HW3 Stelios Giagkos f3352410

Exercise 1 (multiple choices question):

In the central square of a town, there are four thermometers A, B, C and D (of slightly different technology from each other) that measure the air temperature. The aim is to estimate the (mean) temperature at noon for the month July. To this end, we record each day the temperature shown by each thermometer at noon. At the end of the month, we average the measurements taken from thermometer A and, separately, we do the same for the thermometers B, C and D. Which of the following is correct?

1. The four averages are all estimates and, in general, have exactly the same value, since they measure the same quantity.

2. The four averages are all estimates of the mean temperature at noon in July.

3. In practice, we can detect which of the four averages/estimates is better, by comparing it with the true mean temperature value.

4. If the true mean temperature at noon in the July is known, there is no need to estimate it.

Answer:

The correct answer is:

2. The four averages are estimates of the mean temperature at noon in July

Each thermometer provides an estimate of the same quantity, the mean temperature.

Exercise 2 (multiple choices question):

In twenty different medical laboratories throughout the world, the scientists aim to estimate the correlation coefficient, r, between the smoking and heart deceases. To this end, each lab uses the relevant data that it has collected during the last five years (for simplicity, assume that each lab has at its disposal a data set $Y_i, i = 1, ..., 20$, of 10.000 cases) and twenty respective values for r are obtained, namely $r_1, r_2, ..., r_{20}$. Which of the following are correct?

1. The $r_1, r_2, ..., r_{20}$ are specific instances of a random variable, r, corresponding to the correlation coefficient.

2. The $r_1, r_2, ..., r_{20}$ are all estimators of r.

3. The sets $Y_1, Y_2, ..., Y_{20}$ are specific instances of the "random set" Y, which models all the data sets of the form of Y_i's with 10.000 cases each.

4. The $r_1, r_2, ..., r_{20}$ are all estimates of the correlation coefficient and they can be considered as instances of the estimator (random variable) r that corresponds to the correlation coefficient.

Answer:

The correct answers are

(1)The $r1, r2, ... , r20$ are specific instances of a random variable, r, corresponding to the correlation coefficient.

The $r1, r2, ... , r20$ are instances from the random variable r, which represents the true correlation coefficient. Each lab generates its own estimate of the correlation coefficient based on its own data and samples, and these estimates constitute realizations of a random process.

and **(4) The $r1, r2, ... , r20$ are all estimates of the correlation coefficient and they can be considered as instances of the estimator (random variable) r that corresponds to the correlation coefficient.**

As mentioned during the class, each lab is based on its own data , therefore the correlation coefficients vary for every lab , due to the dependence of the data. This leads to multiple estimations of the same parameter.

Exercise 3 (multiple choices question):

Which of the following criteria are suitable to quantify the deviation of an estimator, $\hat\theta$, of the true model value θ_0 in a single-parameter problem?

1. $E[|\hat\theta - \theta_o|]$

2. $E[\hat\theta - \theta_o]$

3. $E\left[(\hat\theta - \theta_o)^2\right]$

4. $E\left[(\hat\theta - \theta_o)^3\right]$

5. $(\hat\theta - \theta_o)^2$

Answer:

The correct answers are:

1. $E[|\hat\theta - \theta_0|]$ which is the the mean absolute error of the estimator from the true parameter value.

3. $E\left[(\hat\theta - \theta_0)^2\right]$ which is the mean squared error (MSE) combining both the variance and bias of the estimator.

Exercise 4 (multiple choices question):

Why, in practice, the quantity $E\left[(\hat\theta - \theta_o)^2\right]$ cannot be computed explicitly?

1. The estimator $\hat\theta$ is unknown.

2. The true value, θ_o, is unknown in practice.

3. While all the involved quantities may be known, the computations are very demanding.

4. The expectation operator prevents the explicit computation.

Answer:

The correct answer is:

2. The true value, θ_0, is unknown in practice.

In most real-world situations, the true parameter θ_0 is unknown. Since the mean squared error (MSE) is given by $MSE = E\left[(\hat\theta - \theta_0)^2\right]$ we cannot calculate this expectation precisely without knowing θ_0. Estimators are employed precisely because the true parameter value is not available.

Exercise 5 (multiple choices question):

In a practical machine learning problem, we have at our disposal only a specific data set Y of size N and, based on this, we compute an estimate $\hat{\theta}$ of the (unknown to us) θ_o parameter of an adopted model. Let $\hat{\theta}$ be the estimator associated with $\hat{\theta}$, for data sets of size N. Assume that it is known that the MSE error, $MSE = \mathrm{E}\left[(\hat{\theta} - \theta_o)^2\right]$ is very small. What implications has this to the obtained $\hat{\theta}$?

1. The estimate $\hat{\theta}$ is very likely to be away from θ_o.

2. Another estimate based on a slightly different data set would also be close to θ_o, with high probability.

3. In general, the specific choice of Y does not play a significant role to the accuracy of the estimate $\hat{\theta}$.

4. The estimate $\hat{\theta}$ of θ_o, based on any data set of size N (for the specific problem) can be adopted with confidence, since, very small MSE implies that estimates are very close to θ_o, with high probability.

