MATH5824 Generalised Linear and Additive Models

Robert G Aykroyd

2024-01-31

Table of contents

Weekly schedule

Items will be added here week-by-week and so keep checking when you need up-to-date information on what you should be doing. Note that items specific to *MATH5824M* will be marked accordingly, otherwise items refer to the material common with *MATH3823*.

ℹ Week 1 (29 January - 2 February)

- **Before first Lecture:** Please read the *Overview*.
- **Lecture on Tuesday:** We will briefly cover all material in *Chapter: Introduction*.
- **Before next Lecture:** Please re-read *Chapter 1* carefully, especially any sections not covered in Lectures.
- **Lecture on Thursday:** Start *Chapter 2: Essentials of Normal Linear Models* with *Section 2.1: Overview* & *Section 2.2: Linear models*.
- **Before next Lecture:** Please read the *MATH5824M Overview*.
- **Lecture on Friday:** Cover the whole of *MATH5824M Chapter 1: Non-parametric Modelling*.
- **Weekly feedback:** Complete the Chapter 1 Quizzes and self-study the Exercises in *Section 1.5* – solutions to be added during Week 1. If you have time, then also self-study the *MATH5824M Exercises* in *Section 1.4*.

ℹ Advanced notice

- **Module Assessment:** Set on 14 March with submission deadline 23 April (that is after the break). You will be expected to write a short report based on an RStudio practical.
- **Computer classes:** 27/28 February for Practice and 19/20 March for Assessment – check your timetable.
- **Generative AI usage within this module:** The assessments for this module fall in the red category for using Generative AI which means you must not use Generative AI tools. The purpose and format of the assessments makes it inappropriate or impractical for AI tools to be used.

Provisional Weekly Lecture Schedule

Week 1	Chapter 1	All
Week 2	Chapter 2	All
Week 3	Exercises	Exercises 1, 2
	Chapter 3	Sections 3.1-3.3
Week 4		Sections 3.4-3.5
	Exercises	Exercises 3
Week 5	Chapter 4	Sections 4.1-4.3
Week 6		Sections 4.4-4.5
Week 7	Chapter 5	Sections 5.1-5.3
Week 8		Sections 5.4-5.6
	Exercises	Exercises 4, 5
Easter		
Week 9	Chapter 6	All
Week 10	Exercises	Exercises 6
Week 11	Revision	

Overview

Preface

These lecture notes are produced for the University of Leeds module "MATH5824 - Generalized Linear and Additive Models" for the academic year 2023-24. They are based on the lecture notes 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 year, again, 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/MATH5824/. 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 am a novice at Quarto/RMarkdown and a complete beginner using Github and hence please be patient if there are hitches along the way.

RG Aykroyd, Leeds, January 22, 2024

Changes since last year

Feedback from the students last year was very positive, but there were consistent comments regarding two issues: (1) a shortage of practice exercises and the opportunity to discuss these in class, and (2) limited RStudio support in preparation for the assessment. For the first of these, additional exercises have been prepared and are included in the learning material. Also, I am trying some short quizzes so that you can check your basic knowledge. Further, I intend to set-aside some lecture time for us to discuss selected exercises. For the second, an additional computer session has been added, in Week 5 (26 February - 1 March), this is 3 weeks before the assessed practice in Week 8 (18 - 22 March). Further, a few new instructional videos will be available addressing some RStudio topics. Together, these represents a considerable about of extra work for me, but I hope that they are helpful and so please give your feedback whenever there is an opportunity.

Generative AI usage within this module

The assessments for this module fall in the red category for using Generative AI which means you must not use Generative AI tools. The purpose and format of the assessments makes it inappropriate or impractical for AI tools to be used.

⚠ Warning

Statistical ethics and sensitive data
Please note that from time to time we will be using data sets from situations which some might perceive as sensitive. All such data sets will, however, be derived from real-world studies which appear in textbooks or in scientific journals. The daily work of many statisticians involves applying their professional skills in a wide variety of situations and as such it is important to include a range of commonly encountered examples in this module. Whenever possible, sensitive topics will be signposted in advance. If you feel that any examples may be personally upsetting then, if possible, please contact the module lecturer in advance. If you are significantly effected by any of these situations, then you can seek support from the Student Counselling and Wellbeing service.

