Introduction to Statistics
(Lectures 7-8: Estimation)

1 Estimation

OVERVIEW: Lectures 7-8 will focus on *estimation*. As we have mentioned before, a strategy in the estimation problem is an *estimator*. The loss function used for this problem is the squared distance between the estimated value and the true parameter. This loss gives rise to the popular *mean squared error* performance criterion.

We will analyze two commonly used estimation strategies—Maximum Likelihood (ML) and the Ridge estimator. We will do this in the context of the homoskedastic Normal regression model with (known variance); in which the conditional distribution of $Y|X$ is modeled as

$$Y|X \sim \mathcal{N}(X\beta, \sigma^2 \mathbb{I}_n), \tag{1.1}$$

where Y is the $n \times 1$ vector of outcome variables, X is the $n \times k$ matrix of regressors, and $\beta \in \mathbb{R}^k$ is the parameter of interest. We denote by P_X the distribution of the regressors, and we assume that this distribution is known and does not depend on β.

We will show that if $k \leq n$ and X has full-column rank, the Maximum Likelihood estimator for the model in (1.1) is the OLS estimator

$$\widehat{\beta}_{\text{OLS}} \equiv (X'X)^{-1}X'Y,$$

(which we introduced in the last problem set). We will compute the risk of this estimator and we will show that this estimator is minimax. We also show that, despite the minimaxity, when $k \geq 3$ the estimator is dominated and we present an "empirical Bayes" estimator that dominates it.

We also show that the *Ridge estimator* is a Bayes decision rule for the prior distribution $\beta \sim \mathcal{N}_k(0, (\sigma^2/\lambda)\mathbb{I}_k)$. The Ridge estimator is well-defined regardless of the number of covariates and it is admissible by construction. The Ridge estimator provides a simple example of "shrinkage" or "regularization". A practical concern is that the Ridge estimator is biased, and the bias depends on the choice of the prior hyperparameters. Moreover, we show that the minimax risk of the Ridge estimator equals infinity.

One result that we will not cover in the notes, but that is important to keep in mind is that despite the inadmissibility of the OLS estimator for the *full* vector of coefficients β, OLS is admissible for the problem in which we are interested in estimating only one regression coefficient but controlling for other variables.

1.1 Mean-squared estimation error

In an estimation problem the action space equals the parameter space. In the context of the linear regression model, we will assume that parameter space simply refers to all possible values that β can take, which we take to be any vector in $\mathbb{R}^k$. This means that we are assuming that both σ^2 and $\mathbb{P}_X$ are known.

The loss function for this problem is the so-called *quadratic loss*, defined as:

$$\mathcal{L}(a,\beta) = ||a-\beta||^2 = (a-\beta)'(a-\beta) = \sum_{j=1}^{k}(a_j-\beta_j)^2,$$

where β_j represents the j-th coordinate of the k-dimensional vector β.

Decision rules in the estimation problem are called estimators. Thus, an estimator—which we will denote as $\widehat{\beta}$—is a map from data (Y,X) to the parameter space. Its risk is given by

$$\mathbb{E}_{\mathbb{P}_\beta}[\mathcal{L}(\widehat{\beta},\beta)] = \mathbb{E}_{\mathbb{P}_\beta}[(\widehat{\beta}-\beta)'(\widehat{\beta}-\beta)]. \tag{1.2}$$

Equation (1.2) is referred to as *the mean-squared estimation error* at β, and denoted $\text{MSE}(\widehat{\beta},\beta)$.

