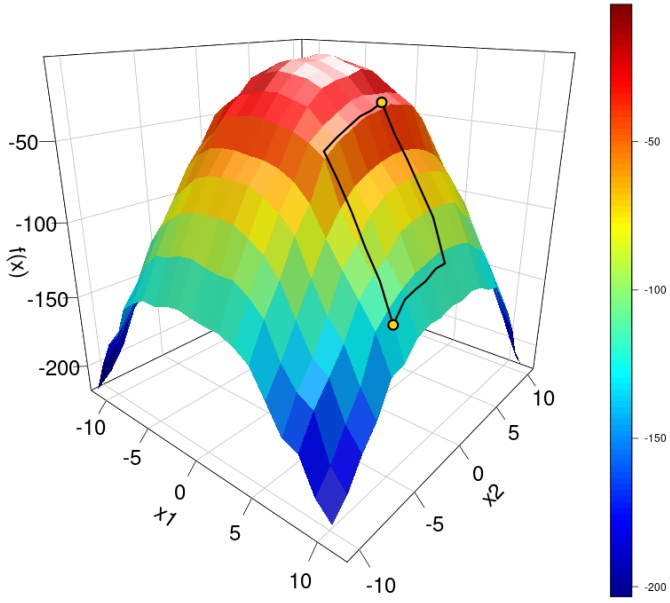
Interpretable Machine Learning

Marginal Effects



Learning goals

- Why parameter-based interpretations are not always possible for parametric models
- How marginal effects can be used in such cases
- Drawbacks of marginal effects
- Model-agnostic applicability

INTERPRETATIONS OF LINEAR MODELS

- The LM can be directly interpreted by evaluating the model coefficients:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \cdots + \epsilon$$

- A change in x_1 by Δx_1 results in a change in y by $\Delta y = \Delta x_1 \cdot \beta_1$.

- Default interpretations correspond to $\Delta x_1 = 1$, i.e., $\Delta y = \beta$.

- All interpretations are done ceteris paribus, i.e., all remaining features are kept constant.

INTERPRETATIONS OF POLYNOMIAL MODELS

- If higher-order terms or interactions are present, parameter-based interpretations are not possible anymore:

$$y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_{1,2} x_1, x_2 + \epsilon \tag{1}$$

- The isolated main effects of both features vary across different values

- The interaction depends on values of the remaining feature

- The marginal effect (ME) allows us to determine a feature effect nonetheless.

MARGINAL EFFECTS

- The most common definition of the marginal effect (ME) corresponds to the derivative of the prediction function w.r.t. a feature. We refer to this variant as the derivative ME (dME):

$$dME_j(x) = \frac{\partial f(x)}{\partial x_j}$$

- A less commonly known definition corresponds to the change in predicted outcome due to an intervention in the data, e.g., by increasing a feature value by one unit. As this variant corresponds to a forward difference, we refer to it as a forward ME (fME):