Answer

The correct answers are:

2. Another estimate based on a slightly different data set would also be close to θ_0, with high probability.

When the MSE is very small, an estimate from a slightly different data set will also likely be close to θ_0 as the estimator exhibits low variability and minimal bias

3. In general, the specific choice of Y does not play a significant role to the accuracy of the estimate $\hat{\theta}$.

Low variance means the estimate doesn't rely much on the particular data set used, as long as the data set is representative of the overall population.If the MSE is small,the estimator is robust, meaning that different data sets of size N would produce similar estimates. Therefore, the specific choice of Y does not significantly affect the accuracy.

4. The estimate $\hat{\theta}$ of θ_0, based on any data set of size N (for the specific problem) can be adopted with confidence, since, very small MSE implies that estimates are very close to θ_0, with high probability.

A very small MSE indicates that the estimator consistently produces values close to the true parameter θ_0 with high probability. As a result, an estimate $\hat{\theta}$ based on any sufficiently large data set can be trusted, provided the MSE remains small.

Exercise 6

Consider the following figures, each one showing the locations of the true (unknown) θ_o (black circle) and $\mathrm{E}[\hat{\theta}]$ of an estimator (red circle). The arrangement of the blue boxes is indicative of the variance of each estimator around its mean, which is given for each case (a is a common scalar parameter in all cases).
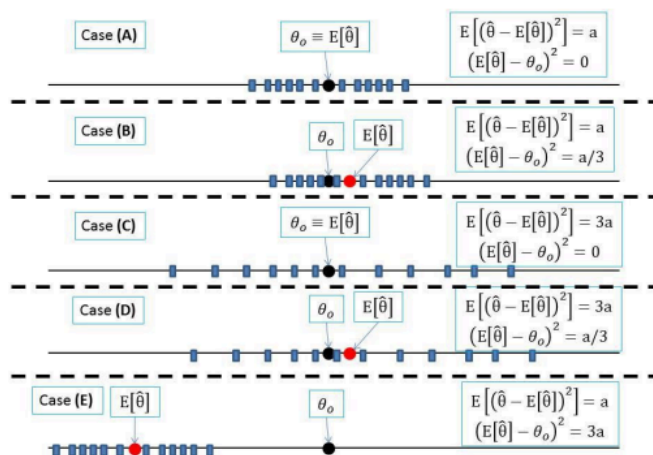
For each of the above cases (a) determine the associated MSE value and (b) state whether or not estimator is biased or unbiased.

Answer:

First of all, it is important to define each one of the above terms that are being used. So,

1. θ_0 is the actual, unknown value that we are trying to estimate.
2. $\hat{\theta}$ is an estimate of θ_0, that can be biased or unbiased.
3. $E[\hat{\theta}]$ is the average value of the estimator.
4. $E[\hat{\theta}] - \hat{\theta}$ is the Bias of the estimator.
5. $MSE = E\left[(\hat{\theta} - \theta_0)^2\right]$ measures the estimators accuracy and it is also equal to $MSE(\hat{\theta}) = Variance(\hat{\theta}) + Bias(\hat{\theta})^2 (1)$

where:

$$Variance(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2$$

$$Bias(\hat{\theta})^2 = (E[\hat{\theta}] - \theta_0)^2$$

Case (A): First of all we can see that the Estimator is **unbiased** as $E[\hat{\theta}] \equiv \theta_0$

$$Variance(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2 = \alpha$$

$$Bias(\hat{\theta})^2 = (E[\hat{\theta}] - \theta_0)^2 = 0$$

So from (1) we get that $MSE(\hat{\theta}) = Variance(\hat{\theta}) + Bias(\hat{\theta})^2 = \alpha + 0 = \alpha$

Case (B): First of all we can see that the Estimator is **biased** as $E[\hat{\theta}] \neq \theta_0$

$$Variance(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2 = \alpha$$

$$Bias(\hat{\theta})^2 = (E[\hat{\theta}] - \theta_0)^2 = \frac{\alpha}{3}$$

So from (1) we get that $MSE(\hat{\theta}) = Variance(\hat{\theta}) + Bias(\hat{\theta})^2 = \alpha + \frac{\alpha}{3} = \frac{4\alpha}{3}$

Case (C): First of all we can see that the Estimator is **unbiased** as $E[\hat{\theta}] \equiv \theta_0$

$$Variance(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2 = 3\alpha$$

$$Bias(\hat{\theta})^2 = (E[\hat{\theta}] - \theta_0)^2 = 0$$

So from (1) we get that $MSE(\hat{\theta}) = Variance(\hat{\theta}) + Bias(\hat{\theta})^2 = 3\alpha + 0 = 3\alpha$

Case (D): First of all we can see that the Estimator is **biased** as $E[\hat{\theta}] \neq \theta_0$

$$Variance(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2 = 3\alpha$$

$$Bias(\hat{\theta})^2 = (E[\hat{\theta}] - \theta_0)^2 = \frac{\alpha}{3}$$

So from (1) we get that $MSE(\hat{\theta}) = Variance(\hat{\theta}) + Bias(\hat{\theta})^2 = 3\alpha + \frac{\alpha}{3} = \frac{10\alpha}{3}$

