The General Linear Test in the Ridge Regression

Whasoo Bae[a], Minji Kim[b], Choongrak Kim[1,b]

[a]Department of Data Science/Institute of Statistical Information, Inje University, Korea
[b]Department of Statistics, Pusan National University, Korea

Abstract

We derive a test statistic for the general linear test in the ridge regression model. The exact distribution for the test statistic is too difficult to derive; therefore, we suggest an approximate reference distribution. We use numerical studies to verify that the suggested distribution for the test statistic is appropriate. A asymptotic result for the test statistic also is considered.

Keywords: Linear test, reference distribution, shrinkage parameter, test statistic.

1. Introduction

In regression analysis, the ridge regression model (Hoerl and Kennard, 1970) is a good alternative to the classical linear model when covariates are highly correlated. The ridge regression model has been studied by many researchers and indicates many good properties. Among them, there exists a ridge transformation parameter with which ridge regression estimator has a smaller mean squared error than the classical linear regression model, and the ridge regression estimator is a Bayes estimator when the prior for the regression coefficients are Gaussian under the squared error loss function. For other properties of the ridge regression estimator, see Seber and Lee (2003) and Kim and Kang (2010). However, relatively few studies are done in the testing problem for the ridge regression coefficients.

The general linear test in the classical linear model is often used. One example is the Cobb-Douglas production function (Chipman and Rao, 1964) in the field of econometrics; however, the general linear test problem in the ridge regression has not been studied. Obenchain (1977) studied interval estimation of the general linear combination of regression coefficients in the ridge regression using the singular values decomposition of the design matrix, and Hoerl and Kennard (1990) proposed a degrees of freedom in the analysis of variance model using the ridge regression. We study the general linear test problem in the ridge regression model and derive a test statistic for the general linear test that suggests an approximate reference distribution. As far as we know, this study was not considered so far since the test statistic under the general linear restriction is quite tedious and the corresponding degrees of freedom for the test statistic is quite different from the classical linear model. Note that the hat matrix in the ridge regression is not idempotent so that the degrees of freedom are not any more an integer but a real number; therefore, a traditional F-distribution in the classical linear model is not exactly suitable.

This paper is organized as follows. In Section 2, the ridge regression is defined and reviews on the general linear test in the classical linear model are given. Derivation of test statistic in the ridge regression model and an appropriate definition for the degrees of freedom is suggested, and an

approximate reference distribution is suggested in Section 3. Numerical studies for the proposed test statistic are given in Section 4, and concluding remarks and further studies are also given in Section 5.

2. Past Works

2.1. The general linear test in the linear model

Consider a multiple linear regression model

$$y = X\beta + \varepsilon,$$

where y is an n-vector of responses, X is an $n \times p$ design matrix of known covariates with 1's in the first column, β is a p-vector of unknown coefficients, and ε is an error term, which is a normally distributed n-vector with mean $\mathbf{0}$ and covariance $\sigma^2 I$. The general linear test is specified as the following hypotheses

$$H_0 : C\beta = m \quad vs. \quad H_1 : C\beta \neq m,$$

where C is a $q \times p$ matrix with the rank of q and m is a q-vector of given constants. Then, it can be easily shown that

$$\hat{\beta}_H = \hat{\beta} - (X^t X)^{-1} C^t \left[C(X^t X)^{-1} C^t \right]^{-1} \left(C\hat{\beta} - m \right),$$

where $\hat{\beta}_H$ is least squares estimator of β under H_0, and $\hat{\beta}$ is least squares estimator of β. The test statistic in the linear model is generally given by the form

$$F = \frac{(\text{SSE}(R) - \text{SSE}(F))/(df_R - df_F)}{\text{SSE}(F)/df_F},$$

where $\text{SSE}(R)$ and $\text{SSE}(F)$ denote error sum of squares under the reduced model and under the full model, respectively, and df_R and df_F denote the degrees of freedom under the reduced model and under the full model, respectively. It can be easily shown that $F \sim F(p - r, n - p)$ under H_0. If we apply this fact to the general linear test case, it can be easily shown that