$$fME_j(x, h_j) = f(x_1, \ldots, x_j + h_j, \ldots, x_p) - f(x)$$

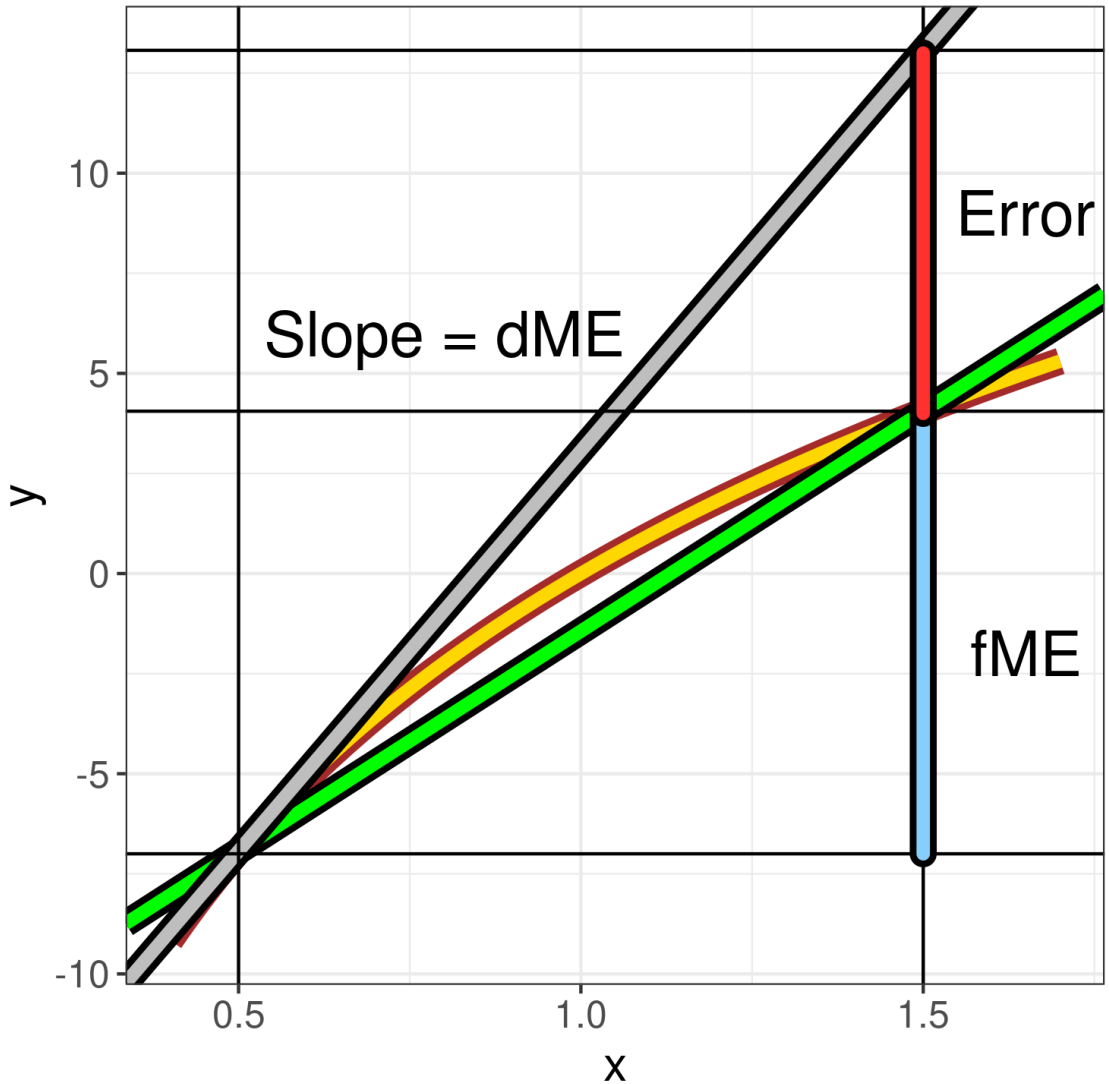
DERIVATIVE VERSUS FORWARD DIFFERENCE

- The dME is not suited to interpret non-linear prediction functions, as the derivative of the prediction function at one point may be substantially different at another.

- The fME is better suited for non-linear prediction functions. It essentially corresponds to a movement on the prediction function, indicating changes in predicted outcome regardless of the function's shape.

- However, with both variants, we lose information about the prediction function along the finite difference.

DERIVATIVE VERSUS FORWARD DIFFERENCE



ADDITIVE RECOVERY

- Due being based on a finite difference, both variants only recover terms within the prediction function that depend on the feature(s) of interest.

- Consider a prediction function $\widehat{f}(x) = ax_1 + bx_2$. It follows that:

$$dME_1(x) = a$$
$$fME_1(x, h_1) = ah_1$$

- The ME removes effects of other features that are linked additively, regardless of their number and effect structure.

MODEL-AGNOSTIC APPLICABILITY