It will be convenient to express the mean squared error in terms of the "bias" and "variance" of $\widehat{\beta}$. Assuming that $\overline{\beta} \equiv \mathbb{E}_\beta[\widehat{\beta}]$ is finite, we define the <u>bias</u> of $\widehat{\beta}$ at β as

$$B_\beta(\widehat{\beta}) = \overline{\beta} - \beta.$$

If the covariance matrix of $\widehat{\beta}$—denoted $\text{V}_\beta(\widehat{\beta})$—is also finite, the mean squared-error can be written as

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\beta}[(\widehat{\beta}-\beta)'(\widehat{\beta}-\beta)] &= \mathbb{E}_{\mathbb{P}_\beta}[(\widehat{\beta}-\overline{\beta}+\overline{\beta}-\beta)'(\widehat{\beta}-\overline{\beta}+\overline{\beta}-\beta)], \\
&= (\overline{\beta}-\beta)'(\overline{\beta}-\beta) + \mathbb{E}_{\mathbb{P}_\beta}[(\widehat{\beta}-\overline{\beta})'(\widehat{\beta}-\overline{\beta})] \\
&= ||B_\beta(\widehat{\beta})||^2 + \text{tr}\left(\text{V}_\beta(\widehat{\beta})\right),
\end{aligned}$$

where $\text{tr}(\cdot)$ is the trace operator. The decomposition is fairly straightforward, but it highlights the fact that the bias and variance of the estimator fully determine its risk whenever the loss is quadratic. Also, the correlation between any of the components of $\widehat{\beta}$ is not relevant for the risk calculation. In the next subsections we will compute the mean-squared estimation error of some popular estimators for model (1.1), including Bayes and minimax estimators.

1.2 The Maximum Likelihood Estimator and its mean-squared error

According to model (1.1), the distribution of $Y|X$ is a multivariate normal with parameters $X\beta$ and $\sigma^2\mathbb{I}_n$ and thus has a p.d.f given by

$$f(Y|\beta, X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right). \tag{1.3}$$

This is true, regardless of whether we have very many covariates or not.

Assume that the distribution of X has a density $f_X(X)$ that does not depend on β. For a fixed realization of the data, define the <u>likelihood function</u>, $L(\beta; (Y, X))$, as the value attained by the p.d.f. of the joint distribution of (Y, X) at different values of the parameter vector β.[1] Thus,

$$L(\beta, (Y, X)) \equiv f(Y|\beta, X) f_X(X).$$

The <u>maximum likelihood estimator</u> of β is then defined as the value of β that maximizes the likelihood; that is

$$\widehat{\beta}_{\text{ML}} \equiv \text{argmax}_{\beta \in \mathbb{R}^k} L(\beta, (Y, X)).$$

The likelihood function implied by (1.3) is decreasing in $(Y - X\beta)'(Y - X\beta)$, a term which is usually referred to as the sum of squared residuals. Therefore, maximizing the likelihood is equivalent to solving the problem

$$\min_{\beta \in \mathbb{R}^k} (Y - X\beta)'(Y - X\beta).$$

The first-order conditions for the program above, which are necessary and sufficient, yield

$$X'(Y - X\widehat{\beta}_{\text{ML}}) = \mathbf{0}_{k\times 1} \iff X'Y = (X'X)\widehat{\beta}_{\text{ML}}.$$

If $k > n$, there are infinitely many solutions that maximize the likelihood. If $k \leq n$, and X has full-column rank there is a unique solution to this problem given by the ordinary least-squares (OLS) estimator:

$$\widehat{\beta}_{\text{ML}} = \widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y. \tag{1.4}$$

Mean-squared Error of the Maximum Likelihood Estimator: We have shown that we can express the mean-squared error of any estimator in terms of its bias and variance. Under the assumptions we have made for the distribution of $Y|X$, it is possible to show that the bias of the

[1] For a textbook definition of the likelihood function for parametric models see Chapter 10.3 of Brucen Hansen's book urlhttps://www.ssc.wisc.edu/ bhansen/probability/Probability.pdf

maximum likelihood estimator is zero, since:

$$\mathrm{B}_\beta(\widehat{\beta}_{\mathrm{ML}}) \equiv \mathbb{E}_{\mathbb{P}_\beta}[(X'X)^{-1}X'Y] - \beta = \mathbf{0}_{k\times 1}. \tag{1.5}$$

for any β. Thus, we say that the OLS estimator of β (which is the ML estimator for problem (1.1)) is unbiased.[2]

