MATH5824 Generalized Linear and Additive Models

Robert G Aykroyd

1/13/23

Table of contents

Weekly schedule

! Important

Our regular class times are:

Tuesday 11-12, Roger Stevens, LT25 (with MATH3823)
Thursday 2-3, Roger Stevens, LT23 (with MATH3823)
Friday 13-14, Roger Stevens, LT22 (MATH5824 only)

i Week 1 (30 January - 3 February)

- **Before next Lecture:** Please read MATH3823 *Preface*.
- **Lecture on Tuesday:** We will briefly cover all material in MATH3823 *Chapter 1: Introduction*.
- **Before next Lecture:** Please re-read MATH3823 *Chapter 1* carefully.
- **Lecture on Thursday:** Start MATH3823 *Chapter 1: Essentials of Normal Linear Models*.

- **Before next Lecture:** Please read MATH5824 *Preface*.
- **Lecture on Friday:** Start MATH5824 *Chapter 1: Introduction to Non-linear Modelling* by briefly considering *1.1 Overview*.

- **Weekly feedback:** Self-study the Exercises in MATH3823 *Section 1.5* – solutions to be posted during Week 1.

i **Week 2 (6 - 10 February)**

- Details will be added during Week 1.

Coursework Practical Sessions (20 - 24 March)

- Details to follow in early March.

Preface

These lecture notes are produced for the University of Leeds module "MATH5824 - Generalized Linear and Additive Models" for the academic year 2022-23. They are based on those used previously for this module and I am grateful to previous module lecturers for their considerable effort: Lanpeng Ji, Amanda Minter, John Kent, Wally Gilks, and Stuart Barber. This is the first year, however, that they have been produced in accessible format and hence some errors might occur during this conversion process. For information, I am using Quarto (a successor to RMarkdown) from RStudio to produce both the html and PDF, and then GitHub to create the website which can be accessed at rgaykroyd.github.io/MATH3823/. Please note that the PDF versions will only be made available on the University of Leeds Minerva system. Although I am a long-term user of RStudio, I have not previously used Quarto/RMarkdown nor Github and hence please be patient if there are hitches along the way.

In the Level 3 component of this module, we extend the simple linear regression model to the generalized linear model which can cope with non-normally distributed response variables, in particular data following binomial and Poisson distributions. However, we still just use linear functions of the predictor variables. A further extension of the linear model is the generalized additive model. Here, we no longer insist on the predictor variables affecting the response via a linear function of the predictors, but allow the response to depend on a more general smooth function of the predictor. In the Level 5 component of this module, we study splines and their use in interpolating and smoothing the effects of explanatory variables in the generalized linear models of the Level 3 component of this module (see separate Lecture Notes accompanying MATH3823).

RG Aykroyd, Leeds, November 22, 2022

1 Introduction to Non-linear Modelling

1.1 Overview

Table 1.1 reports on the depth of a coal seam determined by drilling bore holes at regular intervals along a line. The depth y at location $t = 6$ is missing: could we estimate it?

Table 1.1: Coal-seam depths (in metres) below the land surface at intervals of 1 km along a linear transect.

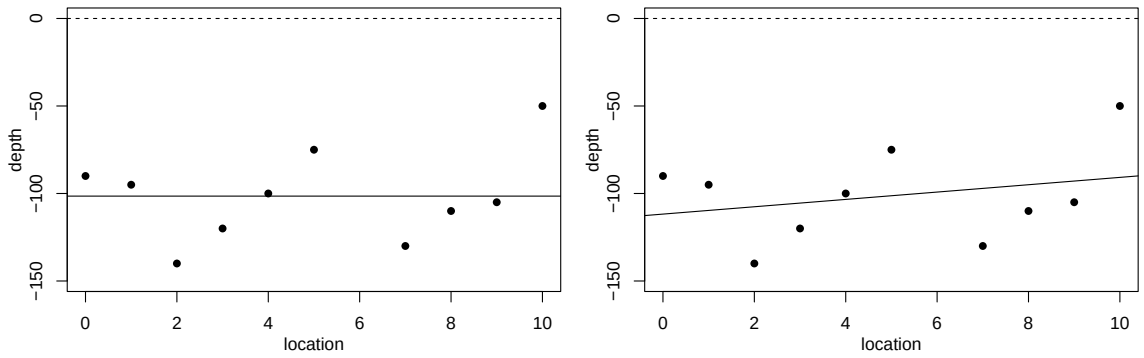
Location, t	0	1	2	3	4	5	6	7	8	9	10
Depth, y	-90	-95	-140	-120	-100	-75	NA	-130	-110	-105	-50

Figure Figure 1.1 plots these data, superimposed with predictions from several polynomial regression models.

