Statistical Learning and Data Mining

Module 12: Nonlinear Modelling

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Statistical Learning and Data Mining: Content

1. Statistical foundations.

2. Regression methods.

3. Classification methods.

4. **Nonlinear modelling**.

5. Tree-based methods.

6. Generalised linear models or unsupervised learning (depending on time and other considerations).

Module 12: Nonlinear Modelling

1. Basis functions
2. Regression splines
3. Smoothing splines
4. Local regression
5. Generalised additive models

Nonlinear modelling

Our general framework for regression problems is the additive error model

$$Y = \beta_0 + f(X) + \varepsilon,$$

where $f(\cdot)$ is an unknown regression function.

- In module, we move beyond linear specifications for $f(X)$ to study methods that can approximate arbitrary "well-behaved" functions $f(X)$ within the observed range of X.

- We consider the case with one predictor for most of this module, before extending these methods to multiple predictors in the last section.

Nonlinear modelling

For logistic regression, we consider the model

$$Y|X = x \sim \mathrm{Ber}(p(x)),$$

where

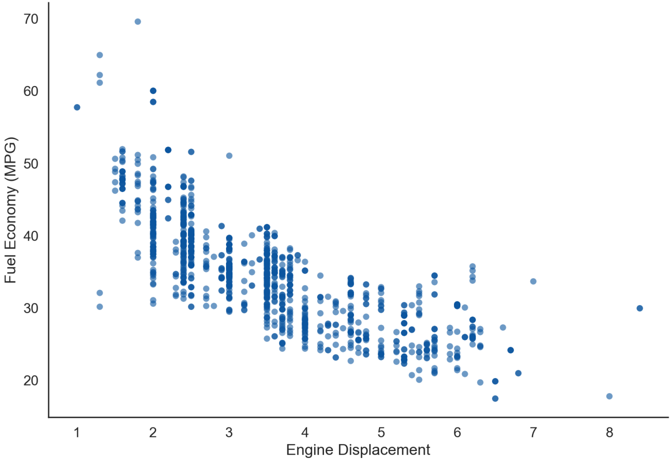
$$p(x) = \frac{\exp(\beta_0 + f(X)}{1 + \exp(\beta_0 + f(X))},$$

thus moving beyond linearity to model possibly nonlinear decision boundaries.

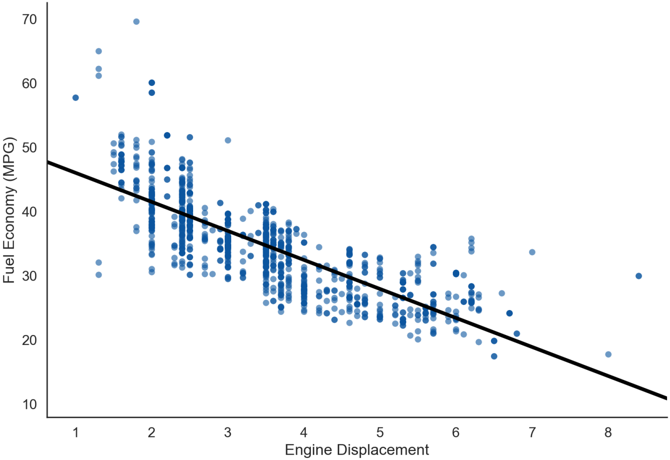
Example: Fuel Economy

- This example uses data extracted from the `fueleconomy.gov` website run by the US government, which lists different estimates of fuel economy for passenger cars and trucks.
- For each vehicle in the dataset, we have information on various characteristics such as engine displacement and number of cylinders, along with laboratory measurements for the city and highway miles per gallon (MPG) of the car.
- We here consider the unadjusted highway MPG for 2010 cars as the response variable, and a single predictor, engine displacement.

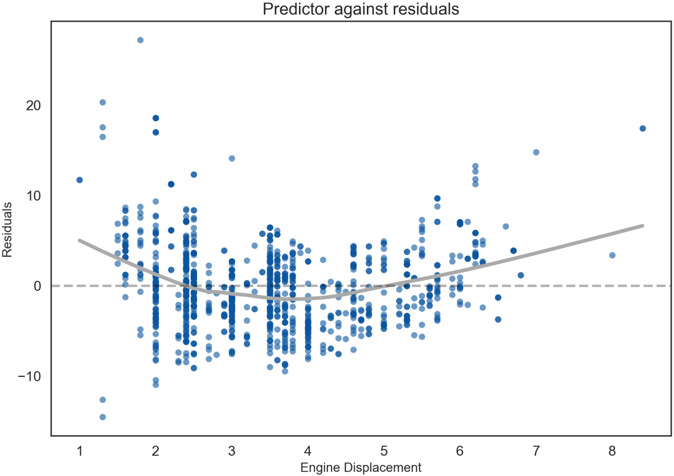
Example: Fuel Economy



Example: Fuel Economy



Example: Fuel Economy



Polynomial regression

Earlier in the unit, we studied the polynomial regression model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_d X^d + \varepsilon,$$