$$\text{SSE}(R) - \text{SSE}(F) = \left(\hat{\beta} - \hat{\beta}_H \right) X^t X \left(\hat{\beta} - \hat{\beta}_H \right)$$
$$= \left(C\hat{\beta} - m \right)^t \left[C(X^t X)^{-1} C^t \right]^{-1} \left(C\hat{\beta} - m \right),$$

and

$$\text{SSE}(F) = y^t (I - H) y,$$

where $H = X(X^t X)^{-1} X^t$. Now, since, $\text{SSE}(R) - \text{SSE}(F)$ and $\text{SSE}(R)$ are independent and $df_R = n - (p - q)$ and $df_F = n - p$,

$$\frac{\text{SSE}(R) - \text{SSE}(F)}{\sigma^2} \sim \chi^2(q)$$

under H_0 and

$$\frac{\text{SSE}(F)}{\sigma^2} \sim \chi^2(n - p).$$

As a result, the test statistic is

$$F = \frac{(\text{SSE}(R) - \text{SSE}(F))/q}{\text{SSE}(F)/(n - p)} \sim F(q, n - p),$$

under H_0 and we reject it if $F > F_\alpha$ with significance level α. Note that when $X^t X$ is singular, the distribution of test statistics is also known as F. In the case, a generalized inverse of $X^t X$ is used.

2.2. The ridge regression model

When multicollinearity exists in the linear regression model, the least squares estimator $\hat{\beta} = (X^t X)^{-1} X^t y$ is very unstable. To overcome this problem, Hoerl and Kennard (1970) suggested the ridge regression estimator defined as

$$\hat{\beta}_R = (X^t X + \theta I)^{-1} X^t y,$$

where θ is a shrinkage parameter that should be estimated based on the data. To estimate θ, the generalized cross validation criterion, suggested by Wahba *et al.* (1979), is often used, and it is given by

$$\text{GCV}_\theta = \sum_i e_{i,\theta}^2 / \left[1 - \frac{1}{n} \text{tr}(H_\theta) \right]^2,$$

where

$$e_{i,\theta} = y_i - \hat{y}_{i,\theta}$$

is a ridge regression residual when θ is given and

$$H_\theta = X(X^t X + \theta I)^{-1} X^t$$

is hat matrix in the ridge regression.

3. The General Linear Test in the Ridge Regression

3.1. Problem set-up

Consider the multiple linear regression model

$$y = X\beta + \varepsilon,$$

and we assume that two or more covariates are highly correlated; therefore, we are required to estimate β via the ridge regression method. Under this situation, we wish to test the general linear hypotheses

$$H_0 : C\beta = m \quad vs. \quad H_1 : C\beta \neq m.$$

Recall that the ridge regression estimator can be interpreted as minimizing $\varepsilon^t \varepsilon$ under the restriction of $\beta^t \beta \leq k^2$ for some k which is related with the shrinkage parameter θ. Therefore, the estimator of β in the ridge regression model under the general linear hypothesis $H_0 : C\beta = m$ reduces to minimizing $\varepsilon^t \varepsilon$ under two restrictions;

$$\beta^t \beta \leq k^2$$

and

$$C\beta = m.$$

3.2. Estimation of regression coefficients

Based on the previous arguments, the problem reduces to minimizing

$$
\begin{aligned}
Q &= \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon} + \theta \left(\boldsymbol{\beta}^t \boldsymbol{\beta} - k^2 \right) + \lambda (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{m}) \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^t (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \theta \left(\boldsymbol{\beta}^t \boldsymbol{\beta} - k^2 \right) + \lambda (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{m}) \\
&= \boldsymbol{y}^t \boldsymbol{y} - 2\boldsymbol{\beta}^t \boldsymbol{X}^t \boldsymbol{y} + \boldsymbol{\beta}^t \boldsymbol{X}^t \boldsymbol{X} \boldsymbol{\beta} + \theta \left(\boldsymbol{\beta}^t \boldsymbol{\beta} - k^2 \right) + \lambda (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{m}),
\end{aligned}
$$

