Sixth essay Raul Adell

This essay may contain ideas given in class complemented by the book given on the references of the course as well as content from Stanford's online ITSL course given by R. Tibshirani and T. Hastie. Book section: 7. Moving beyon linearity. Along with reading the book chapter, and contrasting it with online videos and class notes, this is what I considered most important.

Linear models can be very powerful tools for prediction and inference, as seen in past essays. However, in many cases, these models may be too restrictive to capture the complexity of the underlying data. This can lead to underfitting, where the model fails to capture the true relationship between the predictors and the response, or overfitting, where the model fits the noise in the data and performs poorly on new data.

The authors note that there are several ways to address these limitations, such as by including interaction terms or polynomial terms in the model. However, these approaches can quickly become unwieldy as the number of predictors increases. As a result, they introduce the idea of using more flexible models, such as generalized additive models (GAMs), which can capture nonlinear relationships between the predictors and response without requiring complex interactions or polynomials.

One should note that many of these more flexible models are "black boxes," meaning that it can be difficult to interpret how they arrive at their predictions. In this essay we are going to review some strategies to relax the lienearity assumption while still attempting to maintain as much interpretability as possible. The main strategies are mentioned below:

- Polynomial regression: In polynomial regression, a polynomial function is fit to the data in order to capture nonlinear relationships between the predictors and response. This approach can be useful when the relationship is not linear, but can become cumbersome as the degree of the polynomial increases.
- Step function: A step function is a piecewise constant function that approximates the relationship between the predictors and response by dividing the range of the predictors into intervals and assigning a constant value to each interval. This approach can be useful when the relationship is not continuous.
- Regression splines: A regression spline is a piecewise polynomial function that approximates the relationship between the predictors and response by dividing the range of the predictors into intervals and fitting a separate polynomial function to each interval. This approach is more flexible than polynomial regression and step functions, but can still become cumbersome as the number of intervals increases.
- Smoothing splines: A smoothing spline is a curve that minimizes the sum of squared residuals subject to a smoothness penalty. This approach can be useful when the relationship between the predictors and response is not well approximated by a polynomial or piecewise function.
- Local regression: Local regression is a nonparametric approach to regression that fits a separate linear regression model to each observation in the data, with the weights assigned based on the distance from the observation to the predictor values of interest. This approach can be useful when the relationship between the predictors and response varies across the range of the predictors.

- Generalized additive models: Generalized additive models (GAMs) are a flexible class of models that combine multiple smooth functions of the predictors to approximate the relationship between the predictors and response. This approach can capture nonlinear relationships and interactions between the predictors without requiring complex interactions or polynomials.

Lets go one by one in order to have a good understanding of each of the approaches.

Polynomial regression is a type of regression analysis in which a polynomial function is fit to the data in order to capture nonlinear relationships between the predictors and response. The basic idea is to expand the linear model with polynomial terms. For example, a quadratic polynomial model would include a term for the square of the predictor variable, while a cubic polynomial model would include terms for the square and cube of the predictor variable.

The polynomial regression model can be expressed mathematically as:

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_m x^m + \varepsilon$

where y is the response variable, x is the predictor variable, β_0, β_1, ..., β_m are the coefficients of the model, m is the degree of the polynomial, and ε is the error term.

The coefficients can be estimated using least squares regression, which involves finding the values of β_0, β_1, ..., β_m that minimize the sum of squared residuals:

$SSR(\beta) = \Sigma_i (y_i - f(x_i, \beta))^2$

where SSR(β) is the sum of squared residuals, y_i is the observed response value for the ith observation, $f(x_i, \beta)$ is the predicted response value for the ith observation based on the polynomial model, and β is the vector of polynomial coefficients.

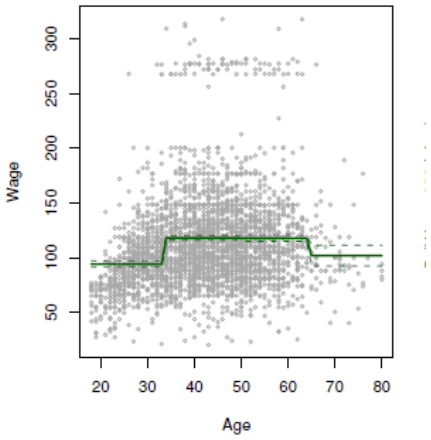
One potential issue with polynomial regression is that as the degree of the polynomial increases, the model becomes increasingly complex and may overfit the data. This can lead to poor performance on new, unseen data. To address this issue, it is often necessary to choose an appropriate degree for the polynomial based on the data. Anova tests are specially useful for this purpose.