The variance of the ML estimator is

$$\begin{aligned}
\mathbb{V}_\beta(\widehat{\beta}_{\mathrm{ML}}) &= \mathbb{E}_{\mathbb{P}_\beta}[(\widehat{\beta}_{\mathrm{ML}} - \beta)(\widehat{\beta}_{\mathrm{ML}} - \beta)'] \\
&= \mathbb{E}_{f_X}\left[\mathbb{E}_{\mathbb{P}_\beta}[(\widehat{\beta}_{\mathrm{ML}} - \beta)(\widehat{\beta}_{\mathrm{ML}} - \beta)'|X]\right] \\
&= \mathbb{E}_{f_X}\left[(X'X)^{-1}X'\mathbb{E}_{\mathbb{P}_\beta}[(Y - X\beta)(Y - X\beta)']X(X'X)^{-1}\right] \\
&= \sigma^2\mathbb{E}_{f_X}\left[(X'X)^{-1}\right].
\end{aligned}$$

This means that the mean squared error at β—henceforth denoted $\mathrm{MSE}(\beta;\widehat{\beta}_{\mathrm{ML}})$—equals

$$\sigma^2\mathrm{tr}\left(\mathbb{E}_{f_X}\left[(X'X)^{-1}\right]\right).$$

for any β. Consequently, an interesting property of the ML/OLS estimator is that its risk function is constant over the parameter space.

The decision-theoretic approach to statistics that was presented last class is devoted to the problem of finding rules that, in one way or another, are optimal. Thus, in such a framework there is only room for maximum likelihood estimation insofar as such procedure has good risk properties (at least within some class of decision rules).

It is possible to provide such justification for the maximum likelihood estimator of model (1.1). In particular, it is possible to show that the OLS estimator minimizes risk among the class of all unbiased estimators, provided some regularity conditions are met. This is a version of the celebrated *Gauss-Markov* theorem. There is a proof of this statement in Section 4.8 of Bruce Hansen's book.[3]. I encourage you to glance at this section. My focus on these lecture notes is to relate the OLS estimator to the Bayes and Minimax procedures we have discussed previously.

[2] An estimator $\widehat{\beta}$ is unbiased if $\mathbb{E}_\beta[\widehat{\beta}] = \beta$ for all β in the parameter space.

[3] https://www.ssc.wisc.edu/ bhansen/econometrics/Econometrics.pdf

1.3 Bayesian and Minimax estimation of β

1.3.1 Bayes Estimation of β under a Normal Prior

In this subsection we analyze the Bayes estimator of the parameter β. We have already showed that any Bayes rule can be obtained by minimizing posterior loss for each data realization. Let π denote a prior over the parameter β. In our set-up the posterior loss of an action a is

$$\mathbb{E}_\pi[\, ||a-\beta||^2 \mid (Y,X)\,].$$

Using the same argument that we used to decompose the mean squared-error in terms of bias and variance we can show that

$$\begin{aligned}\mathbb{E}_\pi[\, ||a-\beta||^2 \mid (Y,X)\,] &= \mathbb{E}[\, ||a-\mathbb{E}_\pi[\,\beta \mid (Y,X)\,] + \mathbb{E}_\pi[\beta \mid (Y,X)] - \beta||^2 \mid (Y,X)\,], \\ &= ||a-\mathbb{E}_\pi[\beta \mid (Y,X)]||^2 + \operatorname{tr}\left(\mathbb{V}_\pi(\beta \mid (Y,X))\right).\end{aligned}$$

This shows that, regardless of the specific prior we pick, the Bayes estimator is the posterior mean of β

$$\widehat{\beta}_{\text{Bayes}} = \mathbb{E}_\pi[\,\beta \mid (Y,X)\,].$$

Thus, reporting the posterior mean of β is a Bayes decision rule.