where θ and λ are Lagrangian multipliers. By the partial differentiating Q with respect to by unknown parameters $\boldsymbol{\beta}, \theta$ and λ, we have

$$
\boldsymbol{0} = \frac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^t \boldsymbol{y} + 2\boldsymbol{X}^t \boldsymbol{X} \boldsymbol{\beta} + 2\theta \boldsymbol{\beta} + \boldsymbol{C}^t \lambda, \tag{3.1}
$$

$$
\boldsymbol{0} = \frac{\partial \boldsymbol{Q}}{\partial \theta} = \boldsymbol{\beta}^t \boldsymbol{\beta} - k^2, \tag{3.2}
$$

$$
\boldsymbol{0} = \frac{\partial \boldsymbol{Q}}{\partial \lambda} = \boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{m}. \tag{3.3}
$$

From (3.1), we have

$$
\left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right) \boldsymbol{\beta} = \boldsymbol{X}^t \boldsymbol{y} - \frac{1}{2} \boldsymbol{C}^t \lambda .
$$

Therefore,

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_L &= \left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} \left(\boldsymbol{X}^t \boldsymbol{y} - \frac{1}{2} \boldsymbol{C}^t \lambda \right) \\
&= \hat{\boldsymbol{\beta}}_R - \frac{1}{2} \left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} \boldsymbol{C}^t \lambda ,
\end{aligned}
$$

where $\hat{\boldsymbol{\beta}}_R$ is ridge estimator and $\hat{\boldsymbol{\beta}}_L$ is the estimator of $\boldsymbol{\beta}$ of unknown coefficients under H_0. Also, from (3.3), $\boldsymbol{C}\hat{\boldsymbol{\beta}}_L = \boldsymbol{m}$. Since, we know that

$$
\frac{1}{2} \boldsymbol{C} \left(\boldsymbol{X}^T \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} \boldsymbol{C}^T \lambda = \boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}
$$

and

$$
\lambda = \left[\frac{1}{2} \boldsymbol{C} \left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} \boldsymbol{C}^t \right]^{-1} \left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m} \right),
$$

we can finally get the estimator of coefficients

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_L &= \hat{\boldsymbol{\beta}}_R - \frac{1}{2} \left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} \boldsymbol{C}^t \left[\frac{1}{2} \boldsymbol{C} \left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} \boldsymbol{C}^t \right]^{-1} \left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m} \right) \\
&= \hat{\boldsymbol{\beta}}_R - \left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} \boldsymbol{C}^t \left[\boldsymbol{C} \left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} \boldsymbol{C}^t \right]^{-1} \left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m} \right) \\
&= \hat{\boldsymbol{\beta}}_R - \boldsymbol{K}\boldsymbol{C}^t \left[\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}^t \right]^{-1} \left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m} \right) ,
\end{aligned} \tag{3.4}
$$

where

$$
\boldsymbol{K} = \left(\boldsymbol{X}^t \boldsymbol{X} + \theta \boldsymbol{I} \right)^{-1} .
$$

For a relevant discussion, see Seber and Lee (2003), for example.

3.3. A suggested test statistic

Motivated by the general test statistic in the linear model

$$F = \frac{(\text{SSE}(R) - \text{SSE}(F))/q}{\text{SSE}(F)/(n - p)} \sim F(q, n - p),$$

under $H_0 : \boldsymbol{C\beta} = \boldsymbol{m}$, we wish to compute $\text{SSE}(R) - \text{SSE}(F)$ and $\text{SSE}(F)$. First, by noting that

$$\text{SSE}(R) = \boldsymbol{e}_L{}^t \boldsymbol{e}_L = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right),$$

we have

$$\begin{aligned}
\text{SSE}(R) &= \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_R + \boldsymbol{X}\hat{\boldsymbol{\beta}}_R - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_R + \boldsymbol{X}\hat{\boldsymbol{\beta}}_R - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right) \\
&= \left[\boldsymbol{e}_R + \boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)\right]^t \left[\boldsymbol{e}_R + \boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)\right],
\end{aligned}$$

where

$$\boldsymbol{e}_R = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_R,$$

and therefore,

$$\begin{aligned}
\text{SSE}(R) &= \boldsymbol{e}_R{}^t \boldsymbol{e}_R + 2\left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \boldsymbol{e}_R + \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \boldsymbol{X} \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right) \\
&= \boldsymbol{e}_R{}^t \boldsymbol{e}_R + 2\left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_R\right) + \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \boldsymbol{X} \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right) \\
&= \boldsymbol{e}_R{}^t \boldsymbol{e}_R + \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_R\right) + \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right)\right].
\end{aligned}$$

