STAT40810 — Stochastic Models

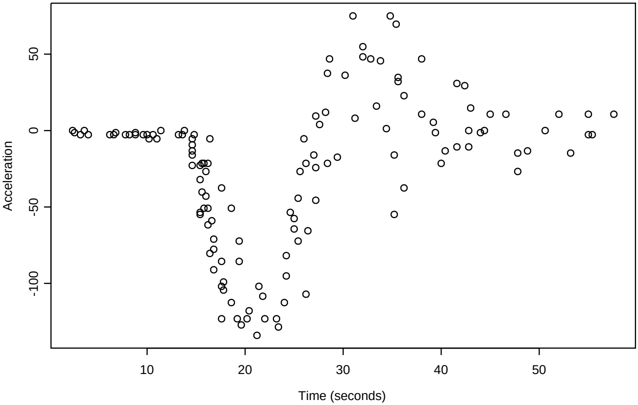
Brendan Murphy

Week 3

Nonparametric Regression

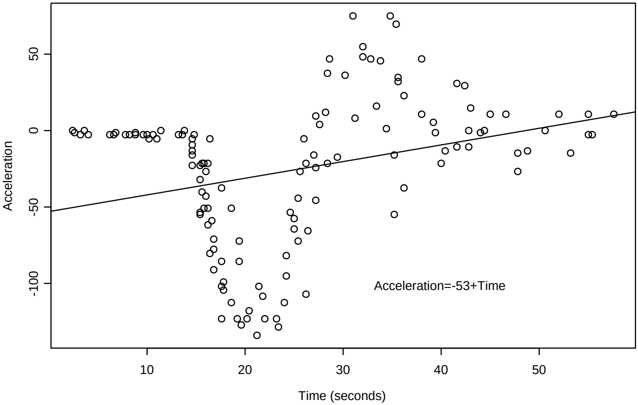
Example: Motorcycle Crash Simulation

- Data were collected recording the acceleration versus time for a simulated motorcycle accident. Here is a scatter plot of the data.



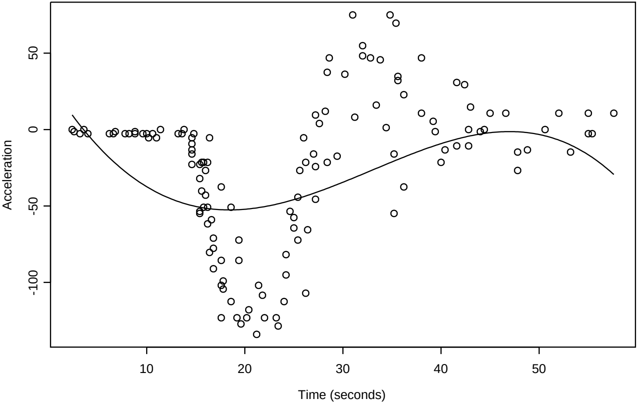
Linear Regression

- A line was fitted to the data using least squares and we got the following fit:



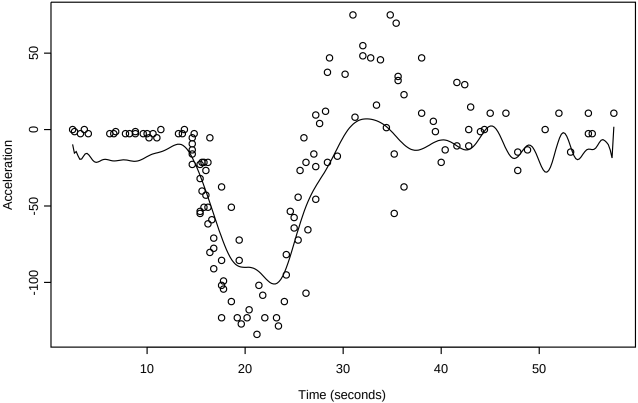
Polynomial Regression

- A quartic was fitted to the data using least squares and we got the following fit:



Polynomial Regression

- A polynomial of degree 40 was fitted to the data using least squares and we got the following fit:



Comments and Potential Problems

- We can immediately see that the linear regression fails miserably for this data.
- The polynomial regression appears to do much better, provided that we use a high enough order polynomial.
- We could use a polynomial of degree 132 (there were 133 data points) and we would get a perfect fit.
- There are numerical problems with fitting high order polynomials because we need to invert a matrix with very small determinant.
- The numerical problems can be reduced if we use orthogonal polynomials, for example, Hermite polynomials.
- Polynomial regression can give terrible results when we extrapolate.

