7.2 Step Functions

Zongyi Liu

2023-06-02

7.2 Step Functions

Using polynomial functions in a linear model imposes a **global** structure on the non-linear function of X. To avoid making such a global structure, we can use step functions, which breaks the range of X into **bins**, and fit a different constant in each bin. This can convert a continuous variable into ordered categorical variables.

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

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

Here I is an **indicator function** that returns a 1 if the condition is true, and 0 if the condition is false.

For example, $I(c_K \leq X)$ equals 1 if $c_K \leq X$, and equals 0 otherwise.

These are sometimes called dummy variables.

For any value of X, $C_0(X) + C_1(X) + ... + C_K(X) = 1$, since X must be in exactly one of the $K+1$ intervals. We then use least squares to fit a linear model using $C_1(X), C_2(X), ..., C_K(X)$ as predictors:

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

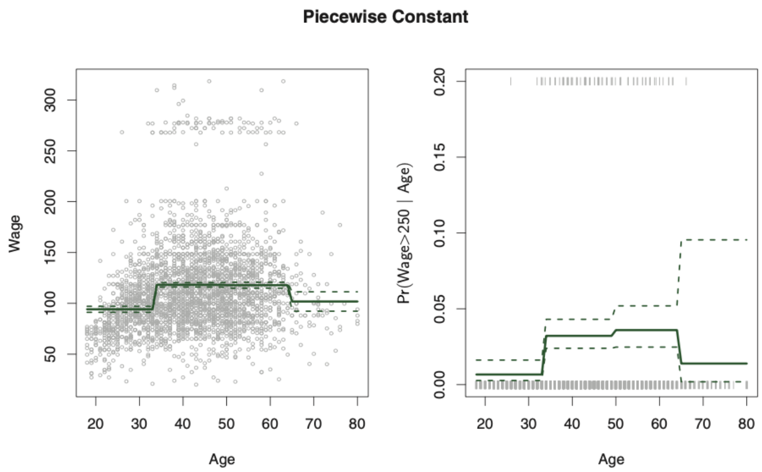
For a given value of X, at most one of C_1, C_2, …, C_K can be non-zero.

When $X < c_1$, all of the predictors in the equation above are zero, so β_0 can be interpreted as the mean value of Y for $X < c_1$.

We can fit the logistic regression model

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \ldots + \beta_K C_K(x_i))}$$

Unless there are natural breakpoints in the predictors, piecewise-constant functions might miss the action.



Left: The solid curve displays the fitted value from a least squares regression of `wage` (in thousands of dollars) using step functions of age. The dotted curves indicate an estimated 95 % confidence interval.

Right: We model the binary event `wage`>250 using logistic regression, again using step functions of age. The fitted posterior probability of `wage` exceeding $250,000 is shown, along with an estimated 95 % confidence interval.