Posterior Mean under a Normal Prior: Consider then the following prior on β:

$$\beta \sim \pi(\beta) \equiv \mathcal{N}_k(\beta_0\,,\, \sigma^2 V^{-1}). \tag{1.6}$$

The prior assumes that all the coefficients are approximately normal with values close to the vector β_0 and covariance matrix given by $\sigma^2 V^{-1}$. There is no magical recipe to select a prior. More often than not, the selection of a prior trades-off interpretation and convenience in its implementation.

We derive the posterior distribution of β. One way of deriving this posterior distribution is by an application of Bayes Theorem

$$\pi(\beta \mid \sigma^2, y, X) = \frac{f(Y \mid \beta, X)\pi(\beta)}{\int_\Theta f(Y,\, \mid \beta, X)\pi(\beta)d\beta}.$$

The posterior is thus proportional to the likelihood times the prior, both of which are Gaussian. Consequently, $f(Y|X,\beta,\sigma^2)\,\pi(\beta|\sigma^2)$ is—up to a constant that does not depend on β—proportional to

$$\exp\left(-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right)\exp\left(-\frac{1}{2\sigma^2}(\beta-\beta_0)'V(\beta-\beta_0)\right). \tag{1.7}$$

The expression above equals:

$$\exp\left(-\frac{1}{2\sigma^2}Y'Y\right)\exp\left(-\frac{1}{2\sigma^2}\beta\left(V+X'X\right)\beta+\frac{1}{\sigma^2}(Y'X+V\beta_0)\beta\right).$$

Completing the square and ignoring all the terms that do not have β on them, gives the posterior distribution as a constant times the exponential of:

$$-\frac{1}{2\sigma^2}\left(\beta-\left(V+X'X\right)^{-1}\left(X'y+V\beta_0\right)\right)\left(V+X'X\right)\left(\beta-(V^{-1}+X'X)^{-1}(X'Y+V\beta_0)\right).$$

This implies that:

$$\beta|Y,X\sim\mathcal{N}_k\left(\left(V+X'X\right)^{-1}\left(X'Y+V\beta_0\right),\,\sigma^2\left(V+X'X\right)^{-1}\right). \tag{1.8}$$

This means that the Bayesian Estimator of β given the Gaussian prior $\pi(\beta)$ is:

$$\widehat{\beta}_{\text{Bayes}}\equiv\left(V+X'X\right)^{-1}\left(X'Y+V\beta_0\right).$$

The posterior mean estimator is well defined regardless the number of covariates.[4] The posterior mean estimator for $V=\lambda\mathbb{I}_k$ and $\beta_0=0$ is called the Ridge estimator:

$$\widehat{\beta}_{\text{Ridge}}\equiv(X'X+\lambda\mathbb{I}_k)^{-1}X'Y,$$

which, by construction, is admissible.[5]

Posterior Mean as a Penalized/Regularized OLS estimator: Equation (1.7) is decreasing as a function of the *penalized* sum of squared residuals

$$(y-X\beta)'(y-X\beta)+(\beta-\beta_0)'V(\beta-\beta_0).$$

Under the Gaussian posterior the posterior mean and posterior mode are the same. Therefore, another way of deriving the posterior mean estimator is by finding a solution to the problem:

$$\min_{\beta}(y-X\beta)'(y-X\beta)+(\beta-\beta_0)'V(\beta-\beta_0)$$

The F.O.C are

$$X'(Y-X\beta)+V(\beta-\beta_0)=\mathbf{0}_{n\times k}\iff X'Y+V\beta_0=(X'X+V)\beta.$$

[4] If V is positive definite, then the matrix $V+X'X$ is positive definite as well, and thus invertible.

[5] See Chapters 29.5-29.7 of Bruce Hansen's book for a definition of Ridge regression and some of its properties.

Solving for β gives the estimator $\widehat{\beta}_{\text{Bayes}}$.

The Ridge estimator is biased (conditionally and unconditionally):

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_\beta}[\widehat{\beta}_{\text{Ridge}}|X] &= (X'X + \lambda\mathbb{I}_k)^{-1}X'X\beta, \\ &= (X'X + \lambda\mathbb{I}_k)^{-1}(X'X + \lambda\mathbb{I}_k)\beta - \lambda(X'X + \lambda\mathbb{I}_k)^{-1}\beta, \\ &= (\mathbb{I}_k - \lambda(X'X + \lambda\mathbb{I}_k)^{-1})\beta.\end{aligned}$$

The magnitude of the bias depends on λ and β.

