M50 Homework 7

Alex Craig

Exercise 1 (MSE)

Prove that

$$\text{MSE}_{\hat{y}(x|D)} = \sigma^2 + \text{var}(\hat{y}(x|D)) + E[\hat{y}(x|D) - f(x)]^2$$

where $\text{MSE}_{\hat{y}(x|D)}$ is defined in Equation 3 in the class notes.

Solution

The mean squared error of an estimator can be decomposed into two main parts: variance and bias squared. The general formula for MSE is:

$$\text{MSE} = \text{variance} + \text{bias}^2$$

In the context of this problem, the MSE of the estimator $\hat{y}(x|D)$ is the expected value of the squared difference between the estimator and the true value, which can be written as:

$$\text{MSE}_{\hat{y}(x|D)} = E[(\hat{y}(x|D) - f(x))^2]$$

Expanding this, we get:

$$\text{MSE}_{\hat{y}(x|D)} = E[(\hat{y}(x|D) - E[\hat{y}(x|D)] + E[\hat{y}(x|D)] - f(x))^2]$$
$$E[(\hat{y}(x|D) - E[\hat{y}(x|D)])^2] + 2E[(\hat{y}(x|D) - E[\hat{y}(x|D)])(E[\hat{y}(x|D)] - f(x))] + E[(E[\hat{y}(x|D)] - f(x))^2]$$
$$\text{var}(\hat{y}(x|D)) + 0 + (E[\hat{y}(x|D)] - f(x))^2$$
$$\text{var}(\hat{y}(x|D)) + \text{bias}^2_{\hat{y}(x|D)}$$

Here, $\text{var}(\hat{y}(x|D))$ is the variance of the estimator, and $\text{bias}^2_{\hat{y}(x|D)}$ is the square of the bias, which is $(E[\hat{y}(x|D)] - f(x))^2$.

The equation can be rewritten as:

$$\text{MSE}_{\hat{y}(x|D)} = \sigma^2 + \text{var}(\hat{y}(x|D)) + (E[\hat{y}(x|D)] - f(x))^2$$

This completes the proof, showing that the mean squared error of the estimator is the sum of its variance, squared bias, and an additional error term σ^2, which could represent the inherent noise in the data or model.

Exercise 2 (Laplace's Rule)

Consider a Bernoulli random variable

$$X \sim \text{Bernoulli}(q).$$

(You can assume this means $P(X = 1) = q$). If we have samples $X_1, \ldots, X_N$ for X then we have seen that a consistent and unbiased estimator of q is

$$\hat{q} = \frac{\sum_{i=1}^{N} X_i}{N}.$$

An alternative estimator called Laplace's rule of succession is

$$\hat{q}_L = \frac{Y + 1}{N + 2}.$$

The motivation for defining $\hat{q}_L$ is as follows: Think of X as a biased coin. If we know that it is possible for to roll a heads or a tails then we should include this information in our estimator. However, using the original estimator, if we roll a sequence

of only heads (or tails), we will estimate that $q = 1$ (or $q = 0$). To correct for this, we pretend we have two additional observations (hence the $N + 2$ in the denominator) and that one is heads and one is tails (hence $Y + 1$ in the numerator). This is a simple example where we are incorporating prior information into our estimator – that is, information beyond what is present in the data and our model. In this case, that prior information is that the coin has two sides and could land on either one, however unlikely that might be.

Part A

Derive formula for the mean-squared error

$$\text{MSE}_{\hat{q}_L} = E\left[(\hat{q}_L - q)^2\right]$$

and decompose it into the variance and the squared bias. Your formulas should be in terms of q and N.

Solution

1. **Variance** of $\hat{q}_L$:

   The variance of an estimator is its squared deviation from its own mean:

   $$\text{Var}(\hat{q}_L) = E[(\hat{q}_L - E[\hat{q}_L])^2]$$

2. **Bias** of $\hat{q}_L$:

   The bias is the difference between the expected value of the estimator and the true value:

   $$\text{Bias}(\hat{q}_L) = E[\hat{q}_L] - q$$

   The squared bias is then:

   $$(\text{Bias}(\hat{q}_L))^2 = (E[\hat{q}_L] - q)^2$$

The MSE can then be represented as:

$$\text{MSE}(\hat{q}_L) = \text{Var}(\hat{q}_L) + (\text{Bias}(\hat{q}_L))^2$$

Part B

Is $\hat{q}_L$ unbiased and consistent?

Solution

Exercise 2b asks whether Laplace's rule of succession estimator $\hat{q}_L$ is unbiased and consistent. Let's analyze these two properties:

Unbiasedness

An estimator is unbiased if its expected value equals the true parameter value. From the previous calculation, we have:

$$E[\hat{q}_L] = \frac{Nq + 1}{N + 2}$$

To check for unbiasedness, we need to see if $E[\hat{q}_L] = q$. Clearly, since $E[\hat{q}_L]$ depends on N, it does not equal q for all N. Therefore, $\hat{q}_L$ is not an unbiased estimator of q.

