1

Model selection: data examples and introduction

This book is about making choices. If there are several possibilities for modelling data, which should we take? If multiple explanatory variables are measured, should they all be used when forming predictions, making classifications, or attempting to summarise analysis of what influences response variables, or will including only a few of them work equally well, or better? If so, which ones can we best include? Model selection problems arrive in many forms and on widely varying occasions. In this chapter we present some data examples and discuss some of the questions they lead to. Later in the book we come back to these data and suggest some answers. A short preview of what is to come in later chapters is also provided.

1.1 Introduction

With the current ease of data collection which in many fields of applied science has become cheaper and cheaper, there is a growing need for methods which point to interesting, important features of the data, and which help to build a model. The model we wish to construct should be rich enough to explain relations in the data, but on the other hand simple enough to understand, explain to others, and use. It is when we negotiate this balance that model selection methods come into play. They provide formal support to guide data users in their search for good models, or for determining which variables to include when making predictions and classifications.

Statistical model selection is an integral part of almost any data analysis. Model selection cannot be easily separated from the rest of the analysis, and the question 'which model is best' is not fully well-posed until supplementing information is given about what one plans to do or hopes to achieve given the choice of a model. The survey of data examples that follows indicates the broad variety of applications and relevant types of questions that arise.

Before going on to this survey we shall briefly discuss some of the key general issues involved in model selection and model averaging.

(i) *Models are approximations:* When dealing with the issues of building or selecting a model, it needs to be realised that in most situations we will not be able to guess the 'correct' or 'true' model. This true model, which in the background generated the data we collected, might be very complex (and almost always unknown). For working with the data it might be of more practical value to work instead with a simpler, but almost-as-good model: 'All models are wrong, but some are useful', as a maxim formulated by G. E. P. Box expresses this view. Several model selection methods start from this perspective.

(ii) *The bias–variance trade-off:* The balance and interplay between variance and bias is fundamental in several branches of statistics. In the framework of model fitting and selection it takes the form of balancing simplicity (fewer parameters to estimate, leading to lower variability, but associated with modelling bias) against complexity (entering more parameters in a model, e.g. regression parameters for more covariates, means a higher degree of variability but smaller modelling bias). Statistical model selection methods must seek a proper balance between overfitting (a model with too many parameters, more than actually needed) and underfitting (a model with too few parameters, not capturing the right signal).

(iii) *Parsimony:* 'The principle of parsimony' takes many forms and has many formulations, in areas ranging from philosophy, physics, arts, communication, and indeed statistics. The original Ockham's razor is 'entities should not be multiplied beyond necessity'. For statistical modelling a reasonable translation is that only parameters that really matter ought to be included in a selected model. One might, for example, be willing to extend a linear regression model to include an extra quadratic term if this manifestly improves prediction quality, but not otherwise.

(iv) *The context:* All modelling is rooted in an appropriate scientific context and is for a certain purpose. As Darwin once wrote, 'How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service'. One must realise that 'the context' is not always a precisely defined concept, and different researchers might discover or learn different things from the same data sets. Also, different schools of science might have different preferences for what the aims and purposes are when modelling and analysing data. Breiman (2001) discusses 'the two cultures' of statistics, broadly sorting scientific questions into respectively those of prediction and classification on one hand (where even a 'black box' model is fine as long as it works well) and those of 'deeper learning about models' on the other hand (where the discovery of a non-null parameter is important even when it might not help improve inference precision). Thus S. Karlin's statement that 'The purpose of models is not to fit the data, but to sharpen the questions' (in his R. A. Fisher memorial lecture, 1983) is important in some contexts but less relevant in others. Indeed there are differently spirited model selection methods, geared towards answering questions raised by different cultures.

(v) *The focus:* In applied statistics work it is often the case that some quantities or functions of parameters are more important than others. It is then fruitful to gear model building and model selection efforts towards criteria that favour good performance precisely for those quantities that are more important. That different aims might lead to differently selected models, for the same data and the same list of candidate models, should not be considered a paradox, as it reflects different preferences and different loss functions. In later chapters we shall in particular work with focussed information criteria that start from estimating the mean squared error (variance plus squared bias) of candidate estimators, for a given focus parameter.

(vi) *Conflicting recommendations:* As is clear from the preceding points, questions about ‘which model is best’ are inherently more difficult than those of the type ‘for a given model, how should we carry out inference’. Sometimes different model selection strategies end up offering different advice, for the same data and the same list of candidate models. This is not a contradiction as such, but stresses the importance of learning how the most frequently used selection schemes are constructed and what their aims and properties are.

