Linear regression is a fundamental tool taught in introductory statistics to build predictive models, but it is limited by its assumption of linear data. Oftentimes, data does have a predictable pattern or shape, that just isn't well described by a linear regression line.

For our project, we looked at methodologies to relax the linearity assumption, and build stronger predictive models.

Linear regression uses calculus to minimize the sum of squared residuals in a model, by selecting optimal intercepts and linear coefficients. Polynomial regression extends this simple idea, by finding optimal coefficients for a polynomial equation. High degree polynomials can be selected, but there is a risk of overfitting. Generally, cross-validation can find an optimal degree.

Alternatively, if different regions of the data have different patterns, step functions can be used. They have basis functions of 1 or 0 times a coefficient, at different knots. Knots can be uniformly assigned or picked individually.

In place of step functions, the basis functions can be different polynomial equations with optimized coefficients. This predicts the data well but creates disjoint connections between equations, which aren't smooth.

Regression splines are basis functions that apply a polynomial (usually cubic) equation to each region divided by knots. Power basis functions are used to ensure the second derivative between two adjacent equations line up, so ensure the line is smooth. The coefficients are optimized to minimize the sum of squared residuals. They are highly flexible, but a lot of decision making is involved in terms of the number of knots to use and where to put them.

Smoothing Splines are equations with a loss penalty methodology. It has a knot at every data point, but a penalty to ensure it is not overfitting. It creates a very smooth line and can be adjusted with a tuning parameter, that can be tweaked to an optimal fit.

MARS is an algorithm that is easy to use, and highly effective for creating models. It systematically goes through the data and adds hinge basis functions to fit the data as well as possible. The hinge is a series of lines that go up and down, and the coefficients are optimized to fit the data as well as possible. This overfits it, so it does a backward prune, and removes as many functions as it can while using generalized cross-validation to create an optimal fit. The result is a series of straight lines, that predicts very strongly.

For our project, we analyzed a dataset describing bike rentals and tried to predict the hourly number of rentals based on a few predictors.

Using cross-validation, we built a generalized additive model, involving two polynomial regressions, and three regression splines. We compared this to a MARS model, and ultimately determined through cross-validation that for this dataset MARS made the best predictive model.