Each of these models would predict a different value for the missing observation y_6. We do not know the accuracy of the depth measurements, so in principle any of these curves could be correct. Clearly, the residual variance is largest for the constant-depth model in Figure 1.1a, and smallest for the cubic polynomial in Figure 1.1c. However, none of these models produces a convincingly good fit. Moreover, these models are not particularly believable, since we know that geological pressures exerted over very long periods of time cause the landscape and its underlying layers of rock to undulate and fracture. This suggests we need a different strategy.

Next, consider the simulated example in Figure 1.2. At first look we might be happy with the fitted curves in Figure 1.2a or Figure 1.2b. The data, however, are created with a *change-point* at $x = 0.67$ where the relationship changes from linear with slope 0.6 to a constant value of 0.75. This description is completely lost with these two models.

Figure 1.2c shows the result of fitting one linear function to the data below 0.67 and a second linear function above. Clearly, this fits well but it has assumed that the change-point location is known – which is unrealistic. Finally, Figure 1.2d shows a fitted *cubic smoothing spline* to the data – we will studies these models later. This shows an excellent fit and leads to appropriate conclusions. That is, the relationship is approximately linear for small values, then there is a rapid increase, and finally a near constant value for high values. Of course, this is not exactly as the true relationship with a discontinuity at $x = 0.67$ but



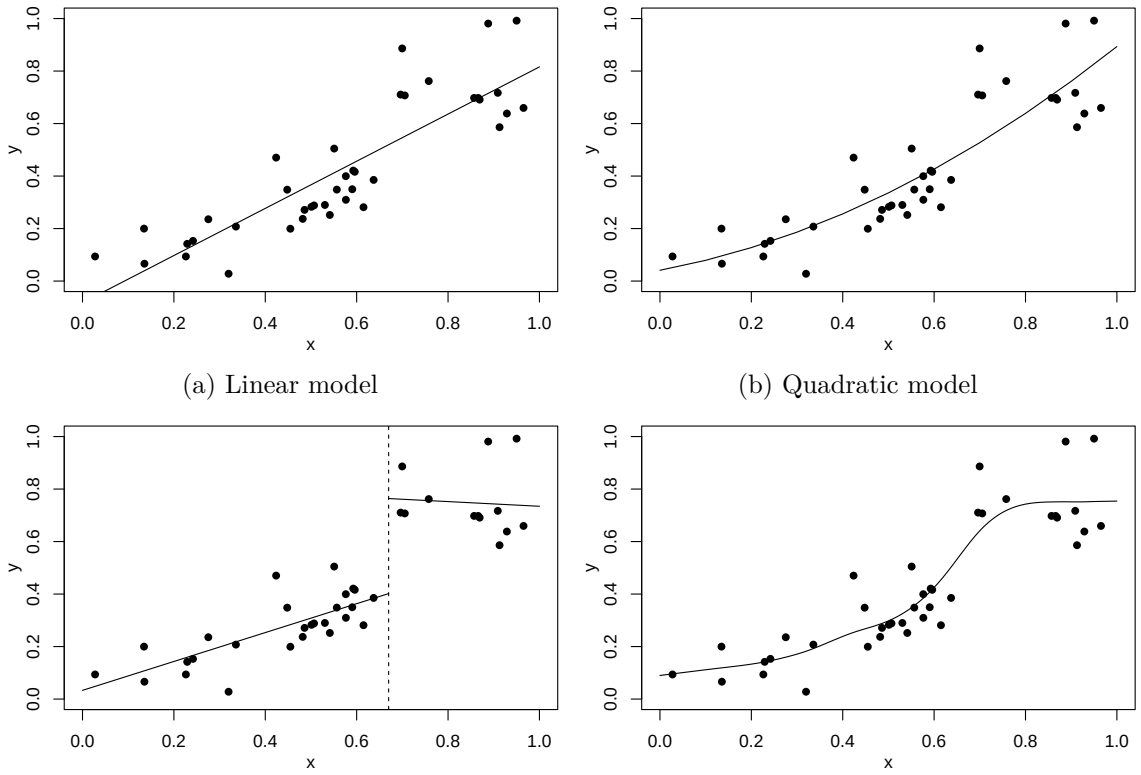
(a) Constant model

(b) Linear model

(c) Quadratic model

(d) Cubic model

Figure 1.1: The coal-seam data superimposed with predictions from polynomial regression models.



