THE UNIVERSITY OF SYDNEY

Lecture 4: Review of Regression & Introduction to Smoothing

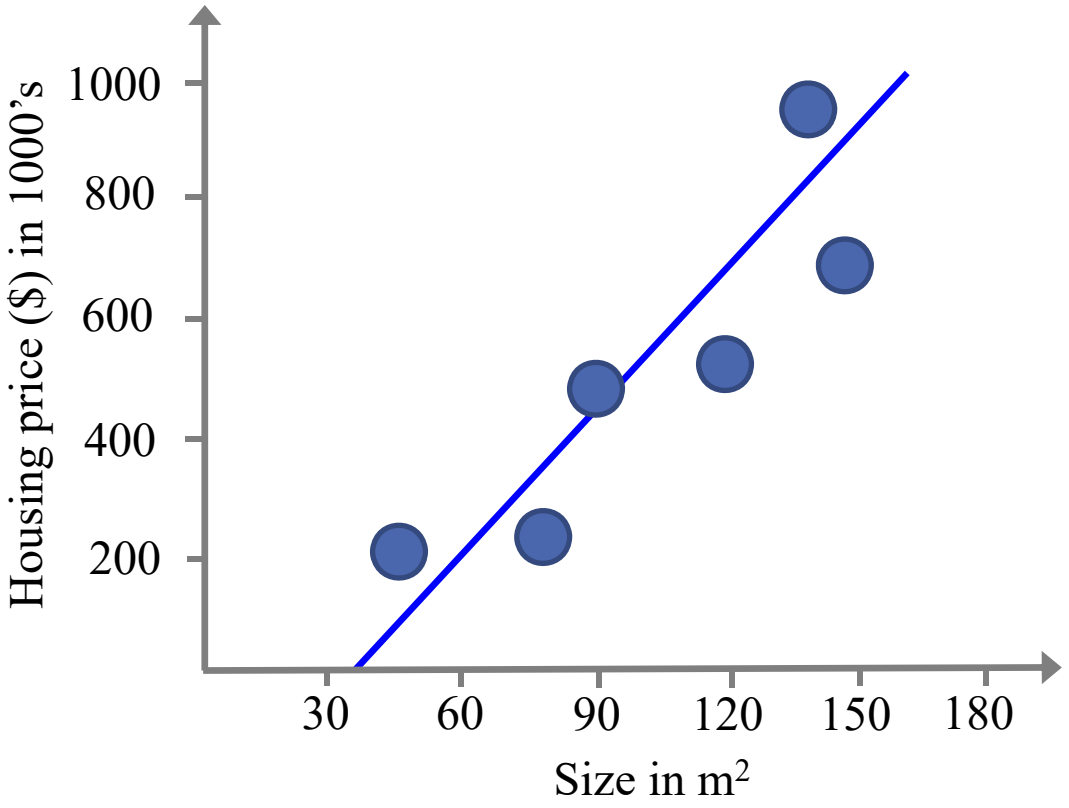
STAT5003

Pengyi Yang

Review of Simple Regression

Motivating example

Suppose your friend wants to sell his house. Given the size of the house, how much should he expect to get?



The linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

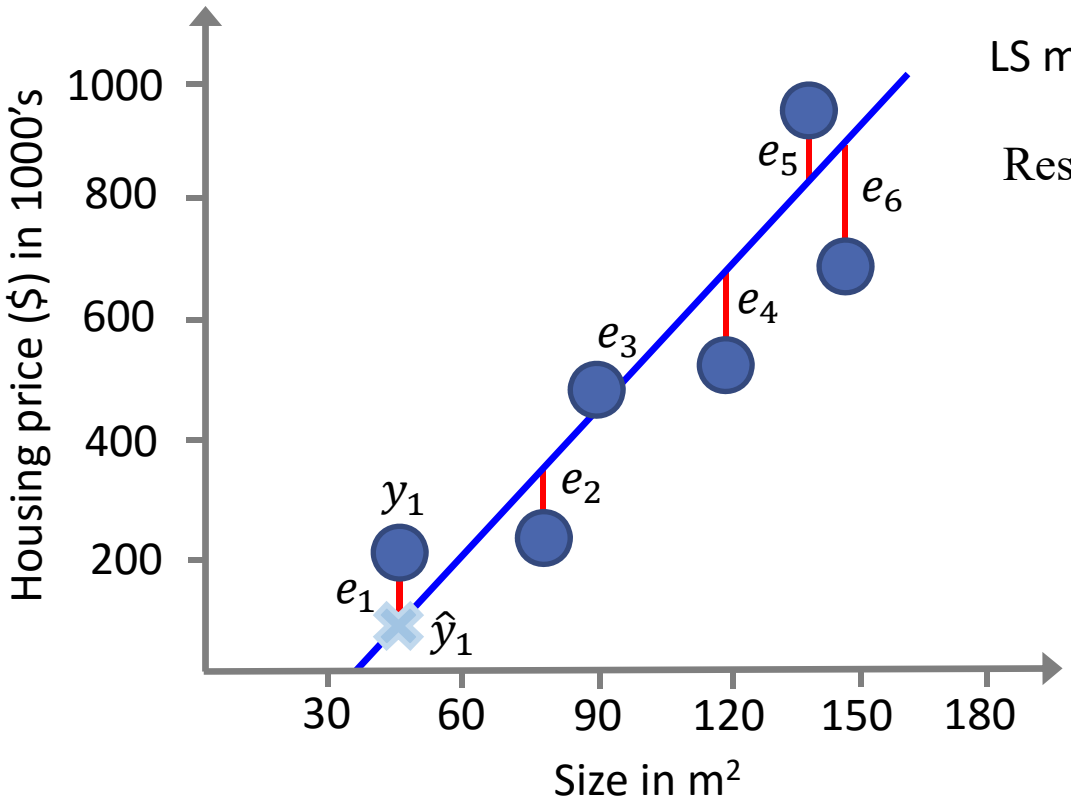
$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{(data; } i = 1 \ldots n\text{)}$$

where:

- X is predictor (independent variable) and Y is response (dependent variable)
- β_0 is the intercept of the regression line (the expected value of Y when $X = 0$)
- β_1 is the slope of the regression line (the average increase in Y associated with a one-unit increase in X)
- ϵ is a residual (error); (random with zero mean and finite variance)

Least squares regression line

There are multiple ways to determine "optimal" line through data points.



