Chapter 14
Stein-Rule Estimation

The ordinary least squares estimation of regression coefficients in linear regression model provides the estimators having minimum variance in the class of linear and unbiased estimators. The criterion of linearity is desirable because such estimators involve less mathematical complexity, they are easy to compute, and it is easier to investigate their statistical properties. The criterion of unbiasedness is attractive because it is intuitively desirable to have an estimator whose expected value, i.e., the mean of the estimator should be the same as the parameter being estimated. Considerations of linearity and unbiased estimators sometimes may lead to an unacceptably high price to be paid in terms of the variability around the true parameter. It is possible to have a nonlinear estimator with better properties. It is to be noted that one of the main objectives of estimation is to find an estimator whose values have high concentration around the true parameter. Sometimes it is possible to have a nonlinear and biased estimator that has smaller variability than the variability of a best linear unbiased estimator of the parameter under some mild restrictions.

In the multiple regression model

$$\underset{n\times 1}{y} = \underset{n\times k}{X}\underset{k\times 1}{\beta} + \underset{n\times 1}{\varepsilon},\ E(\varepsilon)=0,\ V(\varepsilon)=\sigma^2 I\ ,$$

the ordinary least squares estimator (OLSE) of β is $b=(X'X)^{-1}X'y$ which is the best linear unbiased estimator of β in the sense that it is linear in y, $E(b)=\beta$ and b has smallest variance among all linear and unbiased estimators of β. Its covariance matrix is

$$V(b)=E(b-\beta)(b-\beta)'=\sigma^2(X'X)^{-1}.$$

The weighted mean squared error of an estimator $\hat{\beta}$ is defined as

$$E(\hat{\beta}-\beta)'W(\hat{\beta}-\beta)=\sum_i\sum_j w_{ij}\,E(\hat{\beta}_i-\beta_i)(\hat{\beta}_j-\beta_j)$$

where W is $k\times k$ fixed positive definite matrix of weights w_{ij}. The two popular choices of weight matrix W are

(i) W is an identity matrix, i.e. $W=I$ then $E(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$ is called as the **total mean squared error** (MSE) of $\hat{\beta}$.

(ii) $W = X'X$, then

$$E\left(\hat{\beta}-\beta\right)'X'X\left(\hat{\beta}-\beta\right)=E\left(X\hat{\beta}-X\beta\right)'\left(X\hat{\beta}-X\beta\right)$$

is called as the predictive mean squared error of $\hat{\beta}$. Note that $E(y)=X\hat{\beta}$ is the predictor of average value $E(y)=X\beta$ and $\left(X\hat{\beta}-X\beta\right)$ is the corresponding prediction error.

There can be other choices of W and it depends entirely on the analyst how to define the loss function so that the variability is minimum.

If a random vector with k elements $(k>2)$ is normally distributed as $N(\mu, I)$, μ being the mean vector, then Stein established that if the linearity and unbiasedness are dropped, then it is possible to improve upon the maximum likelihood estimator of μ under the criterion of total MSE. Later, this result was generalized by James and Stein for linear regression model. They demonstrated that if the criteria of linearity and unbiasedness of the estimators are dropped, then a nonlinear estimator can be obtained which has better performance than the best linear unbiased estimator under the criterion of predictive MSE. In other words, James and Stein established that OLSE is inadmissible for $k>2$ under predictive MSE criterion, i.e., for $k>2$, there exists an estimator $\hat{\beta}$ such that

$$E\left(\hat{\beta}-\beta\right)'X'X\left(\hat{\beta}-\beta\right)\leq E(b-\beta)'X'X(b-\beta)$$

for all values of β with strict inequality holding for some values of β. For $k\leq 2$, no such estimator exists and we say that "b can be beaten" in this sense. Thus it is possible to find estimators which will beat b in this sense. So a nonlinear and biased estimator can be defined which has better performance than OLSE. Such an estimator is **Stein-rule estimator** given by

$$\hat{\beta}=\left[1-c\frac{\sigma^2}{b'X'Xb}\right]b \text{ when } \sigma^2 \text{ is known}$$

and

$$\hat{\beta}=\left[1-c\frac{e'e}{b'X'Xb}\right]b \text{ when } \sigma^2 \text{ is unknown.}$$

Here c is a fixed positive characterizing scalar, $e'e$ is the residuum sum of squares based on OLSE and $e=y-Xb$ is the residual. By assuming different values to c, we can generate different estimators. So a class of estimators characterized by c can be defined. This is called as a family of Stein-rule estimators.