Hence, we get

$$\text{SSE}(R) = \text{SSE}(F) + \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_R\right) + \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right)\right].$$

Now, we can express $\text{SSE}(R) - \text{SSE}(F)$ as a function of $\boldsymbol{\beta}_R$ by using (3.4), *i.e.*, since,

$$\begin{aligned}
&\left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \left[\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_R\right) + \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right)\right] \\
&= \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \left(\boldsymbol{e}_R + \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right) \\
&= \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \left[\boldsymbol{e}_R + \boldsymbol{y} - \boldsymbol{X}\left[\hat{\boldsymbol{\beta}}_R - \boldsymbol{K}\boldsymbol{C}^t\{\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}^t\}^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)\right]\right] \\
&= 2\left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \boldsymbol{e}_R + \left(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_L\right)^t \boldsymbol{X}^t \boldsymbol{X}\boldsymbol{K}\boldsymbol{C}^t \left[\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}^t\right]^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right),
\end{aligned}$$

and therefore,

$$\begin{aligned}
&\text{SSE}(R) - \text{SSE}(F) \\
&= 2\left[\boldsymbol{K}\boldsymbol{C}^t\left\{\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}^t\right\}^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)\right]^t \boldsymbol{X}^t \boldsymbol{e}_R + \left[\boldsymbol{K}\boldsymbol{C}^t\left\{\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}^t\right\}^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)\right]^t \boldsymbol{X}^t \boldsymbol{X}\boldsymbol{K}\boldsymbol{C}^t\left\{\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}^t\right\}^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right).
\end{aligned}$$

Further, let

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{K}\boldsymbol{C}^t \left(\boldsymbol{C}\boldsymbol{K}\boldsymbol{C}^t\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right),$$

then

$$\text{SSE}(R) - \text{SSE}(F)$$
$$= 2\boldsymbol{T}^t\boldsymbol{e}_R + \boldsymbol{T}^t\boldsymbol{T}$$
$$= 2\left[\boldsymbol{XKC}^t\left(\boldsymbol{CKC}^t\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)\right]^t \boldsymbol{e}_R + \left[\boldsymbol{XKC}^t\left(\boldsymbol{CKC}^t\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)\right]^t \boldsymbol{XKC}^t\left(\boldsymbol{CKC}^t\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right).$$

Based on the result in the classical linear model, we may consider a test statistic for testing $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{m}$ in the ridge regression as

$$G = \frac{(\text{SSE}(R) - \text{SSE}(F))/(df_R - df_F)}{\text{SSE}(F)/df_F}, \tag{3.5}$$

where

$$\text{SSE}(R) - \text{SSE}(F)$$
$$= 2\left[\boldsymbol{XKC}^t\left(\boldsymbol{CKC}^t\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)\right]^t \boldsymbol{e}_R + \left[\boldsymbol{XKC}^t\left(\boldsymbol{CKC}^t\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)\right]^t \boldsymbol{XKC}^t\left(\boldsymbol{CKC}^t\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right),$$

and

$$\text{SSE}(F) = \boldsymbol{e}_R{}^t\boldsymbol{e}_R.$$

Remark 1. $\text{SSE}(R) - \text{SSE}(F)$ is sum of two terms; one is linear in $\boldsymbol{T}$, and the other is quadratic in $\boldsymbol{T}$. Note that the linear term will be close to zero because it is a linear combination of $\boldsymbol{T}$ and the residual vector $\boldsymbol{e}_R$, and the quadratic term will be large compared to the linear term. The quadratic term will be small under H_0 and will be large under H_1 since $\boldsymbol{T}$ is a function of $\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}$.