LS minimises $\sum_{i=1}^{n} e_i^2$

Residual sum of squares (RSS)

Least squares regression line

There are multiple ways to determine "optimal" line through data points.

$$\widehat{\text{RSS}} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

intercept slope

Theorem 1. The least squares approach minimize the RSS by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$ such that:

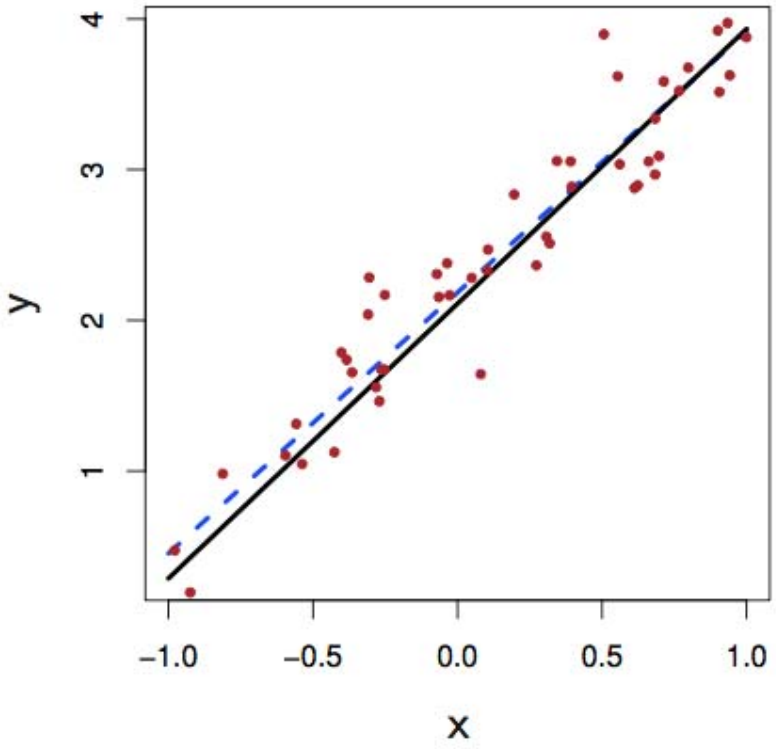
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\widehat{\text{Cov}}\,(x, y)}{\widehat{\text{Var}}\,(x)},$$

where:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

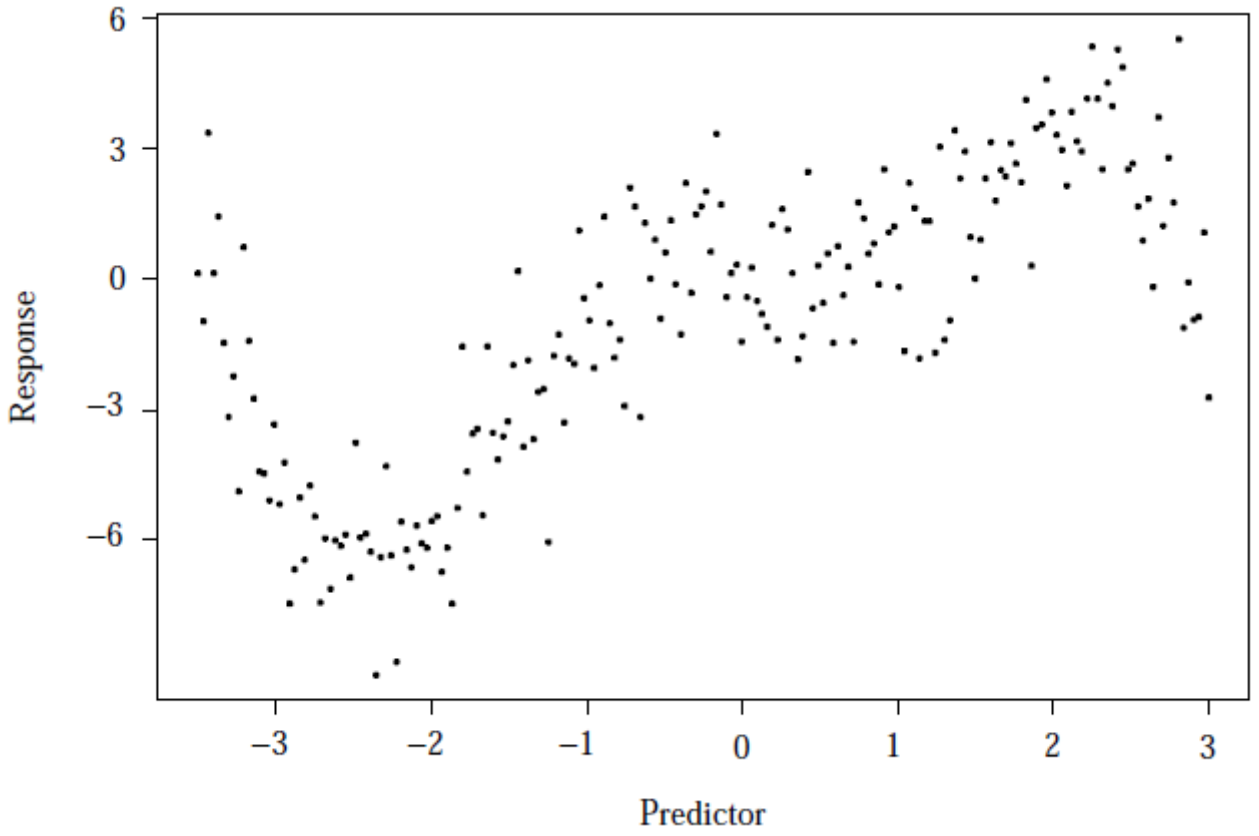
Linear Regression Fit

- Linear regression often fits well to data that have linear relationship.
- Various theoretical analyses of linear regression fit are available for further inference of the data and the relationship of predictor and response variables.



Scatterplot smoothing

- How would we find a fit to this curve?



Introduction to Smoothing

Predictor (X) – response (Y) data

- With *predictor–response data*, the random response variable Y is assumed to be a stochastic function of the value of a predictor variable X.
- A typical model for *predictor–response data* is

$$Y_i = s(x_i) + \varepsilon_i,$$

- where

  ε_i are zero-mean stochastic noise and
  s is a smooth function.