Official Module Description

Module summary

Linear regression is a tremendously useful statistical technique but is limited to normally distributed responses. Generalised linear models extend linear regression in many ways - allowing us to analyse more complex data sets. In this module we will see how to combine continuous and categorical predictors, analyse binomial response data and model count data.A further extension is the generalised 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.

Objectives

On completion of this module, students should be able to:

- carry out regression analysis with generalised linear models including the use of link functions, deviance and overdispersion;
- fit and interpret the special cases of log linear models and logistic regression;
- compare a number of methods for scatterpot smoothing suitable for use in a generalised additive model;
- use a backfitting algorithm to estimate the parameters of a generalised additive model;
- interpret a fitted generalised additive model;
- use a statistical package with real data to fit these models to data and to write a report giving and interpreting the results.

Syllabus

Generalised linear model; probit model; logistic regression; log linear models; scatterplot smoothers; generalised additive model.

University Module Catalogue

For any further details, please see MATH5824 Module Catalogue page

1 Non-parametric Modelling

1.1 Introduction

Here is a short video [3 mins] to introduce the chapter.

In the Level 3 component of this module, we extend the simple linear regression model to the generalised 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 generalised 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 generalised linear models of the Level 3 component of this module (see separate Lecture Notes accompanying MATH3823). Towards the end of the material, we will learn that the fitting of generalised additive models is a straightforward extension of what is learnt in MATH3823.

Outline of the additional material in MATH5824 compared to MATH3823:

1. Interpolating and smoothing splines.
2. Cross-validation and fitting splines to data.
3. The generalised additive model.

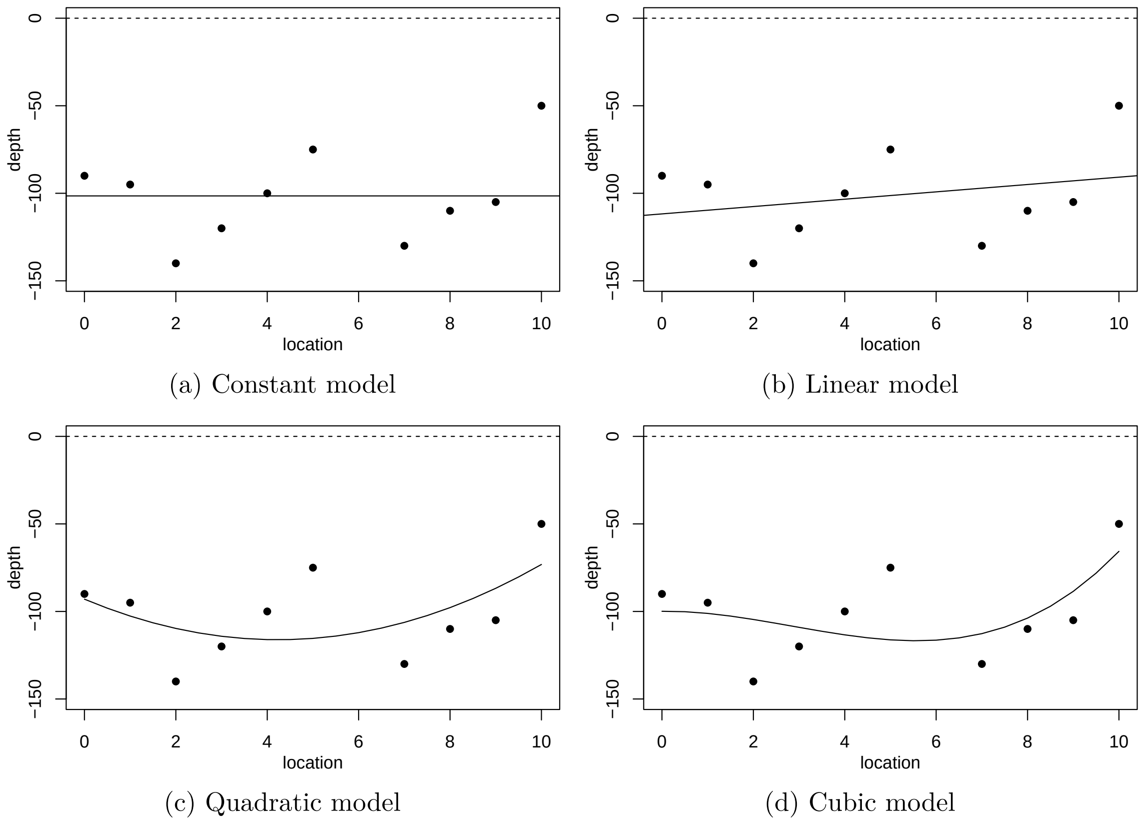
1.2 Motivation

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 $x = 6$ 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, x	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 1.1 plots these data, superimposed with predictions from several polynomial regression models.



