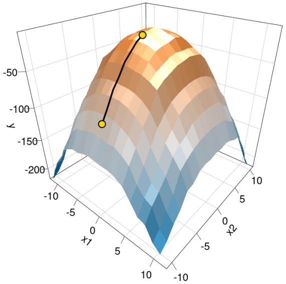
Interpretable Machine Learning

Feature Effects

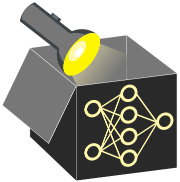
Marginal Effects



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

INTERPRETATION OF SIMPLE MODELS

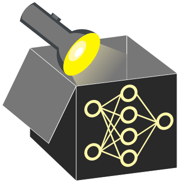


- **Linear Models**:
  - Change in x_j by Δx_j results in change in y by $\Delta y = \Delta x_j \cdot \theta_j$
  - Model equation:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

  - Default interpretations correspond to $\Delta x_j = 1$, i.e., $\Delta y = \theta_j$
  - Assumes "ceteris paribus" (all other features held constant)

INTERPRETATION OF SIMPLE MODELS



- **Linear Models**:
  - Change in x_j by Δx_j results in change in y by $\Delta y = \Delta x_j \cdot \theta_j$
  - Model equation:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \epsilon$$

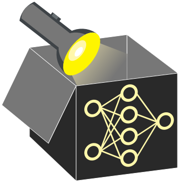
  - Default interpretations correspond to $\Delta x_j = 1$, i.e., $\Delta y = \theta_j$
  - Assumes "ceteris paribus" (all other features held constant)
- **Non-Linear Models with Interactions**:
  - For models with higher-order or interaction terms, single coefficients are not sufficient:

$$y = \theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2 + \theta_{1,2} x_1 x_2 + \epsilon$$

  - Marginal effect of x_1 varies with different values of x_2 (and vice versa)
  - Interactions depend on the values of other features

MARGINAL EFFECTS (ME)

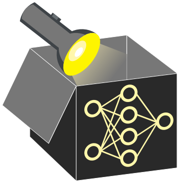
▸ "Bartus" 2005 ▸ "Scholbeck" 2024



- MEs measures prediction changes due to varying *one/several* features.
- **How to compute it?**
  1. **Derivative MEs (dMEs)**: *numeric deriv.* (slope of tangent)
     $\rightsquigarrow$ needs differentiability, fails for step-wise models.
  2. **Forward MEs (fMEs)**: *forward difference* $\hat{f}(\boldsymbol{x} + \boldsymbol{h}) - \hat{f}(\boldsymbol{x})$
     $\rightsquigarrow$ works for *any* model, any feature type.
- **Caveat**: dMEs can mislead when the prediction surface is non-smooth (e.g., decision trees); fMEs remain well-defined (due to finite differences).

MARGINAL EFFECTS (ME)

▸ "Bartus" 2005 ▸ "Scholbeck" 2024



- MEs measures prediction changes due to varying *one/several* features.
- **How to compute it?**
  1. **Derivative MEs (dMEs)**: *numeric deriv.* (slope of tangent)
     $\rightsquigarrow$ needs differentiability, fails for step-wise models.
  2. **Forward MEs (fMEs)**: *forward difference* $\hat{f}(\boldsymbol{x} + \boldsymbol{h}) - \hat{f}(\boldsymbol{x})$
     $\rightsquigarrow$ works for *any* model, any feature type.
- **Caveat**: dMEs can mislead when the prediction surface is non-smooth (e.g., decision trees); fMEs remain well-defined (due to finite differences).
- **Local instantiations (one number per data point)**
  - **ME** (at observed point $\boldsymbol{x}^{(i)}$): Individual, obs.-specific "what-if" effect.
  - **MEM** (at mean $\bar{\boldsymbol{x}}$): Effect at artificial profile ("average obs.").
  - **MER** (at representative value $\boldsymbol{x}^*$): Effect at a user-defined profile.
- **Global summary – Average Marginal Effect (AME)**:
  Expectation of the (d/f)MEs; captures the *global overall* effect.