Case (E): First of all we can see that the Estimator is **biased** as $E[\hat{\theta}] \neq \theta_0$

$$Variance(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2 = \alpha$$

$$Bias(\hat{\theta})^2 = (E[\hat{\theta}] - \theta_0)^2 = 3\alpha$$

So from (1) we get that $MSE(\hat{\theta}) = Variance(\hat{\theta}) + Bias(\hat{\theta})^2 = \alpha + 3\alpha = 4\alpha$

Exercise 7

Consider a single-parameter problem with θ_o being the unknown parameter. Let $\hat{\theta}_1$ be an estimator of θ_o, which minimizes the criterion associated with a specific loss function

and is based on a randomly selected data set comprising N_1 data points. Let $\hat{\theta}_2$ be another estimator of θ_o, which minimizes the criterion associated with the same loss function as $\hat{\theta}_1$, and is based on another randomly selected data set of N_2 ($\neq N_1$) data points. Which of the following statements are true?

1. The two estimators are identical, since the both result from the same loss function.

2. Both estimators coincide with θ_o, since they both minimize a criterion associated with the specific loss function.

3. The two estimators, although they are expressed by the same mathematical formula, they differ since they apply on data sets of different sizes.

4. Any two estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ_o associated with the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively, coincide with θ_o, since both of them are instances of estimators that minimize the criterion associated with the specific loss function.

Answer

The correct answer is:

3. The two estimators, although they are expressed by the same mathematical formula, they differ since they apply on data sets of different sizes.

Even if the estimators utilize the same formula, their reliance on different data sets, with varying sizes in this case, can result in distinct estimates.This is a reason for the two estimators to differ.

Exercise 8

Consider the problem of estimating an unknown parameter θ, based on a set of noisy observations of it, of the form $y_n = \theta + \eta_n$, $n = 1, 2, \dots$, where η_n's are instances of the noise, whose mean value is zero ($E[\eta] = 0$). Which of the following corresponding estimators of θ are unbiased?

1. $\frac{1}{3}\sum_{n=1}^{3} y_n$

2. $\frac{1}{3}\sum_{n=1}^{3} y_n + 0.2$

3. $\frac{1}{100}\sum_{n=1}^{100} y_n{}^3$

4. $\frac{1}{100}\sum_{n=1}^{100} y_n$

Answer:

The unbiasedness condition requires that the expected value of the estimator equals the true parameter θ

For Option 1

$\theta_1 = \frac{1}{3}\sum_{n=1}^{3} y_n$

The expected value is: $E\left[\frac{1}{3}\sum_{n=1}^{3} y_n\right] = \frac{1}{3}\sum_{n=1}^{3} E[y_n] = \frac{1}{3}\sum_{n=1}^{3}(\theta + E[\eta_n]) = \frac{1}{3}\sum_{n=1}^{3}\theta = \theta$

So, This estimator is unbiased.

For Option 2

$\theta_2 = \frac{1}{3}\sum_{n=1}^{3} y_n + 0.2$

The expected value is: $E\left[\frac{1}{3}\sum_{n=1}^{3} y_n + 0.2\right] = E\left[\frac{1}{3}\sum_{n=1}^{3} y_n\right] + 0.2 = \theta + 0.2$

This estimator is biased because the expected value is $\theta + 0.2$, not θ.

For Option 3

$\theta_3 = \frac{1}{100}\sum_{n=1}^{100} y_n^3$

The expected value is: $E\left[\frac{1}{100}\sum_{n=1}^{100} y_n^3\right] = \frac{1}{100}\sum_{n=1}^{100} E[y_n^3]$

Here, y_n^3 includes higher-order terms of the noise η_n, making this estimator not linear in y_n. As a result, the expected value of y_n^3 is not simply θ, so this estimator is biased.

For Option 4

$\theta_4 = \frac{1}{100}\sum_{n=1}^{100} y_n$

The expected value is: $E\left[\frac{1}{100}\sum_{n=1}^{100} y_n\right] = \frac{1}{100}\sum_{n=1}^{100} E[y_n] = \frac{1}{100}\sum_{n=1}^{100}\theta = \theta$

This estimator is unbiased.

Options 1 and 4 are unbiased estimators.

Even though these estimators involve fractions, they are unbiased because these fractions apply equally to both the θ and noise terms, maintaining the expected value as θ.

Exercise 9

Consider the problem of estimating an unknown parameter θ, based on a set of noisy observations of it, of the form $y_n = \theta + \eta_n$, $n = 1,2,\dots$, where η_n's are instances of the noise, which is assumed white and Gaussian with zero mean ($E[\eta] = 0$) and variance equal to σ_η^2. In this case, it is known that $\frac{1}{N}\sum_{n=1}^{N} y_n$ is the minimum variance unbiased

estimator (MVUE). Consider the estimators $\hat{\theta}_4 = \frac{1}{4}\sum_{n=1}^{4} y_n$ and $\hat{\theta}_{200} = \frac{1}{200}\sum_{n=1}^{200} y_n$, which are of the above form. Since they are both unbiased, the MSE for each one of them is $\frac{\sigma_\eta^2}{4}$ and $\frac{\sigma_\eta^2}{200}$, respectively. Clearly, the second estimator exhibits lower MSE. How this matches with the general conclusion that $\frac{1}{N}\sum_{n=1}^{N} y_n$ is MVUE?

