Moving Beyond Linearity

Sravani Vadlamani

Linear Models

- Simple to describe and implement
- Easy to interpret and infer in comparison to other methods
- Limited in terms of predictive power as the assumption of linearity is always an approximation and may not always hold
- Linear models can be improved only so far by using ridge, lasso and PCR techniques
- Need methods that relax the linearity assumption but maintain the interpretability

Extensions to Linear Models

- Polynomial regression
  - Add extra predictors obtained by raising each of the original predictors to a power
- Step Functions
  - Split a continuous variable into k distinct regions to produce a qualitative variable. This has the effect of fitting a piecewise function
- Regression Splines
  - Extension of polynomial regression and step function and are more flexible
  - Divide the range of X into k distinct regions and fit a polynomial function within each region
  - The polynomials are constrained so that they join smoothly at the region boundaries called knots
  - When there are sufficient regions, the splines can result in an extremely flexible fit

Extensions to Linear Models

- Smoothing splines
  - Similar to regression splines but they minimize a residual sum of squares criterion subject to a smoothness penalty
- Local regression
  - Similar to smoothing splines but the regions are allowed to overlap in a smooth way
- Generalized Additive Models
  - Extend above methods to deal with multiple predictors

Polynomial Regression

- Extend linear regression to accommodate non-linear relation between the predictors and the response variable
- Standard Linear Model $y = \beta_0 + \beta_1 x_1 + \varepsilon$
- Polynomial Model $y = \beta_0 + \beta_1 x_1 + \beta_1 x_1^2 + \beta_1 x_1^3 + \ldots\ldots + \beta_1 x_1^d + \varepsilon$
- Large values of d will produce an extremely non-linear curve
- Not usual to use d greater than 3 or 4 as the polynomial curve becomes overly flexible and may take on strange shapes
- The higher order predictors are transformations of the original predictor x_1 and the coefficients for the polynomial model can be estimated using least squares

Polynomial Regression

- The interpretation of the regression coefficients is difficult and not important in comparison to the overall fit of the model.
- The key is to be able to get a better prediction that will be useful

Step Function

- Polynomial functions impose a global structure on the non-linear function of X. Step functions can be used to avoid imposing a global structure
- Step functions split the range of X into bins and fit a different constant to each bin. This translates to converting a continuous variable into an ordered categorical variable.
- Create k cut points c_1, c_2, …. c_k in the range of X and construct k + 1 categorical variables

Step Function

- Create k cut points c_1, c_2, …. c_k in the range of X and construct k + 1 categorical variables

$$C_0\ (X) = I(X < c_1)$$

$$C_1\ (X) = I(c_1 < X < c_2)$$

$$C_2\ (X) = I(c_2 < X < c_3)$$

.

.