- The conditional distribution of $Y|X$ describes how Y depends on X.
- One sensible smooth curve through the data would connect the conditional means of $Y|X$ for the range of predictor values observed $(x_{1, \ldots,} x_n)$.

Local averaging

- Most smoothers (smoothing function) rely on the concept of *local averaging*.
- The Y_i whose corresponding x_i are near x should be averaged in some way to glean information about the appropriate value of the smooth at x.
- A generic local-averaging smoother can be written as

$$\hat{s}(x) = \text{avg } \{Y_i | \, x_i \in N(x)\}$$

- for
  - some generalised average function "avg" and
  - some neighbourhood of x, say $N(x)$.

Smoothers

- Different smoothers result from different choices for the *averaging function* (e.g., mean, weighted mean, median, or M-estimate),

- and the *neighbourhood* (e.g., the nearest few neighbouring points, or all points within some distance).

- The form of $N(x)$ may even vary with x so that different neighbourhood sizes or shapes may be used in different regions of the dataset.

Span (and related concepts)

- The most important characteristic of a neighbourhood is its *span*, which is represented by the smoothing parameter λ.
- The *span* of a neighbourhood measures its *inclusiveness:*
  - Neighbourhoods with small span are strongly local, including only very nearby points,
  - Neighbourhoods with large span have wider membership.

- Ways to measure a neighbourhood's inclusiveness include

✓ *size* (number of points),

✓ *span* (proportion of sample points that are members),

✓ *bandwidth* (physical length or volume of the neighborhood)

Constant-Span Running Mean (k-nearest neighbours)

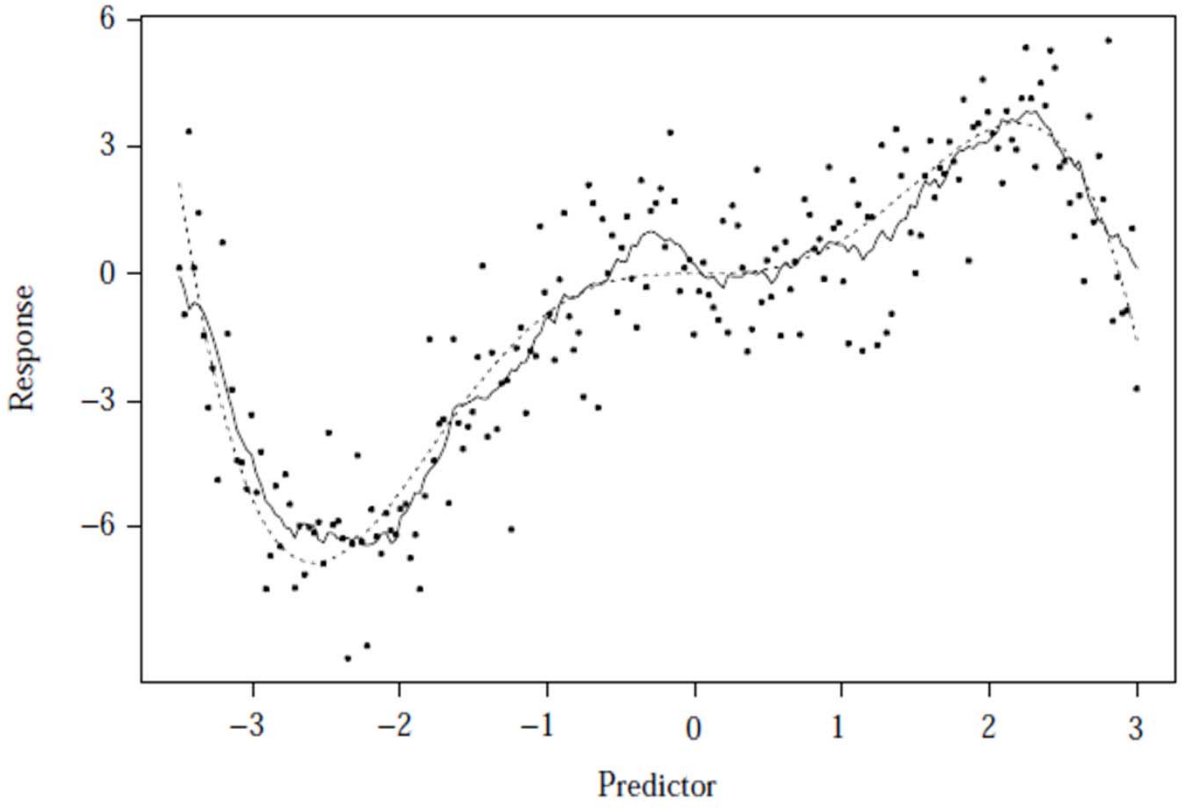
- A simple smoother takes the sample mean of k nearby points:
- We define $\mathcal{N}(x_i)$ as x_i itself, the $(k-1)/2$ points whose predictor values are nearest below x_i , and the $(k-1)/2$ points whose predictor values are nearest above x_i .
- This $\mathcal{N}(x_i)$ is termed the *symmetric nearest neighbourhood*, and the smoother is called a *moving average* or *a k-nearest neighbours* (kNN) smoother.
- Without loss of generality, assume hereafter that the data pairs have been sorted so that the x_i are in increasing order. Then the constant-span running-mean smoother can be written as

$$\hat{s}_k(x_i) = \text{mean}\left\{Y_j \text{ for } \max\left(i - \frac{k-1}{2}, 1\right) \le j \le \min\left(i + \frac{k-1}{2}, n\right)\right\}.$$

- For the purposes of graphing or prediction, one can compute $\hat{s}$ at each of the x_i and *interpolate linearly in between*.