(a) Constant model

(b) Linear model

(c) Quadratic model

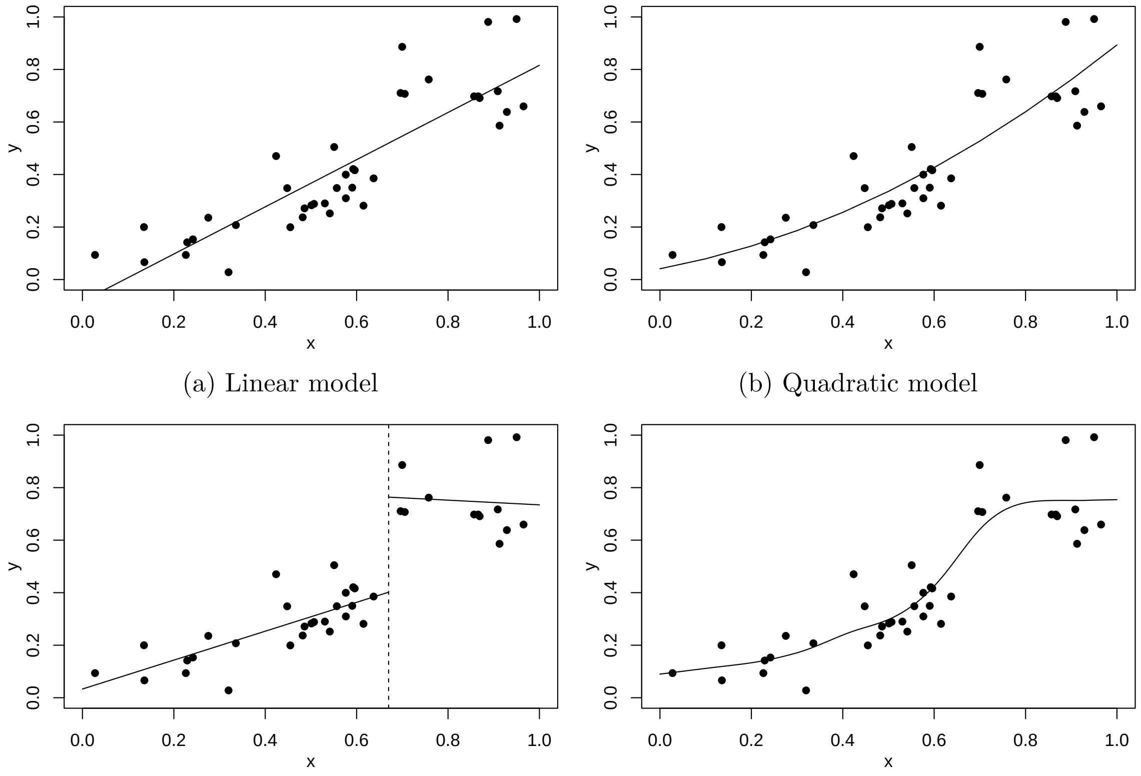
(d) Cubic model

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

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*



(a) Linear model

(b) Quadratic model

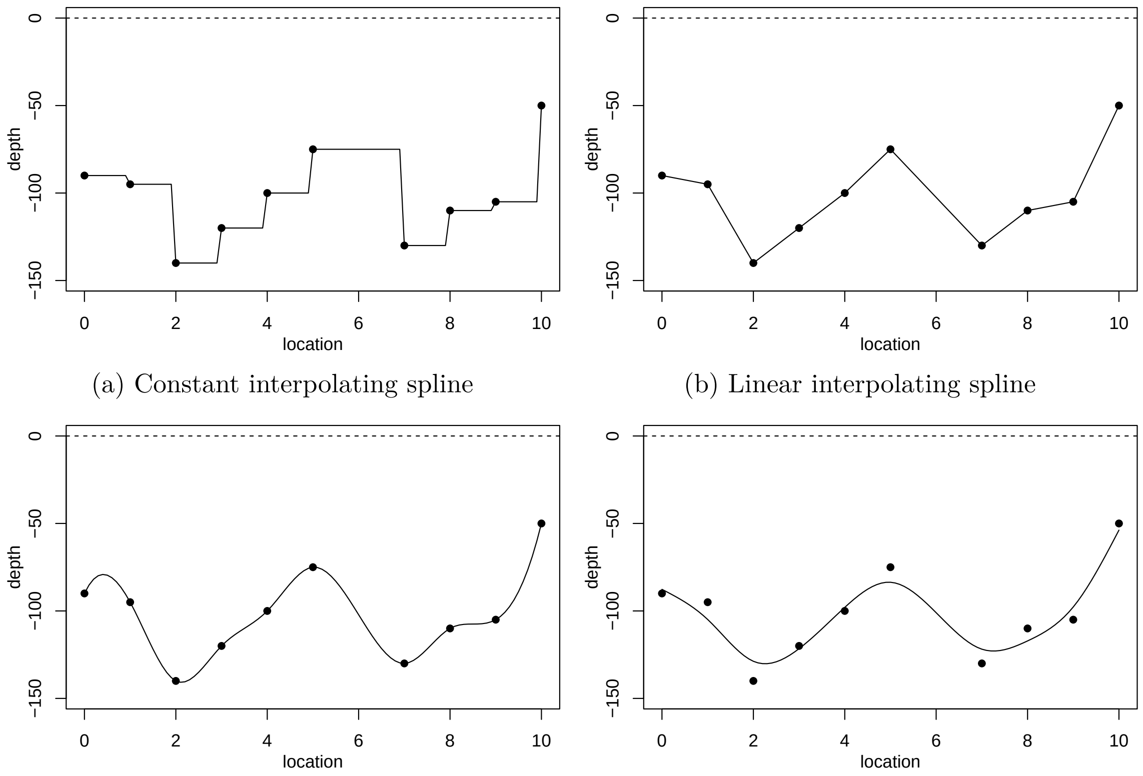
(c) Piecewise linear model

(d) Cubic smoothing spline

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

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 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.

Focus on polynomials quiz

Test your knowledge recall and comprehension to reinforce basic ideas about continuity and differentiability.

1. Which of the predicted curves in Figure 1.3 are continous?

- (A) None of the models

- (B) Model (a) only

- (C) Models (c) and (d) only

- (D) All models

2. Which of the predicted curves in Figure 1.3 have a continous first derivative?

- (A) None of the models

- (B) Model (b) only

- (C) Models (c) and (d) only

- (D) All models

3. Which of the predicted curves in Figure 1.3 have a continous second derivative?

- (A) None of the models

- (B) Model (b) only

- (C) Models (c) and (d) only

- (D) All models

4. Which of the predicted curves in Figure 1.3 have a continous third derivative?

- (A) None of the models

- (B) Model (b) only

- (C) Models (c) and (d) only

- (D) All models

4. Which of the predicted curves in Figure 1.3 has the highest residual sum of squares?

- (A) None of the models

- (B) Model (a)

- (C) Model (b)

- (D) Model (c)

- (E) Model (d)

- (F) All model

1.3 General modelling approaches

We wish to model the dependence of a response variable y on an explanatory variable x, where y and x are both continuous. We observe y_i at each time x_i, for $i = 1, \dots, n$, where the observation locations are ordered: $x_1 < x_2 < ... < x_n$. We imagine that the y's are noisy versions of a smooth function of x, say $f(x)$. That is,

$$y_i = f(x_i) + \epsilon_i, \tag{1.1}$$

where the $\{\epsilon_i\}$ are i.i.d:

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2). \tag{1.2}$$

We suppose we do not know the correct form of function f: how can we estimate it?

It is useful to divide modelling approaches into two broad types: parametric and non-parametric.