1. The noise variance varies for the two estimators, so that $\frac{\sigma_\eta^2}{4}$ and $\frac{\sigma_\eta^2}{200}$ to become equal.

2. The general conclusion is valid only for specific N.

3. The general result holds only for the mean of the two estimators.

4. Estimator $\hat{\theta}_{200}$ is biased, so that the addition of the bias term to the variance term, gives the MSE of $\hat{\theta}_4$.

Answer:

The correct answer is :

2. The general conclusion is valid only for specific N.

In this case, the general conclusion (that $\frac{1}{N}\sum y_n$ is MVUE) holds for each specific N. For $N = 200$, the variance (and thus the MSE) is smaller than for $N = 4$, which explains why the estimator $\hat{\theta}_{200}$ has a lower MSE than $\hat{\theta}_4$. The statement implies the general result applies to specific values of N, and as N increases, the MSE decreases.

Exercise 10

Consider a multi-parameter regression problem, where $\boldsymbol{\theta} = [\theta_0, \ \theta_1, ..., \theta_l]^T$ denotes the parameter vector that contains all the parameters of the associated model and let $\boldsymbol{\theta}_o = [\theta_{o0}, \ \theta_{o1}, ..., \theta_{ol}]^T$ be the true value of $\boldsymbol{\theta}$. Also, let $\hat{\boldsymbol{\theta}} = [\hat{\theta}_0, \ \hat{\theta}_1, ..., \hat{\theta}_l]^T$ be an estimator of $\boldsymbol{\theta}_o$. Which of the following statements is correct?

1. If the MSE per parameter $\hat{\theta}_l$ is known, the MSE of $\hat{\boldsymbol{\theta}}$, is obtained through the multiplication of the MSEs of all the parameters.

2. If $\hat{\boldsymbol{\theta}}$ is unbiased, its MSE equals to the sum of the variances of $\hat{\theta}_l$'s around their means.

3. In the (utopic) case where the variances of all $\hat{\theta}_l$'s are equal to 0, the MSE of $\hat{\boldsymbol{\theta}}$ equals to the sum of the bias terms of all $\hat{\theta}_l$'s.

4. Although $\hat{\boldsymbol{\theta}}$ may be biased, the biased terms corresponding to $\hat{\theta}_l$'s are not taken into account in the computation of the MSE of $\hat{\boldsymbol{\theta}}$.

Answer:

The correct answers are:

Keeping in mind the equation: $MSE(\hat{\theta}) = Variance(\hat{\theta}) + Bias(\hat{\theta})^2$

2. If $\hat{\theta}$ is unbiased, its MSE equals the sum of the variances of $\hat{\theta}_i$'s around their means.

For an unbiased estimator, the MSE is simply the sum of the variances of the parameters, because the bias term is zero: $MSE(\hat{\theta}_i) = Var(\hat{\theta}_i)$ So,the MSE equals the sum of the variances of the $\hat{\theta}_i$'s.

and **3. In the (utopic) case where the variances of all $\hat{\theta}_i$'s are equal to 0, the MSE of $\hat{\theta}$ equals the sum of the bias terms of all $\hat{\theta}_i$'s.**

If the variances of all $\hat{\theta}_i$'s are zero, the MSE would consist entirely of the bias terms: $MSE(\hat{\theta}_i) = Bias^2(\hat{\theta}_i)$ So, the total MSE would be the sum of the squared bias terms.

Exercise 11

Consider a multi-parameter regression problem, where $\boldsymbol{\theta} = [\theta_0, \theta_1, ..., \theta_l]^T$ denotes the parameter vector that contains all the parameters of the associated model and let

$\boldsymbol{\theta}_o = [\theta_{o0}, \theta_{o1}, ..., \theta_{ol}]^T$ be the true value of $\boldsymbol{\theta}$. Also, let $\hat{\boldsymbol{\theta}} = [\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_l]^T$ be the minimum variance unbiased estimator (MVUE) of $\boldsymbol{\theta}_o$. Which of the following statements is correct?

1. One way to get an unbiased estimator with variance less than that of $\hat{\boldsymbol{\theta}}$, is to magnify the norm of $\hat{\boldsymbol{\theta}}$.

2. The estimator that results from the shrinkage of the norm of $\hat{\boldsymbol{\theta}}$ is also an unbiased one.

3. All biased estimators that result from the shrinkage of the norm of $\hat{\boldsymbol{\theta}}$, have variance less than that of $\hat{\boldsymbol{\theta}}$.

4. One way to get a biased estimator with variance less than that of $\hat{\boldsymbol{\theta}}$, is to shrink the norm of $\hat{\boldsymbol{\theta}}$.

Answer:

The correct answer is :

4. One way to get a biased estimator with variance less than that of $\hat{\theta}$ is to shrink the norm of $\hat{\theta}$.