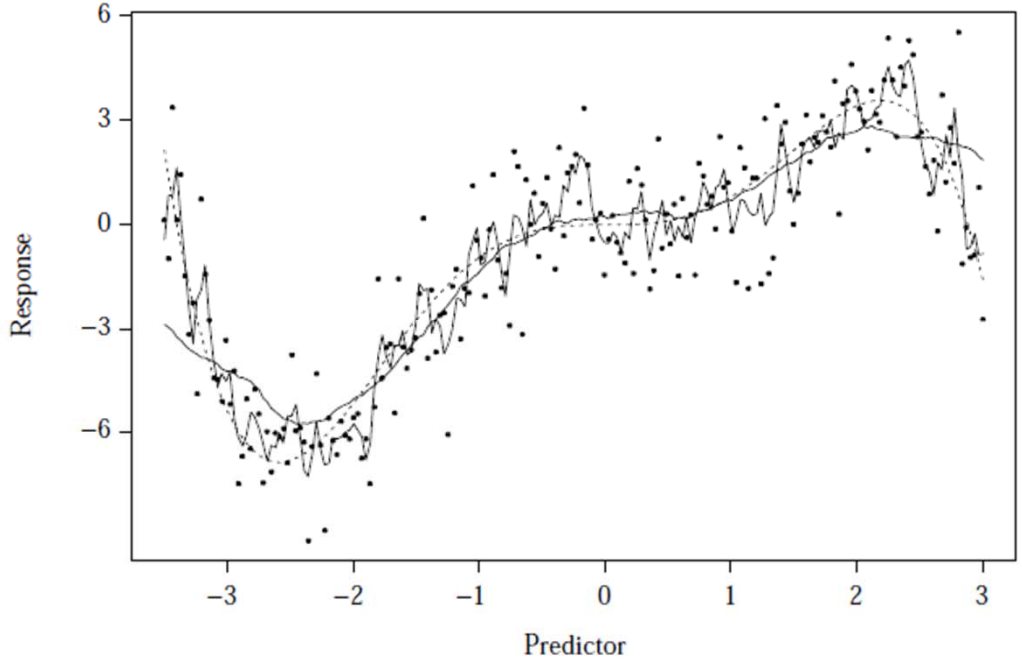
Running-mean smoothing example

- Results from a constant-span running-mean smoother with k=13 (solid line), compared with the true underlying curve (dotted line).
- The true relationship, $s(x) = x^3 \sin\{(x + 3.4)/2\}$, is shown with a dotted line; the estimate $\hat{s}_k(x)$ is shown with the solid line.



Effect of Span

- A natural smoothing parameter for the constant-span running-mean is $\lambda = k$.
- As for all smoothers, this parameter controls wiggliness, here by directly controlling the number of data points contained in any neighbourhood.
- The figure bellow shows the results from a constant-span sunning-mean smoother with k=3 (wigglier solid line) and k=43 (smoother solid line). The underlying true curve is shown with a dotted line.



Demonstration

Performance of Smoothers

What estimator is the best?

- For a given point x, let $\hat{s}(x)$ be an estimator of $s(x)$. *Which estimator is best?*

- We can assess the quality of $\hat{s}(x)$ as an estimator of $s(x)$ at x using mean squared error at x:

$$\mathrm{MSE}(\hat{s}(x)) = E\{[\hat{s}(x) - s(x)]^2\},$$

- Which can be further decomposed to bias and variance

$$\mathrm{MSE}(\hat{s}(x)) = (\mathrm{bias}\{\hat{s}(x)\})^2 + \mathrm{var}\{\hat{s}(x)\}$$

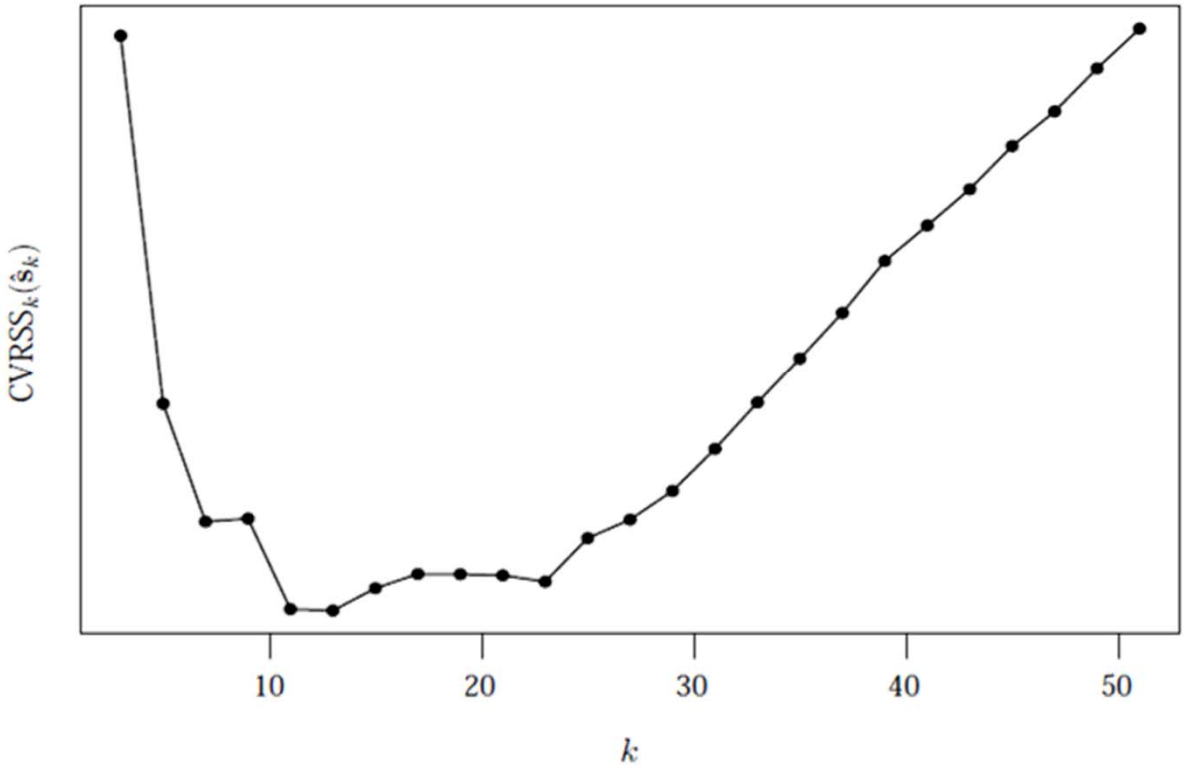
Span Selection for Linear Smoothers

- The best choice for k must balance a trade-off between bias and variance.
  - For small k, the estimated curve will be wiggly but exhibit more flexibility to fit the data.
  - For large k, the estimated curve will be smooth but exhibit substantial bias in some regions.
- The role of the smoothing parameter is to control this trade-off between bias and variance.
- One way to select best k is to use leave-one-out cross-validation which gives *cross-validated residual sum of squares*, $\text{CVRSS}_k(\hat{\boldsymbol{s}}_k)$

$$\frac{\text{CVRSS}_k(\hat{\mathbf{s}}_k)}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{s}_k^{(-i)}(x_i)\right)^2$$

Cross-validated residual sum of squares

- Typically, $\text{CVRSS}_k(\hat{\boldsymbol{s}}_k)$ is plotted against k.



