Moving Beyond Linearity

Rosalba Radice

Analytics Methods for Business

Non-Linear Models

- Linear models are relatively simple, and have advantages over other approaches in terms of interpretation and inference.
- More sophisticated models which account for non-linearity are polynomial regression and generalized additive models (GAMs).
- Polynomial regression extends the linear model by adding extra predictors, obtained by raising each of the original predictors to a power. For example, a cubic regression uses three variables, X, X^2, and X^3, as predictors. This approach provides a simple way to provide a nonlinear fit to data.
- GAM is an additive modeling technique where the impact of the predictive variables is captured through smooth functions.

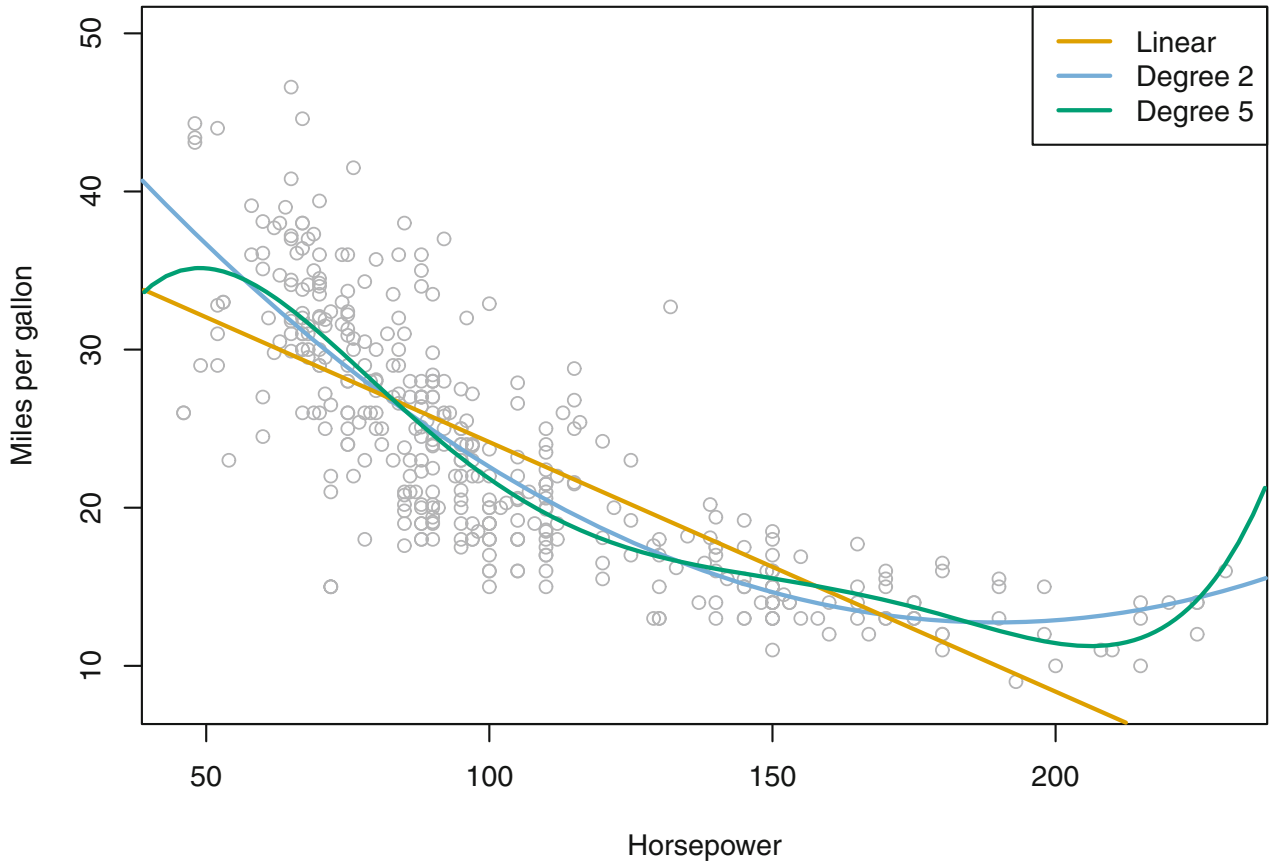
Polynomial Regression

In general a polynomial regression can be written as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$