The variance of the Ridge estimator can be computed using the law of total variance:

$$\mathbb{V}_\beta(\widehat{\beta}_{\text{Ridge}}) = \mathbb{V}_{f_X}\left(\mathbb{E}_{\mathbb{P}_\beta}[\widehat{\beta}_{\text{Ridge}}|X]\right) + \mathbb{E}_{f_X}\left[\mathbb{V}_\beta(\widehat{\beta}_{\text{Ridge}}|X)\right],$$

where

$$\mathbb{V}_\beta(\widehat{\beta}_{\text{Ridge}}|X) = \sigma^2(X'X + \lambda\mathbb{I}_k)^{-1}X'X(X'X + \lambda\mathbb{I}_k)^{-1}.$$

Now that we have computed the bias and variance: what can we say about the mean-squared error (MSE) of the Ridge estimator vis-à-vis that of OLS?

Since the Ridge estimator is admissible, it is not possible for the MSE of OLS to be smaller everywhere in the parameter space (otherwise, the Ridge estimator would not be admissible!). This means there are points in the parameter space in which the MSE of the Ridge estimator has to be below that of OLS.

This observation is true for any distribution of covariates for which the risk is well-defined. In particular, for distributions that are degenerate at some parameter value this means that variance of the Ridge estimator has to be smaller than that of the ML/OLS estimator.

Finally, it is worth mentioning that although—by construction—the Ridge estimator minimizes average risk (with respect to the prior in (1.6)), its worst-case risk is infinity. To see this, it suffices to look at the expression for the bias of the Ridge estimator, and notice that if $||\beta||$ is unbounded, so is the bias.

1.4 Minimax Estimation of β.

In the context of the regression model, a minimax estimator solves the following problem.

$$\inf_{\widehat{\beta}} \sup_{\beta} MSE(\widehat{\beta}, \beta).$$

In this subsection we will show that the OLS estimator is minimax when $k < n$. This is, we will show that:

$$\inf_{\widehat{\beta}} \sup_{\beta} \text{MSE}(\widehat{\beta}, \beta) = \sup_{\beta} \text{MSE}(\widehat{\beta}_{\text{OLS}}, \beta) \tag{1.9}$$

The argument is constructive, and will actually use the Ridge estimator (and its risk) to establish the minimaxity of OLS. Note first that for any estimator (not only OLS):

$$\inf_{\widehat{\beta}} \sup_{\beta} \text{MSE}(\widehat{\beta}, \beta) \leq \sup_{\beta} \text{MSE}(\widehat{\beta}_{\text{OLS}}, \beta).$$

Thus, it is sufficient to show that

$$\inf_{\widehat{\beta}} \sup_{\beta} \text{MSE}(\widehat{\beta}, \beta) \geq \sup_{\beta} \text{MSE}(\widehat{\beta}_{\text{OLS}}, \beta).$$

The following inequality relating Bayes risk and minimax risk holds for any Bayesian estimator:

$$\inf_{\widehat{\beta}} \sup_{\beta} \text{MSE}(\widehat{\beta}, \beta) \geq \inf_{\widehat{\beta}} \mathbb{E}_{\pi} \left[\text{MSE}(\widehat{\beta}, \beta) \right].$$

Let π_λ denote the prior in (1.6) and let $\widehat{\beta}_\lambda$ denote the Ridge estimator corresponding to that prior. The equation in the previous display then implies that for any λ:

$$\inf_{\widehat{\beta}} \sup_{\beta} \text{MSE}(\widehat{\beta}, \beta) \geq \mathbb{E}_{\pi_\lambda} \left[\text{MSE}(\widehat{\beta}_\lambda, \beta) \right].$$

