Chapter 7 Notes: Moving Beyond Linearity

Statistical Learning with R

Jon Geiger

March 24, 2022

All of the regression techniques used thus far have been linear. Starting with Least Squares regression, we moved onto improvements such as PCR, PLS, Ridge Regression, and the Lasso, but all of these techniques are still linear by nature. This chapter will introduce the following statistical learning methods to help us move beyond linearity:

1. **Polynomial Regression**: extends the linear model by adding extra predictors, the original predictor raised to powers.
2. **Step Functions**: cuts the range of a variable into distinct regions to produce a qualitative variable.
3. **Regression Splines**: divides the range of the predictor into distinct regions, and within each region, a polynomial function is fit. This provides an extremely flexible fit.
4. **Smoothing Splines**: a different formulation of regression splines, fit by minimizing an RSS criterion.
5. **Local Regression**: similar to splines, the predictor is separated into regions, but these regions can overlap.
6. **Generalized Additive Models (GAMs)**: extends the methods above to allow for multiple predictors.

7.1: Polynomial Regression

While the standard linear model takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

the polynomial function takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i,$$

where ϵ_i is the error term for a particular data point x_i, and d is an integer value representing the maximum polynomial degree. These coefficients can be estimated very easily using simple least squares regression, as y is still a linear combination of various predictors. Typically, d will be, at most, 3 or 4.

Let's say we fit a degree-four polynomial to a set of data, and we look at the fit at a certain value of x_0:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

What is the variance of this fit? Least squares gives us the variance for each coefficient $\hat{\beta}_j$, which we can use to estimate the variance of $\hat{f}(x_0)$, and the standard error at this point (pointwise standard error) is the square root of this variance. Assuming normal errors, plotting twice the standard error along with the polynomial fit gives an approximate 95% confidence interval. It turns out that these standard errors blow up as we get to either extreme of the x variable.

7.2: Step Functions

One issue with polynomial regression is that it imposes a certain *global* structure onto the non-linear function of X. Rather than imposing a global structure, we can use *step functions* to break X into *bins* and fit a *constant value* within each bin. This converts a continuous variable into an *ordered categorical* variable.

We create cutpoints $c_1, c_2, \ldots, c_K$ in the range of X, then construct $K+1$ new variables:

$$\begin{aligned} C_0(X) &= I(X < c_1) \\ C_1(X) &= I(c_1 \leq X < c_2) \\ C_2(X) &= I(c_2 \leq X < c_3) \\ &\vdots \\ C_{K-1}(X) &= I(c_{K-1} \leq X < c_K) \\ C_K(X) &= I(X \geq c_K) \end{aligned}$$

where I is an indicator function returning 1 if the condition is true, and 0 if it's false. We can notice that since an individual observation x_i can only be in one of these $K+1$ intervals, it must be the case that $C_0(X) + C_1(X) + \cdots + C_K(X) = 1$. We can use least squares to fit a linear model of these step functions with $C_1(X), C_2(X), \ldots, C_K(X)$ as predictors:

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \epsilon_i.$$

We can notice that when all of the C predictors are zero (when x_i falls in the range less than c_1), the intercept β_0 can be interpreted as the mean of Y for values where $X \leq c_1$. By contrast, when X falls into a different range, the equation will predict a response of $\beta_0 + \beta_j$ for $c_j \leq X < c_{j+1}$, so β_j woudl represent the average increase in the response for X relative to β_0 (the mean response for values where $X \leq c_1$).

Unless there are natural breakpoints in the data, piecewise-constant functions can miss the action. These sorts of functions are used fairly often in biostatistics and epidemiology, as well as other disciplines.

7.3: Basis Functions

7.4: Regression Splines

7.5: Smoothing Splines

7.6: Local Regression

7.7: Generalized Additive Models