Non-parametric Regression

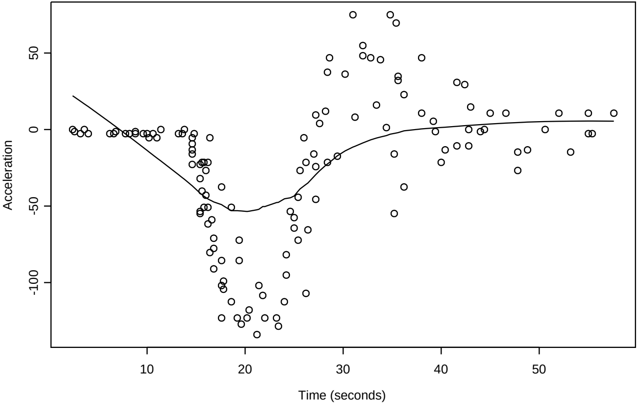
- Non-parametric regression includes a vast number of methods which are used to fit relationships while making few assumptions.
- Non-parametric regression is commonly used to smooth data, making it less noisy than the original data.
- Methods which we will describe are:
    - Lowess (Locally Weighted Regression)
    - Kernel Smoothing
    - Smoothing Splines
- Parts of these names may be familiar.

Lowess Regression

- Lowess regression is a variant on linear regression, where it does a different regression to predict each point.
- The regression used to predict each data point uses only a fraction of the data (the fraction is usually called the *span*).
- The regression also weights the observations according to how far they are away from the point being predicted.
- This will be described in more detail later.

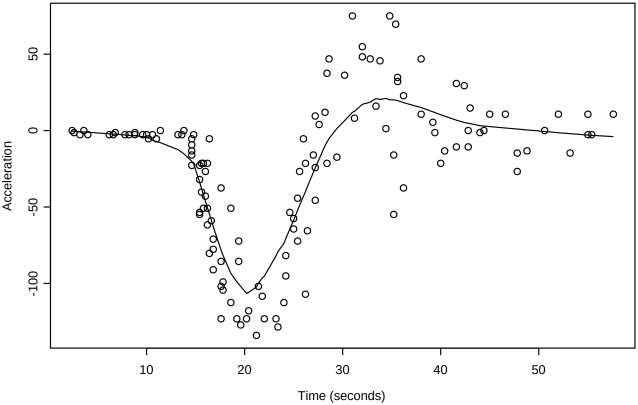
Lowess Regression Results

- A lowess regression was completed and the following curve was fitted:



Lowess Regression Results II

- Another lowess regression was completed with a different smoothing parameter and the following curve was fitted:



Kernel Smoothing

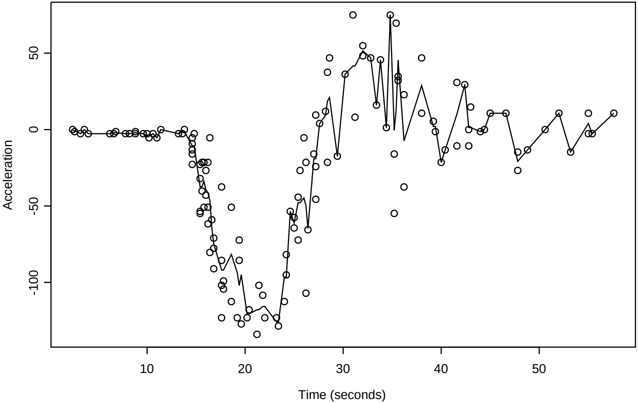
- Kernel smoothing is an alternative method of using weights to get a good local fit.
- The method requires specifying a kernel and bandwidth parameter. The choice of bandwidth is very important, whereas the choice of kernel is less important.
- The value of the response variable is predicted using $\hat{y}_j = \sum_{i=1}^{n} w_{ij} y_i$ where

$$w_{ij} = \frac{\frac{1}{h} K\left(\frac{x_i - x_j}{h}\right)}{\frac{1}{h}\sum_{k=1}^{n} K\left(\frac{x_i - x_k}{h}\right)}$$

- More on this later...