(a) Linear model

(b) Quadratic model

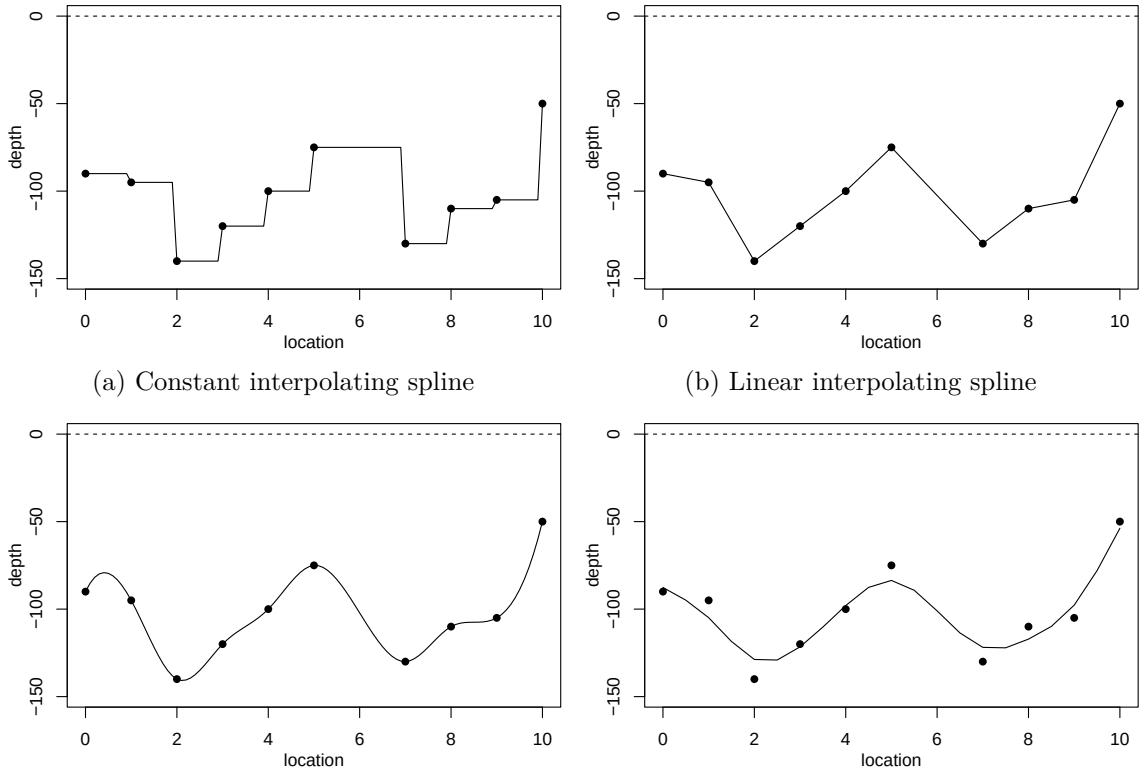
(c) Piecewise linear model

(d) Cubic smoothing spline

Figure 1.2: Simulated data superimposed with predictions from various models.

it would definitely suggest something extreme occurs between about 0.6 to 0.7. Full details will follow later, but the cubic spline fits local cubic polynomials which are constrained to create a continuous curve.

Now returning to the coal seam data. Figure 1.3 shows the data again, superimposed with predictions from methods which are not constrained to produce such smooth curves.



(a) Constant interpolating spline

(b) Linear interpolating spline

(c) Cubic interpolating spline

(d) Cubic smoothing spline

Figure 1.3: The coal-seam data superimposed with predictions from various spline models.

The simplest method, *constant-spline interpolation*, assumes that the dependent variable remains constant between successive observations, with the result shown in Figure 1.3a. However, the discontinuities in this model make it quite unreliable. A better method, whose results are shown in Figure 1.3b, is *linear-spline interpolation*, which fits a straight line between successive observations. Even so, this method produces discontinuities in the *gradient* at each data point. A better method still, shown in Figure 1.3c, is *cubic spline interpolation*, which fits a cubic polynomial between successive data points such that both the gradient and the curvature at each data point is continuous.

A feature of all these interpolation methods is that they fit the data exactly. Is this a good thing? The final method assumes that there may be some measurement error in the observations, which justifies fitting a smoother cubic spline than the cubic interpolating spline, but as we see in Figure 1.3d which does not reproduce the data points exactly. Is

this a bad thing? We will see during this module how to construct and evaluate these curves. Here, the results are presented only for motivation.