Consistency

An estimator is consistent if it converges in probability to the true parameter value as the sample size N goes to infinity. We need to check if $\hat{q}_L$ converges to q as $N \to \infty$.

$$\lim_{N\to\infty} E[\hat{q}_L] = \lim_{N\to\infty} \frac{Nq+1}{N+2}$$

As N becomes very large, the $+1$ and $+2$ terms become negligible compared to the terms involving N. Thus, the limit simplifies to:

$$\lim_{N\to\infty} \frac{Nq+1}{N+2} = \lim_{N\to\infty} \frac{Nq}{N} = q$$

Therefore, $\hat{q}_L$ is a consistent estimator of q.

Summary

- The Laplace's rule of succession estimator $\hat{q}_L$ is **not unbiased** because its expected value does not equal q for all sample sizes N.
- The estimator $\hat{q}_L$ is **consistent**, as it converges to the true value q as the sample size N goes to infinity.

Part C

Now compute $\text{MSE}_{\hat{q}}$. Note that in this case the bias is zero, so it should be straightforward to derive from the standard error. Surprisingly, $\text{MSE}_{\hat{q}} > \text{MSE}_{\hat{q}_L}$ for some values of N and q. For which values is this the case? Note that this is quite surprising since it seems like $\hat{q}$ should be the best guess of q!

Solution

Computing MSE for $\hat{q}$

The original estimator $\hat{q}$ is given by:

$$\hat{q} = \frac{Y}{N}$$

where Y is the number of successes in N trials, and q is the true probability of success.

1. Bias of $\hat{q}$ The bias of an estimator is defined as $E[\hat{q}] - q$. For the estimator $\hat{q}$, the expected value $E[\hat{q}]$ is q because $E[Y] = Nq$ (since Y follows a binomial distribution with parameters N and q):

$$E[\hat{q}] = E\left[\frac{Y}{N}\right] = \frac{E[Y]}{N} = \frac{Nq}{N} = q$$

Therefore, the bias of $\hat{q}$ is 0, making it an unbiased estimator.

2. Variance of $\hat{q}$ The variance of $\hat{q}$ can be calculated as:

$$\text{Var}(\hat{q}) = \text{Var}\left(\frac{Y}{N}\right) = \frac{1}{N^2}\text{Var}(Y)$$

Since Y follows a binomial distribution, $\text{Var}(Y) = Nq(1-q)$, so:

$$\text{Var}(\hat{q}) = \frac{Nq(1-q)}{N^2} = \frac{q(1-q)}{N}$$

3. Mean Squared Error of $\hat{q}$ The mean squared error is the sum of the variance and the square of the bias. Since the bias is 0, the MSE is just the variance:

$$\text{MSE}(\hat{q}) = \text{Var}(\hat{q}) = \frac{q(1-q)}{N}$$

Comparison with MSE of $\hat{q}_L$

To compare $\text{MSE}(\hat{q})$ with $\text{MSE}(\hat{q}_L)$, recall the result from Exercise 2a and 2b. We need to consider the values of N and q for which $\text{MSE}(\hat{q}) > \text{MSE}(\hat{q}_L)$.

Given that $\text{MSE}(\hat{q}) = \frac{q(1-q)}{N}$ and the expression for $\text{MSE}(\hat{q}_L)$ involves both variance and squared bias, the comparison depends on how these quantities change with different values of N and q. Specifically, for small sample sizes N or values of q close to 0 or 1, $\hat{q}_L$ can have a lower MSE due to its adjustment for extreme outcomes, despite being biased.

Therefore, while $\hat{q}$ might seem like the best estimator due to its unbiasedness, there can be scenarios (especially with small N or extreme q values) where $\hat{q}_L$ provides a lower mean-squared error, demonstrating a trade-off between bias and variance in statistical estimation.

Exercise 3 (Impulse function features)

Consider the model

$$Y|X \sim \text{Normal}$$

Suppose that $X \in [0, 1)$ and define the intervals

$$I_i = \left[\frac{i-1}{K}, \frac{i}{K}\right)$$

for $i = 1, \ldots, K$. Notice that

$$[0, 1) = I_1 \cup I_2 \cup \ldots \cup I_K.$$

That is, each x in $[0, 1)$ is in one of these disjoint intervals. Now introduce the features

$$\phi_i(x) = \begin{cases} 1 & \text{if } x \in \left[\frac{i-1}{K}, \frac{i}{K}\right) \\ 0 & \text{if } x \notin \left[\frac{i-1}{K}, \frac{i}{K}\right) \end{cases}$$

Part A

Are ϕ_i orthogonal with respect to a random variable X taking values in $[0, 1)$? Does it depend on the distribution of the random variable?