Let

$$\left[1-c\frac{\sigma^2}{b'X'Xb}\right]=\delta$$

be a scalar quantity. Then

$$\hat{\beta}=\delta b.$$

So essentially we say that instead of estimating $\beta_1,\beta_2,...,\beta_k$ by $b_1,b_2,...,b_k$ we estimate them by $\delta b_1,\delta b_2,...,\delta b_k$, respectively. So in order to increase the efficiency, the OLSE is multiplied by a constant δ. Thus δ is called the **shrinkage factor**. As Stein-rule estimators attempt to shrink the components of b towards zero, so these estimators are known as **shrinkage estimators.**

First, we discuss a result which is used to prove the dominance of Stein-rule estimator over OLSE.

Result: Suppose a random vector Z of order $(k\times 1)$ is normally distributed as $N(\mu,I)$ where μ is the mean vector and I is the covariance matrix.

Then

$$E\left[\frac{Z'(Z-\mu)}{Z'Z}\right]=(k-2)E\left(\frac{1}{Z'Z}\right).$$

An important point to be noted in this result is that the left-hand side depends on μ , but the right-hand side is independent of μ.

Now we consider the Stein-rule estimator $\hat{\beta}$ when σ^2 is known. Note that

$$\begin{aligned}E\left(\hat{\beta}\right)&=E(b)-c\sigma^2E\left(\frac{b}{b'X'Xb}\right)\\&=\beta-\text{(In general, a non-zero quantity)}\\&\neq 0,\end{aligned}$$

in general.

Thus the Stein-rule estimator is biased while OLSE b is unbiased for β.

The predictive risk of b and $\hat{\beta}$ are

$$PR(b) = E(b-\beta)'X'X(b-\beta)$$

$$PR(\hat{\beta}) = E(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta).$$

The Stein-rule estimator $\hat{\beta}$ is better them OLSE b under the criterion of predictive risk if

$$PR(\hat{\beta}) < PR(b).$$

Solving the expressions, we get

$$\begin{aligned}
b-\beta &= (X'X)^{-1}X'\varepsilon \\
PR(b) &= E\left[\varepsilon'X(X'X)^{-1}X'X(X'X)^{-1}X'\varepsilon\right] \\
&= E\left[\varepsilon'X(X'X)^{-1}X'\varepsilon\right] \\
&= E\left[tr(X'X)^{-1}X'\varepsilon\varepsilon'X\right] \\
&= tr(X'X)^{-1}X'E(\varepsilon\varepsilon')X \\
&= \sigma^2 tr(X'X)^{-1}X'X \\
&= \sigma^2 \;\; trI_k \\
&= \sigma^2 k.
\end{aligned}$$

$$\begin{aligned}
PR(\hat{\beta}) &= E\left[(b-\beta) - \frac{c\sigma^2}{b'X'Xb}b\right]' X'X\left[(b-\beta) - \frac{c\sigma^2}{b'X'Xb}\right] \\
&= E(b-\beta)'X'X(b-\beta) - E\left[\frac{c\sigma^2}{b'X'Xb}\left\{b'X'X(b-\beta) + (b-\beta)'X'Xb\right\}\right] \\
&\quad + E\left[\frac{c^2\sigma^4}{(b'X'Xb)^2}b'X'Xb\right] \\
&= \sigma^2 k - 2E\left[\frac{c\sigma^2(b-\beta)'X'Xb}{b'X'Xb}\right] + E\left[\frac{c^2\sigma^4}{b'X'Xb}\right].
\end{aligned}$$