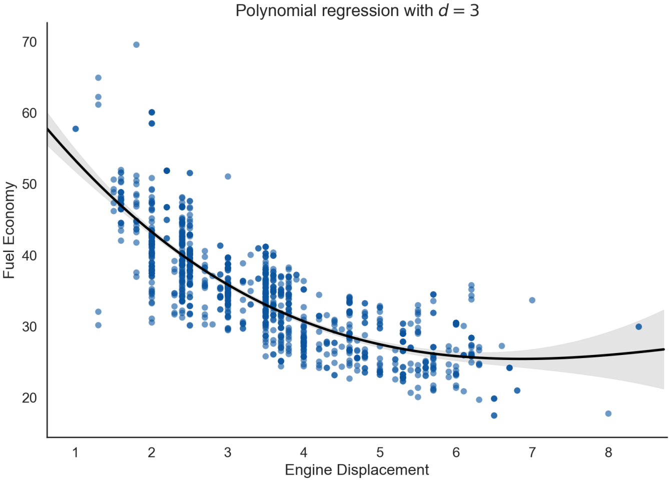
where d is the polynomial degree.

This simple approach can work well in some cases, and is a useful benchmark.

Example: Cubic Polynomial

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     FE   R-squared:                       0.682
Model:                            OLS   Adj. R-squared:                  0.681
Method:                 Least Squares   F-statistic:                     786.9
Date:                                   Prob (F-statistic):          1.84e-273
Time:                                   Log-Likelihood:                -3167.1
No. Observations:                1107   AIC:                             6342.
Df Residuals:                    1103   BIC:                             6362.
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept                66.3000      2.040     32.499      0.000      62.297      70.303
EngDispl                -14.5304      1.652     -8.797      0.000     -17.771     -11.289
np.power(EngDispl, 2)     1.6312      0.415      3.929      0.000       0.816       2.446
np.power(EngDispl, 3)    -0.0556      0.033     -1.697      0.090      -0.120       0.009
==============================================================================
Omnibus:                      111.321   Durbin-Watson:                   0.952
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              323.151
Skew:                           0.511   Prob(JB):                     6.74e-71
Kurtosis:                       5.442   Cond. No.                     1.99e+03
==============================================================================
```

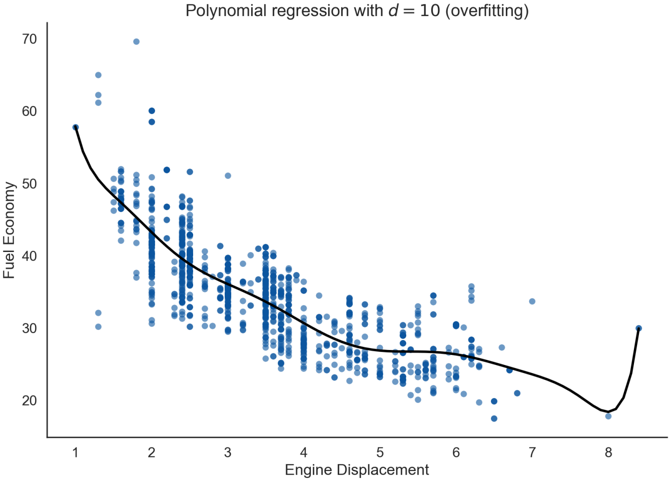
Example: Fuel Economy



Limitations of polynomial regression

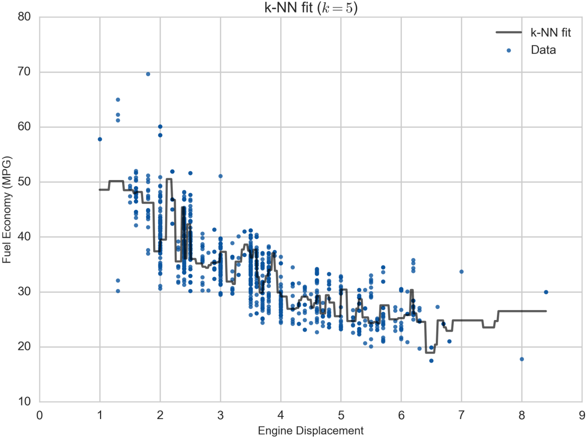
- Polynomials can overfit. A polynomial of degree d can approximate d points exactly, so that increasing d produces a wiggly curve that gets close to the data, but predicts poorly.