Take a sequence $\lambda_m \to 0$. Under any such sequence $\widehat{\beta}_{\lambda_m}$ converges (for every data realization) to $\widehat{\beta}_{\text{OLS}}$. It can be shown that under some regularity conditions on f_X it then follows:

$$\mathbb{E}_{\pi_\lambda} \left[\text{MSE}(\widehat{\beta}_\lambda, \beta) \right] \to \text{MSE}(\widehat{\beta}_{\text{OLS}}, \beta).$$

Since the MSE of $\widehat{\beta}_{\text{OLS}}$ is constant, then the result follows. The argument used to establish the minimaxity of OLS is standard. A textbook reference for it can be found in Theorem 1.12 of Section 5 in Lehmann and Casella (1998).

1.5 Suboptimality of the OLS estimator

Even though the OLS estimator is minimax, it is well known that if $k \geq 3$, then the estimator is dominated. The result is usually attributed to Charles Stein and his 1956 paper "Inadmissibility of the usual estimator of the mean of a multivariate distribution", which focuses on a multivariate

normal model and not explicitly on the OLS. The estimator that dominates the usual mean is also often attributed to Willard James and Charles Stein in 1961 and referred to as the James-Stein estimator.

This section we provide a superficial presentation of the estimator that dominates OLS. For simplicity, we will consider the case in which f_X is degenerate at a particular point.

Assume that $X'X = \mathbb{I}_n$. Start with a Bayesian estimator for β under the normal prior $\beta \sim \mathcal{N}_k(0, v\mathbb{I}_k) = \mathcal{N}_k(0, \sigma^2(v/\sigma^2)\mathbb{I}_k)$. We have shown that such Bayes estimator is

$$\begin{aligned}
\widehat{\beta}_{\text{Bayes}} &= \left(\mathbb{I}_k + \frac{\sigma^2}{v}\mathbb{I}_k\right)^{-1} \widehat{\beta}_{\text{OLS}}, \\
&= \left(\frac{v}{v+\sigma^2}\right) \widehat{\beta}_{\text{OLS}}, \\
&= \left(1 - \frac{\sigma^2}{v+\sigma^2}\right) \widehat{\beta}_{\text{OLS}}
\end{aligned}$$

The hyperparameter v "shrinks" the OLS estimator towards the prior mean. Instead of picking v *a priori*, it is possible to use data to estimate it. Such an approach is usually called "empirical Bayes".

The distribution of the data conditional on the parameter is

$$\widehat{\beta}_{\text{OLS}} \sim \mathcal{N}_k(\beta, \sigma^2\mathbb{I}_k).$$

This conditional distribution, along with the prior, specify a full joint distribution over $(\widehat{\beta}_{\text{OLS}}, \beta)$. The marginal distribution of the data is

$$\widehat{\beta}_{\text{OLS}} \sim \mathcal{N}_k\left(0, (\sigma^2 + v)\mathbb{I}_k\right).$$

We can use this statistical model to estimate the "shrinkage" factor that appears in the Bayes estimator. Note that

$$||\widehat{\beta}_{\text{OLS}}||^2 \sim (\sigma^2 + v)\chi^2_k,$$

Therefore, standard results for the mean of the inverse of a chi-square distribution yield

$$\mathbb{E}\left[\frac{k-2}{||\widehat{\beta}_{\text{OLS}}||^2}\right] = \frac{1}{v+\sigma^2},$$

provided $k \geq 3$. An unbiased estimator for the "shrinkage" factor that appears in the formula of

$\widehat{\beta}_{\text{Bayes}}$ is thus:

$$\left(1 - \frac{\sigma^2(k-2)}{||\widehat{\beta}_{\text{OLS}}||^2}\right).$$

We now show that the "empirical Bayes" estimator

$$\widehat{\beta}_{JS} \equiv \left(1 - \frac{\sigma^2(k-2)}{||\widehat{\beta}_{\text{OLS}}||^2}\right)\widehat{\beta}_{\text{OLS}},$$

dominates OLS.