(vii) *Model averaging:* Most selection strategies work by assigning a certain score to each candidate model. In some cases there might be a clear winner, but sometimes these scores might reveal that there are several candidates that do almost as well as the winner. In such cases there may be considerable advantages in combining inference output across these best models.

1.2 Egyptian skull development

Measurements on skulls of male Egyptians have been collected from different archaeological eras, with a view towards establishing biometrical differences (if any) and more generally studying evolutionary aspects. Changes over time are interpreted and discussed in a context of interbreeding and influx of immigrant populations. The data consist of four measurements for each of 30 skulls from each of five time eras, originally presented by Thomson and Randall-Maciver (1905). The five time periods are the early predynastic (around 4000 B.C.), late predynastic (around 3300 B.C.), 12th and 13th dynasties (around 1850 B.C.), the ptolemaic period (around 200 B.C.), and the Roman period (around 150 A.D.). For each of the 150 skulls, the following measurements are taken (all in millimetres): $x_1 =$ maximal breadth of the skull (MB), $x_2 =$ basibregmatic height (BH), $x_3 =$ basialveolar length (BL), and $x_4 =$ nasal height (NH); see Figure 1.1, adapted from Manly (1986, page 6). Figure 1.2 gives pairwise scatterplots of the data for the first and last time period, respectively. Similar plots are easily made for the other time periods. We notice, for example, that the level of the x_1 measurement appears to have increased while that of the x_3 measurement may have decreased somewhat over time. Statistical modelling and analysis are required to accurately validate such claims.

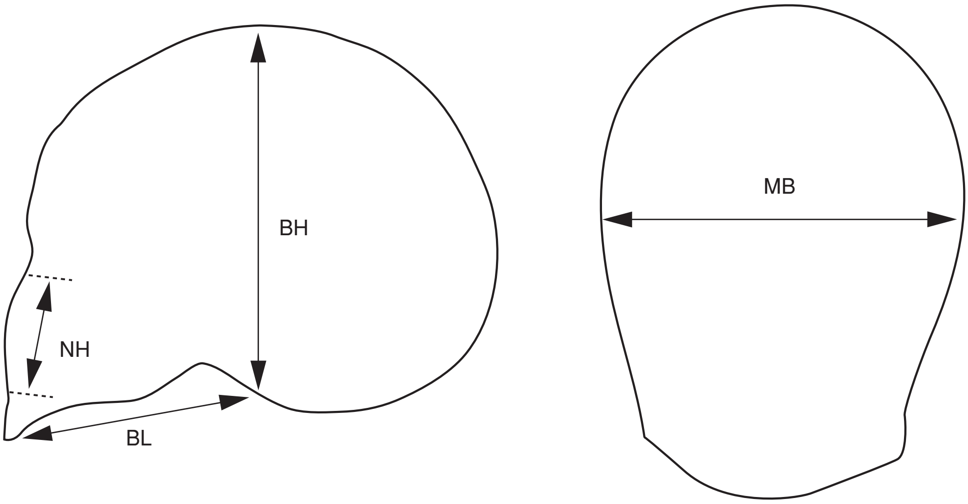

Fig. 1.1. The four skull measurements $x_1 = \mathrm{MB}$, $x_2 = \mathrm{BH}$, $x_3 = \mathrm{BL}$, $x_4 = \mathrm{NH}$; from Manly (1986, page 6).

There is a four-dimensional vector of observations $y_{t,i}$ associated with skull i and time period t, for $i = 1, \ldots, 30$ and $t = 1, \ldots, 5$, where $t = 1$ corresponds to 4000 B.C., and so on, up to $t = 5$ for 150 A.D. We use $\bar{y}_{t,\bullet}$ to denote the four-dimensional vector of averages across the 30 skulls for time period t. This yields the following summary measures:

$$\begin{aligned}
\bar{y}_{1,\bullet} &= (131.37, 133.60, 99.17, 50.53),\\
\bar{y}_{2,\bullet} &= (132.37, 132.70, 99.07, 50.23),\\
\bar{y}_{3,\bullet} &= (134.47, 133.80, 96.03, 50.57),\\
\bar{y}_{4,\bullet} &= (135.50, 132.30, 94.53, 51.97),\\
\bar{y}_{5,\bullet} &= (136.27, 130.33, 93.50, 51.37).
\end{aligned}$$

Standard deviations for the four measurements, computed from averaging variance estimates over the five time periods (in the order MB, BH, BL, NH), are 4.59, 4.85, 4.92, 3.19. We assume that the vectors $Y_{t,i}$ are independent and four-dimensional normally distributed, with mean vector ξ_t and variance matrix Σ_t for eras $t = 1, \ldots, 5$. However, it is not given to us how these mean vectors and variance matrices could be structured, or how they might evolve over time. Hence, although we have specified that data stem from four-dimensional normal distributions, the model for the data is not yet fully specified.