- Polynomials display a highly **nonlocal fit**: observations in one region, especially outliers, can seriously affect the fit in another region.
- Polynomials are unstable near the boundaries of the data. You should never extrapolate polynomial regressions to generate predictions outside the observed range of the predictor.

Example: polynomial overfitting



K-Nearest Neighbours

Another nonlinear model that we have seen is the KNN method.



The KNN method can lead to an unnecessarily discontinuous and noisy fit.

Overview

Regression splines divide the range of X into different regions and fit polynomial regressions within each, with suitable restrictions for smoothness when crossing between regions.

Smoothing splines are similar to regression splines, but arise from regularised risk minimisation where we minimise the empirical loss subject to a smoothness penalty.

Local regression fits a weighted linear regression in the neighbourhood of a prediction point x_0, where training points x_i further away from x_0 have decreasing weight.

Generalised additive models extend these methods too multiple predictors.

Basis functions

Basis functions

The key idea of several nonlinear methods (including polynomial regression, regression splines, and smoothing splines) is to augment X with additional variables which are transformations of X, and then use linear models in this new space of derived features.

Basis functions (key concept)

Let $h_m(X)$ be the deterministic mth transformation of X, $m = 1, \ldots, M$. We call h_m a **basis function**. The model

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

is a **linear basis expansion** in X.

Since the specification is linear in the basis functions h_m, all the estimation and inference tools for the linear model immediately apply to the basis function model.

Basis functions: examples

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