In addition to choosing an appropriate degree for the polynomial, it may also be necessary to consider other issues such as multicollinearity, outliers, and nonconstant variance of the error term. Overall, while polynomial regression can be a useful tool for capturing nonlinear relationships between the predictors and response, it is important to carefully consider the limitations and potential issues associated with this approach.

In this section leverage is also described, together with the confidence intervals for regression. Leverage measures how far away an observation is from the average of the predictor variables, and how much influence it has on the estimated regression coefficients. Observations with high leverage can have a large impact on the estimated regression line, particularly if they are outliers or have extreme values on one or more of the predictor

variables. That's why tails are bad for extrapolation because they tend to have observations that are further away from the center of the predictor variable distribution, which means that these observations can have higher leverage and influence on the estimated regression line. When there is scarce information at the end of the domain, the standard deviation tends to get wider, which means that the model may be less accurate when making predictions outside the range of the predictor variables.

A **step function** is a function that takes a constant value on a set of intervals, or steps, and changes abruptly from one constant to another at the boundaries between steps. Notice that using polynomial functions of features in a linear model imposes a global structure on the non-linear function of X. With step functions we make a more local approach because a point only affects the fit in the region it is. This approach can be useful when there is evidence to suggest that the relationship between a predictor variable and the response variable is not linear, but rather changes abruptly at certain thresholds or cutpoints.



The step function approach involves dividing the range of a predictor variable into a set of intervals, or bins, and fitting a separate constant value to the response variable within each bin. The number and size of the bins can be chosen based on prior knowledge or empirical evidence, or they can be selected using a search algorithm that seeks to minimize the residual sum of squares of the resulting model.

One important consideration when using step functions is the choice of bin boundaries. If the boundaries are chosen poorly, then the resulting step function may not accurately capture the true underlying relationship between the predictor and response variables. One way to address this issue is to use cross-validation techniques to select the optimal number and placement of bins.

Another limitation of the step function approach is that it may be too rigid for some applications, as it assumes that the relationship between the predictor and response variables is piecewise-constant. In cases where the true relationship is more complex, or changes continuously across the range of the predictor variable, other methods such as regression splines or smoothing splines may be more appropriate.

Basis functions involve transforming the original predictor variable into a set of new variables, which are then used as inputs to a linear regression model.

One common approach to using basis functions is to use polynomials. For example, a second-order polynomial can be used by transforming the predictor variable x into x^2, and a third-order polynomial can be used by transforming x into x^2 and x^3. By adding these transformed variables as additional predictors in a linear regression model, we can capture non-linear relationships between the predictor and response variables.

Another approach to using basis functions is to use a set of pre-specified functions, such as sine and cosine functions or Gaussian functions. These functions are often referred to as radial basis functions. By choosing an appropriate set of basis functions, we can model complex non-linear relationships between the predictor and response variables.

One advantage of using basis functions is that they provide a flexible approach to modeling non-linear relationships without requiring the user to specify the form of the relationship a priori. However, an important consideration is the choice of basis functions and the order of the polynomial, as overfitting can occur if too many basis functions are used.

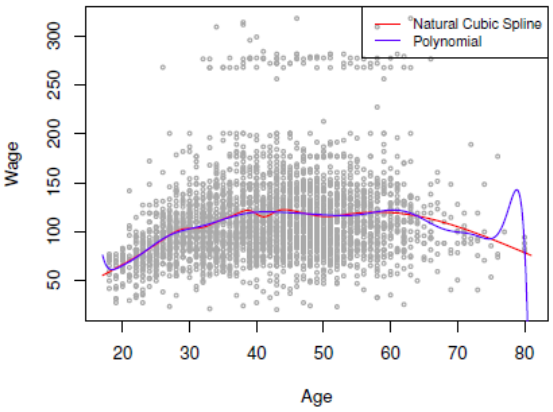
Piecewise polynomials are polynomials of different degrees that are fitted to different regions of the predictor variable. For example, we could fit a second-order polynomial to the first region of the predictor variable and a linear function to the second region. This results in a continuous function that changes polynomial degree at certain points, known as knots.

The most common type of piecewise polynomial used is a cubic polynomial. A cubic polynomial is a polynomial of degree 3, and has the form:

$f(x) = \beta0 + \beta1x + \beta2x^2 + \beta3x^3$

To fit a piecewise cubic polynomial, we first divide the predictor variable into a set of regions or intervals, each with its own set of coefficients β0, β1, β2, and β3. We then estimate the values of these coefficients by minimizing the residual sum of squares of the resulting model, subject to the constraint that the function is continuous and has continuous first and second derivatives at the knots.