Demonstration

Running Lines and Running Polynomials

Running Lines (Running Polynomials)

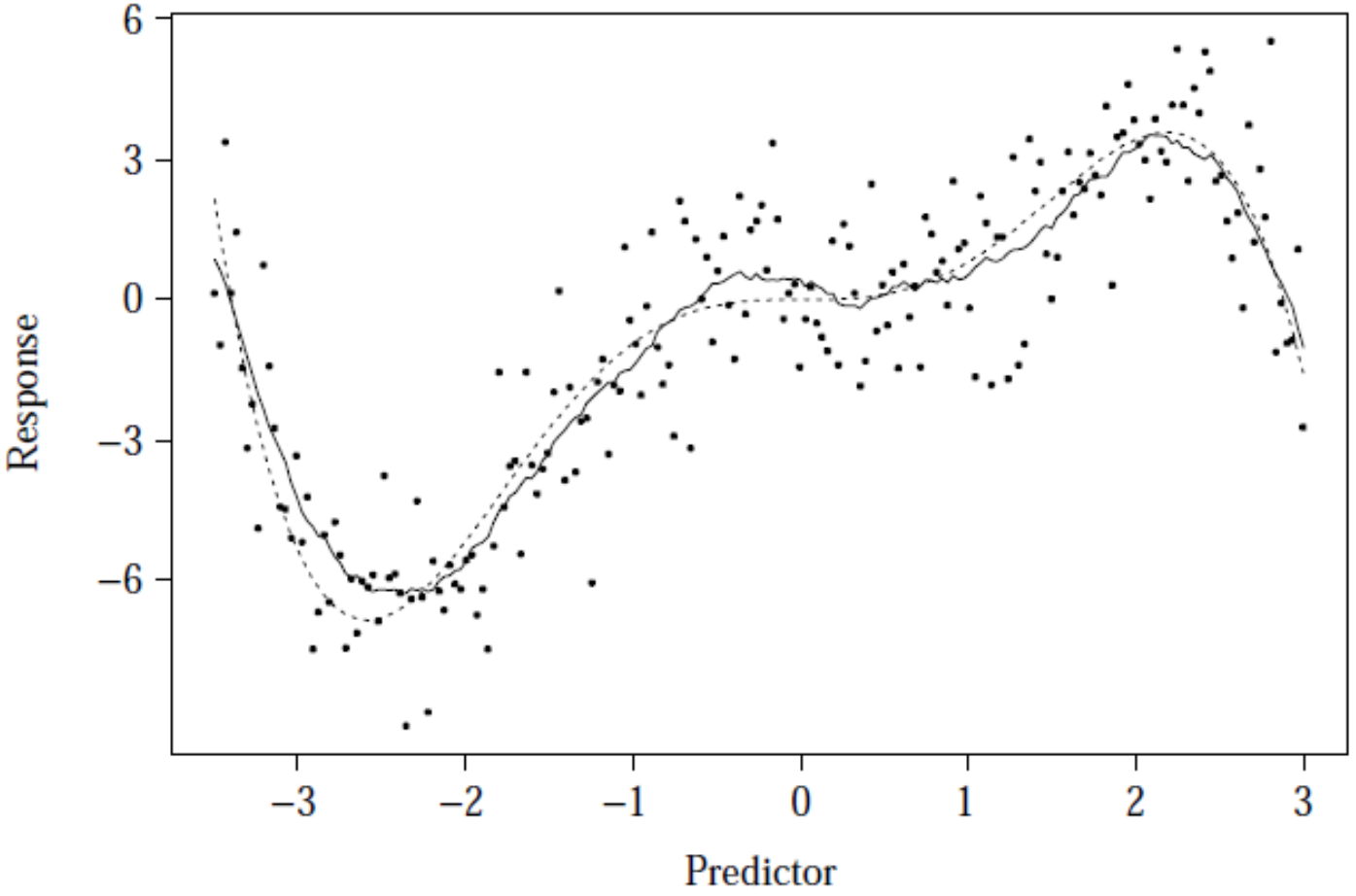
- The constant-span running-mean smoother exhibits visually unappealing wiggliness for any reasonable k. It also can have strong bias at the edges because it fails to recognize the local trend in the data.
- The running-line smoother can mitigate both problems.
- Consider fitting a linear regression model to the k data points in $N(x_i)$. Then the least squares linear regression prediction at x is

$$l_i(x) = \overline{Y}_i + \widehat{\beta}_i (x - \overline{x}_i),$$

- where $\overline{Y}_i$, $\overline{x}_i$, $\widehat{\beta}_i$, are the mean response, the mean predictor, and the estimated slope of the regression line, respectively, for the data in $N(x_i)$.
- The *running-line smooth* at x_i is $\hat{s}_k(x_i) = l_i(x_i)$.

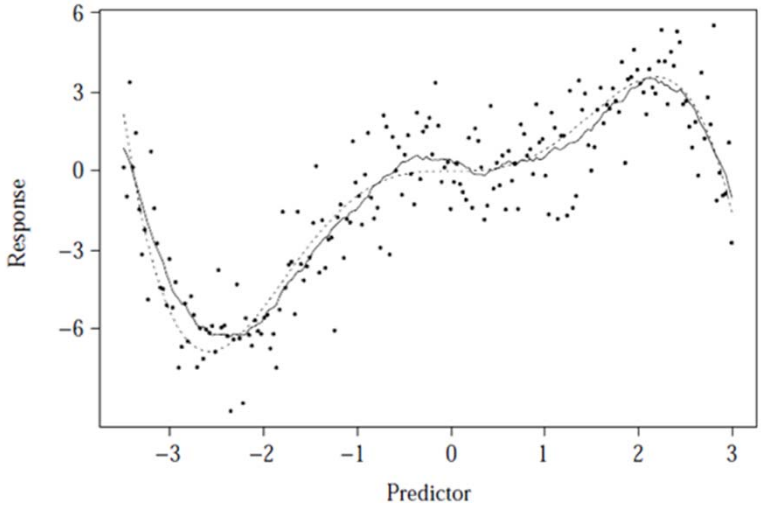
Running Line fit

Results from a running-line smooth for k=23 (solid line), compared with the true underlying curve (dotted line).



