MATH5824 Generalized Linear and Additive Models

Robert G Aykroyd

2/8/23

Table of contents

Weekly schedule

Week 3 (13 - 17 February)

- Details to be added soon.

Week 2 (6 - 10 February)

- **Before next Lecture:** Please re-read MATH3823 *Section 2.1: Overview* and read MATH3823 *Section 2.2: Types of normal linear model*.
- **Lecture on Tuesday:** We will briefly cover all remaining material in MATH3823 *Chapter 2: Essentials of Normal Linear Models*.
- **Lecture on Thursday:** Cancelled due to UCU strike.
- **Lecture on Friday:** Cancelled due to UCU strike.
- **Before next Lecture:** Please re-read MATH3823 *Chapter 2* carefully.
- **Before next Lecture:** Please self-study MATH5824 *Chapter 2: Introducing Splines*.
- **Weekly feedback:** Self-study the Exercises in MATH3823 *Section 2.6* and MATH5824 *Section 2.2* – solutions to be posted by the end of Week 3.

Week 1 (30 January - 3 February)

- **Before next Lecture:** Please read MATH3823 *Overview*.
- **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 2: Essentials of Normal Linear Models* with *Section 2.1: Overview*.
- **Before next Lecture:** Please read MATH5824 *Overview*.
- **Lecture on Friday:** Start MATH5824 by briefly considering *Chapter 1: Introduction to Non-parametric Modelling*.
- **Weekly feedback:** Self-study the Exercises in MATH3823 *Section 1.5* – solutions to be posted during Week 1.

Coursework Practical Sessions (20 - 24 March)

- Coursework for this module involves a single written report worth 20% of the module grade. This will mainly involve investigating different models using R and interpreting the results.
- Tasks are expected to be handed out before 16 March with hand-in deadline expect to be after 28 March. Further details to follow in early March.

Overview

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

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 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.

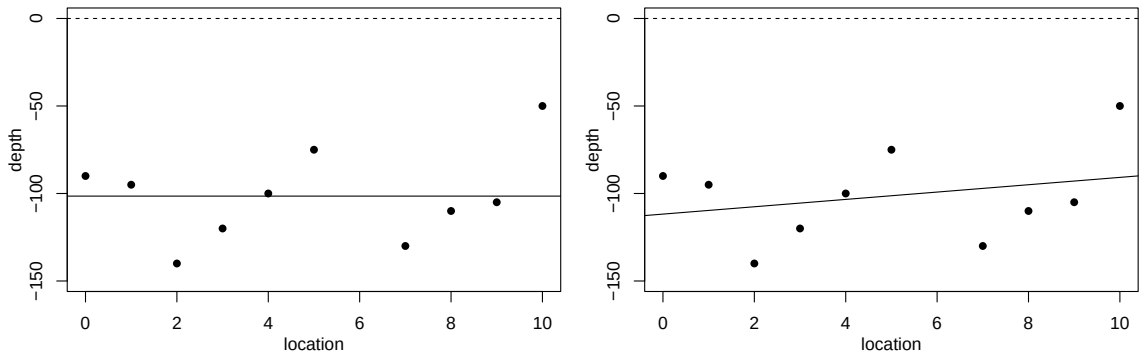
Location, x	0	1	2	3	4	5	6	7	8	9	10
Depth, y	-90	-95	-140	-120	-100	-75	NA	-130	-110	-105	-50

