Local regression and `loess`

Local regression

- If $f(x)$ is a derivable function in x_0 then, the Taylor's approximation says that it is locally approximated by a line passing through $(x_0, f(x_0))$, i.e.,

$$f(x) = \underbrace{f(x_0)}_{\alpha} + \underbrace{f'(x_0)}_{\beta}(x - x_0) + \text{error}$$

- we introduce the weighted least squares by weighting observations x_i with their distance from x_0:

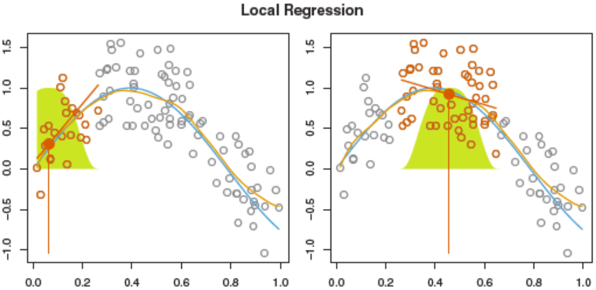
$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left\{ y_i - \alpha - \beta(x_i - x_0) \right\}^2 w_h(x_i - x_0)$$

  - h $(h > 0)$ is a scale factor, called bandwidth or smoothing parameter, and
  - $w_h(\cdot)$ is a symmetric density function around 0, said kernel.

Local regression

- By varying x_0, we obtain a whole estimated curve $\hat{f}(x)$.
- the most important component is h, which regulates the smoothness of the curve, while the choice of w is less relevant.
- we could think to w as the density of the normal distribution $N(0, 1)$

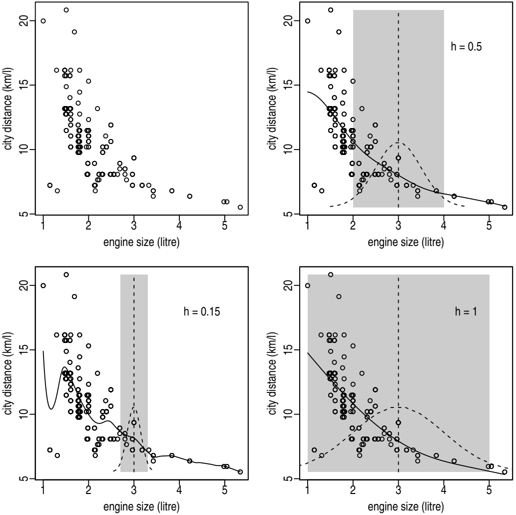
Local regression



Local regression: blue curve represents the real $f(x)$, orange curve corresponds to the local regression estimate $\hat{f}(x)$. The orange points are local to the target point x_0, represented by the orange vertical line. The green bell-shape superimposed on the plot indicates weights assigned to each point, decreasing to zero with distance from the target point. The fit $\hat{f}(x)$ at x_0 is obtained by fitting a weighted linear regression (orange line segment), and using the fitted value at x_0 (orange solid dot) as the estimate $\hat{f}(x_0)$

Local regression

The effect of h is relevant



Variable bandwidths and `loess`

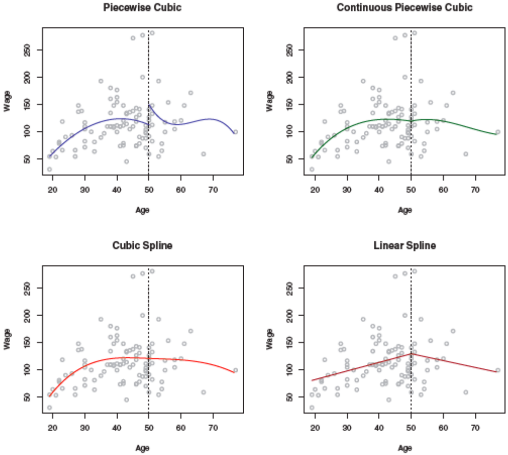
- In many cases, there is an advantage in using a non constant bandwidth along the x-axis, according it to the level of sparseness of observed points
- variable bandwidth: it is reasonable to use larger values of h when x_i are more scattered
- How do we modify h?
- loess: express the smoothing parameter defining the fraction of effective observations for estimating $f(x)$ at a certain point x_0 on the x-axis;
- this fraction is kept constant
- this implies automatically a setting of the bandwidth related to the sparsity of data

Splines

Interpolating splines

- 'Spline' is a mathematical tool useful in many contexts finalised to approximate functions or to interpolate data.
- we choose K points $\xi_1 < \xi_2 < \cdots < \xi_K$, called knots, along the x-axis.
- a function $f(x)$ is constructed, so that it passes exactly through the knots and is free at the other points
- we look for "smooth" functions
- between two successive knots, in the interval (ξ_i, ξ_{i+1}), curve $f(x)$ coincides with a suitable polynomial, of prefixed degree d
- these sections of polynomials meet at point ξ_i $(i = 2, \ldots, K-1)$
- in the sense that the resulting function $f(x)$ has a continuous derivative from degree 0 to degree $d-1$ in each of the ξ_i.