Remark 2. Recall that df_F in the classical linear model is $n - p$ which is, in fact, $n - \text{tr}(\boldsymbol{H})$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ is the hat matrix. Therefore, it is quite reasonable to define the degree of freedom under full model as $n - \text{tr}(\boldsymbol{H}_\theta)$, where $\boldsymbol{H}_\theta = \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X} + \theta\boldsymbol{I})^{-1}\boldsymbol{X}^t$ is the hat matrix in the ridge regression.

By using arguments in Remark 1 and Remark 2, we may also consider an approximate version of test statistic

$$\tilde{G} = \frac{\boldsymbol{T}^t\boldsymbol{T}/(df_R - df_F)}{\boldsymbol{e}_R{}^t\boldsymbol{e}_R/df_F}$$
$$= \frac{\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)^t\left(\boldsymbol{CKC}^t\right)^{-1}\boldsymbol{CKX}^t\boldsymbol{XKC}^t\left(\boldsymbol{CKC}^t\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R - \boldsymbol{m}\right)/q}{\boldsymbol{e}_R{}^t\boldsymbol{e}_R/(n - \text{tr}(\boldsymbol{H}_\theta))}. \tag{3.6}$$

To use G or $\tilde{G}$ as a test statistic we need to verify the distribution of $\text{SSE}(R) - \text{SSE}(F)$ and $\text{SSE}(F)$ under H_0, however, it is a very difficult task. Independence between $\text{SSE}(R) - \text{SSE}(F)$ and $\text{SSE}(F)$ is required. Here, we conjecture that the proposed test statistic follows $F(q, (n - \text{tr}(\boldsymbol{H}_\theta)))$-distribution under H_0, because, if $\theta = 0$, then the proposed test statistic reduces to the test statistic in the classical linear model. In general, $\text{tr}(\boldsymbol{H}_\theta)$ is not an integer, so that we may use the closest integer to $\text{tr}(\boldsymbol{H}_\theta)$.

Table 1: The average of $T^t e_R$, $T^t T$, G, $\tilde{G}$ based on 100 replications, and θ.

m	$T^t e_R$	$T^t T$	G	$\tilde{G}$	θ
0	0.044	0.010	1.029	1.023	0.013
1	0.032	0.407	42.621	42.604	
2	0.052	1.580	165.459	165.430	
3	0.072	3.529	369.541	369.501	
4	0.092	6.254	654.869	654.818	
5	0.112	9.756	1021.441	1021.380	

3.4. Joint confidence region

For the interval estimation of the general linear combination of regression coefficients $C\beta$ in the ridge regression, Obenchain (1977) proposed a method using a singular values decomposition of the design matrix. But, the suggested interval by Obenchain (1977) is for the orthogonal transformation of regression coefficients.

Based on the arguments in Section 3.3, we suggest an approximate joint confidence region for general linear function $C\beta$. First, we note that

$$E\left(C\hat{\beta}_R\right) = C(I - \theta K)\beta$$

and

$$\mathrm{Cov}\left(C\hat{\beta}_R\right) = KX^T XK\sigma^2.$$

Therefore, an approximate $100 \times (1 - \alpha)\%$ joint confidence region for the general linear function $C\beta$, based on the arguments in Section 3.3, is β satisfying the following inequality;

$$\left(\hat{\beta}_R - (I - \theta K)\beta\right)^T C^T \left(CKX^T XKC^T\right)^{-1} C\left(\hat{\beta}_R - (I - \theta K)\beta\right) \le qs^2 F_\alpha\left(q, (n - \mathrm{tr}(H_\theta))\right).$$