Running lines; more interpretations

- Figure below shows a running-line smooth of the data introduced in before, for the $k = 23$ chosen by cross-validation.



- The edge effects are much smaller and the smooth is less wiggly than with the constant span running-mean smoother.

- Since the true curve is usually well approximated by a line even for fairly wide neighbourhoods, k may be increased from the optimal value for the constant-span running-mean smoother.
- This reduces variance without seriously increasing bias.

Demonstration

Kernel Smoothers

Kernel Smoothers

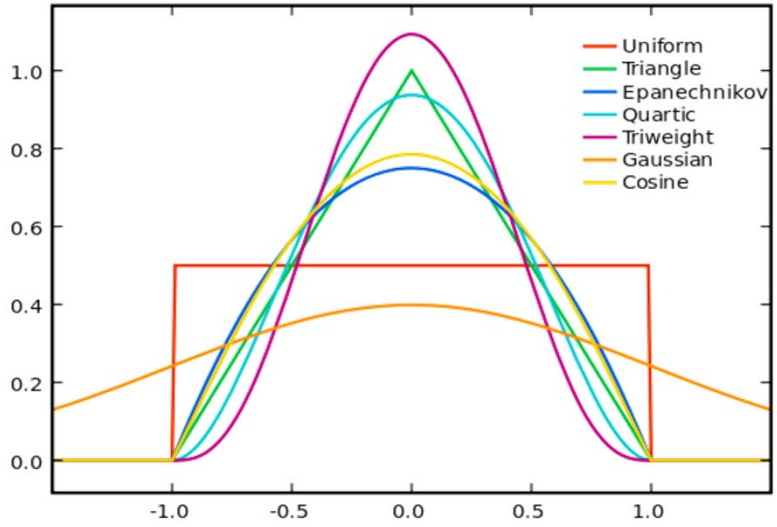
- For the smoothers mentioned so far, there is a discontinuous change to the fit each time the neighbourhood membership changes. Therefore, they tend to fit well statistically but exhibit visually unappealing jitters or wiggles.
- One approach to increasing smoothness is to redefine the neighbourhood so that points only gradually gain or lose membership in it.
- Let *K* be a symmetric kernel centred at 0.
- A kernel is essentially a weighting function—in this case it weights neighbourhood membership.
- One reasonable kernel choice would be the standard normal density, $K(z) = (\frac{1}{\sqrt{2\pi}}) \exp\{-z^2/2\}$.

Kernel Smoother

- Then let

$$\hat{s}_h(x) = \sum_{i=1}^{n} Y_i \frac{K((x - x_i)/h)}{\sum_{j=1}^{n} K((x - x_j)/h)}$$

- where the smoothing parameter h is called the *bandwidth*.
- Notice that for many common kernels such as the normal kernel, all data points are used to calculate the smooth at each point, but very distant data points receive very little weight.



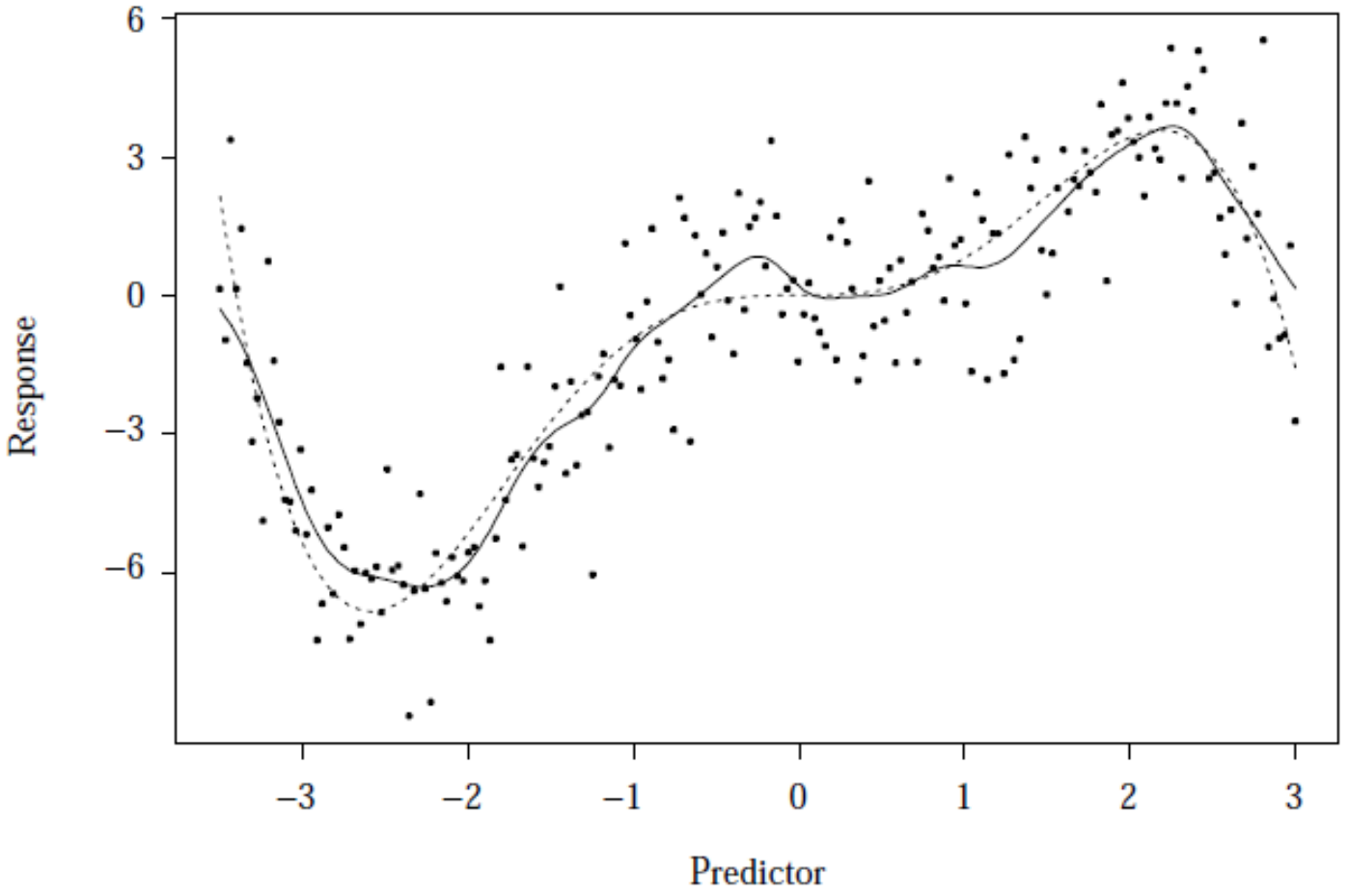
Distant data points receive very little weight

- Proximity increases a point's influence on the local fit; in this sense the *concept of local averaging remains*.
  - A large bandwidth yields a quite smooth result because the weightings of the data points change little across the range of the smooth.
  - A small bandwidth ensures a much greater dominance of nearby points, thus producing more wiggles.