Suppose

$$Z = \frac{1}{\sigma}\left(X'X\right)^{1/2} b$$

or $b = \sigma\left(X'X\right)^{-1/2} Z$

$$\mu = \frac{1}{\sigma}\left(X'X\right)^{1/2} \beta$$

or $\beta = \sigma\left(X'X\right)^{-1/2} \mu$

and $Z \sim N\left(\mu, I\right)$, i.e., $Z_1, Z_2, ..., Z_k$ are independent. Substituting these values in the expressions for $PR\left(\hat{\beta}\right)$, we get

$$\begin{aligned}
PR\left(\hat{\beta}\right) &= \sigma^2 k - 2E\left[c\sigma^2 \frac{\sigma^2\left(Z-\mu\right)'Z}{\sigma^2 Z'Z}\right] + E\left[\frac{c^2\sigma^4}{\sigma^2 Z'Z}\right] \\
&= \sigma^2 k - 2E\left[c\sigma^2 \frac{Z'\left(Z-\mu\right)}{Z'Z}\right] + E\left[\frac{c^2\sigma^2}{Z'Z}\right] \\
&= \sigma^2 k - 2c\sigma^2 E\left[\frac{Z'\left(Z-\mu\right)}{Z'Z}\right] + c^2\sigma^2 E\left(\frac{1}{Z'Z}\right) \\
&= \sigma^2 k - c\sigma^2\left[2\left(k-2\right)-c\right]E\left(\frac{1}{Z'Z}\right) \quad \left(\text{using the result}\right) \\
&= PR\left(b\right) - c\sigma^2\left[2\left(k-2\right)-c\right]E\left(\frac{1}{Z'Z}\right).
\end{aligned}$$

Thus

$$PR\left(\hat{\beta}\right) < PR\left(b\right)$$

if and only if

$$c\sigma^2\left[2\left(k-2\right)-c\right]E\left(\frac{1}{Z'Z}\right) > 0.$$

Since $Z \sim N\left(\mu, I\right)$, $\sigma^2 > 0.$ So $Z'Z$ has a non-central Chi-square distribution. Thus

$$\begin{aligned}
&E\left(\frac{1}{Z'Z}\right) > 0 \\
&\Rightarrow c\left[2\left(k-2\right)-c\right] > 0.
\end{aligned}$$

Since $c > 0$ is assumed, so this inequality holds true when

$$2(k-2)-c>0$$

or $0<c<2(k-2)$ provided $k>2$.

So as long as $0<c<2(k-2)$ is satisfied, the Stein-rule estimator will have smaller predictive risk them OLSE. This inequality is not satisfied for $k=1$ and $k=2$.

To find the value of *c* for which $PR(\hat{\beta})$ is minimum, we differentiate

$$PR(\hat{\beta}) = PR(b) - c\sigma^2\left[2(k-2)-c\right]E\left(\frac{1}{Z'Z}\right)$$

with respect to *c* and it gives as follows:

$$\frac{d\left(PR(\hat{\beta})\right)}{dc} = \frac{d\left(PR(b)\right)}{dc} - \sigma^2 E\left(\frac{1}{Z'Z}\right)\frac{d\left[2(k-2)c-c^2\right]}{dc} = 0$$

$$\Rightarrow 2(k-2)-2c=0.$$

or $c=k-2$.

Further,

$$\left.\frac{d^2\left(PR(\hat{\beta})\right)}{dc^2}\right|_{c=k-2} = 2 > 0.$$

The largest gains efficiency arises when $c=k-2$. So if the number of explanatory variables are more than two, then it is always possible to construct an estimator which is better than OLSE.

The optimum Stein-rule estimator or James-Stein rule estimator of β in this case is given by

$$\hat{\beta} = \left[1-\frac{(k-2)\sigma^2}{b'X'Xb}\right]b \text{ when } \sigma^2 \text{ is known.}$$

To avoid the change of sign in this estimator, the "positive part" version of this estimator called as Positive part Stein-rule estimator is given by

$$\hat{\beta}^{+} = \begin{cases} \left[1 - \dfrac{(p-2)\sigma^2}{b'X'Xb}\right]b & \text{when} \quad 0 < \dfrac{(p-2)\sigma^2}{b'X'Xb} < 1 \\ 0 & \text{when} \quad \dfrac{(p-2)\sigma^2}{b'X'Xb} > 1. \end{cases}$$

When σ^2 is unknown then it can be shown that the Stein-rule estimator

$$\hat{\beta} = \left[1 - c\frac{e'e}{b'X'Xb}\right]b$$

is better than OLSE b if and only if

$$0 < c < \frac{2\left(k-2\right)}{n-k+2};\ k > 2.$$

The optimum choice of c giving the largest gain is efficiency is

$$c = \frac{k-2}{n-k+2}.$$