```
> table(cut(age, 4))

(17.9,33.5]   (33.5,49]   (49,64.5] (64.5,80.1]
        750        1399         779          72
> fit <- lm(wage ~ cut(age, 4), data = Wage)
> coef(summary(fit))
                        Estimate Std. Error   t value     Pr(>|t|)
(Intercept)            94.158392   1.476069 63.789970 0.000000e+00
cut(age, 4)(33.5,49]   24.053491   1.829431 13.148074 1.982315e-38
cut(age, 4)(49,64.5]   23.664559   2.067958 11.443444 1.040750e-29
cut(age, 4)(64.5,80.1]  7.640592   4.987424  1.531972 1.256350e-01
```

Here cut() automatically picked the cutpoints at 33.5, 49, and 64.5 years of age. We could also have specified our own cutpoints directly using the breaks option. The function cut() returns an ordered categorical variable; the lm() function then creates a set of dummy variables for use in the regression.

The age<33.5 category is left out, so the intercept coefficient of \$94.160 can be interpreted as the average salary for those under 33.5 years of age, and the other coefficients can be interpreted as the average additional salary for those in the other age groups.

We can produce predictions and plots just as we did in the case of the polynomial fit.

Basis Functions

- The idea is to have at hand a family of functions or transformations that can be applied to a variable $X : b_1(X), b_2(X), ..., b_K(X)$. Instead of fitting a linear model in X, we fit the model

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i. \qquad (*)$$

- Note that the basis functions $b_1(\cdot), b_2(\cdot), \ldots, b_K(\cdot)$ are fixed and known. (In other words, we choose the functions ahead of time.)
- Polynomial and piecewise-constant regression models are in fact special cases of a basis function approach.

For the polynomial regression, the basis functions are $b_j(x_i) = x_i^j$, and for piecewise constant functions they are $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$.

we can think of $(*)$ as a standard Linear model with predictors $b_1(x_i), b_2(x_i), \ldots, b_k(x_i)$. Hence, we can use least squares to estimate the unknown regression coefficients in $(*)$.

Regression Splines

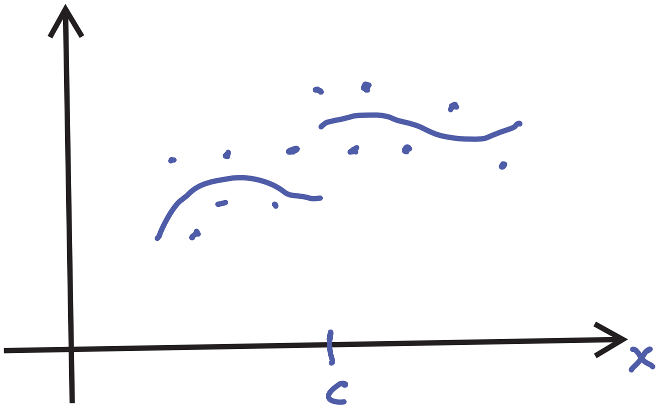
Now we discuss a flexible class of basis functions that extends upon the polynomial regression and piecewise constant regression approaches that we have just seen.

