OLS} - 2k(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\hat{\boldsymbol{\beta}}_{OLS}. \tag{4}$$

The OLS estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ has large sample theory given, for example, by Sen and Singer (1993, p. 280). Let the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and let the ith leverage $h_i = \boldsymbol{H}_{ii}$ be the ith diagonal element of $\boldsymbol{H}$. Consider the multiple linear regression model (1) where the e_i are iid with $E(e_i) = 0$ and $\mathrm{V}(e_i) = \sigma^2$. Assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$ and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \ \boldsymbol{V}). \tag{5}$$

Note that $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{V}$, and if $k/n \to 0$, then

$$\left(\frac{\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}}{n}\right)^{-1} = n(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1} \to \boldsymbol{V}. \tag{6}$$

Knight and Fu (2000) derived the large sample theory for ridge regression and the Tibshirani (1996) lasso estimator with p fixed. The following section derives some large sample theory for the Liu-type estimator $\hat{\boldsymbol{\beta}}_{k,d}$ and for $\hat{\boldsymbol{\beta}}_{KL}$.

2 LARGE SAMPLE THEORY

The large sample theory assumes that p is fixed and that Equation (5) holds for the OLS estimator. Then $\hat{\boldsymbol\beta}_{k,d} = (\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}(\boldsymbol X^T \boldsymbol Y - d\hat{\boldsymbol\beta}) =$

$$(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}(\boldsymbol X^T \boldsymbol X)(\boldsymbol X^T \boldsymbol X)^{-1}\boldsymbol X^T \boldsymbol Y - d(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}\hat{\boldsymbol\beta} =$$

$$\boldsymbol A_n \hat{\boldsymbol\beta}_{OLS} - d(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}\hat{\boldsymbol\beta}$$

where $\boldsymbol A_n = (\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}(\boldsymbol X^T \boldsymbol X) = \boldsymbol B_n = \boldsymbol I - k(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}$ since $\boldsymbol A_n - \boldsymbol B_n = \boldsymbol 0$. This identity appears in Gunst and Mason (1980, p. 332) and was used by Pelawa Watagoda and Olive (2021) to simplify ridge regression large sample theory. Thus

$$\hat{\boldsymbol\beta}_{k,d} = [\boldsymbol I - k(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}]\hat{\boldsymbol\beta}_{OLS} - d(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}\hat{\boldsymbol\beta} =$$

$$\hat{\boldsymbol\beta}_{k,d} = \hat{\boldsymbol\beta}_{OLS} - \frac{k}{n} n(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}\hat{\boldsymbol\beta}_{OLS} - \frac{d}{n} n(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}\hat{\boldsymbol\beta}. \tag{7}$$

Theorem 1. Assume Equations 5) and 6) hold, and that $\hat{\boldsymbol\beta}$ is a consistent estimator of $\boldsymbol\beta$. a) If $k/\sqrt{n} \to 0$ and $d/\sqrt{n} \to 0$, then $\hat{\boldsymbol\beta}_{k,d}$ is asymptotically equivalent to $\hat{\boldsymbol\beta}_{OLS}$ with

$$\sqrt{n}(\hat{\boldsymbol\beta}_{k,d} - \boldsymbol\beta) \xrightarrow{D} N_p(\boldsymbol 0, \sigma^2\ \boldsymbol V).$$

b) If $k/\sqrt{n} \to \tau \geq 0$ and $d/\sqrt{n} \to \delta$, then

$$\sqrt{n}(\hat{\boldsymbol\beta}_{k,d} - \boldsymbol\beta) \xrightarrow{D} N_p(-(\tau + \delta)\boldsymbol V\boldsymbol\beta, \sigma^2\ \boldsymbol V).$$

c) If $k/n \to 0$ and $d/n \to 0$, then $\hat{\boldsymbol\beta}_{k,d}$ is a consistent estimator of $\boldsymbol\beta$.

Proof. a) follows from b).

b) By Equation (7),

$$\sqrt{n}(\hat{\boldsymbol\beta}_{k,d} - \boldsymbol\beta) = \sqrt{n}(\hat{\boldsymbol\beta}_{OLS} - \boldsymbol\beta) - \frac{k}{\sqrt{n}}\, n(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}\hat{\boldsymbol\beta}_{OLS} - \frac{d}{\sqrt{n}}\, n(\boldsymbol X^T \boldsymbol X + k\boldsymbol I)^{-1}\hat{\boldsymbol\beta}$$

$$\xrightarrow{D} N_p(\boldsymbol 0, \sigma^2\ \boldsymbol V) - \tau\boldsymbol V\boldsymbol\beta - \delta\boldsymbol V\boldsymbol\beta \sim N_p(-(\tau + \delta)\boldsymbol V\boldsymbol\beta, \sigma^2\ \boldsymbol V).$$

c) By Equation (7), $\hat{\boldsymbol\beta}_{k,d} \xrightarrow{P} \boldsymbol\beta - 0\boldsymbol V\boldsymbol\beta - 0\boldsymbol V\boldsymbol\beta = \boldsymbol\beta$.