Proposition 1. *Suppose $X'X = \mathbb{I}_k$ and $\sigma^2 = 1$. If $k \geq 3$, the James-Stein estimator dominates the ML/OLS estimator.*

Proof.

$$||\widehat{\beta}_{\text{JS}} - \widehat{\beta}_{\text{OLS}}||^2 = ||\widehat{\beta}_{\text{JS}} - \beta||^2 + ||\beta - \widehat{\beta}_{\text{OLS}}||^2 + 2(\widehat{\beta}_{\text{JS}} - \beta)'(\beta - \widehat{\beta}_{\text{OLS}})$$

implies

$$||\widehat{\beta}_{\text{JS}} - \beta||^2 = ||\widehat{\beta}_{\text{JS}} - \widehat{\beta}_{\text{OLS}}||^2 - ||\beta - \widehat{\beta}_{\text{OLS}}||^2 + 2(\widehat{\beta}_{\text{JS}} - \beta)'(\widehat{\beta}_{\text{OLS}} - \beta).$$

Taking expectation on both sides yields

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\beta}\left[||\widehat{\beta}_{\text{JS}} - \beta||^2\right] &= \mathbb{E}_{P_\beta}\left[||\widehat{\beta}_{\text{JS}} - \widehat{\beta}_{\text{OLS}}||^2\right] \\
&- \mathbb{E}_{\mathbb{P}_\beta}\left[||\widehat{\beta}_{\text{OLS}} - \beta||^2\right] \\
&+ 2\mathbb{E}_{\mathbb{P}_\beta}\left[(\widehat{\beta}_{\text{JS}} - \beta)'(\widehat{\beta}_{\text{OLS}} - \beta)\right].
\end{aligned}$$

Algebra shows that

$$\widehat{\beta}_{JS} - \widehat{\beta}_{\text{OLS}} \equiv \left(\frac{(k-2)}{||\widehat{\beta}_{\text{OLS}}||^2}\right)\widehat{\beta}_{\text{OLS}}.$$

Therefore,

$$\mathbb{E}_{\mathbb{P}_\beta}\left[||\widehat{\beta}_{JS} - \beta||^2\right] = \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{(k-2)^2}{||\widehat{\beta}_{\text{OLS}}||^2}\right] - k + 2\mathbb{E}_{\mathbb{P}_\beta}\left[(\widehat{\beta}_{\text{JS}} - \beta)'(\widehat{\beta}_{\text{OLS}} - \beta)\right]..$$

Also

$$\mathbb{E}_{\mathbb{P}_\beta}\left[(\widehat{\beta}_{\text{JS}} - \beta)'(\widehat{\beta}_{\text{OLS}} - \beta)\right] = \sum_{j=1}^{k} \text{Cov}(\widehat{\beta}^j_{\text{JS}}, \widehat{\beta}^j_{OLS}).$$

and the last term can be shown to equal

$$\begin{aligned}\sum_{j=1}^{k} \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{\partial \widehat{\beta}_{JS}^{j}}{\partial \widehat{\beta}_{\text{OLS}}^{j}}\right] &= K - \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{k(k-2)}{||\widehat{\beta}_{\text{OLS}}||^2}\right] + \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{2(k-2)}{||\widehat{\beta}_{\text{OLS}}||^2}\right]' \\ &= K - \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{(k-2)^2}{||\widehat{\beta}_{\text{OLS}}||^2}\right]\end{aligned}$$

Therefore,

$$\mathbb{E}_{\mathbb{P}_\beta}\left[||\widehat{\beta}_{JS} - \beta||^2\right] = k - \mathbb{E}_{\mathbb{P}_\beta}\left[\frac{(k-2)^2}{||\widehat{\beta}_{\text{OLS}}||^2}\right] \leq k = \mathbb{E}_{\mathbb{P}_\beta}\left[||\widehat{\beta}_{OLS} - \beta||^2\right].$$

□

References

Lehmann, E. L. and G. Casella (1998): *Theory of point estimation*, vol. 31 of *Springer Texts in Statistics*, Springer, New York, 2nd ed.