- $h_1(X) = X$, $M = 1$ recovers the original linear model.
- $h_1(X) = X$, $h_2(X) = X^2$, $h_3(X) = X^3$, $M = 3$ leads to a polynomial regression.
- $h_m(X) = \log(X)$, $h_m(X) = \sqrt{X}$, etc, permit other nonlinear transformations of the predictor.

Regression splines

Linear spline

A **linear spline** is a linear regression with changing slopes at discrete points. If there is one split point (called a knot) at ξ, the model is:

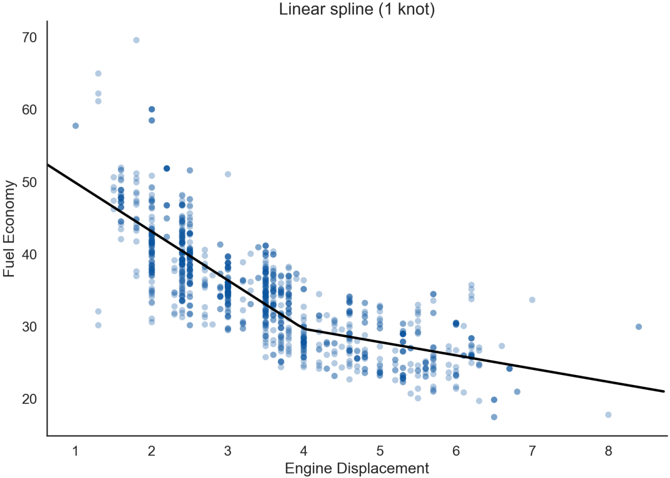
$$Y = \beta_0 + \beta_1 X + \beta_2 (X - \xi)_+ + \varepsilon,$$

where we define $(X - \xi)_+$ as

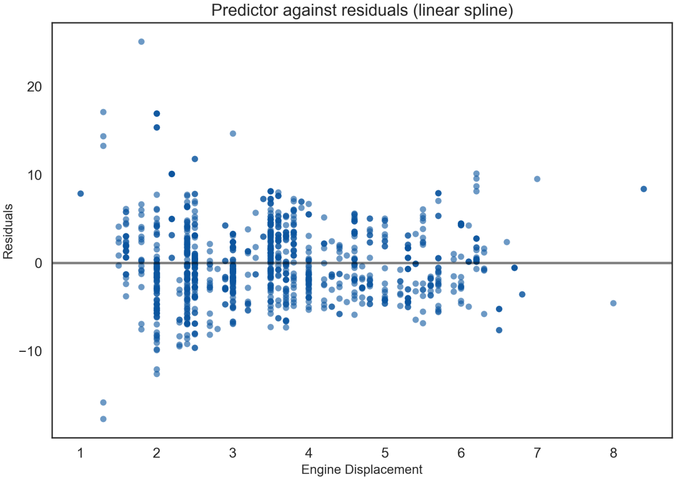
$$(X - \xi)_+ = \begin{cases} 0 & \text{if } X - \xi \leq 0 \\ X - \xi & \text{if } X - \xi > 0 \end{cases}$$

Note that $(X - \xi)_+ \equiv I(X > \xi)(X - \xi)$. The $+$ symbol means the positive part.

Example: linear spline with one knot



Example: linear spline with one knot



Regression Splines

The model

$$Y = \beta_0 + \beta_1 X + \beta_2 (X - \xi)_+ + \varepsilon$$

has interpretation

$$Y = \begin{cases} \beta_0 + \beta_1 X + \varepsilon & \text{if } X \leq \xi \\ \beta_0 + \beta_1 X + \beta_2 (X - \xi) + \varepsilon & \text{if } X > \xi \end{cases}$$

Piecewise linear regression

Our linear spline is a special case of a **piecewise linear regression** model

$$Y = \begin{cases} \beta_{01} + \beta_{11} X + \varepsilon & \text{if } X \leq \xi \\ \beta_{02} + \beta_{12} X + \varepsilon & \text{if } X > \xi \end{cases}$$

However, a linear spline restricts β_{02} and β_{12} such that the regression function is continuous (the two parts have the same value at ξ).

Piecewise linear regressions

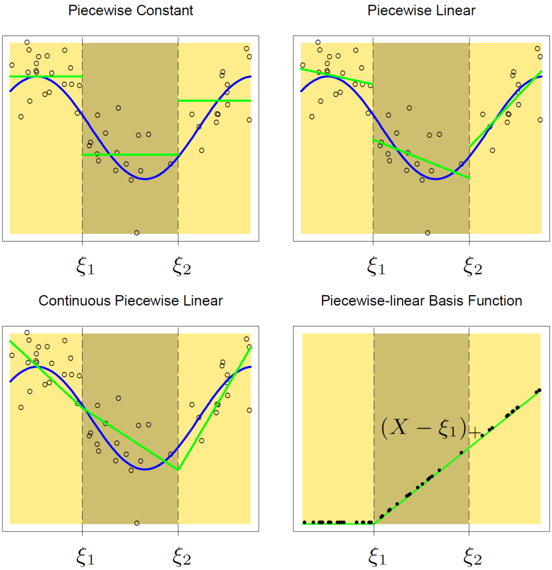
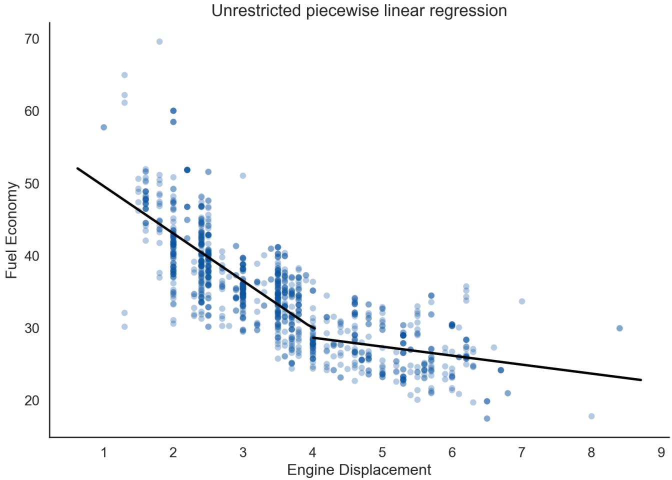


Figure from ESL

Example: piecewise linear regression

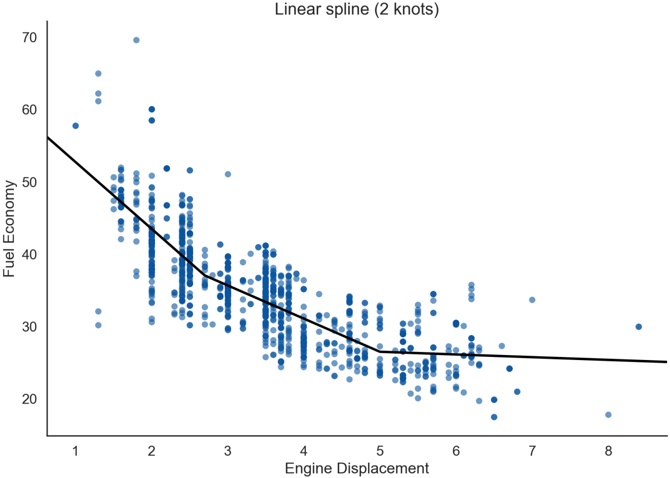


Regression Splines

The linear spline model with K knots $\xi_1, \xi_2, \ldots, \xi_K$ is:

$$Y = \beta_0 + \beta_1 X + \beta_2 (X - \xi_1)_+ + \ldots + \beta_{K+1}(X - \xi_K)_+ + \varepsilon,$$