Unlike polynomial regression, cubic splines provide a more flexible way of modeling non-linear relationships between the predictor and response variables, as they can approximate a wide range of curve shapes without the risk of overfitting that high-degree polynomials can have. One reason cubic splines make a good fit is that they are able to capture non-linear relationships without requiring a large number of parameters, which can reduce the risk of overfitting. Cubic splines are also continuous and smooth, which can help to avoid problems with discontinuities or sharp changes in the fitted function.



The number of degrees of freedom in a cubic spline depends on the number of knots used to divide the predictor variable into different regions. If we have K knots, then we have K+4 degrees of freedom, as we need 4 coefficients to define each cubic polynomial and there are K intervals between the knots. The choice of the number and location of knots can be critical to the performance of the cubic spline model. Too few knots can result in an oversimplified model that does not capture the true underlying relationship between the predictor and response variables, while too many knots can lead to overfitting.

For a given set of K knots t_1, t_2, ..., t_K, let B_1(x), B_2(x), ..., B_{K+3}(x) be a set of basis functions, where:

$$
\begin{aligned}
B_1(x) &= 1 \\
B_2(x) &= x \\
B_3(x) &= x^2 \\
B_4(x) &= x^3 \\
B_{k+3}(x) &= (x - t_k)^3_+ \text{ for } k = 1, 2, ..., K
\end{aligned}
$$

The truncated power basis function has the property that it imposes continuity of the spline function and its first k-1 derivatives at the knots.

$$
(x - \xi_j)^k_+ = \begin{cases} (x - \xi_j)^k & \text{if } x > \xi_j \\ 0 & \text{otherwise} \end{cases}
$$

Using these basis functions, any cubic spline function can be represented as a linear combination of the basis functions with coefficients c_1, c_2, ..., c_{K+3}:

$$
f(x) = c_1 B_1(x) + c_2 B_2(x) + c_3 B_3(x) + c_4 B_4(x) + \sum_{k=1}^{K} c_{k+3} B_{k+3}(x)
$$

The question left to answer is how many and where to place this knots. One approach to selecting the number and placement of knots is to use cross-validation techniques. We can try different numbers of knots and select the number that results in the lowest cross-validation error. Another approach is to use a search algorithm, such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC), which seek to balance the goodness of fit of the model with its complexity.

One limitation of piecewise polynomials is that they can be sensitive to the choice of knot locations. If the knots are not placed correctly, then the resulting model may not accurately capture the true underlying relationship between the predictor and response variables. However, when used appropriately, piecewise polynomials can provide a flexible approach to modeling non-linear relationships that may not be captured by simpler methods such as linear or polynomial regression.

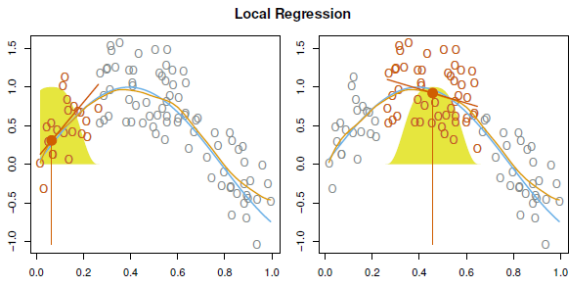
In order to avoid the knot-selection issue, this following technique reduces the problem into the selection of a single parameter. The goal of **smoothing splines** is to find a function f(x) that fits the data points (x_1, y_1), (x_2, y_2), …, (x_n, y_n), while minimizing a penalty term that controls the smoothness of the function. The smoothness penalty is controlled by a tuning parameter λ, similarly that the approach taken in optimal control theory or regularization.

Mathematically, the objective function for smoothing splines is:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 \, dx$$

The first term in the objective function measures the fit of the function to the data points, while the second term measures the smoothness of the function. The tuning parameter λ controls the amount of smoothing in the spline function. A smaller value of λ will result in a more flexible spline function that fits the data more closely, while a larger value of λ will result in a smoother spline function that is less affected by noise in the data.

Unlike other regression techniques, **local regression** focuses on fitting a model to a subset of data points around a given point of interest, rather than trying to fit a single global model to the entire data set.The key idea behind local regression is to weight each data point based on its proximity to the point of interest. The closer a data point is to the point of interest, the greater the weight assigned to it. This is accomplished using a kernel function, which assigns weights to each data point based on its distance from the point of interest.



The local regression model is then fitted by minimizing the following objective function:

$$\text{minimize } \sum_{i=1}^{n} K\left(x_0, x_i\right)\left(y_i - f(x_i)\right)^2 + \lambda \int f''(t)^2\, dt$$

where y_i is the response variable associated with the i-th data point, f(x_i) is the predicted value of the response variable at x_i, and λ is a tuning parameter that controls the smoothness of the fitted curve. The second term in the objective function is a penalty term that penalizes large values of the second derivative of the fitted curve, which helps to ensure a smooth fit.