When we reduce the norm of $\hat{\theta}$, we introduce some biasbut we also lower the variance. This balance between bias and variance often leads us to a lower overall MSE, even though the estimator becomes biased.

Exercise 12

(a) Use the Lagrangian function of the ridge regression problem

$$\min \quad L(\boldsymbol{\theta}) = \sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 + \lambda\|\boldsymbol{\theta}\|^2$$

and show that the solution satisfies the equation

$$(\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T + \lambda I)\hat{\boldsymbol{\theta}} = \sum_{n=1}^{N} y_n \boldsymbol{x}_n \qquad\text{(A)}$$

Hints: Take the gradient of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ equate to $\boldsymbol{0}$ and solve.

It is $(\boldsymbol{\theta}^T \boldsymbol{x})\,\boldsymbol{x} = (\boldsymbol{x}\boldsymbol{x}^T)\,\boldsymbol{\theta}$

It is $A\boldsymbol{z} - \lambda\boldsymbol{z} = A\boldsymbol{z} - \lambda I\boldsymbol{z} = (A - \lambda I)\boldsymbol{z}$, where A is a matrix, $\boldsymbol{z}$ is a vector and λ is a scalar and I is the identity matrix.

It is $\frac{\partial\|\theta\|^2}{\partial\theta} = \frac{\partial(\theta^T\theta)}{\partial\theta} = 2\theta$

(b) Prove that the above solution can be expressed in matrix form as

$$\hat{\theta} = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}$$

where $X = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1l} \\ 1 & x_{21} & \cdots & x_{2l} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nl} \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

Hint: Prove that $X^T X = \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T$ and $X^T \boldsymbol{y} = \sum_{n=1}^{N} y_n \boldsymbol{x}_n$

Answer:

a) The Lagrangian function for ridge regression is:

$L(\theta) = \sum_{n=1}^{N}(y_n - \theta^T x_n)^2 + \lambda\|\theta\|^2$

To find the optimum θ, I solve gradient of $L(\theta) = 0 \rightarrow \frac{\partial L(\theta)}{\partial \theta} = -2\sum_{n=1}^{N}(y_n - \theta^T x_n)x_n + 2\lambda\theta = 0 \rightarrow \sum_{n=1}^{N} y_n x_n - \sum_{n=1}^{N}(\theta^T x_n)x_n = \lambda\theta$

$\sum_{n=1}^{N}(\theta^T x_n)x_n = (X^T X)\theta$, where X is the matrix of input features.

Thus, $\sum_{n=1}^{N} y_n x_n = (X^T X)\theta + \lambda\theta \rightarrow (X^T X + \lambda I)\theta = \sum_{n=1}^{N} y_n x_n \rightarrow (X^T X + \lambda I)\hat{\theta} = \sum_{n=1}^{N} y_n x_n$

b) $X^T X = \begin{bmatrix} x_1^T & x_2^T & \cdots & x_N^T \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \sum_{n=1}^{N} x_n x_n^T$

$$X^T y = \begin{bmatrix} x_1^T & x_2^T & \cdots & x_N^T \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \sum_{n=1}^N y_n x_n$$

From $(X^T X + \lambda I)\hat{\theta} = X^T y$ by solving for $\hat{\theta}$ gives us: $\hat{\theta} = (X^T X + \lambda I)^{-1} X^T y. \rightarrow \hat{\theta} = (X^T X + \lambda I)^{-1} X^T y.$

Exercise 13

Consider a 1-dimensional parameter estimation problem, where the true parameter value is θ_o. Let $\hat{\theta}_{MVU}$ be a minimum variance unbiased estimator of θ_o. Consider the parametric set F of all estimators of the form

$$\hat{\theta}_b = (1 + \alpha)\hat{\theta}_{MVU}, \qquad (1)$$

with $a \in R$. Recall that

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - E[\hat{\theta}])^2\right] + \left(E(\hat{\theta}) - \theta_o\right)^2 \qquad (2)$$

(a) What can you infer from the fact that $\hat{\theta}_{MVU}$ is an unbiased estimator of θ_o?

(b) Prove that all $\hat{\theta}_b$'s of F, for $\alpha \neq 0$, are biased estimators of θ_o.

(c) Find the $MSE(\hat{\theta}_{MVU})$ using eq. (2). Explain why this value cannot be zero for finite N.

(d) Express the $MSE(\hat{\theta}_b)$ in terms of $MSE(\hat{\theta}_{MVU})$ (Substitute eq. (1) to eq. (2) and after some algebra, utilize the results of (c) above).

(e) Determine the range of values of the parameter α that result to estimators with MSE lower than $MSE(\hat{\theta}_{MVU})$ (Consider the inequality $MSE(\hat{\theta}_b) < MSE(\hat{\theta}_{MVU})$, substitute $MSE(\hat{\theta}_b)$ from (d) and you will end up with a second order polynomial wrt α, which should be negative. Determine its roots and define the range of values of α where polynomial is negative).

(f) Prove that for any value of α in the range defined in (e), it is $|\hat{\theta}_b| < |\hat{\theta}_{MVU}|$ (Show first that $|1 + \alpha| < 1$ and then utilize eq. (1)).