Example: linear spline with two knots



Regression splines (key concept)

A **regression spline** is a piecewise degree-d polynomial regression that restricts the regression function $f(X)$ to be continuous and have continuous first $d-1$ derivatives.

This general approach extends the idea of splines by fitting polynomials instead of linear functions at each region.

Piecewise cubic polynomials

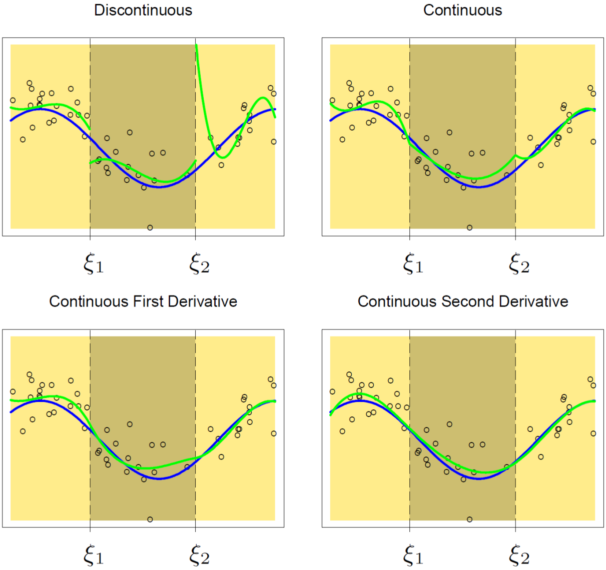


Figure from ESL

Cubic spline (key concept)

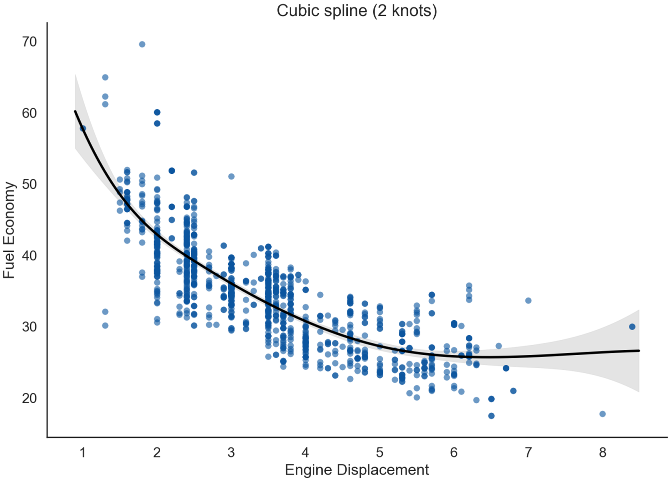
The **cubic spline** model with K knots is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (x - \xi_1)^3_+ \ldots + \beta_{K+3}(x - \xi_k)^3_+ + \varepsilon$$

For example, when there is $K = 1$ knot:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (x - \xi_1)^3_+ + \varepsilon$$

Example: Cubic Spline

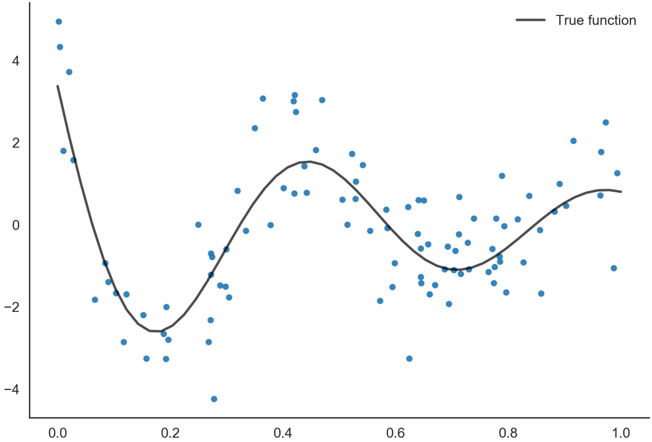


Regression splines vs polynomial regression

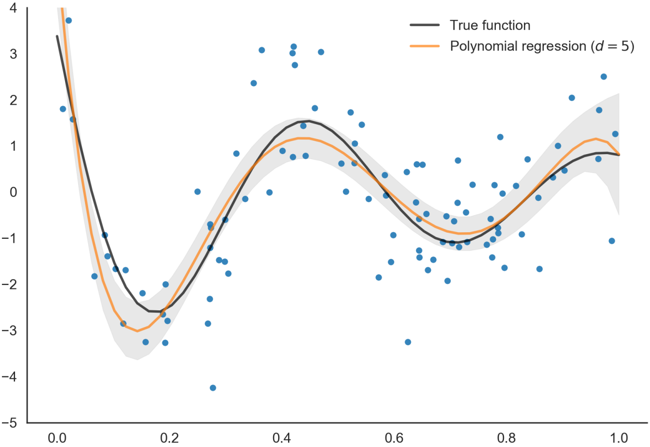
- Regression splines are preferable to polynomial regressions in most applications.

- The reason is that we can increase the flexibility of a spline by increasing the number of knots. This leads to more stable estimates compared to increasing d, which is the only option for polynomial regressions.

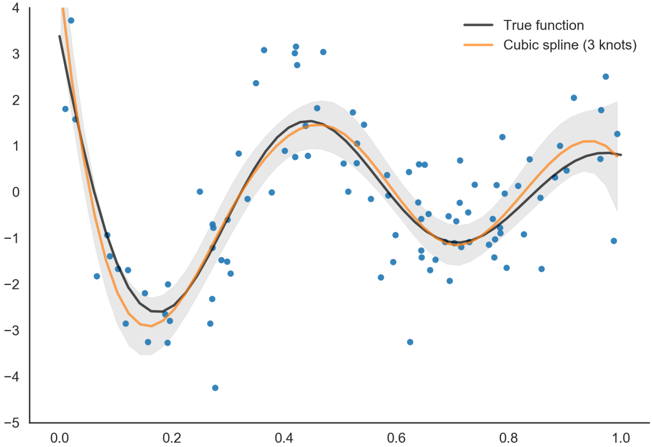
Regression splines vs polynomial regression: illustration



Regression splines vs polynomial regression: illustration



Regression splines vs polynomial regression: illustration



Natural splines

Natural splines add the constraint the regression function is required to be linear at the boundaries, generally producing more stable fits at the outer range of the predictors compared to cubic splines (which tend suffer from high variance in these regions).

Natural splines

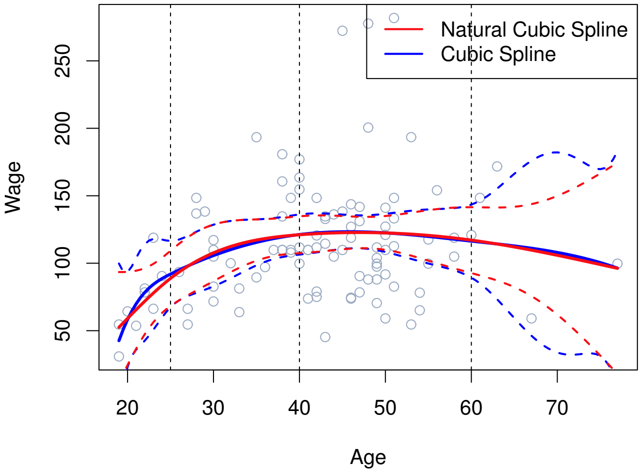


Figure from ISL

Modelling choices

- Polynomial order: linear and cubic splines are the most common choices.