Solution

Two functions ϕ_i and ϕ_j are said to be orthogonal with respect to a random variable X if their expected product is zero, i.e.,

$$E[\phi_i(X) \cdot \phi_j(X)] = 0 \quad \text{for all } i \neq j$$

Since each $\phi_i(x)$ is an indicator function for a distinct interval I_i, and these intervals are disjoint, $\phi_i(x)$ and $\phi_j(x)$ (for $i \neq j$) cannot both be 1 for any x. Therefore, their product will always be 0:

$$\phi_i(x) \cdot \phi_j(x) = 0 \quad \text{for all } i \neq j$$

Hence, for any value of X, the expected product $E[\phi_i(X) \cdot \phi_j(X)] = 0$, confirming that the features ϕ_i are orthogonal to each other.

Dependence on the Distribution of X

The orthogonality of ϕ_i and ϕ_j in this case is independent of the distribution of the random variable X. This is because the orthogonality is derived purely from the non-overlapping nature of the intervals that the ϕ_i functions represent. Regardless of how X is distributed over the interval $[0, 1)$, the fact that the intervals I_i are disjoint ensures that ϕ_i and ϕ_j (for $i \neq j$) cannot be non-zero simultaneously.

Conclusion

The features ϕ_i are orthogonal with respect to a random variable X taking values in $[0, 1)$, and this orthogonality is independent of the distribution of X.

Part B

Using statsmodels, implement fitting the model with these features. You can make up your simulated data set to fit or copy the code I used in class to fit the Fourier and polynomial models. I recommend writing a function $\phi(x_i)$ which takes the array of predictors and outputs an array $[\phi_j(X_1), \ldots, \phi_j(X_N)]$. Use $K = 10$ and $N = 100$.

Solution

```
import numpy as np
import statsmodels.api as sm

# Parameters
N = 100  # Number of data points
K = 10   # Number of intervals

# Step 1: Generate a simulated dataset
np.random.seed(0)  # For reproducibility
X = np.random.uniform(0, 1, N)  # Uniformly distributed X values in [0, 1)
Y = np.sin(2 * np.pi * X) + np.random.normal(0, 0.1, N)  # Y as a function of X with noise

# Step 2: Implement the feature function
def phi(x, k, K):
    """Indicator function for the interval [(k-1)/K, k/K)."""
    return (x >= (k-1)/K) & (x < k/K)

# Creating the design matrix with the features
phi_matrix = np.column_stack([phi(X, k, K) for k in range(1, K+1)])

# Step 3: Fit the model using statsmodels
model = sm.OLS(Y, phi_matrix)
results = model.fit()

# Display the results
# results.summary()
```

Part C

As usual, let $\hat{\beta}_j$ be the fitted value of β_j using least squares, meaning the value that minimizes the squared residuals. Show that in this model $\hat{\beta}_j$ is simply the average value of Y_i among data points where $X_i \in I_j$; that is

Solution

In the model, Y is the dependent variable, and $\phi_i(x)$ are indicator functions that represent whether x falls within a specific interval I_i. The model can be written as:

$$Y = \beta_1\phi_1(X) + \beta_2\phi_2(X) + \ldots + \beta_K\phi_K(X) + \epsilon$$

where ϵ represents the error term.

Approach

To prove this statement, we'll consider the least squares estimation procedure for linear regression, which minimizes the sum of squared residuals:

$$\text{minimize} \quad \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i = \sum_{j=1}^{K} \beta_j \phi_j(X_i)$.

Analysis for a Single Interval I_j

Focusing on a single interval I_j and its corresponding coefficient β_j, the least squares criterion becomes:

$$\text{minimize} \quad \sum_{i=1}^{N} (Y_i - \beta_j \phi_j(X_i))^2$$

However, remember that $\phi_j(X_i)$ is 1 if $X_i \in I_j$ and 0 otherwise. Therefore, only the points X_i in the interval I_j contribute to the sum.

Simplifying the Summation

For data points where $X_i \in I_j$, $\phi_j(X_i) = 1$. For data points outside I_j, $\phi_j(X_i) = 0$, and they don't contribute to the sum. So the summation simplifies to:

$$\text{minimize} \quad \sum_{X_i \in I_j} (Y_i - \beta_j)^2$$

Finding the Minimum

To find the minimum, we differentiate this sum with respect to β_j and set the derivative to zero:

$$\frac{d}{d\beta_j} \sum_{X_i \in I_j} (Y_i - \beta_j)^2 = 0$$

Expanding and simplifying gives:

$$-2 \sum_{X_i \in I_j} (Y_i - \beta_j) = 0$$

Solving for β_j:

$$\sum_{X_i \in I_j} Y_i = \beta_j \times \text{count of } X_i \in I_j$$

Hence,

$$\beta_j = \frac{\sum_{X_i \in I_j} Y_i}{\text{count of } X_i \in I_j}$$

Conclusion

This equation shows that β_j, the fitted value of the coefficient for the j-th interval using least squares, is indeed the average value of Y_i among the data points where X_i falls in the interval I_j. This is a direct result of the least squares minimization procedure and the specific structure of the indicator functions $\phi_i(x)$ used in the model.