$$.\ C_{k-1}\ (X) = I(c_{k-1} < X < c_k)$$

$$C_k\ (X) = I(c_k \leq X)$$

I is an indicator function that returns 1 if the condition is true

Step Function

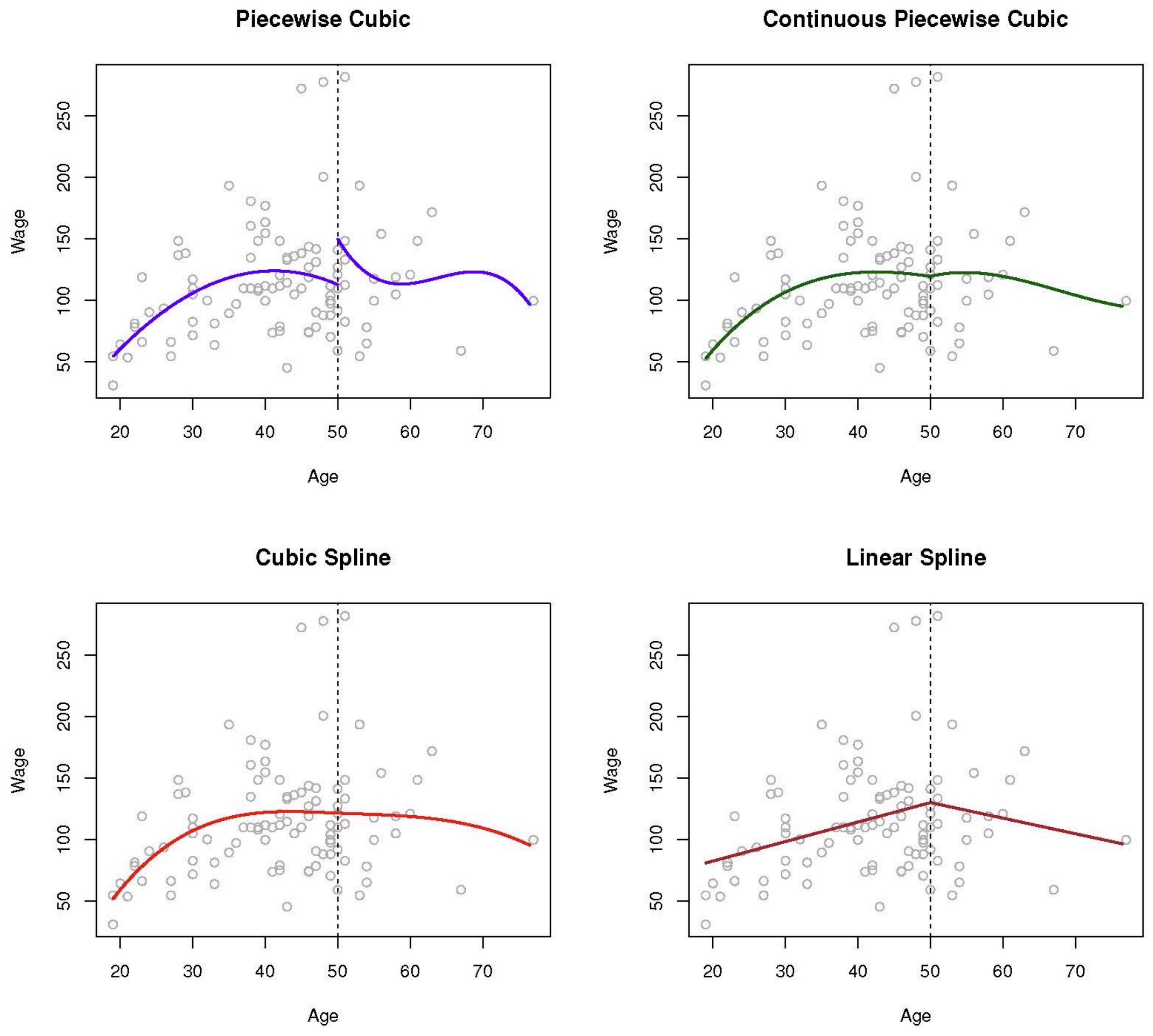
- X must be in exactly one of the K+1 intervals and hence

$$C_0\ (X) + C_1\ (X) + C_2\ (X) + \ldots\ldots.. C_k\ (X) = 1$$

- Once the categorical variables have been created, a linear model is fit using these categorical variables as predictors
- For a given value of X, only one of c_1, c_2, …. c_k can be zero. When X < c_1, all the predictors are zero and hence β_0 is interpreted as the mean value of y.
- Unless there are natural break points in the predictors, step functions can miss interesting trends in the data

Splines

- Types
  - Regression splines
  - Smoothing splines
- Give superior results to polynomial regression
- Useful when the function seems to be changing rapidly



Generalized Additive Models

- Extend simple linear regression models to multiple linear regression i.e., allow more than one predictor
- Provide a general framework to extend a standard linear model by allowing non-linear functions of each of the variables while maintaining additivity. GAMs can be applied to both quantitative and qualitative responses.

Pros and Cons of GAMs

- Allow fitting non-linear functions for each of the predictor variables simultaneously. We need not manually try the different transformations on each variable individually
- The non-linear fits can help in making more accurate predictions for the response variable
- Since the model is additive, the effect of each X on Y can be examined individually while holding all the other variables fixed.
- Since GAMs are additive models, the biggest limitation is that important interactions among variables are not accounted for.