4. Numerical Study
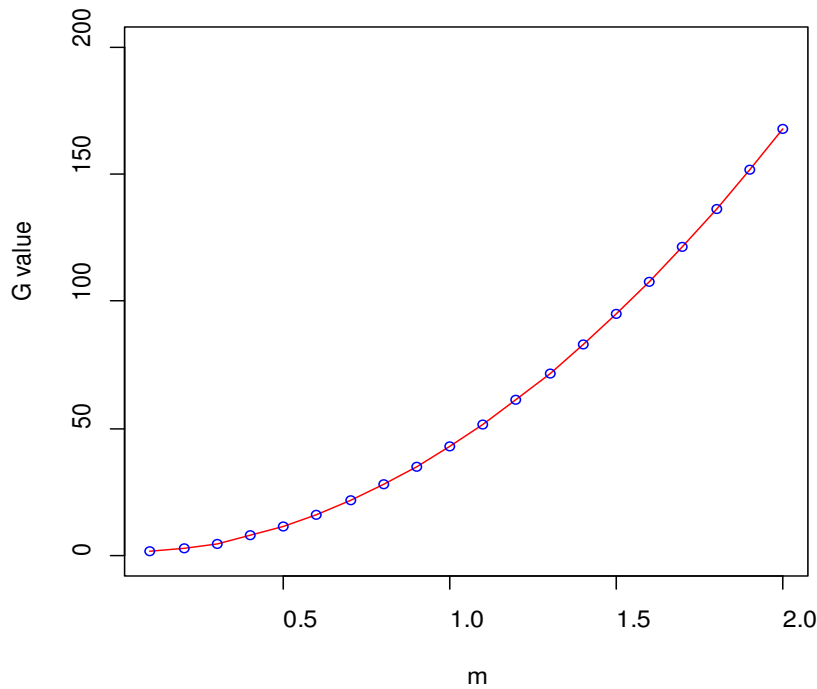
4.1. Independence set-up

First, we generate X_{i1}, X_{i2} and X_{i3} from Uniform$(0, 1)$ independently, and also generate ϵ_i from $N(0, 0.1)$. By supposing $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, we can obtain y_i's. Next, we estimate θ using GCV method and compute the values of $T^t e_R$ and $T^t T$ under $H_0 : C\beta = m$ where $C = (0, 1, -2, 1)$ and $m = 0$. Lastly, compute the proposed test statistic G in (3.5) and the approximate version of test statistic $\tilde{G}$ in (3.6) to check $T^t e_R$ is negligible compared to $T^t T$. By Remark 2, we can define the degree of freedom as

$$df_F = n - \mathrm{tr}(H_\theta) \quad \text{and} \quad df_R = n - \mathrm{tr}(H_\theta) + q.$$

We repeat this procedure for 100 times when $m = 0$, (*i.e.*, under H_0), and $m = 1, 2, 3, 4, 5$ under H_1. Table 1 shows the average values of $T^t e_R$, $T^t T$, G, and $\tilde{G}$ when $m = 0, 1, \ldots, 5$ after 100 replications. We see that $T^t e_R$ is very close to zero as expected, so that $\tilde{G}$ statistic is a reliable substitute for G statistic. Also, we see that G statistic increases as m increases. In this case, the mean of $\mathrm{tr}(H_\theta)$ is 3.977, which is close to $p = 4$, because θ is close to 0. Since, we get $F_{0.05}(1, 26) = 4.225$ by rounding the degree of freedom, and reject H_0 when $m \ge 1$.

Table 2: The average of $T^t e_R$, $T^t T$, G, $\tilde{G}$ based on 100 replications, and θ.

m	$T^t e_R$	$T^t T$	G	$\tilde{G}$	θ
0	0.000	0.011	1.266	1.225	0.024
1	0.001	0.148	16.486	16.375	
2	0.002	0.562	61.830	61.648	
3	0.002	1.253	137.297	137.044	
4	0.003	2.220	242.887	242.564	
5	0.004	3.464	378.601	378.208	

Figure 1: *G(solid) and $\tilde{G}$(dotted) for m = 0.0(0.1)2.0 under independent set-up.*

4.2. Multicollinearity set-up

Note that ridge regression is useful when covariates are correlated. Again, we consider a multiple linear regression model in (4.1). In this case, we generate X_{i2} and X_{i3} from Uniform(0, 1) independently, however, let $X_{i1} = X_{i2} + \delta_i$, where δ_i is generated from $N(0, 0.5)$. Finally, generate ε_i from $N(0, 0.1)$. By supposing $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, we can obtain y_i's. Based on these random numbers, we make the same analysis as in Section 4.1. In this case, the correlation between covariates X_1 and X_2 is 0.7538, and the average of tr(H_θ) is 3.942. Since $F_{0.05}(1, 26) = 4.225$ and $F_{0.10}(1, 26) = 2.909$ by rounding the degree of freedom, we observe that G and $\tilde{G}$ are larger than $F_{0.05}$ when $m \geq 1$. Therefore, we reject H_0 when $m \geq 1$ with significance level $\alpha = 0.05$.