The local regression model is fit separately for each point of interest, resulting in a curve that varies as a function of the predictor variable. It is sometimes referred to as a memory-based procedure, because like nearest-neighbors, we need all the training data each time we wish to compute a prediction.The final fitted curve is obtained by connecting the local fits, resulting in a smooth curve that captures the underlying trend in the data.

Local regression is a flexible and powerful method for fitting curves to data, and is particularly useful in cases where the relationship between the predictor and response variables is complex or non-linear. However, it can be computationally intensive and requires careful selection of the kernel function and tuning parameters. The most important one is the span s, the proportion of points used to compute the local regression at each x_i . The smaller the span, the more wiggly and local the result of the fit. Once again, cross-validation can be used to select the best span parameter.

Finally, we describe the most general method. **GAMs** are an extension of linear regression models that allow for the modeling of non-linear relationships between variables. In a GAM, the relationship between the response variable Y and the predictor variables X1, X2, ..., Xp is modeled as:

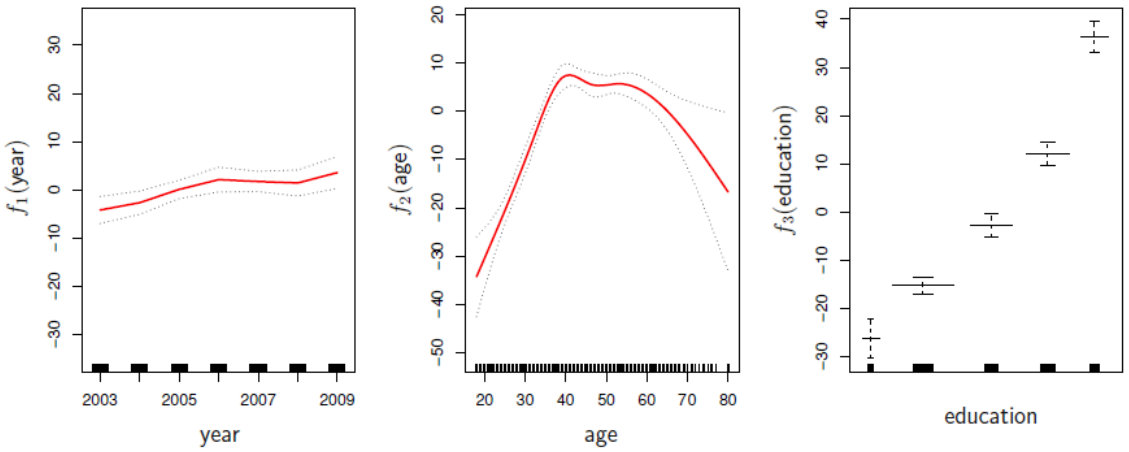
$$Y = f_1(X_1) + f_2(X_2) + ... + f_p(X_p) + \epsilon$$

where ach function f_j is fit independently using a nonparametric regression method such as smoothing splines or local regression.

The GAM can be fitted using the following steps:

1. Choose a family of distributions for the response variable Y, such as normal, binomial, or Poisson.
2. Specify the form of each smooth function f_j(X_j) using a set of basis functions.

3. Use maximum likelihood or another optimization method to estimate the parameters of the model, including the smoothing parameters that control the complexity of the smooth functions.



The first two functions are natural splines in year and age, with four and five degrees of freedom, respectively. The third function is a step function, fit to the qualitative variable education.

The GAM can be used to model complex relationships between the response and predictor variables, while still allowing for straightforward interpretation of the effect of each predictor variable. The smooth functions f_j can reveal the shape of the relationship between each predictor and the response, and can capture non-linearities, interactions, and other complex patterns. Additionally, the GAM allows for the incorporation of both continuous and categorical predictor variables.

GAMs can be seen as a compromise between linear models, which are often too restrictive, and fully nonparametric models, which can be too complex and difficult to interpret. They offer a flexible and interpretable way to model complex relationships in data. Also the model is restricted to be additive, meaning that the effects of each predictor variable on the response variable are independent of the values of the other predictor variables. This assumption limits the ability of the model to capture interactions between variables.

Fitting a Generalized Additive Model (GAM) involves estimating the smooth functions of the predictor variables that best predict the response variable. This is not as easy as fitting a single spline. Tne popular method for estimating these smooth functions is backfitting, which iteratively updates the smooth functions while holding the others constant. The backfitting algorithm starts by initializing the smooth functions to some initial value, typically a linear relationship between the predictor variable and the response variable. Then, the algorithm iteratively updates the smooth functions for each predictor variable by minimizing a penalized residual sum of squares. This minimization involves finding the optimal smoothing parameter for each smooth function, which balances the fit to the data with the smoothness of the function.

GAMs can be used in regression and in classification problems.