DERIVATIVE VS. FORWARD DIFFERENCE

dME (tangent, green)

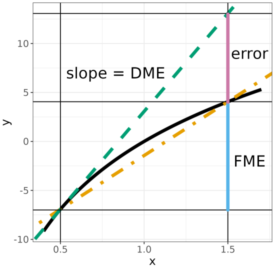
- slope of the tangent at x;
- delivers a *rate* of change $\frac{\partial \widehat{f}}{\partial x}$.

fME (secant, orange)

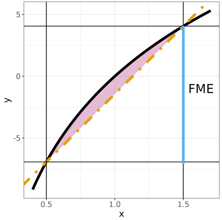
- vertical gap between two model evaluations;
- always *exact* change in predicted outcome.
- Non-linearity measure (pink band, bottom): quantifies deviation of secant and true curve

When the two differ

- Curvature makes the tangent overshoot or undershoot $\Rightarrow$ dME may be badly biased.
- fME is robust to kinks, plateaus, trees, . . .



black = non-lin. function
blue = fME; pink = dME error



DERIVATIVE VS. FORWARD DIFFERENCE

dME (tangent, green)

- slope of the tangent at x;
- delivers a *rate* of change $\frac{\partial \widehat{f}}{\partial x}$.

fME (secant, orange)

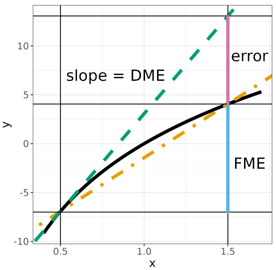
- vertical gap between two model evaluations;
- always *exact* change in predicted outcome.
- Non-linearity measure (pink band, bottom): quantifies deviation of secant and true curve

When the two differ

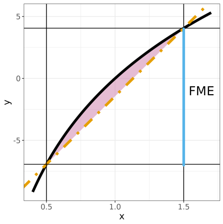
- Curvature makes the tangent overshoot or undershoot $\Rightarrow$ dME may be badly biased.
- fME is robust to kinks, plateaus, trees, . . .

Recommendations

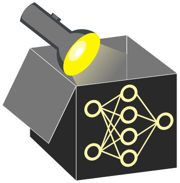
- Use fME for any non-linear / non-smooth model
- Use dME for lin. func.-s or analytic convenience



black = non-lin. function
blue = fME; pink = dME error



ME FOR CONTINUOUS FEATURES



- **Derivative Marginal Effect (dME)**:

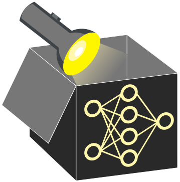
$$\text{dME}_j(\mathbf{x}) = \frac{\partial \hat{f}(\mathbf{x})}{\partial x_j} \approx \frac{\hat{f}(x_1, \ldots, x_j + h_j, \ldots, x_p) - \hat{f}(x_1, \ldots, x_j - h_j, \ldots, x_p)}{2h_j}$$

- **Forward Marginal Effect (fME)**:

$$\text{fME}_j(\mathbf{x}, h_j) = \hat{f}(x_1, \ldots, x_j + h_j, \ldots, x_p) - \hat{f}(\mathbf{x})$$

- **Note:** fME is not scale-invariant – halving the step size does not halve the effect.
- **Additive Recovery:** dME and fME isolate terms involving the target feat.
  - **Example:** For $\widehat{f}(\mathbf{x}) = ax_1 + bx_2$: $\text{dME}_1(\mathbf{x}) = a, \quad \text{fME}_1(\mathbf{x}, h_1) = ah_1$
  - Effects from additively linked features (e.g., x_2) are canceled.
  - Enables focus on direct feature-specific influence in $\widehat{f}$.

ME FOR CATEGORICAL FEATURES

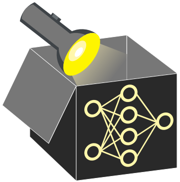


- **Traditional Approach**:
  - Choose a baseline category for the categorical feature x_j
    $\rightsquigarrow$ Either the observed value x_j or a fixed reference x_j^{ref}
  - Replace x_j with an alternative category x_j^{new}
  - Compute the change in prediction, keeping all other feat. $\mathbf{x}_{-j}$ fixed
- **fME Definition for Categorical Features**:

$$\text{fME}_j(\mathbf{x}; x_j^{\text{new}}) = \widehat{f}(x_j^{\text{new}}, \mathbf{x}_{-j}) - \widehat{f}(x_j, \mathbf{x}_{-j})$$

  - x_j: original category of feature j in obs. $\mathbf{x}$ (or reference category x_j^{ref})
  - x_j^{new}: new category to evaluate
  - $\mathbf{x}_{-j}$: all other features held fixed
- **Advantages**:
  - Mirrors continuous feature fME: measures discrete change in pred.
  - Any level can act as baseline - no fixed reference needed.

