Summary:

This week, we expanded our toolbox for modeling non-linear relationships using basis expansions and smoothers, such as polynomial regression, splines, and generalized additive models (GAMs). These methods provide flexibility beyond traditional linear models while preserving interpretability. We examined the structure and behavior of step functions, piecewise polynomials, and smoothing splines, and learned how GAMs allow additive nonlinear effects for multiple predictors.

Concepts:

- Polynomial Regression: Fits higher-degree polynomial terms of a predictor to model nonlinear trends. Can suffer from poor extrapolation and overfitting at the boundaries.

- Step Functions: Discretize a predictor into intervals, modeling constant behavior within each range. Simple but can lead to abrupt changes between intervals.

- Piecewise Polynomials: Use different polynomial functions across intervals defined by knots. Better than step functions but can still be discontinuous.

- Splines:

    - Linear Splines: Piecewise linear with continuity at knots.

    - Cubic Splines: Piecewise cubic polynomials that are smooth up to second derivatives.

    - Natural Cubic Splines: Like cubic splines but constrained to be linear beyond boundary knots, improving extrapolation.

    - Smoothing Splines: Fit a smooth curve with a penalty on curvature (second derivative), controlled by a smoothing parameter λ\lambdaλ. Avoid knot selection problems.

- Local Regression (LOESS): Fits separate weighted regressions at each point. Effective for capturing local patterns but computationally intensive.

- Generalized Additive Models (GAMs): Flexible models that express the response as a sum of smooth functions of predictors. Can mix linear and nonlinear terms and be used for both regression and classification tasks.

Uncertainties:

While I understand how splines are constructed and applied, I’m still uncertain about how to choose the number and placement of knots in practice—especially when balancing flexibility and overfitting. How do cross-validation and effective degrees of freedom guide these decisions in real-world datasets?