(g) Determine the value α^* of the parameter α that corresponds to the estimator giving the lowest MSE (Consider the expression of $MSE(\hat{\theta}_b)$ derived in (d), take the derivative wrt α, equate to 0 and solve).

(h) Explain why in practice α^* cannot be determined.

Answer

a) The term "unbiased estimator" means that the expected value of the estimator equals the true parameter. $E[\hat{\theta}_{MVU}]) = \theta_0$

b) We know that $\hat{\theta}_b = (1 + \alpha)\hat{\theta}_{MVU}$

The expectation of $\hat{\theta}_b$ is: $E[\hat{\theta}_b] = \mathbb{E}[(1 + \alpha)\hat{\theta}_{MVU}] = (1 + \alpha)\mathbb{E}[\hat{\theta}_{MVU}] = (1 + \alpha)\theta_0$

So, the bias is: $Bias(\hat{\theta}_b) = \mathbb{E}[\hat{\theta}_b] - \theta_0 = (1 + \alpha)\theta_0 - \theta_0 = \alpha\theta_0$

Therefore, $\hat{\theta}_b$ is biased for $\alpha \neq 0$, since the bias term $\alpha\theta_0$ is non-zero.

c) $\hat{\theta}_{MVU}$, is unbiased, so $E[\hat{\theta}_{MVU}] = \theta_0$. Therefore, the MSE simplifies to the variance:

$$\mathrm{MSE}(\hat{\theta}_{MVU}) = E[(\hat{\theta}_{MVU} - \theta_0)^2] = \mathrm{Var}(\hat{\theta}_{MVU})$$

This value cannot be zero for finite N because there will always be some variability (variance) in the estimator for finite sample sizes.

d) $MSE(\hat{\theta}_b) = E[(\hat{\theta}_b - E[\hat{\theta}_b])^2] + (E[\hat{\theta}_b] - \theta_0)^2 = Var(\hat{\theta}_b) + ((1 + \alpha)\theta_0 - \theta_0)^2 = (1 + \alpha)^2 \, \mathrm{Var}(\hat{\theta}_{MVU}) + (\alpha\theta_0)^2$

e) $MSE(\hat{\theta}_b) < MSE(\hat{\theta}_{MVU}) \rightarrow (1 + \alpha)^2 Var(\hat{\theta}_{MVU}) + (\alpha\theta_0)^2 < Var(\hat{\theta}_{MVU}) \rightarrow (1 + \alpha)^2 + \frac{(\alpha\theta_0)^2}{Var(\hat{\theta}_{MVU})} < 1 \rightarrow 1 + 2\alpha + \alpha^2 + \frac{\alpha^2\theta_0^2}{Var(\hat{\theta}_{MVU})} < 1$

$\rightarrow 2\alpha + \alpha^2(1 + \frac{\theta_0^2}{Var(\hat{\theta}_{MVU})})) < 0 \rightarrow \alpha((1 + \frac{\theta_0^2}{Var(\hat{\theta}_{MVU})})\alpha + 2) < 0$

$\alpha((1 + \frac{\theta_0^2}{Var(\hat{\theta}_{MVU})})\alpha + 2) = 0$

$\rightarrow \alpha = 0$ and $\alpha = -\dfrac{2}{(1 + \frac{\theta_0^2}{Var(\hat{\theta}_{MVU})})}$

For $-2(1 + \frac{\theta_0^2}{Var(\hat{\theta}_{MVU})}) < \alpha < 0$, α is negative, and $(1 + \frac{\theta_0^2}{Var(\hat{\theta}_{MVU})})\alpha + 2$ is positive, so the product is negative. So the range is

$\dfrac{-2}{(1 + \frac{\theta_0^2}{Var(\hat{\theta}_{MVU})})} < \alpha < 0$

f) $\hat{\theta}_b = (1 + \alpha)\hat{\theta}_{MVU}$ We have that $|1 + \alpha| < 1 \quad (1)$ which holds when $-2 < \alpha < 0$ Also $|\hat{\theta}_b| = |1 + \alpha| |\hat{\theta}_{MVU}| \rightarrow$ From (1) $|\hat{\theta}_b| < |\hat{\theta}_{MVU}|$

g) $\frac{d}{da}((1 + \alpha)^2 Var(\hat{\theta}_{MVU}) + (\alpha\theta_0)^2) \rightarrow 2(1 + \alpha)\mathrm{Var}(\hat{\theta}_{MVU}) + 2\alpha\theta_0^2 = 0 \rightarrow \alpha^* = \frac{-Var(\hat{\theta}_{MVU})}{Var(\hat{\theta}_{MVU}) + \theta_0^2}$

h) In practice, α^* cannot be determined because it depends on unknown parameters θ_0 and $Var(\theta_{M\hat{V}U})$, which are not knwown in real-world cases.