Theorem 2. Assume Equations 5) and 6) hold. a) If $k/\sqrt{n} \to 0$, then $\hat{\boldsymbol\beta}_{KL}$ is asymptotically equivalent to $\hat{\boldsymbol\beta}_{OLS}$ with

$$\sqrt{n}(\hat{\boldsymbol\beta}_{KL} - \boldsymbol\beta) \xrightarrow{D} N_p(\boldsymbol 0, \sigma^2\ \boldsymbol V).$$

b) If $k/\sqrt{n} \to \tau \geq 0$, then

$$\sqrt{n}(\hat{\boldsymbol\beta}_{KL} - \boldsymbol\beta) \xrightarrow{D} N_p(-2\tau\boldsymbol V\boldsymbol\beta, \sigma^2\ \boldsymbol V).$$

c) If $k/n \to 0$, then $\hat{\boldsymbol\beta}_{KL}$ is a consistent estimator of $\boldsymbol\beta$.

Proof. a) follows from b).

b) By Equation (4),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{KL} - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - \frac{2k}{\sqrt{n}} \, n(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\hat{\boldsymbol{\beta}}_{OLS}$$

$$\xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \, \boldsymbol{V}) - 2\tau\boldsymbol{V}\boldsymbol{\beta} \sim N_p(-2\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2 \, \boldsymbol{V}).$$

c) By Equation (4),

$$\hat{\boldsymbol{\beta}}_{KL} = \hat{\boldsymbol{\beta}}_{OLS} - \frac{2k}{n} \, n(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{P} \boldsymbol{\beta} - 2(0)\boldsymbol{V}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

3 CONCLUSIONS

Theorems 1 and 2 gave some large sample theory for many ridge-type estimators. Taking $d = -k$ is interesting in Theorem 1. Several of the ridge-type estimators can be computed if $k > 0$ even if $\boldsymbol{X}^T\boldsymbol{X}$ is singular, and such estimators can be useful if $p > n$. Li and Yang (2012) gave a Liu-type estimator that replaced $\hat{\boldsymbol{\beta}}$ by a vector $\boldsymbol{b}$ that represents prior information.

For many regression estimators, a method is needed so that everyone who uses the same units of measurement for the predictors and Y gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. Let the nontrivial predictors $\boldsymbol{u}_i^T = (x_{i,2}, ..., x_{i,p})$ where $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T$. A common method is to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W} = (W_{ij})$. For $j = 1, ..., p-1$, let W_{ij} denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Note that the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is $\boldsymbol{R}_{\boldsymbol{u}} = \boldsymbol{W}^T\boldsymbol{W}/n$. Then regression through the origin is used for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. Large sample theory could be given for $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, as in Pelawa Watagoda and Olive (2021), or for $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, as in this paper.

4 References

Gunst, R.F., and Mason, R.L. (1980), *Regression Analysis and Its Application,* Marcel Dekker, New York, NY.

Hoerl, A.E., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

Kibria, B.M.G., and Lukman, A.F. (2020), "A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications," *Scientifica*, 2020, online.

Knight, K., and Fu, W.J. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356-1378.

Kurnaz, F.S., and Akay, K.U. (2015), "A New Liu-Type Estimator," *Statistical Papers*, 56, 495-517.

Li, Y., and Yang, H. (2012), "A New Liu-Type Estimator in Linear Regression Model," *Statistical Papers*, 53, 427-437.

Liu, K. (1993), "A New Class of Biased Estimate in Linear Regression," *Communications in Statistics: Theory and Methods*, 22, 393-402

Liu, K. (2003), "Using Liu-Type Estimator to Combat Collinearity," *Communications in Statistics: Theory and Methods*, 32, 1009-1020.

Özkale, M.R., and Kaçiranlar, S. (2007), "The Restricted and Unrestricted Two-Parameter Estimators," *Communications in Statistics: Theory and Methods*, 36, 2707-2725.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021), "Comparing Six Shrinkage Estimators with Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers*, 62, 2407-2431.

Sakallioğlu, S., and Kaçiranlar, S. (2008), "A New Biased Estimator Based on Ridge Estimation," *Statistical Papers*, 49, 669-689.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

Yang, H., and Chang, X. (2010), "A New Two-Parameter Estimator in Linear Regression," *Communications in Statistics: Theory and Methods*, 39, 923-934.