UNIVERSITY of WASHINGTON

Spring Quarter 2021

This paper introduces the concepts that I learned on Nonlinear Regression throughout the SPA DRP program with my Directed Reading Mentor, Michael Pearce. The following documentation demonstrate my understanding and thoughts, including an application of the models learned to the data of COVID-19 in King County, Washington.

Linear and Nonlinear Regression

A linear regression could be interpreted as an approach of predicting a quantitative response *Y* based on the predictor variable *X* where the changes of the response variable can be explained by the constant increase in the predictor variable. The correlation between the variables could be projected linearly with taking into consideration of the error in the predicted values. This relationship could be explained mathematically by the model $Y \approx \beta_0 + \beta_1 X + \epsilon$, where β_0, β_1,and ϵ represent the *intercept, slope,* and *error terms* respectively of the simple linear model. A key indication to use a nonlinear regression model is the distribution of the residuals after fitting a linear model. This could be analyzed in the carrying error of the predicted values. In the case where the residual is not homoscedastic through the data, a nonlinear model could be applied and would be the more suitable approach.

Beyond Linearity

Next, moving on with nonlinear model, the simple approach of nonlinear model is by extending the normal predictor variable to *nth* degree polynomials $x_i, x_i^2, x_i^3, \ldots., x_i^d$, this approach is called **Polynomial Regression**. This allows us to produce an extremely non-linear curve with no limitation of the degree polynomials to fit a wide range of curvature. However, disadvantages of the model with higher degree of predictor are that outliers in the predictors data which could affect the results of nonlinear analysis and that incompatible polynomial degrees could present bias and overfitting.

In the case where our data have different regions and each have different patterns, **Step Functions** can provide a better analysis with the methods of breaking the range of variable into regions and fit a different constant in each region. This will convert the continuous variable into an ordered categorical variable. The next approach is called **Basis Function**, the idea involves applying a family of functions or transformation towards variable X. This method can be interpreted as a framework as it may consist of different polynomial equations with optimized coefficients. As a result, the predicted data will create a disjoint connection between different equations, and thus be not smooth.

Regression Splines is a flexible class of basis function with a similar approach where each bin or region is divided by knots and fit a function (called piecewise function) is fitted in each bin. There are various models in the family of Splines with same basic approach such as **Cubic Splines, Natural Splines,** and **Smoothing Splines**. The difference between the models are the constraints that are applied within them. For example, cubic splines use piecewise third-order polynomial with a set of constraints taking into account the continuity, continuity of first derivative, and continuity of second derivative. Commonly, the second derivative is set to zero at the endpoints to provide a boundary condition, while the natural cubic spline with same approach extrapolates linearly beyond the boundary knots.

There are many decisions to be considered in order to find an optimal fit such as tuning parameter, numbers of knots, location of knots and optimal splines model. One of the algorithms that I learned which is very useful and highly effective for creating a model is **Multivariate Adaptive Regression Splines, MARS**. In general, MARS analyze the data and uses hinge function to add related basis functions to fits the data. The hinge functions automatically divide the input data so that the effect of outliers is contained. Since it will be overfitting, MARS will do a backward pass to eliminate the unnecessary functions and repeat the process until it reaches a predefined limit of terms. MARS generalized the model by removing the terms based of the generalized cross validation. This will produce an optimal fit model that predicts well.

The last model that I learned is **Generalized Addictive Models (GAM**). As stated in the name, GAM is a general framework derived from the standard linear model that allows for non-linear relationship between each feature and response by replacing the linear components with a nonlinear function, all while maintaining additivity. This method will automatically produce a non-linear relationship model which we would not need to manually try out different transformation on each variable. Besides, the additivity factor will allow us to examine the effect of predictor variable on Y. However, the additivity's main limitation is that we could miss out on important interactions as many variables would be used.

Project

For my project, I applied various nonlinear regression models towards the data of COVID-19 in King County, Washington. The data was derived from the official government website of King County, ("COVID-19 data dashboard - King County", 2021). The main objective is to present the best compatible model to project the response variable of the weekly rate of positive cases based on various predictor variables such as weekly number of tests, hospitalizations, and deaths over time. The data was divided into several time point, and they were fitted with different model. This process was repeated until the whole dataset was covered while we record the residuals. I concluded that the multiple regression model is the most compatible with least value of RSS