Figure 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



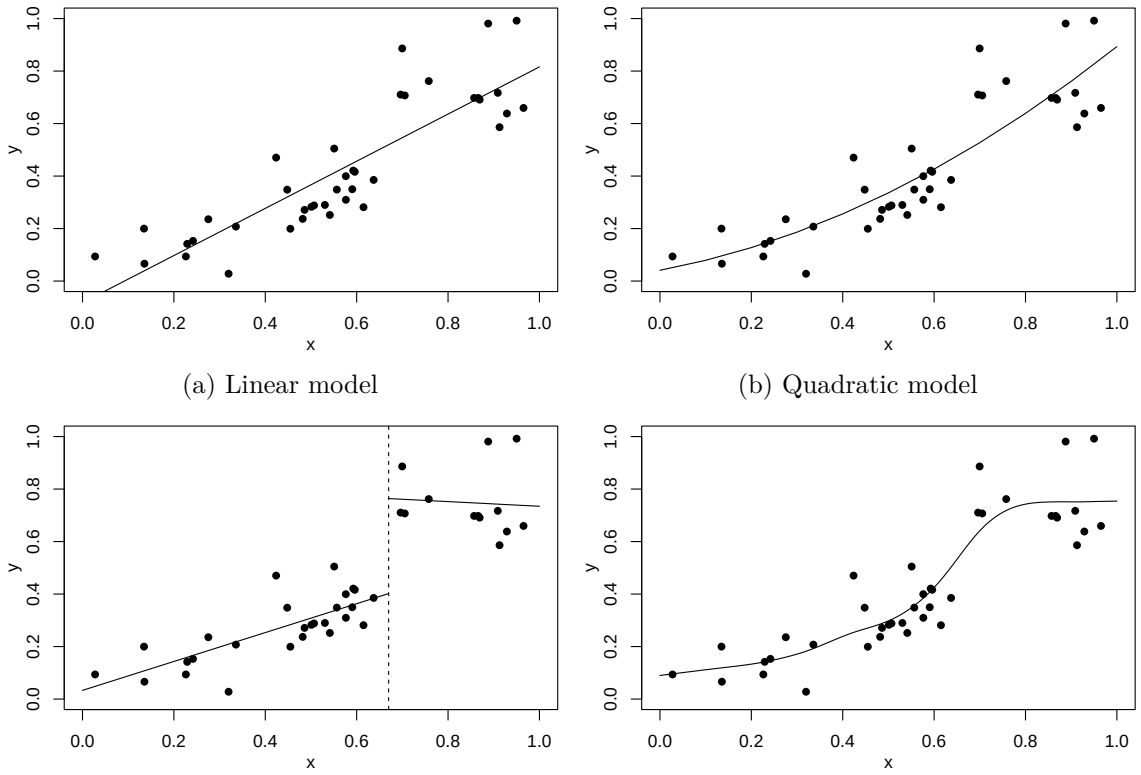
(a) Constant model

(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

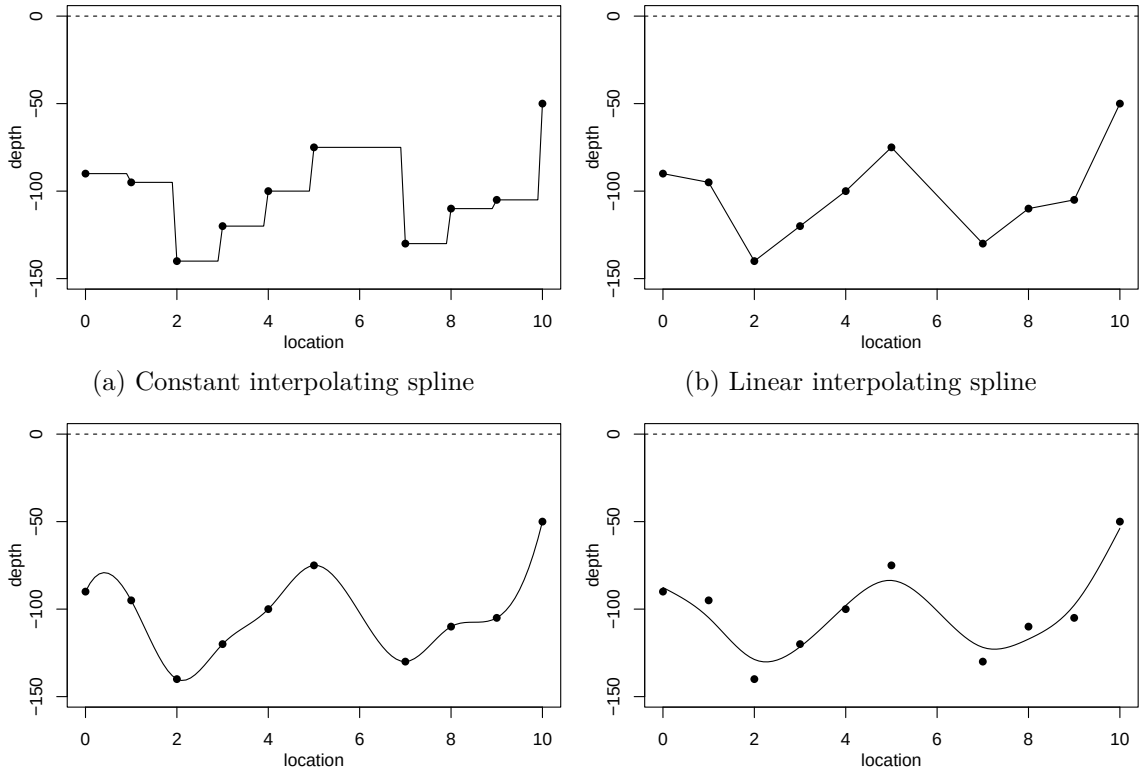
(c) Piecewise linear model

(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.

1.2 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 $i = 1, \ldots, n$, where the observation locations are ordered: $x_1 < x_2 < \ldots < x_n$. 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 (Generalized 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 characterize 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 generalized 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.

2 Introducing Splines

2.1 Basic definitions

Let $t_1 < t_2 < \ldots < t_m$ be a fixed set of *sites* or *knots* which need not correspond to observation locations, as in Figure 2.1.

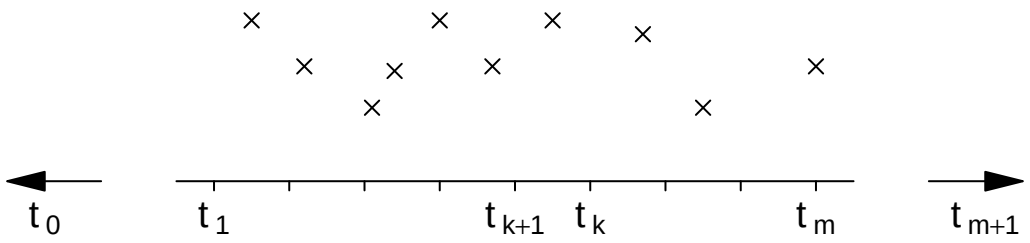


Figure 2.1: Diagram of knots and data points.

Note that we use the symbol t, rather than x, so that we do not confuse knots and observation locations.