- For large enough degree d, a polynomial regression allows us to produce an extremely non-linear curve.
- The coefficients can be easily estimated using least squares linear regression.
- Generally speaking, it is unusual to use d greater than 3 or 4 because for large values of d, the polynomial curve can become overly flexible and can take on some very strange shapes.
- Polynomials do not seem to have good properties in terms of approximating an underlying function over its whole domain.

The Auto Data Example



The `Auto` data set. For a number of cars, `mpg` and `horsepower` are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes $\texttt{horsepower}^2$ is shown as a blue curve. The linear regression fit for a model that includes all polynomials of `horsepower` up to fifth-degree is shown in green.

Introducing a Smooth Function

Let's rewrite the simple regression model in a slight more general way.

So replace

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

with

$$y_i = \beta_0 + f(x_i) + \epsilon_i,$$

where f is a smooth function of x. We can approximate $f(x_i)$ by

$$f(x_i) = \sum_{k=1}^{d} \beta_k b_k(x_i),$$

where the b_k are known basis functions, and β_k unknown regression parameters.

Introducing a Smooth Function - continued

If $f(x_i)$ is believed to be a 3^{rd} order polynomial (so $d = 3$) then

$$b_1(x_i) = x_i, \;\; b_2(x_i) = x_i^2, \;\; b_3(x_i) = x_i^3,$$

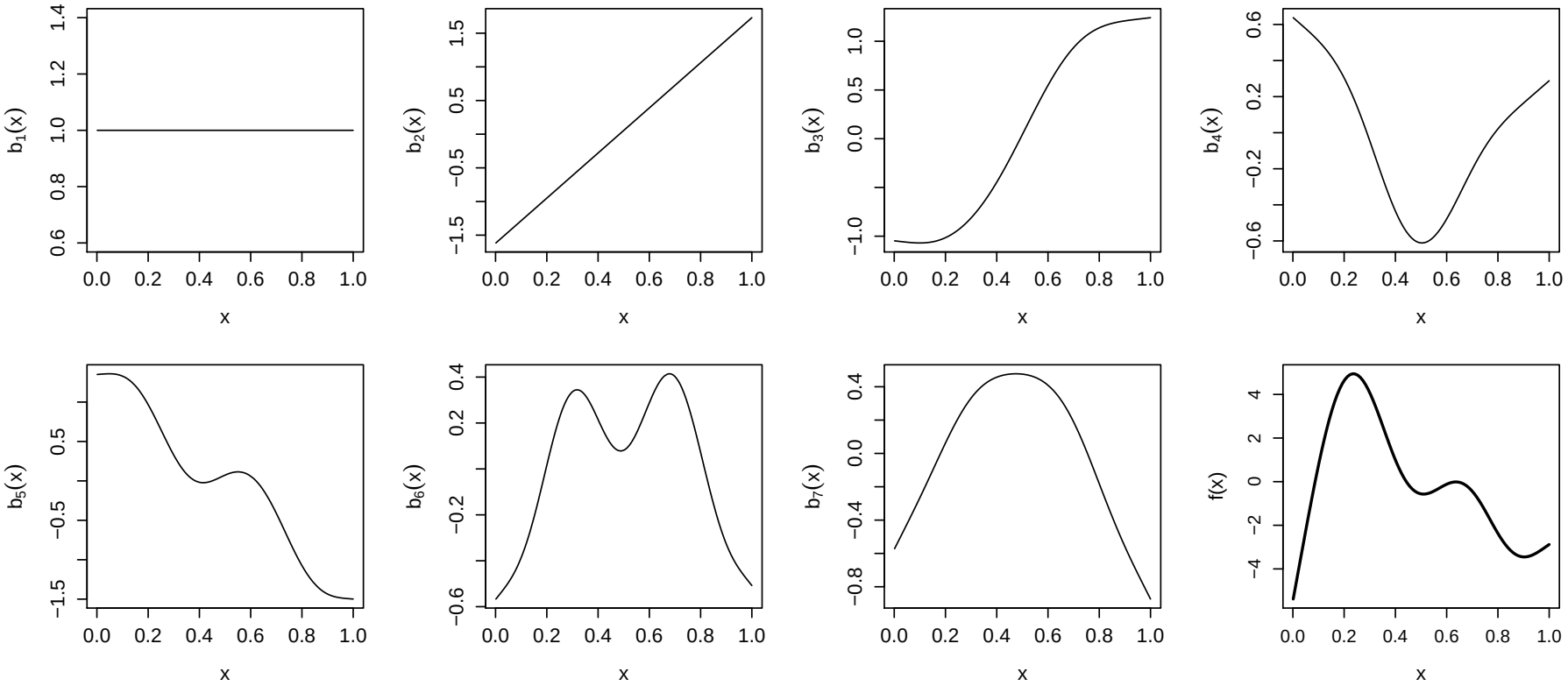
which brings us back to a polynomial model;