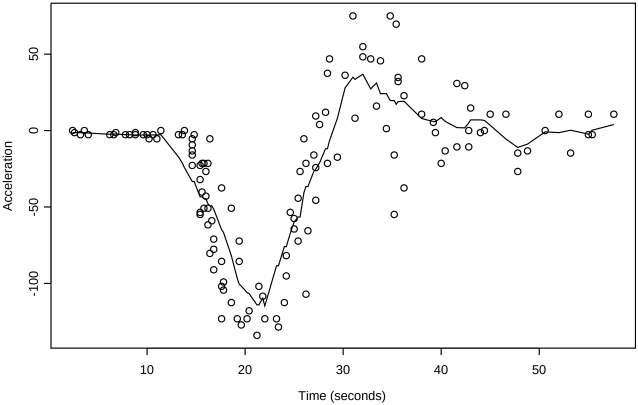
Kernel Smoothing Results

- The kernel smoother with a rectangular kernel and bandwidth 0.5 gave the following fitted curve.



Kernel Smoothing Results

- The kernel smoother with a rectangular kernel and bandwidth 5 gave the following fitted curve.



Smoothing Splines

- We could find the function, $f(x)$, that minimizes

$$\sum_{j=1}^{n}[y_j - f(x_j)]^2,$$

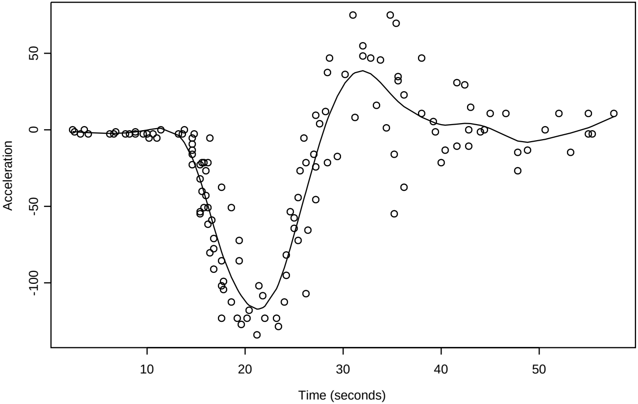
  but the resulting function would just interpolate the points (x_j, y_j).
- The penalised sum of squares criterion chooses a function $f(x)$ which minimizes

$$\sum_{j=1}^{n}[y_j - f(x_j)]^2 + \lambda \int_{\mathcal{X}} f''(x)^2 dx, \text{ where } \lambda > 0.$$

- The λ value controls how smooth we want $f(x)$ to be.
- $\lambda = 0$ would give a function that interpolates (x_j, y_j) whereas $\lambda = \infty$ would give a linear function.

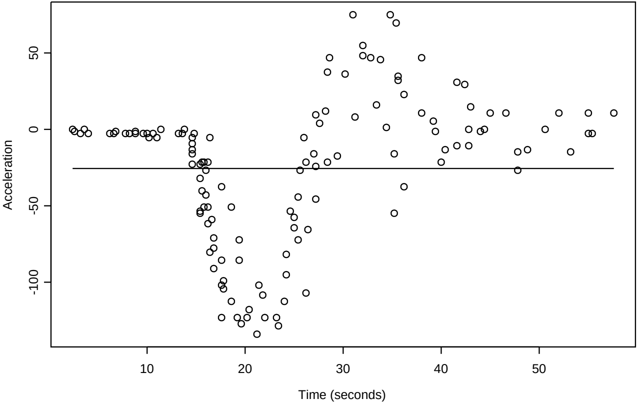
Smoothing Spline Results

- The smoothing spline when the smoothing parameter was chosen by cross-validation.



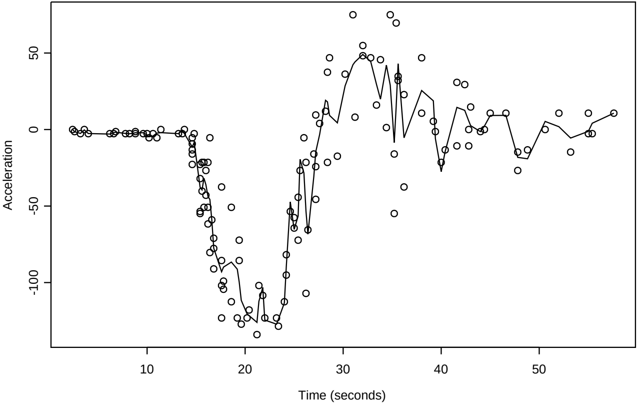
Smoothing Spline Results

- The smoothing spline when the smoothing parameter was very large.