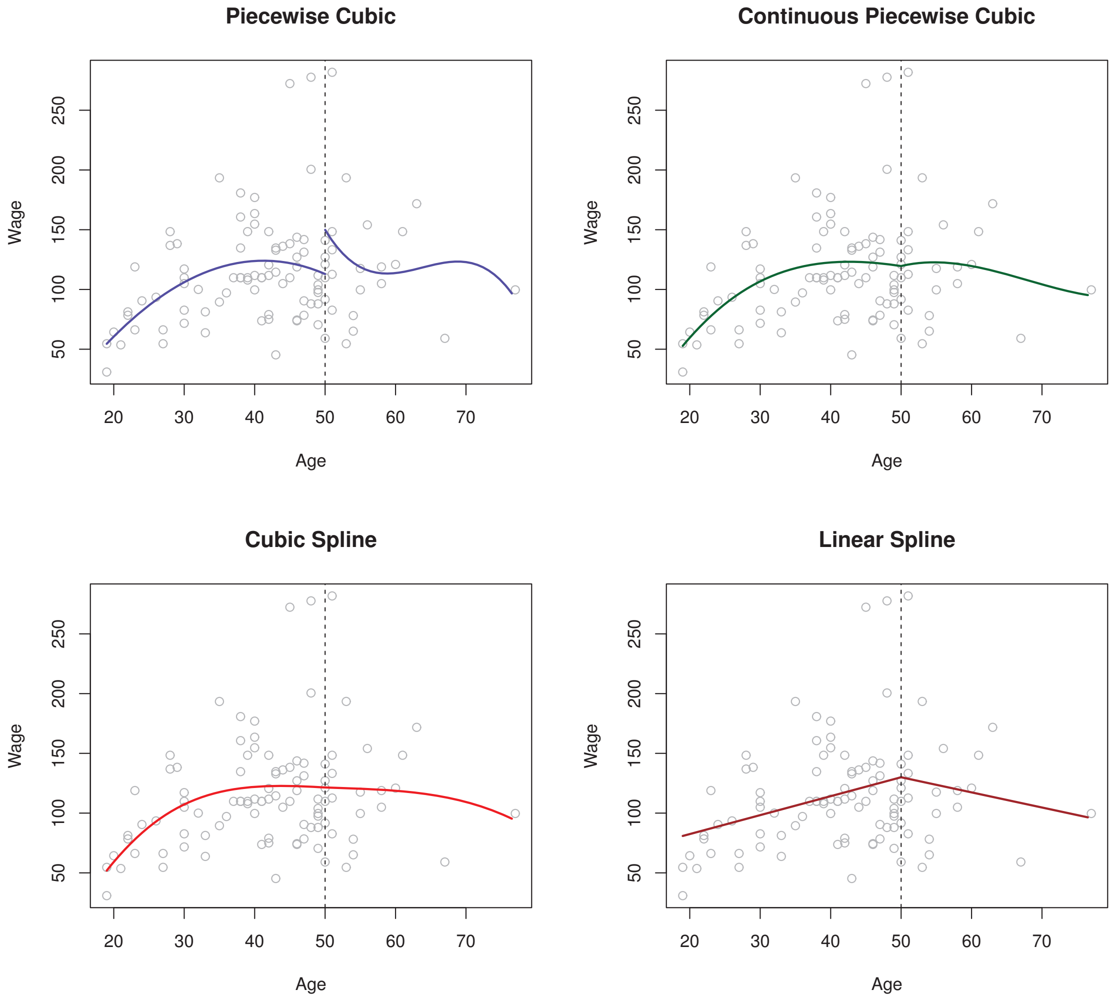
Piecewise Polynomials:

- Instead of fitting a high-degree polynomial over the entire range of X, piecewise polynomial regression involves fitting separate low-degree polynomials over different regions of X defined by knots.
- For example, a piecewise cubic polynomial with a single knot at a point c takes the form

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geqslant c \end{cases}$$

- Each of these polynomial functions can be fit using least squares applied to simple functions of the original predictor.
- Using more knots leads to a more flexible piecewise polynomial. In general, if we place K different knots throughout the range of X, then we will end up fitting $K + 1$ different cubic polynomials.





Various piecewise polynomials are fit to a subset of the Wage data, with a knot at age=50. Top Left: The cubic polynomials are unconstrained. Top Right: The cubic polynomials are constrained to be continuous at age=50. Bottom Left: The cubic polynomials are constrained to be continuous, and to have continuous first and second derivatives. Bottom Right: A linear spline is shown, which is constrained to be continuous.

- In order for the fitted curve to be continuous and look natural, we need to add constraints to the polynomials.
- *Splines* have the “maximum” amount of continuity.