Winter Quarter 2020-2021

N.B.: The following documentation, explanations, and illustrations are used to keep progress of the ongoing research project on non linear regression my Directed Reading Mentor Michael Pierce. This paper summarizes many of integral ideas in linear regression and it's applications to leukemia incidence rates in females.

Summary

To understand the fundamental concepts of nonlinear regression, a retracing of linear regression and its assumptions is very helpful. For example, during the beginning weeks of the guided reading, we defined the assumptions like linearity, additivity, and normality in our own words to define what constituents a model that isnât linear. The largest change from nonlinear regression and linear regression is in the assumption of linearity: that one variableâs variance (the response variable) can be explained by the constant increase of another variable (a predictor variable). The significance lies in the fact that, in certain intervals of the data, the relationship between the two can be explained linearly, but afterward, its predicted values carry increasing error. In running diagnostic tests of the two numerical variables, if the errors are not uniform in their distribution, then this indicates applying a nonlinear model. One key component Michael Pierce and I talked about through the program is about making these distinctions to use linear vs nonlinear, as linear regression is very popular for a reason, data likes to be normally distributed.

Now that the distinctions between when to use the models were discussed, Michael and I read and discussed the various nonlinear models there are the differences they bring. The main idea that was highlighted in our discussion centered around extending the linear regression theory to the nonlinear field. The primary example of this was with a model called the Polynomial Regression. This regression takes a regular predictor from linear regression and raises its power to a higher degree than 1. For example, taking the predictor variable Year to predict the rate of cancer incidence of females to the second power (year2). There are no limits to the degree you bring the predictor, but even at a degree of 5, overfitting begins to make the model unpredictable with new data, a key reason for these models.

The next set of nonlinear regressions bind together as a family all stemming from a similar theory and idea called splines: partition the predictor variable into separate sections(indicated as Knots), wherein each, we fit a model a relationship. This connects nears the ideas of integral calculus and piecewise functions. In addition, this essential idea can be used as a focal point to derived models like the Truncated Power Bases Functions, Spline, Natural Cubic Spline, Smoothing Spline, M.A.R.S. The names vary, the models that make up the family of splines mainly differ from each other just by adding additional constraints. For example, the Spline and the Natural Cubic Spline different in that the Natural spline constraints the curved line to be straight at all its knot points and boundaries, making it much more predictable. And even further, the smoothing spline adds an additional constraint to this model such that it punished fast changes of steepness in the curve, essential âsmoothing out the jaggedness of a curveâ.

The main motivation of the many nonlinear models learned this quarter was to apply into the leukemia incidence data in females briefly mentioned before. Took from the National Institute of Health, and National Cancer Institute, I ran an exploratory data analysis on year and age as covariates to predict incidence rates. After validating through ANOVA forwards selection for the polynomial regressions and general cross-validation for the splines, the knot points converged on an interval between 1985 to 1986 where the incidence rate of change went from horizontal to polynomial or exponential cancer indicate rate growth. With these preliminary results, further medical research into this time frame for causation and inferences will need to be investigated and conducted.