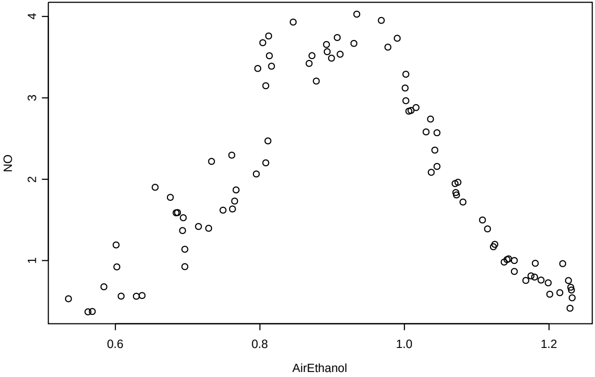
Smoothing Spline Results

- The smoothing spline when the smoothing parameter was very small.



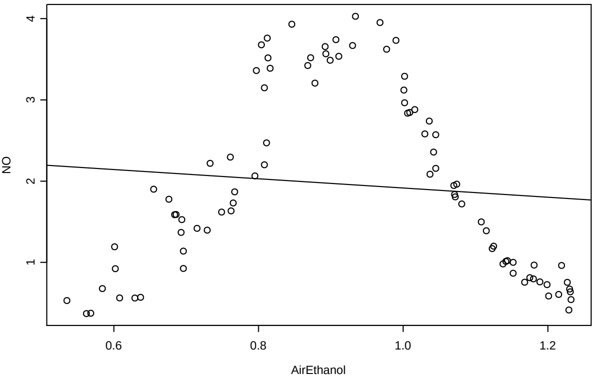
Example: Nitric Oxide Emissions

- Data were collected recording the amount of nitric oxide (NO) emitted from engines, where the air/ethanol mix of the fuel was varied.
- The relationship between the air/ethanol mix and the nitric oxide emissions was of interest.



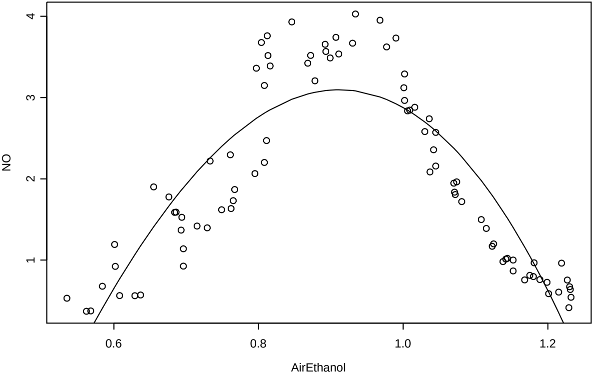
Linear Regression

A line was fitted, but the fitted relationship was very poor.



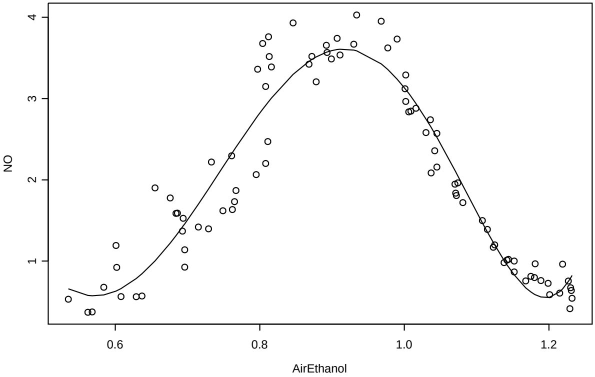
Cubic Regression

A cubic equation was fitted, but the fitted relationship was (again!) very poor.



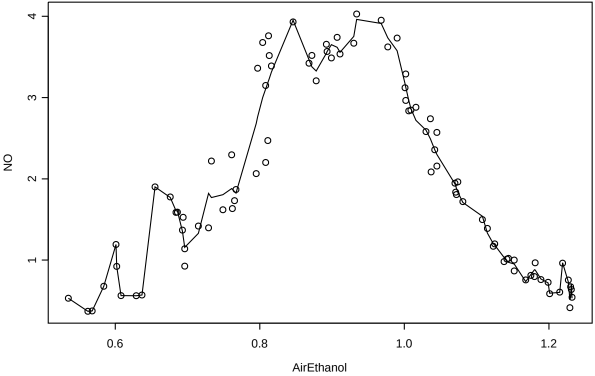
Quintic Regression

A quintic equation was fitted, but the fitted relationship is getting better, but gives poor extrapolations.