Parametric models

By far the most common parametric model is simple linear regression, for example, $f(x) = \alpha + \beta x$, where parameters α and β are to be estimated. This is, of course, the simplest example of the polynomial model family, $f(x) = \alpha + \beta x + \gamma x^2 + \cdots + \omega\ x^p$, where p is the *order* of the polynomial and where all of $\alpha, \beta, \gamma, \ldots, \omega$ are to be estimated. This has as special cases: quadratic, cubic, quartic, and quintic polynomials models. Also common are exponential models, for example $f(x) = \alpha e^{-\beta x}$, where α, β are to be estimated – do not confuse this with the exponential probability density function.

Note that the polynomial models are all linear functions *of the parameters*. They are standard forms in regression modelling, as studied in MATH3714 (Linear regression and Robustness) and MATH3823 (Generalised linear models). The exponential model, however, is an example of a model which is non-linearly in the parameters – it is an example of a *non-linear regression model*.

Although very many parametric models exist, they are all somewhat inflexible in their description of f. They cannot accommodate arbitrary fluctuations in $f(x)$ over x because they contain only a small number of parameters (degrees-of-freedom).

Non-parametric models

In such models, f is assumed to be a smooth function of x, but otherwise we do not know what f looks like. A *smooth function* f is such that $f(x_i)$ is close to $f(x_j)$ whenever x_i is close to x_j. To characterise and fit f we will use an approach based on *splines*. In practice, different approaches to characterizing and fitting smooth f lead to similar fits to the data. The spline approach fits neatly with normal and generalised linear models (NLMs and GLMs), but so do other approaches (for example, kernel smoothing and wavelets). Methods of fitting f based on kernel smoothing and the Nadaraya–Watson estimator are studied in the Level 5 component of MATH5714 (Linear regression, robustness and smoothing) where the choice of bandwidth in kernel methods is analogous to the choice of smoothing parameter value in spline smoothing.

Piecewise polynomial models

A common problem with low-order polynomials is that they can often fit well for part of the data but have unappealing features elsewhere. For example, although none of the models in Figure 1.1 fit the data at all well, we might imagine that three short linear segments might be a good fit to the coal-seam data. Also, the piecewise linear model was a good description of the data in Figure 1.2c. This suggests that local polynomial models might be useful. In some situation, for example when we know that the function f is continuous, jumps in the fitted model, as in Figure 1.2c, are unacceptable. Alternatively, we may require differentiability of f. Such technical issues lead to the use of *splines*, which is introduced in the next chapter.

1.4 Exercises

1.1 Consider the first three models fitted in Figure 1.1 and let the data be denoted, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$. These three models can be written

$$\begin{aligned} (a) &\quad y = \alpha + \epsilon \\ (b) &\quad y = \alpha + \beta x + \epsilon \\ (c) &\quad y = \alpha + \beta x + \gamma x^2 + \epsilon \end{aligned}$$

where ϵ represents normally distributed random error. Use the principle of least squares, or otherwise, to obtain estimates of the model parameters.

Click here to see hints.

For each, start by defining the residual sum of squares (RSS), $RSS = \sum(y_i - \hat{y}_i)^2$ where, in turn (a) $\hat{y}_i = \alpha$, (b) $\hat{y}_i = \alpha + \beta x_i$, (c) $\hat{y}_i = \alpha + \beta x_i + \gamma x_i^2$. Then, find he parameter values which minimise the RSS by (possibly partial) differentiation.

1.2 Discuss possible approaches to fitting an exponential model, $y = \alpha e^{\beta x}$, to data. Note that no actual algebraic derivation, nor numerical coded algorithm is expect.

Click here to see hints.

There is more than one approach. Can least squares be used? Could a simple transformation of the data and the fitted model make solving the problem easier? Is there an algebraic solution? Is there a purely numerical solution?

1.3 In Figure 1.2c, discuss how you might fit a two-part linear model for the case where the change-point is *unknown*. Note that no actual algebraic derivation, nor numerical coded algorithm is expect.

Click here to see hints.

With many similar problems, imaging breaking the problem down into steps. If you know the location of the change-point then what should you do? Can you then try different possible change-point locations?

1.4 Discuss the four fitted models in Figure 1.3. Can you give positive and negative properties of each model? Which do you think is best and which worst? Do you think which is best/worst, depends on the data? Justify your answers.

Click here to see hints.

Don't get too stuck on the data used here, but think of general issues: ease of use, reliability of the data, is there error with the data or is it very reliable? What if the response it discrete?

Note

Exercise 1 Solutions can be found here.