Estimation of out-of-sample prediction error in regression

Assumptions about students

When designing this lesson plan, I've assumed that students

- have taken introductory courses in probability, calculus, and linear algebra.
- have taken courses in statistical learning but have mostly focused on inference rather than prediction.

- are comfortable writing code in Python and have used libraries such as `numpy`, `pandas`, `matplotlib`, and `sklearn` in previous courses (e.g., to fit a linear regression model) but have not used them for more advanced tasks such as cross-validation.
- are comfortable with running Python code in a jupyter lab environment on their own.
- are familiar with the matrix notation used in statistical learning.

Learning outcomes

By the end of this lesson, students should be able to:

- Identify the difference between **inference** and **prediction** in regression.
- Understand the **training/test/validation data split**.
- Define out-of-sample prediction error and **explain** its importance.
- Use Python to estimate out-of-sample prediction error via **cross-validation** given training data.
- Understand the **Bias-Variance Trade-Off** and its relationship to prediction error.