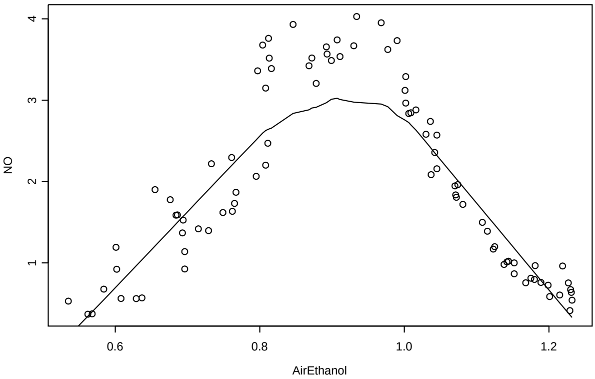
Polynomial Regression

A degree 50 polynomial equation was fitted, but the fitted relationship is getting better, but we are definitely overfitting.



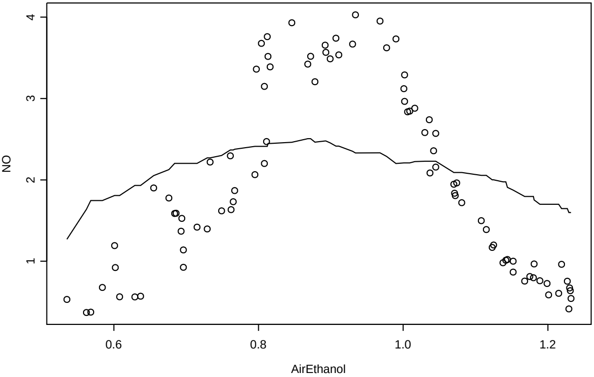
Lowess

A LOWESS regression was completed. The LOWESS curve finds it hard to adjust to the changing curvature.



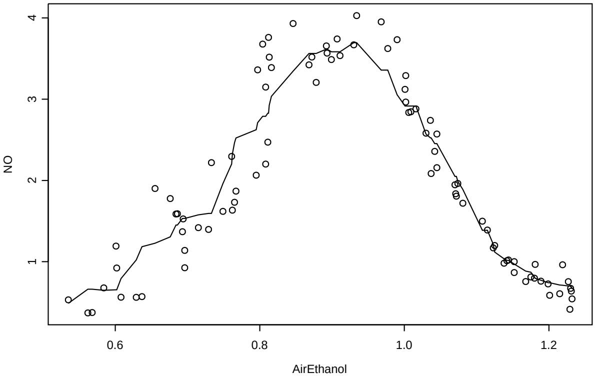
Kernel Smoothing

A kernel smoother was used. The results are very poor.



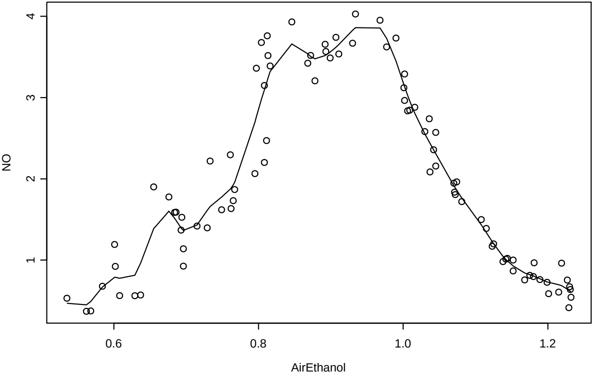
Kernel Smoothing

A kernel smoother was used, with a narrower bandwidth. The results much better.



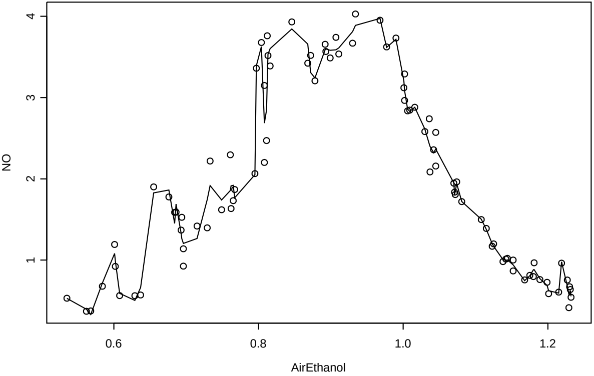
Splines

A spline regression was completed. The fitted curve seems to find the ideal balance between fit and overfitting.



Splines

A spline regression was completed, with small roughness penalty. The fitted curve almost interpolates the data!


NO
AirEthanol