A spline of order $p \geq 1$ is a piecewise-polynomial of order p which is $(p-1)$ times differentiable at the knots. Thus there are coefficients $\{a_{k\ell},\ k = 0, \ldots, m,\ \ell = 0, \ldots, p\}$ such that

$$f(t) = \sum_{\ell=0}^{p} a_{k\ell}\, t^{\ell}, \qquad \text{for } t_k \leq t < t_{k+1}, \tag{2.1}$$

where we take $t_0 = -\infty$ and $t_{m+1} = +\infty$.

If we are using cubic polynomials, $(p = 3)$, then f is given by the following equations:

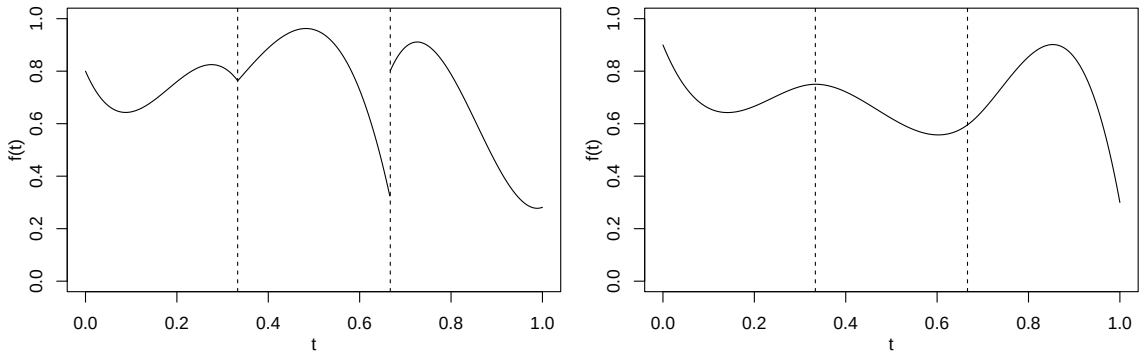
$$f(t) = a_{00} + a_{01}t + a_{02}t^2 + a_{03}t^3, \quad t_0 \leq t < t_1$$

to the left of the first knot,

$$f(t) = a_{10} + a_{11}t + a_{12}t^2 + a_{13}t^3, \quad t_1 \leq t < t_2$$

between the first and second knots, and so on until

$$f(t) = a_{m0} + a_{m1}t + a_{m2}t^2 + a_{m3}t^3, \quad t_m \leq t < t_{m+1}$$



(a) No smoothness constraints

(b) With smoothness constraints

Figure 2.2: Piecewise-cubic functions in three intervals with knot positions indicated with vertical lines.