AVERAGE MARGINAL EFFECTS



Definition (based on fMEs with step h_S, can also be based on dMEs):

$$\text{AME}_S = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(\mathbf{x}_S^{(i)} + \mathbf{h}_S, \mathbf{x}_{-S}^{(i)}) - \hat{f}(\mathbf{x}^{(i)})\right]$$

Why they work in GLMs:

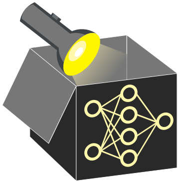
- Link function is monotonic $\Rightarrow$ direction of effect stable.
- Averaging gives sensible results (e.g., logit, probit).

Why they fail on non-parametric models:

- AMEs assume a consistent effect across the feature space.
- Non-parametric models can model complex, non-linear relationships.
- Averaging effects can obscure important heterogeneities.

Takeaway: AMEs can be useful summaries for smooth, monotonic models.
For black-boxes, use **local fMEs** and support them with non-linearity measure.

WHY MARGINAL EFFECTS *STILL* MATTER

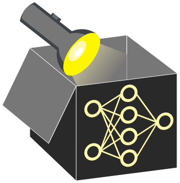


- **Single, formal number:** One scalar per observation; can be averaged (AME), reported with CIs, audited, stored easily.
- **Multivariate changes** Simultaneously perturb multiple *continuous/categ.* feat. Still yields a scalar (unlike PD/ICE, which require multivar. plots).
- **Model-faithful, assumption-light** Measured at the *actual data point*. Captures interactions, no indep. or surrogate-model assumptions (LIME).
- **Non-Linearity Measure:** Quantifies how well local linear approximation holds (e.g., via a normalized squared deviation from the secant).
  $\rightsquigarrow$ Local reliability measure, something PD/ICE plots cannot quantify.
- **Computationally cheap** Just two forward passes (or $k-1$ for a k-level factor) per observation vs. grid$\times n$ for PD/ICE.

Conclusion:

Plots let you *see* the landscape; ME give numbers you can *use*.

USE-CASE: SCALAR VS. VISUAL ESTIMATION



Setting: A clinical model predicts heart attack risk from patient features, e.g., x_1 : systolic blood pressure (BP), x_2 : LDL cholesterol, x_3 : age, ...

Clinician's questions

- *"What if this patient's systolic BP increases by 10 mmHg?"*
- *"What if BP increases by 10 mmHg & LDL by 15 mg/dL?"*

Route A – ICE / PD

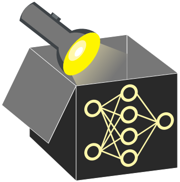
- Plot prediction as a function of BP (1-D) or BP+LDL (2-D) on a grid.
- Manual interpretation of change by looking at curve/surface.

$\rightarrow$ Visual and local; limited to 1–2 features at a time.

Route B – Forward Marginal Effect: $\mathrm{fME} = \widehat{f}(\mathbf{x} + \mathbf{h}) - \widehat{f}(\mathbf{x})$

- **1-D case:** $\mathbf{h} = (10, 0, 0, \dots) \Rightarrow$ risk increases by **+3 % points**
- **2-D case:** $\mathbf{h} = (10, 15, 0, \dots) \Rightarrow$ risk increases by **+4.1 % points**
- One scalar answer per query, extensible to higher dimensions.

RELATION TO ICE AND PD



- **Individual Conditional Expectation (ICE)**:
  - Visualizes predictions for an obs. across a range of feature values.
  - fME corresponds to vertical diff. between points on an ICE curve.
- **Partial Dependence (PD)**:
  - Shows average predictions across a range of feature values.
  - AME is equivalent to vertical differences on PD for linear models.
- **Advantages of fMEs**:
  - Provide exact change in prediction.
  - Applicable to high-dimensional feature changes.
  - Quantifiable and not limited to visual interpretation.