Exercise 14

Consider a set N pairs $(y_n, \boldsymbol{x}_n)$, $n = 1, ..., N$, satisfying the equation

$$y_n = \boldsymbol{\theta}_o{}^T \boldsymbol{x}_n + \eta_n, \qquad\qquad (3)$$

where η_n is normally distributed **zero mean** i.i.d. noise. As it is known, the LS estimator satisfies the equation

$$\left(\textstyle\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n{}^T\right)\boldsymbol{\theta} = \textstyle\sum_{n=1}^{N} y_n \boldsymbol{x}_n \qquad\qquad (4)$$

Consider now the special case where the $\boldsymbol{\theta}$ is a scalar and $\boldsymbol{x}_n = 1$ for all n. In this case, eq. (3) becomes

$$y_n = \theta_o \pm \eta_n. \qquad\qquad (3)^1$$

Let y_n denote the rv that models y_n. Let also $\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$ be the mean of N statistically independent rv's that follow the same pdf. Each one of the y_n's models a y_n.

(a) Derive the LS estimator of θ_o from eq. (4) for this case (Note that in eq. (4) all x_n's are now scalars equal to 1).
(b) Prove that y_n is an unbiased estimator of θ_o (Show that $[y_n] = \theta_o$).
(c) Prove that $\bar{y}$ is an unbiased estimator of θ_o (Show that $E[E[\bar{y}] = \theta_o$).
(d) Since $\bar{y}$ is the LS estimator for the 1-dim. case, it is known that it is minimum variance unbiased estimator. From now on it will be denoted by $\hat{\theta}_{MVU}$.
(e) Prove that the ridge regression estimator for the present 1-dim. case is expressed as $\hat{\theta} = \frac{\sum_{n=1}^{N} y_n}{N+\lambda}$ (use the formula (A) proved in exercise 1a).
(f) Denoting as $\hat{\theta}$ the ridge regression estimator, express it in terms of $\hat{\theta}_{MVU}$.
(g) Prove that $\hat{\theta}$ is a biased estimator (Show that $E[\bar{y}] \neq \theta_o$).
(h) Verify that $|\hat{\theta}| < |\hat{\theta}_{MVU}|$.

Answer

a) $\left(\sum_{n=1}^{N} x_n x_n^T\right)\theta = \sum_{n=1}^{N} y_n x_n \rightarrow \left(\sum_{n=1}^{N} 1\right)\theta = \sum_{n=1}^{N} y_n \rightarrow N\theta = \sum_{n=1}^{N} y_n \rightarrow \hat{\theta}_0 = \frac{1}{N}\sum_{n=1}^{N} y_n = \bar{y}$

b)$E[y_n] = E[\theta_0 + \eta_n] = \theta_0 + E[\eta_n]$ Since $E[\eta_n] = 0$, we have that: $E[y_n] = \theta_0$ Thus, y_n is an unbiased estimator of θ_0.

c) $E[\bar{y}] = E\left[\frac{1}{N}\sum_{n=1}^{N} y_n\right] = \frac{1}{N}\sum_{n=1}^{N} E[y_n]$ From part (b), $E[y_n] = \theta_0 \rightarrow E[y] = \frac{1}{N}\sum_{n=1}^{N} \theta_0 = \theta_0$

So $\bar{y}$ is an unbiased estimator of θ_0.

e)Using the formula from Exercise 1a, $\hat{\theta}_{ridge} = \left(X^T X + \lambda I\right)^{-1} X^T y$

Since $X_n = 1$ for all $n \rightarrow X^T X = N$ and $X^T y = \sum_{n=1}^{N} y_n \rightarrow \hat{\theta}_{ridge} = \frac{\sum_{n=1}^{N} y_n}{N+\lambda}$

f) $\hat{\theta}_{ridge} = \frac{\sum_{n=1}^{N} y_n}{N+\lambda}$ and $\hat{\theta}_{MVU} = \frac{1}{N}\sum_{n=1}^{N} y_n = \bar{y} \rightarrow \hat{\theta}_{ridge} = \frac{N}{N+\lambda}\hat{\theta}_{MVU}$

g) Expectations are $\mathbb{E}[\hat{\theta}_{ridge}] = \frac{N}{N+\lambda}\mathbb{E}[\hat{\theta}_{MVU}]$

$E[\hat{\theta}_{MVU}] = \theta_0 \rightarrow E[\hat{\theta}_{ridge}] = \frac{N}{N+\lambda}\theta_0$, We have that $E[\hat{\theta}_{ridge}] = \theta_0$ only for $\lambda = 0$

h) We proved in f) $\hat{\theta}_{ridge} = \frac{N}{N+\lambda}\hat{\theta}_{MVU}$ also $\frac{N}{N+\lambda} < 1 for \lambda > 0 \rightarrow |\hat{\theta}_{ridge}| = \left|\frac{N}{N+\lambda}\hat{\theta}_{MVU}\right| < |\hat{\theta}_{MVU}| \rightarrow |\hat{\theta}_{ridge}| < |\hat{\theta}_{MVU}|$

From 3, $\hat{\theta}_b = (1+\alpha)\hat{\theta}_{MVU}$, also $\hat{\theta}_{ridge} = \frac{N}{N+\lambda}\hat{\theta}_{MVU} \rightarrow$

$\alpha = \frac{N}{N+\lambda} - 1 = \frac{N-(N+\lambda)}{N+\lambda} = \frac{-\lambda}{N+\lambda} \rightarrow \alpha = \frac{-\lambda}{N+\lambda}$

Exercise 15

Consider a regression problem where both the independent and dependent quantities are scalars and are related via the following linear model

$$y = \theta_o \cdot x + \eta$$

where η follows the zero mean normal distribution with variance σ^2 and $\theta_o = 2$ (thus, the actual model is $y = 2 \cdot x + \eta$).

