In-Class Exercise: Least Squares vs Maximum Likelihood for Gaussian Errors

Statistical Learning, Lecture 1

Consider i.i.d. regression data

$$\begin{pmatrix} Y_1 \\ \boldsymbol{X}_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_N \\ \boldsymbol{X}_N \end{pmatrix},$$

and suppose we use a linear regression model

$$Y = \boldsymbol{X}^\intercal \boldsymbol{\beta} + \epsilon$$

with normally distributed errors $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with variance σ^2. In other words, the conditional probability density of Y given X for parameters $\boldsymbol{\beta}$ is

$$p_{\boldsymbol{\beta}}(Y \mid X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y - \boldsymbol{X}^\intercal \beta)^2}{2\sigma^2}}.$$

Exercise 1. *Show that maximizing the conditional likelihood*

$$\prod_{i=1}^{N} p_{\boldsymbol{\beta}}(Y_i \mid X_i)$$

will lead to exactly the same parameters $\hat{\boldsymbol{\beta}}$ as the least squares criterion, which minimizes the sum of squared errors

$$\sum_{i=1}^{N} (Y_i - \boldsymbol{X}_i^\intercal \boldsymbol{\beta})^2.$$

NB In this course, you are not required to memorize exact derivations (the exam will be open book anyway). The important thing is to remember the conclusions of these derivations.

Hint 1: Use that the maximizer of $\prod_{i=1}^{N} p_{\boldsymbol{\beta}}(Y_i \mid X_i)$ is the same as the maximizer of $\log\Big(\prod_{i=1}^{N} p_{\boldsymbol{\beta}}(Y_i \mid X_i)\Big)$, because log is an increasing function.

Hint 2: Use that $\log(a \times b) = \log a + \log b$, which implies that $\log \prod_{i=1}^{N} p_{\boldsymbol{\beta}}(Y_i \mid X_i) = \sum_{i=1}^{N} \log p_{\boldsymbol{\beta}}(Y_i \mid X_i)$.

Solution to Exercise 1: First some notation: when maximizing a function $f(\boldsymbol{\beta})$, we write $\max_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$ for the maximum value of the function and $\arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$ for the parameters $\hat{\boldsymbol{\beta}}$ where this maximum is achieved. Thus:

$$f(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}} f(\boldsymbol{\beta}).$$

Similarly, $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$ means that $f(\hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$.

Now, to solve the exercise, we start with hints 1 and 2, then plug in the definition of p_β, and finally we simplify the result. The precise argument goes as follows:

$$\begin{aligned}
\arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{N} p_{\boldsymbol{\beta}}(Y_i \mid X_i) &= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \log p_{\boldsymbol{\beta}}(Y_i \mid X_i) \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} -\log p_{\boldsymbol{\beta}}(Y_i \mid X_i) \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} -\log \Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y - \boldsymbol{X}^\intercal \beta)^2}{2\sigma^2}} \Big) \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} -\log \Big(\frac{1}{\sqrt{2\pi\sigma^2}} \Big) - \log \Big(e^{-\frac{(Y - \boldsymbol{X}^\intercal \beta)^2}{2\sigma^2}} \Big) \\
&= \arg\min_{\boldsymbol{\beta}} -N \log \Big(\frac{1}{\sqrt{2\pi\sigma^2}} \Big) + \sum_{i=1}^{N} \frac{(Y - \boldsymbol{X}^\intercal \beta)^2}{2\sigma^2} \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \frac{(Y - \boldsymbol{X}^\intercal \beta)^2}{2\sigma^2} \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (Y - \boldsymbol{X}^\intercal \beta)^2.
\end{aligned}$$

The last two steps as justified, because $-N \log \Big(\frac{1}{\sqrt{2\pi\sigma^2}} \Big)$ does not depend on $\boldsymbol{\beta}$ and therefore does not change where the minimum is achieved. Similarly, the division by $2\sigma^2$ only scales the function, but does not change which parameters achieve the minimum.