Remark 3. To see the performance of the level control and power of the G and $\tilde{G}$ statistics, we compute G and $\tilde{G}$ when $m = 0.0(0.1)2.0$ under independent and multicollinearity set-up, respectively. These are given in Figure 1 and Figure 2, respectively.
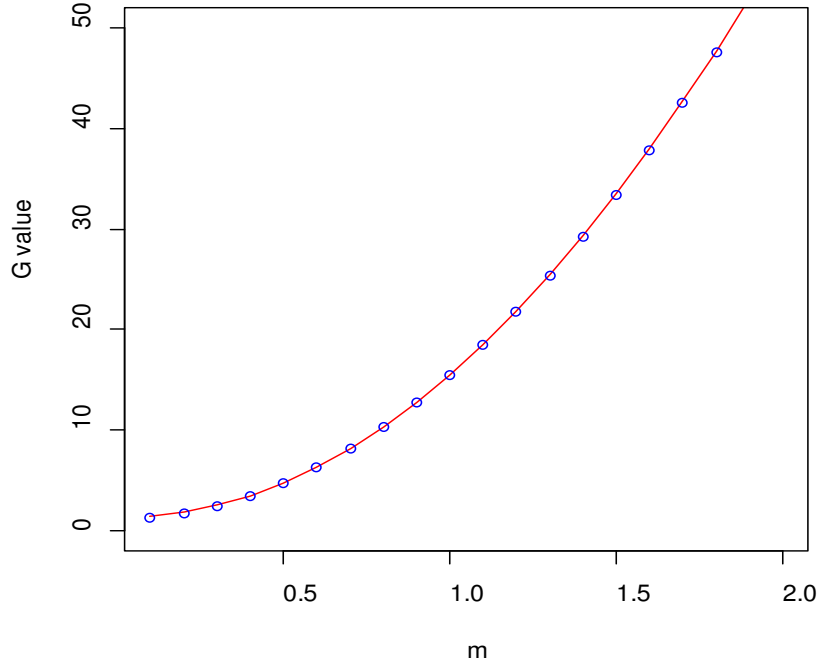
Figure 2: *G(solid) and $\tilde{G}$(dotted) for m = 0.0(0.1)2.0 under multicollinearity set-up.*

4.3. Asymptotic consideration

There is a well-known theorem (Harville, 2008, for example) for the asymptotics of matrices.

Theorem 1. *Let A and B are $n \times n$ matrix, B is non-singular, and $F = B^{-1}A$, then $B^{-1}(I + F + F^2 + \cdots)$ converges iff*

$$\lim_{k \to \infty} F^k = 0$$

and

$$(B - A)^{-1} = \sum_{k=0}^{\infty} F^k B^{-1} = B^{-1}(I + F + F^2 + \cdots),$$

where $F^0 = I$.

By Theorem 1, we first note that

$$K = \left(X^t X + \theta I\right)^{-1} = \left(X^t X - (-\theta I)\right)^{-1}$$

and we put $B = X^t X$ and $A = -\theta I$, then

$$K = \left(X^t X\right)^{-1} \left(I - \theta \left(X^t X\right)^{-1} + \theta^2 \left(X^t X\right)^{-2} - \cdots\right)$$
$$= \left(X^t X\right)^{-1} - \theta \left(X^t X\right)^{-2} + \theta^2 \left(X^t X\right)^{-3} - \cdots. \tag{4.1}$$

Table 3: θ and $\|X^tX\|$ in the independence set-up.

	independence case
θ	0.013
$\|X^tX\|$	7.317
$\|X^tX\|^{-1}$	0.137

Table 4: θ and $\|X^tX\|$ in the multicollinearity set-up, where $\delta_i \sim N(0, \sigma^2)$.

