# Self-Notes on [Bias Variance Dilemma, Probabilistic Interpretation of Regression and Multivaria]

[Unat Tekşen] [504241592]

May 2, 2025

## 1 Bias Variance Dilemma, Probabilistic Interpretation of Regression and Multivaria

### 1.1 Measuring Quality of Estimation - MSE Analysis

If we have estimations, we can write weights like that (Prof. Ethem Alpaydın's notation) and we can judge our estimations and measure the quality of the estimation for different datasets:

$$0.7 = d_1 = \hat{\theta}_1$$
$$0.8 = d_2 = \hat{\theta}_2$$

We can calculate MSE with fixed $\theta$ and random variable d:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(d-\theta)^2 \tag{1}$$

We can compute 2 statistics from estimations:
1) Expectation value:

$$= E(d) \tag{2}$$

2) Variance:

$$= E[(d - E(d))^2] \tag{3}$$

We can write bias in terms of $\theta$ and d:

$$= \theta - E[d] \tag{4}$$

Aim: Writing MSE in terms of 2 terms:

$$E[(d-\theta)^2] = E[d^2] - 2E[d\theta] + E[\theta^2] \tag{5}$$

$$E[(d-\theta-E[d]+E[d])^2] = E[(d-E[d]+E[d]-\theta)^2] \tag{6}$$

$$= E[(d-E[d])^2 + 2(d-E[d])(E[d]-\theta) + (E[d]-\theta)^2] \tag{7}$$

In this notation, $\theta$ is fixed, expectation is not changing!
The first term is variance:

$$= E[(d-E[d])^2] \tag{8}$$

The second term is 0:

$$2(d-E[d])(E[d]-\theta) = 0 \tag{9}$$

Steps:

$$= E[dE[d] - d\theta - E[d]^2 + \theta E[d]] \tag{10}$$

$$= E[dE[d]] \tag{11}$$

$$= E[d]^2 - \theta E[d] - E[d]^2 + \theta E[d] = 0 \tag{12}$$

The third term is $bias^2$:

$$= (E[d]-\theta)^2 \tag{13}$$

So, MSE = variance + $bias^2$.

## 1.2 Measuring Goodness of Fitness

We want to simulate bias variance plot with cross-validation.

## 1.3 Probabilistic Interpretation of Least Square Regression

$$P(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{14}$$

Expressing with likelihood:

$$P(y_1, \ldots, y_N) = \prod_{i=1}^{N} \frac{L}{\sqrt{2\pi\sigma^2}} \left(-\frac{(y^T - \theta^T x^i)^2}{2\sigma^2}\right) \tag{15}$$

We can write log-likelihood, we want to maximize second term:

$$\mathcal{L}(\theta) = \log(l(\theta)) = \sum_{i=1}^{N} \frac{1}{2\pi\sigma^2} + \sum_{i=1}^{N} -\frac{1}{2\sigma^2}(y^T - \theta^T x^i)^2 \tag{16}$$

ML estimation:

$$\underset{\theta}{\text{argmax}}\mathcal{L}(\theta) = \underset{\theta}{\text{argmin}} - \mathcal{L}(\theta) = \underset{\theta}{\text{argmin}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \theta^T x^i)^2 - \sum_{i=1}^{N} \ln \frac{1}{2\pi\sigma^2} \quad (17)$$

- **LSR parameters are the parameters that maximize the likelihood of y's.**
- If $E[y] = 0$ and we add constant $+5$ for each term, only mean will change and variance wont change, so $E[y] = 5$.
- $y^i$ observations come from $\theta^\tau x^i$ where:
- Gaussian Noise $\mathcal{N}(0, \sigma^2)$ is added.

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ x_1 \end{bmatrix} \quad (18)$$

Every $y^i$ comes from a Gaussian distribution:

$$y^i = \theta^\tau x^i + N(0, \sigma^2) \quad (19)$$

$$y^i \sim N(\theta^\tau x^i, \sigma^2) \quad (20)$$

## 1.4 Multivariate Methods

We have d-dimensional dataset:

$$X \in \mathbb{R}^d \quad (21)$$

$$\mu \in \mathbb{R}^d \quad \text{(mean vector)} \quad (22)$$

$$\Sigma \in \mathbb{R}^{d \times d} \quad \text{(covariance matrix)} \quad (23)$$

Mean vector:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_d] \end{bmatrix} \quad (24)$$

Covariance matrix (transpose of covariance matrix is equal to itself):