- Placement of the knots: typically at uniformly spaced quantiles. For example, if there is one knot then we would place it a the sample median. If there are three, we would place them at the sample quantiles 0.25, 0.5, and 0.75.
- Number of knots: model selection.

Degrees of freedom

In the tutorials, we studied the concept of the **degrees of freedom** or **effective number of parameters** of a regression model,

$$\mathsf{df}(\widehat{Y}) = \frac{\sum_{i=1}^{N} \mathsf{Cov}(\widehat{Y}_i, Y_i)}{\sigma^2}.$$

It is very useful to parameterise regression splines by their degrees of freedom. A cubic spline with K interior knots uses $4 + K$ degrees of freedom (including the constant) while a natural cubic spline uses K.

We perform model selection by varying the degrees of freedom.

Smoothing splines

Smoothing splines (key concept)

A **smoothing spline** solves the regularised risk minimisation problem

$$\min_{f} \sum_{i=1}^{N} [y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 dx,$$

where $\lambda \geq 0$ is a fixed smoothness parameter, $f''(x)$ is second derivative of the function, and $\int [f''(x)]^2 dx$ measures the curvature of the function.

This problem is defined in an infinite dimensional function space. The minimisation is in terms of a function f, rather than parameters.

Special cases

Smoothing spline:

$$\min_f \sum_{i=1}^{N} [y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 dx,$$

- $\lambda = 0$: the solution f can be any function that interpolates (passes) the data points. There are many such solutions.

- $\lambda = \infty$: the simple least squares line fit, since no second derivative can be tolerated ($f''(x) \equiv 0$, hence the model is linear).

- $\lambda \in (0, \infty)$ allows functions that vary from very rough to very smooth, and our hope is that in between there is an interesting class of functions that lead to accurate predictions.

Smoothing spline

