FYS-STK4155 Week 37

Janita Ovidie Sandtrøen Willumsen

(Dated: September 17, 2023)

EXERCISE 1

We have assumed that our data can be described by the continous function $f(\boldsymbol{x})$, and an error term $\boldsymbol{\epsilon}$ $N(0, \sigma^2)$. If we approximate the function with the solution derived from a model $\tilde{\boldsymbol{y}} = X\boldsymbol{\beta}$ the data can be described with $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The expectation value

$$
\begin{aligned}
\mathbb{E}(\boldsymbol{y}) &= \mathbb{E}(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \mathbb{E}(X\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\epsilon}) && \text{where the expected value } \boldsymbol{\epsilon} = 0 \\
\mathbb{E}(y_i) &= \sum_{j=0}^{P-1} X_{i,j}\beta_j && \text{for the each element} \\
&= X_{i,*}\beta_i && \text{where } {}_* \text{ replace the sum over index } i
\end{aligned}
$$

The variance for the element y_i can be found by

$$
\begin{aligned}
\mathbb{V}(y_i) &= \mathbb{E}\big[(y_i - \mathbb{E}(y_i))^2\big] \\
&= \mathbb{E}(y_i^2) - (\mathbb{E}(y_i)^2) \\
&= \mathbb{E}((X_{i,*}\beta_i + \epsilon_i)^2) - (X_{i,*}\beta_i)^2 \\
&= \mathbb{E}((X_{i,*}\beta_i)^2 + 2\epsilon_i X_{i,*}\beta_i + \epsilon^2) - (X_{i,*}\beta_i)^2 \\
&= \mathbb{E}((X_{i,*}\beta_i)^2) + \mathbb{E}(2\epsilon_i X_{i,*}\beta_i) + \mathbb{E}(\epsilon^2) - (X_{i,*}\beta_i)^2 \\
&= (X_{i,*}\beta_i)^2 + \mathbb{E}(\epsilon^2) - (X_{i,*}\beta_i)^2 \\
&= \mathbb{E}(\epsilon^2) = \sigma^2
\end{aligned}
$$

The expression for the optimal parameter

$$
\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}
$$

We find the expected value of $\hat{\boldsymbol{\beta}}$

$$
\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}(\boldsymbol{y}) && \text{using that } \boldsymbol{X} \text{ is a non-stochastic variable} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} && \text{using } \mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}
$$

we can find the variance by

$$
\begin{aligned}
\mathbb{V}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\big[(\hat{\boldsymbol{\beta}} - \mathbb{E}(\hat{\boldsymbol{\beta}}))^2\big] \\
&= \mathbb{E}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T) - \mathbb{E}(\hat{\boldsymbol{\beta}})^2 \\
&= \mathbb{E}(((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y})((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y})^T) - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T \\
&= \mathbb{E}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T + \sigma^2)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T \\
&= \boldsymbol{\beta}\boldsymbol{\beta}^T + \sigma^2((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}
$$

Knowing the expectation value and the variance of $\hat{\boldsymbol{\beta}}$, we can define a confidence intervall for each $\hat{\beta}_j \pm std(\hat{\beta}_j)$ for $j = 1, 2, \ldots, P-1$.

EXERCISE 2

Last week we showed that the optimal $\hat{\boldsymbol{\beta}}^{Ridge}$ can be derived from MSE, and is defined as

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

The expectation value is then

$$
\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}^{Ridge}) &= \mathbb{E}((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}) \\
&= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\mathbb{E}(\boldsymbol{y}) && \text{since } \boldsymbol{X} \text{ and } \lambda\boldsymbol{I} \text{ are non-stochastic variables} \\
&= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} && \text{using } \mathbb{E}(\boldsymbol{y}) \text{ from exercise 1}
\end{aligned}
$$

For $\lambda = 0$ we have $\mathbb{E}(\hat{\boldsymbol{\beta}}^{OLS})$. The variance

$$
\begin{aligned}
\mathbb{V}(\hat{\boldsymbol{\beta}}^{Ridge}) &= \mathbb{E}(\hat{\boldsymbol{\beta}}_R\hat{\boldsymbol{\beta}}_R^T) - (\mathbb{E}(\hat{\boldsymbol{\beta}}_R))^2 \\
&= \mathbb{E}(((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y})(((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}))^T) - (\mathbb{E}(\hat{\boldsymbol{\beta}}_R))^2 \\
&= \mathbb{E}((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1})^T) - (\mathbb{E}(\hat{\boldsymbol{\beta}}_R))^2 \\
&= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\mathbb{E}(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1})^T - (\mathbb{E}(\hat{\boldsymbol{\beta}}_R))^2 \\
&= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T + \sigma^2)\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1})^T - (\mathbb{E}(\hat{\boldsymbol{\beta}}_R))^2 \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1})^T + (\mathbb{E}(\hat{\boldsymbol{\beta}}_R))^2 - (\mathbb{E}(\hat{\boldsymbol{\beta}}_R))^2 \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1})^T
\end{aligned}
$$