Interpolating splines



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.

Regression splines

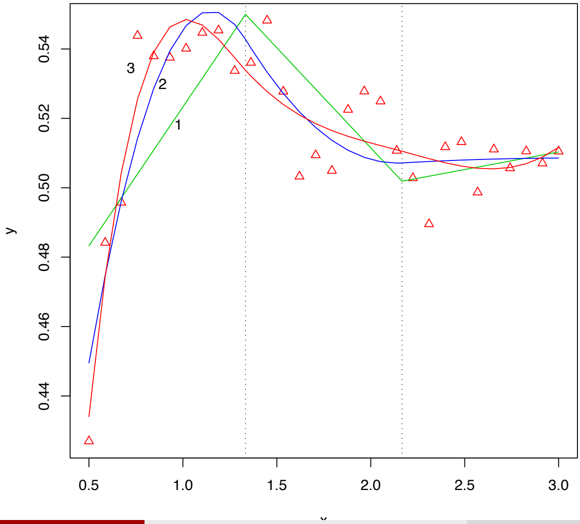
- We have n observed points (x_i, y_i) for $i = 1, \ldots, n$ that we want to interpolate
- we apply these ideas to parametric regression, by fitting a cubic spline ($d = 3$) to the n points
- we divide the x-axis into $K + 1$ intervals separated by K knots, $\xi_1, \ldots, \xi_K$, and interpolate the n points with the least squares criterion
- the obtained function is called regression spline

Regression splines

- The number K of knots and their position along the x-axis need to be chosen
- Because K is a tuning parameter regulating the complexity of the model, we need to perform a model selection according to bias-variance trade-off
- Once the number K has been set, a reasonable choice for knots position is uniformly along the x_i range.

Regression splines

Interpolated functions for $d = 1, 2, 3$



Smoothing splines

- Let us consider the penalized least squares criterion

$$D(f,\lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_{-\infty}^{\infty} \{f''(t)\}^2 \, \mathrm{d}t$$

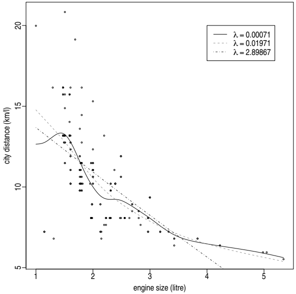
  where λ is a positive penalisation parameter of the roughness degree of curve f (quantified by the integral of $f''(x)^2$), and therefore acts as a smoothing parameter.
- Loss and Penalty formulation
- the term $\sum_{i=1}^{n} [y_i - f(x_i)]^2$ is a loss function that encourages $f(x)$ to fit the data well
- the term $\lambda \int_{-\infty}^{\infty} \{f''(t)\}^2 \, \mathrm{d}t$ is a penalty term that penalizes the variability in $f(x)$

Smoothing splines

- When $\lambda = 0$, then the penalty term has no effect, and so the function f will be very jumpy and will exactly interpolate the training observations.
- When $\lambda = \infty$, f will be perfectly smooth- it will just be a straight line that passes as closely as possible to the training points.
- In this case, f will be the linear least squares line, since the loss function amounts to minimizing the residual sum of squares.
- For an intermediate value of λ, f will approximate the training observations but will be somewhat smooth.
- So λ controls the bias-variance trade-off of the smoothing spline.

Smoothing splines

Estimate of `city distance` according to `engine size` by a smoothing spline, for three choices of λ



A noteworthy mathematical result shows that the solution to that minimization problem is represented by a natural cubic spline whose knots are distinct points x_i.

Summarizing. . .

We have relaxed the linearity assumption while still attempting to maintain as much interpretability as possible. To this end, we consider approaches such as splines and local regression.

- Regression splines involve dividing the range of X into K distinct regions. Within each region, a polynomial function is fit to the data. However, these polynomials are constrained so that they join smoothly at the region boundaries, or knots. Provided that the interval is divided into enough regions, this can produce an extremely flexible fit.
- Smoothing splines are similar to regression splines, but arise in a slightly different situation. Smoothing splines result from minimizing a residual sum of squares criterion subject to a smoothness penalty.
- Local regression is similar to splines, but differs in an important way. The regions are allowed to overlap, and indeed they do so in a very smooth way.
- Generalized additive models allow us to extend the methods above to deal with multiple predictors.