- Remarkably, the unique minimiser is a natural cubic spline with knots at $x_1, x_2, \ldots, x_N$.

- Even though there are N parameters (one for each observation) the penalty term translates to a penalty on the spline coefficients, which are shrunk towards the linear fit.

- Because the solution is a natural spline, we can write it as

$$f(x) = \sum_{i=1}^{N} \theta_i N_i(x)$$

where $N_i(x)$ is the natural spline basis at the knot x_i.

Degrees of freedom

Because the smoothing spline translates to regularised regression on a linear basis expansion, we can show that the fit is linear in $\boldsymbol{y}$,

$$\widehat{\boldsymbol{y}}_\lambda = \boldsymbol{S}_\lambda \boldsymbol{y},$$

where the $N \times N$ matrix $\boldsymbol{S}_\lambda$ is known as the **smoother matrix**.

The degrees of freedom (effective number of parameters) of the smoothing spline is

$$\mathrm{df}_\lambda = \operatorname{tr} \boldsymbol{S}_\lambda = \sum_{i=1}^{N} \{\boldsymbol{S}_\lambda\}_{ii},$$

where $\{\boldsymbol{S}_\lambda\}_{ii}$ is the i-th diagonal element of $\boldsymbol{S}_\lambda$.

Model selection

We can compute the leave-one-out cross validation error very efficiently as

$$\mathsf{MSE}_{\mathsf{cv}}(\lambda) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{f}^{(-i)})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \widehat{f}_\lambda(x_i)}{1 - \{\boldsymbol{S}_\lambda\}_{ii}}\right)^2$$

where $\widehat{f}^{-i}$ is the estimator when leaving observation i out.

Local regression

Local regression (key concept)

Local regression or **kernel smoothing** methods fit a regression model using only nearby observations, where the training cases are assigned weights that dies off smoothly with distance from the prediction point.

Like the KNN method, local regression is a memory based procedure since we use all the training data to fit a model each time we wish to compute a prediction.

Local linear regression

Algorithm Local linear regression