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1d}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{d1}^2 & \cdots & \sigma_{dd}^2 \end{bmatrix} \quad (25)$$

Data X:

$$X = \begin{bmatrix} x_1^{(1)} & x_d^{(1)} \\ x_1^{(2)} & x_d^{(2)} \\ \vdots & \vdots \\ x_1^{(N)} & x_d^{(N)} \end{bmatrix} \tag{26}$$

The probability density function (PDF) is given by:

$$p(x) = \frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right) \tag{27}$$

We can express the mean vector:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} x_i^j \tag{28}$$

$$\Sigma = E[(x-\mu)(x-\mu)^T] \tag{29}$$

$\hat{\mu}$ represents parameters estimated from data:

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)}-\mu)(x^{(i)}-\mu)^T \tag{30}$$

**Example:**

$$X = \begin{bmatrix} 2 & 9 \\ 4 & 6 \\ 6 & 3 \\ 8 & 1 \\ 10 & 1 \end{bmatrix} \tag{31}$$

1) Mean vector

$$\mu_1 = \frac{1}{5}\sum_{i=1}^{5} x_1^i = 6 \tag{32}$$

$$\mu_2 = \frac{1}{5}\sum_{i=1}^{5} x_2^i = 5 \tag{33}$$

$$\mu^T = \begin{bmatrix} 6 & 5 \end{bmatrix} \tag{34}$$

2) Covariance matrix:

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)}-\mu)(x^{(i)}-\mu)^T \tag{35}$$

$$\Sigma = \frac{1}{5}\left(\begin{bmatrix} 2-6 \\ 9-5 \end{bmatrix}\begin{bmatrix} -4 & 4 \end{bmatrix} + \begin{bmatrix} 4-6 \\ 7-5 \end{bmatrix}\begin{bmatrix} -2 & 2 \end{bmatrix} + \cdots + \begin{bmatrix} 10-6 \\ 1-5 \end{bmatrix}\begin{bmatrix} 4 & -4 \end{bmatrix}\right) \tag{36}$$

$$\Sigma = \begin{bmatrix} 8 & -8 \\ -8 & 8 \end{bmatrix} \tag{37}$$

Also, we can use:

$$\sigma_1^2 = \frac{1}{N} \sum_i (x_1^i - \mu_1) \tag{38}$$

$$\sigma_{12}^2 = \frac{1}{N} \sum_i (x_1^i - \mu_1)(x_2^i - \mu_2) \tag{39}$$

### 1.4.1 Covariance

$$\rho = \text{Cov}(x_i, x_j) = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{40}$$

Example:

$$\text{Ex} \quad \text{Cov} = \frac{-8}{\sqrt{8}\sqrt{8}} = -1 \tag{41}$$

- If features are independent, then correlation coefficient is 0 $\rho = 0$.
- If $x_i, x_j$ are independent, diagonals are non-zero.

### 1.4.2 Estimating Curves

When we divide the plot into squares, count samples, create a histogram and we fit a curve to the histogram, it will be bell-shaped.

- If $x_1$ increases, $x_2$ increases or decreases. (Left-top plot in the Figure 1)

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

- If $x_1$ increases, $x_2$ increases. Figure (Left-bottom plot in the 1)

$$\Sigma = \begin{bmatrix} \sigma^2 & + \\ + & \sigma^2 \end{bmatrix}$$

- If $x_1$ increases, $x_2$ decreases. Figure (Right-bottom plot in the 1)

$$\Sigma = \begin{bmatrix} \sigma^2 & - \\ - & \sigma^2 \end{bmatrix}$$

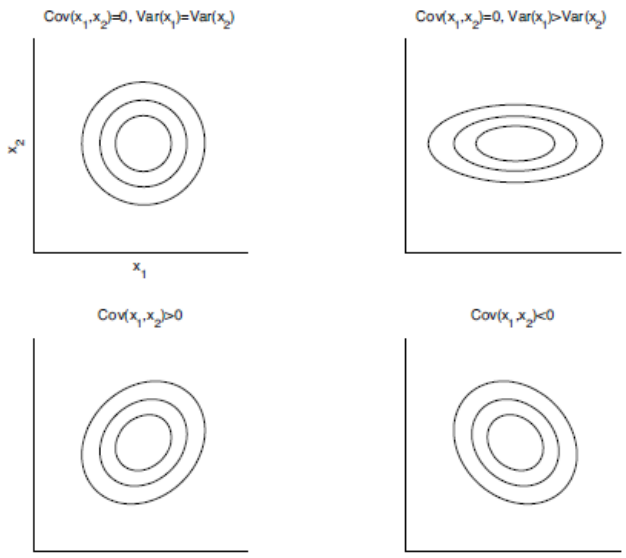


Figure 1: Isoprobability contour plots. Mean is the center, their shapes depend on cov. matrix. Retrieved from [1], page 92.