	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$
θ	0.020	0.048	0.056	0.041	0.024
$\|X^tX\|$	7.428	7.358	7.382	7.433	7.424
$\|X^tX\|^{-1}$	0.135	0.136	0.135	0.135	0.135

In this case, if we show that the term θX^tX is negligibly small compared to $(X^tX)^{-1}$, then, K can be well approximated by $(X^tX)^{-1}$. To compare these two values, we calculate them in the independence set-up and the multicollinearity set-up, but we replace the value of X^tX by its spectral norm, *i.e.*,

$$\left\|X^tX\right\| \equiv \sqrt{\lambda_1(X^tX)}, \tag{4.2}$$

where $\lambda_1(X^tX)$ is the largest eigenvalue of X^tX. Table 3 and Table 4 show the average of θ and spectral norm based on 100 replications.

Since, we don't need to use ridge regression in the independence set-up, it is reasonable to get θ that is close to 0. Therefore, from (4.1), we conclude that

$$K \approx \left(X^tX\right)^{-1}.$$

4.4. An example

As an illustrative example, we apply the proposed test statistic to the body fat data (Neter *et al.*, 1990). In this data set, two covariates X_1 and X_2 are highly correlated (Corr$(X_1, X_2) = 0.92$), so that a ridge regression is used with $\theta = 0.02$ by the GCV criterion. The fitted ridge regression is

$$\hat{y} = 3.485 + 0.909X_1 + 0.080X_2 - 0.376X_3.$$

Here, we like to test $H_0 : C\beta = m$, where $C = (0, 1, 1, -1)$ and $m = 0$. Two proposed test statistics are $G = 4.48$ and $\tilde{G} = 3.46$, respectively, and these are larger than $F_{.1}(1, 17) = 3.03$. Therefore, we reject $H_0 : C\beta = m$. When $C = (0, 1, 1, -2)$ and $m = 0$, two test statistics are $G = 2.36$ and $\tilde{G} = 1.56$, respectively, and the null hypothesis is not rejected.

We also applied the data set to the confidence region for β when $C = (0, 1, 1, -1)$ and $C = (0, 1, 1, -2)$, and we obtained the following two inequalities $(1.205 + 0.019\beta_0 - \beta_1 - \beta_2 + \beta_3)^2 \leq 0.186$ and $(1.581 + 0.034\beta_0 - \beta_1 - \beta_2 + 2\beta_3)^2 \leq 0.437$.

5. Concluding Remarks

We derived an exact and approximate version of test statistic to test a general linear hypothesis in the ridge regression model. The exact distribution of the proposed test statistic was not derived, however, we suggested to use an F-distribution as a reference distribution via numerical studies. The numerical performance of the proposed test statistic was very good, and we can use the proposed test statistic in a real data set. The proposed test statistic has very similar form to the classical linear model.

Future studies are required to derive an exact distribution of the proposed test statistic. If the derivation of an exact distribution is not possible, then we may use the facts given in Section 4.3 to derive an approximate distribution, instead.

References

Chipman, J. S. and Rao, M. M. (1964). The treatment of linear restriction in regression analysis, *Econometrica*, **32**, 198–209.

Harville, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*, Springer.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics*, **12**, 55–67.

Hoerl, A. E. and Kennard, R. W. (1990). Ridge regression: Degrees of freedom in the analysis of variance, *Communications in Statistics - Simulation and Computation*, **19**, 1485–1495.

Kim, C. and Kang, K. (2010). *Regression Analysis*, Kyowoosa, Seoul.

Neter, J., Wasserman, W. and Kutner, M. H. (1990). *Applied Linear Statistical Models*, Irwin, Boston.

Obenchain, R. L. (1977). Classical F-tests and confidence regions for ridge regression, *Technometrics*, **19**, 429–439.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, Wiley, New York.

Wahba, G., Golub, G. H. and Heath, M. T. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, **21**, 215–223.