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 Posterior Mean estimators—in the context of the homoskedastic Normal regression model with (known variance); i.e.,

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

where Y is the $n \times 1$ vector of outcome variables, X is the $n \times k$ matrix of non-stochastic regressors, and $\beta \in \mathbb{R}^k$ is the parameter of interest.

We will show that if $k < n$, 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). We will show this estimator is "best" among all *unbiased* estimators and we will connect this result to the *Cramér-Rao* bound for the variance of estimators in parametric models. We will also argue that "best" among the class of unbiased estimators need not imply admissibility with respect to all estimators. In particular, we show that if $k \geq 3$, the OLS estimator is dominated and we present an "empirical Bayes" estimator that dominates it.

The posterior mean estimator for the regression model assuming $\beta \sim \mathcal{N}_k(0, (\sigma^2/\lambda)\mathbb{I}_k)$ is the popular *Ridge estimator*

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

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 and the true parameter at which we evaluate the bias.

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 Quadratic Loss

Let $\mathcal{A} = \Theta = \mathbb{R}^k$. The loss function we will work with is the so-called quadratic loss

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

which measures the squared distance between the estimator $\widehat{\beta}$ and the parameter β. The risk of any estimator $\widehat{\beta}$ (which is a map from data (Y, X) to Θ) 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. We now present a simple algebraic decomposition of the mean squared error in terms of the "bias" and "variance" of the estimator $\widehat{\beta}$. 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 $\mathrm{V}_\beta(\widehat{\beta})$—is also finite, the mean squared-error can be written as

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

where $\mathrm{tr}(\cdot)$ 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.

1.2 Maximum Likelihood Estimation

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

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

This is true, regardless of whether we have very many covariates or not. For a fixed realization of the data, define the likelihood function, L(β; (Y,X)), as the value attained by the p.d.f. in (1.3) at different values of the data (Y, X); that is

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

The maximum likelihood estimator at data (Y, X) 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 $(Y - XB)'(Y - XB)$, a term which is usually referred to as the sum of square residuals. Therefore, maximizing the likelihood is equivalent to solving the *least-squares* 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 $n \geq k$, there are infinitely many solutions that maximize the likelihood. If $k < n$, there is a unique solution to this problem given by

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

The bias of the maximum likelihood estimator:

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

for any β. Thus, we say that the ML estimator of β is <u>unbiased</u>.[1]

The variance of the ML estimator is

$$\mathbb{V}_\beta(\widehat{\beta}_{\text{ML}}) = \mathbb{E}_{\mathbb{P}_\beta}[\widehat{\beta}_{\text{ML}}\widehat{\beta}_{\text{ML}}`] = (X'X)^{-1}X'\mathbb{E}_{\mathbb{P}_\beta}[YY']X(X'X)^{-1} = \sigma^2(X'X)^{-1}. \tag{1.6}$$

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

$$\sigma^2 \text{tr}\left((X'X)^{-1}\right).$$

[1] An estimator $\widehat{\beta}$ is unbiased if $\mathbb{E}_\beta[\widehat{\beta}] = \beta$ for all $\beta \in \Theta$.

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

1.3 Posterior Mean 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 at 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

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

Before committing to a specific prior π, we will show that under quadratic loss the Bayes estimator is the posterior mean.

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

$$\beta \sim \pi(\beta) \equiv \mathcal{N}_k(\beta_0\,,\, \sigma^2 V^{-1}). \tag{1.7}$$

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 typically 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

$$\exp\left(-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right)\exp\left(-\frac{1}{2\sigma^2}(\beta-\beta_0)'V(\beta-\beta_0)\right). \tag{1.8}$$

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:

$$\beta|Y,X\sim\mathcal{N}_k\left(\left(V+X'X\right)^{-1}\left(X'Y+V^{-1}\beta_0\right),\,\sigma^2\left(V+X'X\right)^{-1}\right). \tag{1.9}$$

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.[2] The posterior mean estimator for $V=\lambda\mathbb{I}_k$ is called the Ridge estimator

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

which, by construction, is admissible. The Ridge estimator is biased:

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

and the magnitude of the bias depends on β. The variance of the Ridge estimator is

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

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

When $k < n$, the trace of this variance has to be smaller than that of the ML/OLS estimator.

Posterior Mean as a Penalized/Regularized OLS estimator: Equation (1.8) 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

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

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

1.4 Optimality of the OLS estimator among unbiased estimators

In this section we will show that the OLS estimator minimizes risk among the class of all unbiased estimators, provided some regularity conditions are met.

Proposition 1. *(Optimality of OLS) Consider the Normal regression model in (1.1). Let $\widehat{\beta}$ be an estimator of β for which*

$$\frac{\partial}{\partial\beta}\int_{\mathbb{R}^k}\widehat{\beta}f(Y|\beta;X)dY = \int_{\mathbb{R}^k}\widehat{\beta}\frac{\partial}{\partial\beta}f(Y|\beta,X)dY,$$

at any β in the parameter space. If $\widehat{\beta}$ is unbiased, then

$$\mathbb{V}_\beta(\widehat{\beta}) - \mathbb{V}_\beta(\widehat{\beta}_{OLS})$$

is positive semi-definite, implying that the mean squared error of $\widehat{\beta}$ is larger than that of the OLS everywhere on the parameter space.