- The choice of smoothing kernel is much less important than the choice of bandwidth and there are few reasons to look beyond a normal kernel.

Normal kernel smoother

Results from a normal kernel smooth using h=0.16 (chosen by cross-validation) compared with the true underlying curve (dotted line).



Normal kernel smoother

- The figure from last slide shows a kernel smooth of the data, using a normal kernel with $h = 0.16$ chosen by cross-validation.
- Since neighbourhood entries and exits are gradual, the result exhibits characteristically rounded features.
- However, note that the kernel smoother does not eliminate systematic bias at the edges, as the running-line smoother does.

Demonstration

Spline Smoothing

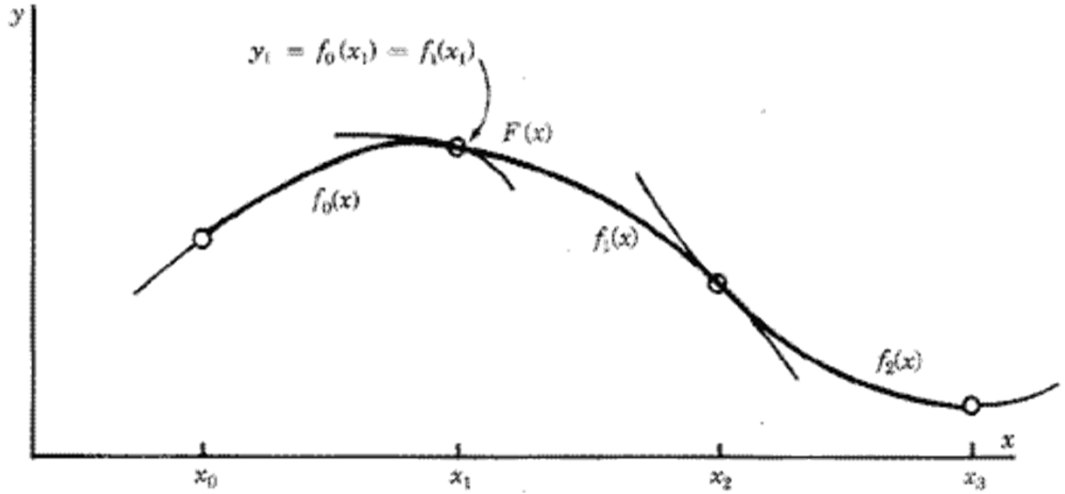
Assume that the data have been sorted in increasing order of the predictor, so x_1 is the smallest predictor value, x_n is the largest.

Define

$$Q_\lambda(\hat{s}) = \sum_{i=1}^{n} (Y_i - \hat{s}(x_i))^2 + \lambda \int_{x_1}^{x_n} \hat{s}''(x)^2 \, dx,$$

where $\hat{s}''(x)$ is the second derivative of $\hat{s}(x)$.

- The summation constitutes a penalty for misfit
- The integral is a penalty for wiggliness.



http://www.aerorocket.com/AeroCAD/AeroRocketCAD.html

Spline Smoothing

$$Q_\lambda(\hat{s}) = \sum_{i=1}^{n} (Y_i - \hat{s}(x_i))^2 + \lambda \int_{x_1}^{x_n} \hat{s}''(x)^2 \, dx,$$

- The parameter λ controls the relative weighting of these two penalties.
- Minimising $Q_\lambda(\hat{s})$ leads to a *cubic smoothing spline* $\hat{s}(x)$.
- This function is a *cubic polynomial* in each interval $[x_i, x_{i+1}]$ for $i = 1, \ldots, n-1$, with these polynomial pieces pasted together twice continuously differentiable at each x_i.

Demonstration

Non-Linear Smoothers

Nonlinear smoothers

- Nonlinear smoothers can be much slower to calculate, and in ordinary cases they offer little improvement over simpler approaches.
- However, the simpler methods can exhibit very poor performance for some types of data.
- The *loess* smoother provides improved robustness to outliers that would introduce substantial noise in an ordinary smoother.

Loess

- Loess is a *Locally weighted scatterplot smoothing* method
- The *loess* ("*LOcal regrESSion*") smoother is a widely used method with good robustness properties.
- It is essentially a weighted running-line smoother, except that each local line is fitted using a robust method rather than least squares.
- As a result, the smoother is nonlinear.
- Loess is fitted iteratively.