### 1.4.3 Multivariate Classification

$$P(c|x) = \frac{P(x|c)P(c)}{P(c)} \tag{42}$$

Numerator $P(x|c=1,\mu_1,\Sigma_1)P(c=1)$ is $g_1(x)$:

$$P(c=1|x) = \frac{P(x|c=1,\mu_1,\Sigma_1)P(c=1)}{P(x)} \tag{43}$$

Numerator $P(x|c=0,\mu_0,\Sigma_0)P(c=0)$ is $g_1(x)$:

$$P(c=0|x) = \frac{P(x|c=0,\mu_0,\Sigma_0)P(c=0)}{P(x)} \tag{44}$$

$$P(c=1|x) \overset{?}{\geq} P(c=0|x) \tag{45}$$

Classifying based on numerators:

$$\text{Class} = \begin{cases} 1 & \text{if } g_1(x) > g_0(x) \\ 0 & \text{if } g_1(x) < g_0(x) \end{cases}$$

Expanding $g_0(x)$ and $g_1(x)$:

$$g_1(x) = \frac{1}{\sqrt{2\pi}|\Sigma_0|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right) P(c=0) \tag{46}$$

$$g_0(x) = \frac{1}{\sqrt{2\pi}|\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right) P(c=1) \tag{47}$$

We can represent $\frac{1}{\sqrt{2\pi}|\Sigma_1|^{1/2}}$ as C.

For example, if we assume that:

$$\Sigma = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \quad x = \begin{bmatrix}6\\ 3\end{bmatrix}, \quad \mu_1 = \begin{bmatrix}6\\ 5\end{bmatrix}, \quad \mu_0 = \begin{bmatrix}0\\ 0\end{bmatrix}$$

$$g_1(x) = C\exp\left(-\frac{1}{2}\begin{bmatrix}0 & -2\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}0\\ -2\end{bmatrix}\right) = C\exp(-2) \tag{48}$$

$$g_0(x) = C\exp\left(-\frac{1}{2}\begin{bmatrix}0 & 3\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}0\\ 3\end{bmatrix}\right) = C\exp\left(-\frac{4.5}{2}\right) \tag{49}$$

Taking the logarithm of both sides:

$$\ln g_0(x) = \ln C - 2 \tag{50}$$

$$\ln g_1(x) = \ln C - \frac{4.5}{2} \tag{51}$$

And it concludes that:

$$\ln g_0(x) > \ln g_1(x) \implies \text{class } 0 \tag{52}$$

If we take the logarithm:

$$g_1(x) = \ln\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\ln|\Sigma_1|^{(1/2)} - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) \tag{53}$$

If we use the distributive property

$$g_1 = -\frac{1}{2}x^T\Sigma_1^{-1}x - x^T\Sigma_1^{-1}\mu_1 - \mu_1\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1 \tag{54}$$

$$g_0 = -\frac{1}{2}x^T\Sigma_0^{-1}x - x^T\Sigma_0^{-1}\mu_0 - \mu_0\Sigma_0^{-1}x + \mu_0^T\Sigma_0^{-1}\mu_0 \tag{55}$$

If we assume these, and try to express $g_1(x) - g_0(x)$ in terms of $x_1$ and $x_2$:

$$\Sigma = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \quad x = \begin{bmatrix}x1\\ x2\end{bmatrix}, \quad \mu_1 = \begin{bmatrix}6\\ 5\end{bmatrix}, \quad \mu_0 = \begin{bmatrix}0\\ 0\end{bmatrix}$$

$$g_0(x) = -\frac{1}{2}\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} - \begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}0\\ 0\end{bmatrix} - \begin{bmatrix}0 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}0\\ 0\end{bmatrix} \tag{56}$$

$$g_1(x) = -\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 6 \\ 5 \end{bmatrix} - \begin{bmatrix} 6 & 5 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 6 & 5 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 6 \\ 5 \end{bmatrix} \tag{57}$$

The difference:

$$g_1(x) - g_0(x) = 6x_1 + 5x_2 - \frac{61}{2} \tag{58}$$

# References


[1] Ethem Alpaydın, *Introduction to Machine Learning, 2nd Edition, MIT Press* 2010.