Proof. We would like show that for any $c \in \mathbb{R}^k$, $c \neq 0$

$$c'\left(\mathbb{V}_\beta(\widehat{\beta}) - \mathbb{V}_\beta(\widehat{\beta}_{\text{OLS}})\right)c \geq 0.$$

□

To do this, we start by computing the variance of $c'\widehat{\beta} - c'\widehat{\beta}_{\text{OLS}}$, which by definition of variance has to be nonegative. Since $\widehat{\beta}$ and $\widehat{\beta}_{\text{OLS}}$ are both unbiased

$$\begin{aligned}
\mathbb{V}_\beta\left(c'\widehat{\beta} - c'\widehat{\beta}_{\text{OLS}}\right) &= \mathbb{V}_\beta(c'\widehat{\beta}) + \mathbb{V}_\beta(c'\widehat{\beta}_{\text{OLS}}) - 2\text{cov}_\beta(c'\widehat{\beta}, c'\widehat{\beta}_{\text{OLS}}), \\
&\quad \text{(since } \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{cov}(X,Y)) \\
&= c'\mathbb{V}_\beta(\widehat{\beta})c + c'\mathbb{V}_\beta(\widehat{\beta}_{\text{OLS}})c - 2\mathbb{E}_{\mathbb{P}_\beta}\left[c'(\widehat{\beta} - \beta)(\widehat{\beta}_{\text{OLS}} - \beta)'c\right] \\
&\quad \text{(using the fact that both estimators are unbiased).} \\
&= c'\mathbb{V}_\beta(\widehat{\beta})c + c'\mathbb{V}_\beta(\widehat{\beta}_{\text{OLS}})c - 2c'\mathbb{E}_{\mathbb{P}_\beta}\left[\widehat{\beta}(\widehat{\beta}_{\text{OLS}} - \beta)'\right]c.
\end{aligned}$$

Where the last term is the covariance between $\widehat{\beta}$ and the OLS residual. Using the definition of OLS

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

and using the definition of the the Gaussian p.d.f. $f(Y|\beta, X)$ we can see verify that

$$\widehat{\beta}_{\text{OLS}} - \beta = \sigma^2(X'X)^{-1}\frac{\partial}{\partial\beta}\ln f(Y|\beta, X).$$

Consequently

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\beta}\left[\widehat{\beta}(\widehat{\beta}_{\text{OLS}} - \beta)'\right] &= \sigma^2\mathbb{E}_{\mathbb{P}_\beta}\left[\widehat{\beta}\left(\frac{\partial}{\partial\beta}\ln f(Y|\beta, X)\right)'\right](X'X)^{-1}, \\
&= \sigma^2\left(\int_{\mathbb{R}^k}\widehat{\beta}\left(\frac{\partial}{\partial\beta}\ln f(Y|\beta, X)\right)' f(Y|\beta, X)dY\right)(X'X)^{-1} \\
&= \sigma^2\left(\int_{\mathbb{R}^k}\widehat{\beta}\left(\frac{\partial}{\partial\beta}f(Y|\beta, X)\right)' dY\right)(X'X)^{-1} \\
&\quad \text{(where we have used the chain rule and } \partial\ln x/\partial x = 1/x) \\
&= \sigma^2\frac{\partial}{\partial\beta}\left(\int_{\mathbb{R}^k}\widehat{\beta}f(Y|\beta, X)\right)(X'X)^{-1} \\
&\quad \text{(by assumption)} \\
&= \sigma^2\left(\frac{\partial}{\partial\beta}\mathbb{E}_{\mathbb{P}_\beta}[\widehat{\beta}]\right)(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1} \\
&\quad \text{(since } \widehat{\beta} \text{ is unbiased).} \\
&= \mathbb{V}_\beta(\widehat{\beta}_{\text{OLS}}).
\end{aligned}$$

Hence,

$$0 \leq \mathbb{V}_\beta \left(c'\widehat{\beta} - c'\widehat{\beta}_{\text{OLS}} \right) = c'\mathbb{V}_\beta(\widehat{\beta})c - c'\mathbb{V}_\beta(\widehat{\beta}_{\text{OLS}})c,$$

for every β. The result then follows.

1.5 Suboptimality of the OLS estimator

If $k \geq 3$, the OLS estimator is dominated. In this section we present an estimator that dominates OLS.

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. This estimator was first proposed by Willard James and Charles Stein in 1961. The estimator is typically referred to as the James-Stein estimator.

Proposition 2. *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

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

and the last term can be shown to equal

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

Therefore,

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

□