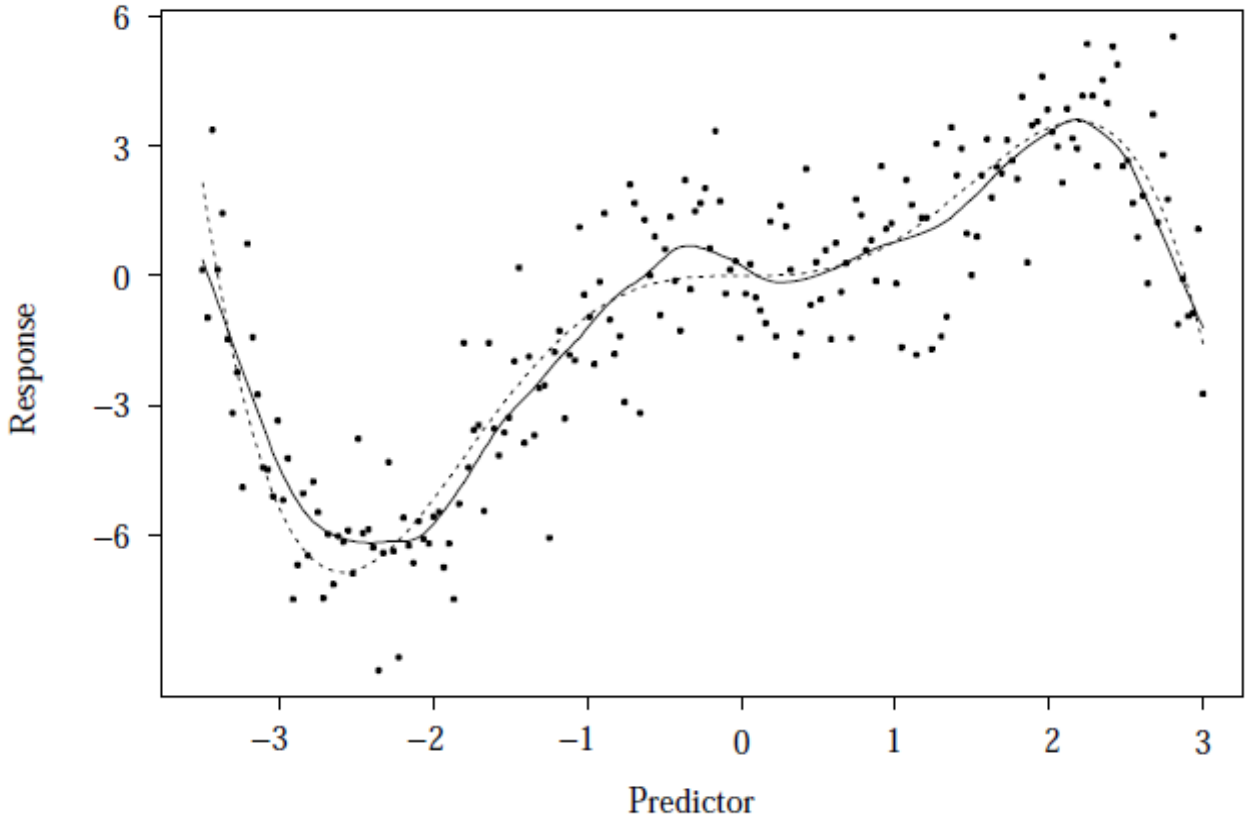
$$K_i(x) = K\left(\frac{x - x_i}{d_k(x_i)}\right),$$

$$K(z) = \begin{cases} (1 - |z|^3)^3 & \text{for } |z| \le 1, \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{j=1}^{n} \left(Y_j - \left(\beta_{0,i}^{(t)} + \beta_{1,i}^{(t)} x_j\right)\right)^2 K_i(x_j).$$

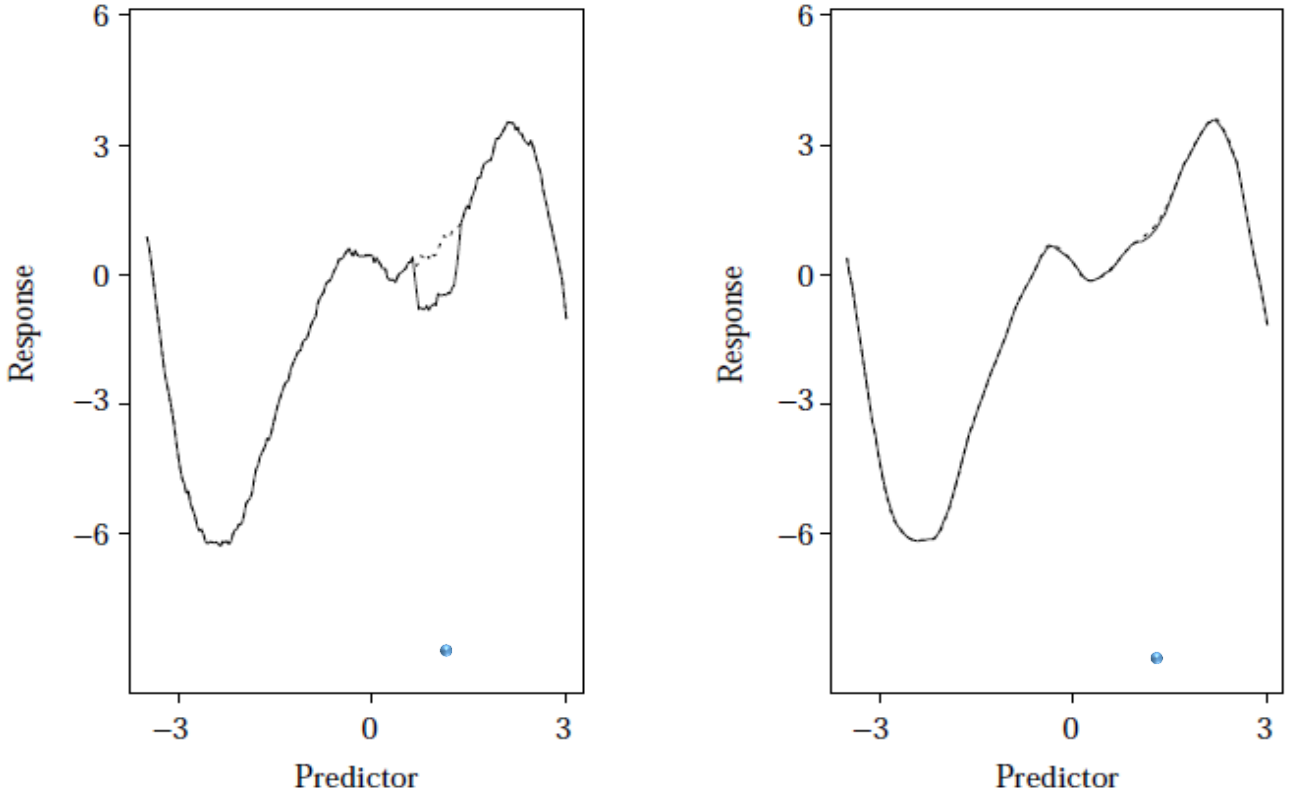
Loess fit

Figure below shows a loess smooth of the data, using $k = 30$ chosen by cross-validation. The results are very similar to the running-line smooth.



Effect of Outliers

- Figure below shows the *effect of outliers*. The dotted line in each panel is the original smooth for loess and running lines; the solid lines are the result when three additional data points at (1, −8) are inserted in the dataset.



Demonstration