(a) Generate $d = 50$ data sets as follows:

- Generate a set D_1 of $N = 30$ data pairs (y_i', x_i), where $y' = 2 \cdot x$.

- Add zero mean and $\sigma^2 = 64$ variance Gaussian noise to the y_i' 's, resulting to y_i's.

- The **observed** data pairs are (y_i, x_i), $i = 1, ...,30$, which constitute the data set D_1.

  Repeat the above procedure $d = 50$ times in order to generate 50 different data sets.

(b) Compute the LS linear **estimates** of θ_o based on $D_1, D_2, ..., D_d$ (thus, $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_d$ numbers/estimates will result).

(c) Consider now the random variable $\hat{\theta}$ that models $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_d$ (that is, $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_d$ can be viewed as instances of the random variable $\hat{\theta})^2$ and

(c1) estimate the $MSE = E\left[(\hat{\theta} - \theta_o)^2\right]$ and

(c2) depict graphically the values $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_d$ and comment on how they are spread around θ_o.

Hint: For (c) approximate MSE as $MSE = \frac{1}{d}\sum_{i=1}^{d}(\hat{\theta}_i - \theta_o)^2$.

Answer:

In [4]:
```python
import numpy as np
import matplotlib.pyplot as plt

# Parameters
d = 50   # number of datasets
N = 30   # number of data pairs in each dataset
theta_0 = 2   # true value of theta
sigma2 = 64   # variance of the noise

# Prepare to store theta estimates
theta_hats = []

# Generate d datasets and compute LS estimates
for _ in range(d):
    # Generate x values uniformly between 0 and 10
    x = np.random.normal(0, 10, N)

    # Generate noise eta ~ N(0, sigma^2)
    eta = np.random.normal(0, np.sqrt(sigma2), N)

    # Generate y' values based on y' = theta_0 * x
    y_new = theta_0 * x

    # Generate y by adding noise to y'
    y = y_new + eta

    # Compute the general LS estimate of theta (theta_hat)
    theta_hat = (x.T @ y) / (x.T @ x)
    theta_hats.append(theta_hat)

# Convert theta_hats to a numpy array for convenience
theta_hats = np.array(theta_hats)

# (c1) Estimate the MSE
MSE = np.mean((theta_hats - theta_0) ** 2)

# (c2) Plot the estimates
plt.figure()
plt.plot(theta_hats, 'bo-', label=r'$\hat{\theta}_i$ values')
plt.axhline(y=theta_0, color='r', linestyle='--', label=r'True $\theta_0 = 2$')
plt.xlabel('Dataset index')
plt.ylabel(r'$\hat{\theta}$')
plt.title(r'LS estimates of $\theta$ over 50 datasets')
plt.legend()
plt.show()

# Print the MSE
print(f"Estimated MSE: {MSE:.5f}")
```
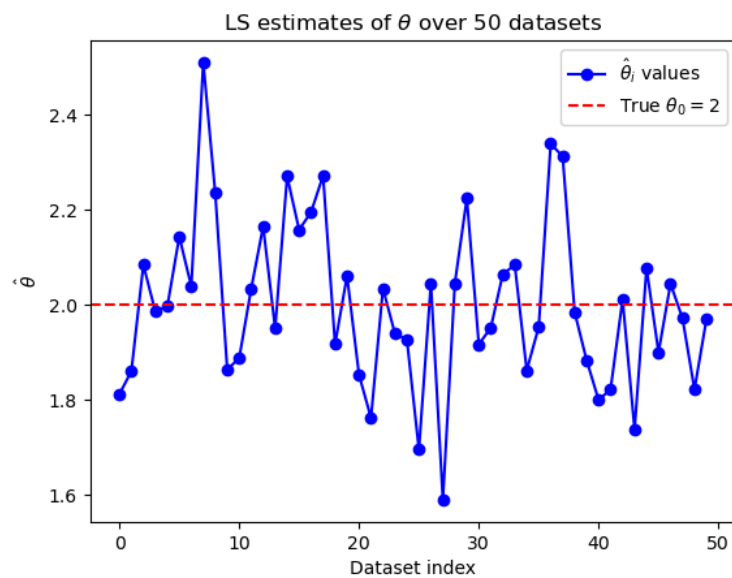
LS estimates of θ over 50 datasets

Estimated MSE: 0.03101

The plot indeed shows variability around the true value $\theta 0=2$, with most estimates ranging between 1.6 and 2.4. Overall, the variability observed is an expected consequence when dealing with estimates from multiple datasets, each of which carries its own noise and uncertainty. The spread of the estimates (between 1.6 and 2.4) indicates that while the estimates are fairly close to the true value on average, there is still noticeable variance.

In [ ]: