Notes on Lecture 2

Lu Sun

March 4, 2020

1 Solution to the Quiz

Suppose a linear relationship with measurement error

$$Y = X^\top \beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

Given that $\mathbf{X}^\top \mathbf{X}$ is invertible, we have following estimation according to least squares

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{2}$$

Thus, the prediction $\hat{y}_0$ on an arbitrary test point x_0 is

$$\begin{aligned}
\hat{y}_0 &= x_0^\top \hat{\beta} \\
&= x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \boldsymbol{\epsilon}) \\
&= x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta + x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon} \\
&= x_0^\top \beta + \sum_{i=1}^{N} \ell_i(x_0)\epsilon_i,
\end{aligned} \tag{3}$$

where $\ell_i(x_0)$ is the i-th element of the row vector $x_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, or equivalently, the column vector $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} x_0$.

2 Derivation of Expected Prediction Error

The following shows the derivation of EPE(x_0) in p.23 of our slides in the lecture 2. Here the problem setting is same with the above section.

First of all, let's try to understand the difference among some notations.

- y_0: the **observed** output value at x_0;
- $f(x_0) = x_0^\top \beta$: the **ground truth** value at x_0;
- $\hat{y}_0 = x_0^\top \hat{\beta}$: the **predicted** value at x_0.

Obviously, according to Eq.(1), we have $y_0 = x_0^\top \beta + \epsilon_0$.

Next we define the squared prediction error (PE) at x_0 by

$$\begin{aligned}
\text{PE} &\coloneqq (y_0 - \hat{y}_0)^2 \\
&= \big((y_0 - x_0^\top \beta) - (\hat{y}_0 - x_0^\top \beta)\big)^2 \\
&= (y_0 - x_0^\top \beta)^2 - 2(y_0 - x_0^\top \beta)(\hat{y}_0 - x_0^\top \beta) + (\hat{y}_0 - x_0^\top \beta)^2 \\
&= A^2 - 2AB + B^2.
\end{aligned} \tag{4}$$

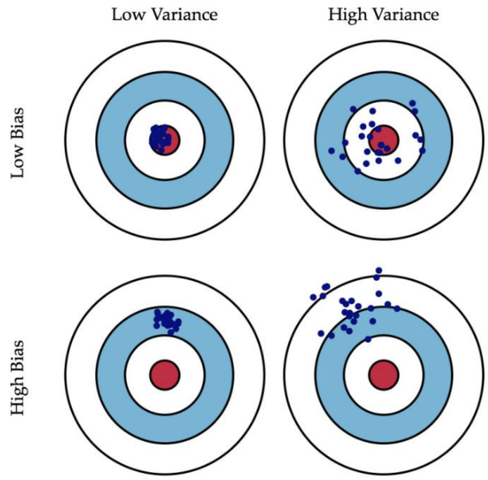

Figure 1: The difference between Bias and Var.

Note that the term A depends only ϵ and thus has expectation 0 ($\mathrm{E}(AB) = 0$), while B depends on the errors in training data $\mathcal{T}$. In this sense, let's take expectation of Eq.(4) conditional on the test point x_0 and training data $\mathcal{T}$,

$$\begin{aligned}\mathrm{E}(\mathrm{PE}|x_0,\mathcal{T}) &= \mathrm{E}\left((y_0 - x_0^\top\beta)^2|x_0,\mathcal{T}\right) + \mathrm{E}\left((\hat{y}_0 - x_0^\top\beta)^2|x_0,\mathcal{T}\right)\\ &= \mathrm{Var}(\epsilon|\mathcal{T}) + \mathrm{E}\left((x_0^\top\hat{\beta} - x_0^\top\beta)^2|x_0,\mathcal{T}\right)\\ &= \sigma^2 + x_0^\top\mathrm{E}\left((\hat{\beta}-\beta)^2|\mathcal{T}\right)x_0, \end{aligned} \tag{5}$$

in which the second term in RHS of (5) equals to

$$\begin{aligned} x_0^\top\mathrm{E}\left((\hat{\beta}-\beta)^2|\mathcal{T}\right)x_0 &= x_0^\top\mathrm{E}\left(((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\beta)^2|\mathcal{T}\right)x_0\\ &= x_0^\top\mathrm{E}\left(((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon)^2|\mathcal{T}\right)x_0\\ &= x_0^\top\mathrm{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon\epsilon^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}|\mathcal{T}\right)x_0. \end{aligned} \tag{6}$$

Since $\mathrm{E}\left(\epsilon\epsilon^\top|\mathcal{T}\right) = \sigma^2\mathbf{I}_N$, Eq.(6) becomes

$$x_0^\top\mathrm{E}\left((\hat{\beta}-\beta)^2|\mathcal{T}\right)x_0 = \sigma^2 x_0^\top\mathrm{E}_{\mathcal{T}}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]x_0. \tag{7}$$

Finally, substituting Eq.(7) intpo Eq.(5) leads to

$$\mathrm{E}(\mathrm{PE}|x_0,\mathcal{T}) = \sigma^2 + \sigma^2 x_0^\top\mathrm{E}_{\mathcal{T}}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\right]x_0. \tag{8}$$

It is worth noting that the definition of EPE in ESL is a little bit confusing, especially on EPE(f) and EPE(x_0). They are actually different and inconsistent. Therefore, we introduce the expectation of PE in the derivation.

3 On the Difference between Bias and Var

Figure 1 illustrates the the difference between Bias and Var. Hope it is helpful for you to understand the concepts intuitively.