```
Gather a fraction $s = k/N$ of training points whose $x_i$ are closest to the prediction point $x_0$.
Assign a weight $K_\lambda(x_i, x_0)$ to each point in the neighbourhood, where $\lambda$ is a tuning parameter. The furthest point from $x_0$ has weight zero, and the closest point has highest weight. All training points outside the neighbourhood have weight zero.
Fit the weighted least squares regression
    $\widehat{\beta}_0, \widehat{\beta}_0 = \underset{\beta_0,\beta_1}{\text{argmin}} \sum_{i=1}^{N} K_\lambda(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i)^2$
The prediction is $\widehat{f}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$.
```

Local average

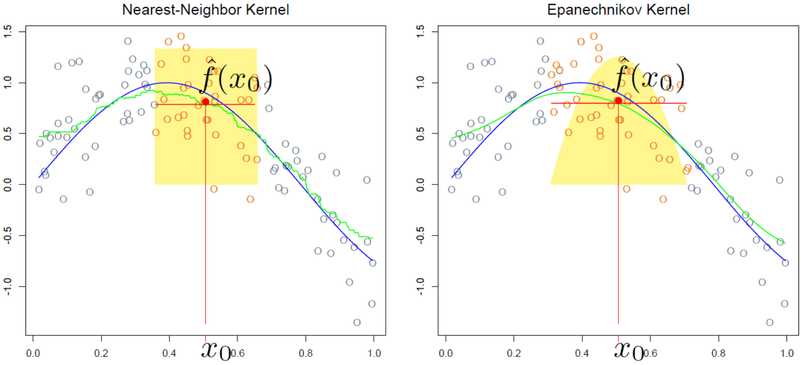


Figure from ESL

Local linear regression

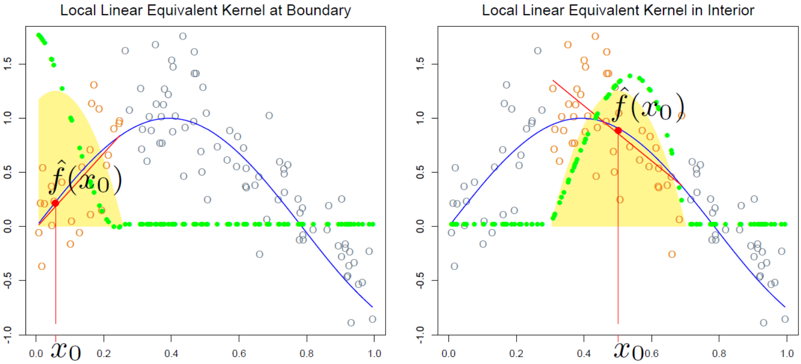


Figure from ESL

Local quadratic regression

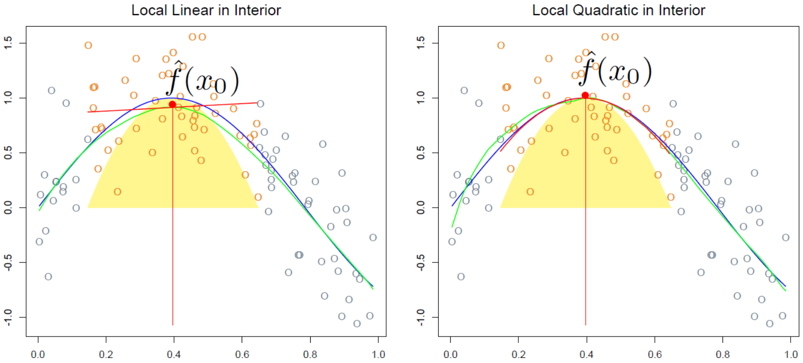


Figure from ESL

Modelling choices

1. The **kernel function** (weight function) $K_\lambda(x_i, x_0)$.

2. Whether to fit an average, linear regression, or polynomial regression in the neighbourhood.

3. The tuning parameter λ, which controls how local the method is. This is the most crucial choice.

Generalised additive models

Generalised additive model (key concept)

The **generalised additive model** (GAM) is

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p) + \varepsilon,$$

where we choose an appropriate fitting function $f_j(X_j)$ for each predictor (linear, polynomial, regression spline, smoothing spline, local regression, etc).

Generalised additive models: advantages

- GAMs allow us to fit nonlinear functions $f_j(X_j)$ for each predictor, so that we can automatically model nonlinear relationships missed by a linear regression. The nonlinear fits can potentially lead to higher predictive accuracy.

- GAMs leverage the advantages of its building blocks into a convenient framework for multiple predictors.

- Because of the additive structure of the model, it is still easy for us to interpret the relationship between each X_j and Y conditional on other predictors.

- We can summarise each function $f_j(X_j)$ by the degrees of freedom.

Generalised additive models: limitation

The main limitation of GAMs is that they do account for interactions. Interactions arise when we consider general multivariate functions such as

$$Y = \beta_0 + f(X_1, X_2) + \varepsilon.$$

One option in this case is to use a bivariate polynomial regression as approximation to $f(X_1, X_2)$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon$$

Interactions

- We could go even further and consider functions of many variables $f(X_1, \ldots, X_p)$, but doing so leads to a **curse of dimensionality**.

- In this setting, the curse of dimensionality refers to problem that the number of parameters in flexible approximations to $f(X_1, \ldots, X_p)$ (such as polynomials) grows exponentially with p.

- Without such a structure, you should focus on the GAM unless domain knowledge suggests specific interaction effects.

Review questions

- What are the main disadvantages of polynomial regression?
- What are regression splines?
- What is a smoothing spline?
- What is a local regression?
- What is a generalised additive model?