$$f(x_i) = \sum_{k=1}^{3} \beta_k b_k(x_i) = \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3.$$

Introducing a Smooth Function - continued

- However, polynomials have problems when approximating the true function.
- In particular their approximations of unknown true functions are concerned with getting good approximations in the vicinity of some particular point of interest, approximation will eventually become very poor as we move away from that point.
- For this reason we can use bases that have better theoretical and numerical properties.
- An example of basis with good theoretical and computational properites are the thin plate regression spline (TPRS) basis.

A Thin Plate Regression Spline (TPRS) Basis



TPRS: seven basis functions and a smooth function of one variable.

Number of Basis

How should we choose d, the number of basis?

A convenient and theoretically founded way of choosing the degree of smoothing is to keep the basis dimension d fixed at a reasonable large number and add a roughness penalty to the least squares fitting objective. That is, minimise

$$\sum_{i=1}^{n} \{y_i - \beta_0 - f(x_i)\}^2 + \lambda \int [f''(x)]^2 dx.$$

The second term measures the roughness of the smooth functions, $\lambda \geq 0$ is the smoothing parameter.

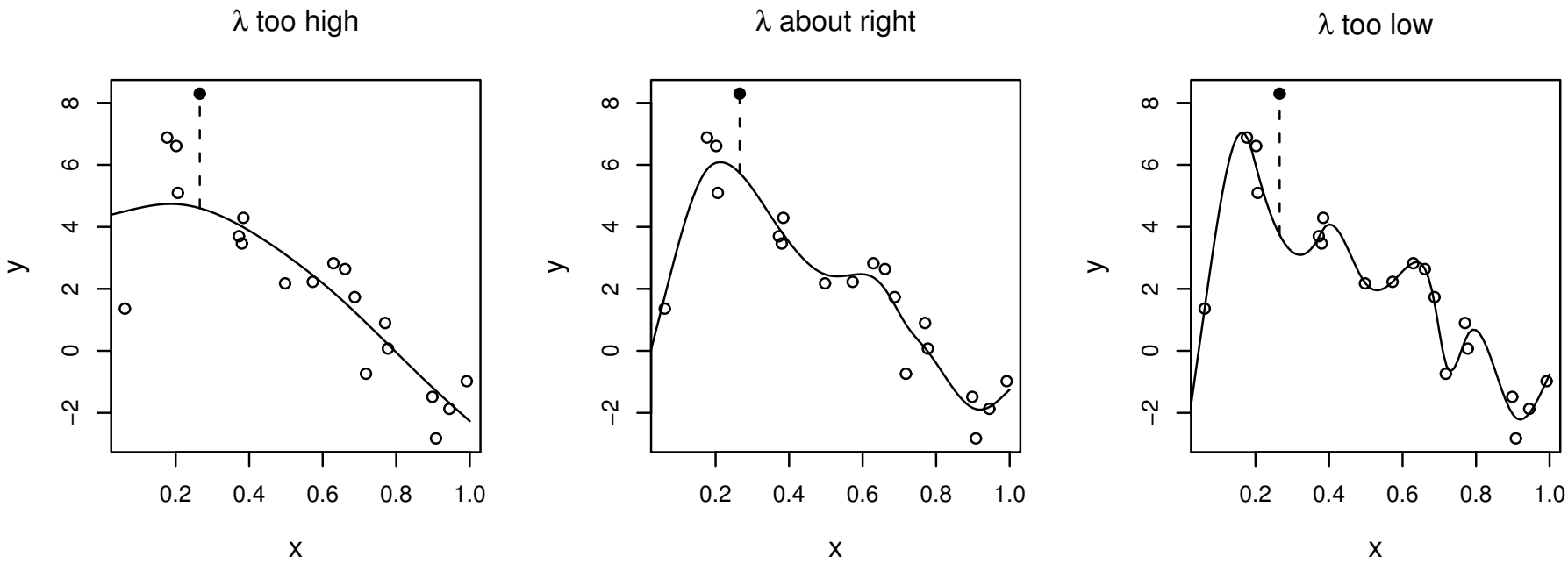
Smoothing Parameter

The trade off between fit and smoothness is controlled by smoothing parameter λ. In particular:

- $\lambda \to \infty$ leads to a straight line estimate; all parameters associated with the wiggly bases will be estimated as 0.
- $\lambda = 0$ gives an un-penalized (potentially very wiggly) regression spline estimate; the parameters of all wiggly bases will be kept in the model.

So it is really crucial to be able to tune λ such that a balance in terms of fit and smoothness can be achieved.

Smoothing Parameter Estimation