to the right of the final knot. This is illustrated in Figure 2.2a with $m = 2$.

Because of the use of polynomials, f is smooth *between* each successive pair of knots. At the knots, however, f might not be continuous and it might not be differentiable – in such cases we would say that the function is not smooth.

To ensure that f is also smooth at each of the knots, we impose smoothness constraints which control continuity of the function and its derivatives at the knots.

Let $f^{(\ell)}$ be the ℓ-th order derivative, with $f^{(0)} = f$ being the function itself, $f^{(1)} = f'$ is the first derivative and $f^{(2)} = f''$ the second derivative. Further, let $f^{(\ell)}(t - \epsilon)$ and $f^{(\ell)}(t + \epsilon)$, for $\epsilon \geq 0$, denote evaluation of the function or its derivative at points just below and just above t – we will be interested in their relative values as $\epsilon \to 0$.

To impose smoothness, we require that

$$\lim_{\epsilon \to 0} f^{(\ell)}(t_k - \epsilon) = \lim_{\epsilon \to 0} f^{(\ell)}(t_k + \epsilon), \tag{2.2}$$

for all $k = 1, \ldots, m$ and for $\ell = 0, \ldots, (p - 1)$.

In other words we say that f is smooth if the limits, from below and from above, of the function and its $(p - 1)$ derivatives exist and are equal.

The meaning of these smoothness constraints is illustrated in Figure 2.2. In Figure 2.2a, a piecewise cubic function with two knots has been plotted. The first derivative, f', is discontinuous at the first knot and the function itself, f, is discontinuous at the second knot. Figure 2.2b shows a similar shaped cubic spline with two knots. This time, the function f and its first two derivatives are continuous at both knots.

The smoothness conditions in Equation 2.2 induce constraints on the coefficients $\{a_{k\ell}\}$. A polynomial of order p has $p + 1$ coefficients, and there are $m + 1$ intervals when we have m

knots. This leads to $(p+1) \times (m+1)$ coefficients but there are p constraints at each of the m knots. Thus the total *degrees of freedom* of the system is

$$\mathrm{df}_{\mathrm{spline}} = (p+1)(m+1) - pm = m + p + 1. \tag{2.3}$$

These degrees of freedom provide the necessary flexibility in the spline.

Note that f is infinitely differentiable everywhere, except at the knots where it is $p-1$ times differentiable. In particular, for $p = 1$, f is a linear spline comprising linear pieces constrained to be continuous at the knots, although the slope of f is discontinuous at the knots. Also, for $p = 3$, f is a cubic spline comprising cubic polynomial pieces continuous at the knots; where the first and second derivatives of f are also continuous, but the third derivative is discontinuous at the knots.

2.2 Exercises

2.1 Why is it not sensible to define a smooth function made-up of constant components? Similarly, why is not sensible to create a differentiable function from linear splines?

2.2 In the situation illustrated in Figure 2.2b, where $p = 3$ and $m = 2$, clearly identify the $(p+1) \times (m+1) = 12$ model parameters and the $pm = 6$ smoothness constraints in terms of the cubic polynomials and their derivatives.

2.3 Further consider the situation illustrated in Figure 2.2b. Suppose now that we require the splines to pass through specified coordinates $(t_1, f(t_1))$ and $(t_2, f(t_2))$. What is the degrees of freedom for this model? How many such cubic splines would satisfy these constraints? Discuss potential additional constraints which would lead to a unique fitted model. Do you think having a unique solution is a positive or negative property?

2.4 For a general problem, what would be the effect of requiring additional constraints of the form of Equation 2.2 but with $\ell = p$? Would this lead to an acceptable fitted cubic spline model? Justify your answer.