In this case the fifth datum (•) has been omitted from fitting and the continuous line shows a penalized regression spline fitted to the remaining data (◦). When the smoothing parameter is too high the spline fits many of the data poorly and does no better with the missing point. When λ is too low the spline fits the noise as well as the signal and the extra variability that this induces causes it to predict the missing datum rather poorly. For the intermediate λ the spline is fitting the underlying signal quite well, but smoothing through the noise: as a result the missing datum is reasonably well predicted.

Some Remarks

- The number of *edf* of a spline estimate is defined in such a way that the role of the penalty in estimation is acknowledged. The higher the *edf* value the higher the complexity of the estimated smooth function. For example, $edf = 1$ indicates that the estimated function is linear.
- Point-wise Bayesian confidence intervals at the points on the estimated function can be easily constructed.
- Testing null hypothesis that a linear model is appropriate, $H_0 : f = 0$, can be achieved by employing an F-test.
- Model comparison can be achieved using AIC
- Model assumptions can be tested using residual analysis.
- Many of the definitions and results showed here can be extended to the case where there are multiple covariates.

Generalized Additive Models

- If the response variable follows an exponential family distribution and we want to model non-linearities with smooth functions, then we can use GAMs.

- All the description and results that we have seen for the additive models can be adapted to GAMs.

- However, because we are dealing with non-Gaussian responses, then the penalised least squares fitting objective

$$\sum_{i=1}^{n} \{y_i - \beta_0 - f(x_i)\}^2 + \lambda \int [f''(x)]^2 dx,$$

can be replaced with

$$\ell + \lambda \int [f''(x)]^2 dx,$$

where ℓ is the likelihood function of the model.

Generalized Additive Models - continued

- Maximization of the penalised likelihood produces estimates which can be used to construct the estimated curve.
- λ can be estimated using modifications of the methods seen previously.
- This approach can be also be generalised to the case with many and different covariates.