We now wish to find a statistical model that provides the clearest explanation of the main features of these data. Given the information and evolutionary context alluded to above, searching for good models would involve their ability to answer the following questions. Do the mean parameters (population averages of the four measurements)

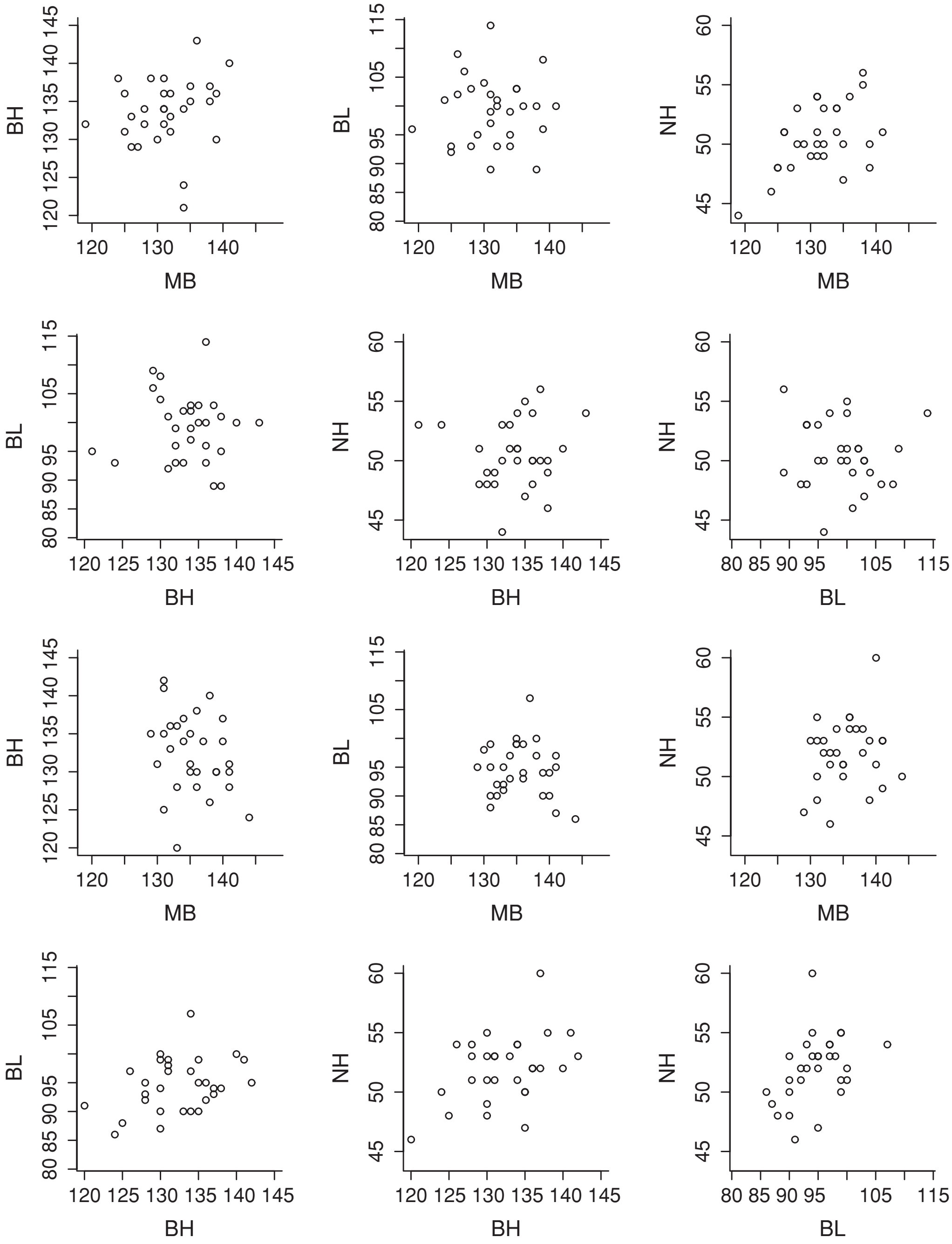

Fig. 1.2. Pairwise scatterplots for the Egyptian skull data. First two rows: early predynastic period (4000 B.C.). Last two rows: Roman period (150 A.D.).

remain the same over the five periods? If not, is there perhaps a linear trend over time? Or is there no clear structure over time, with all mean parameters different from one another? These three questions relate to the mean vector. Each situation corresponds to a different model specification:

(i) If all mean measurements remain constant over the five time periods, we can combine all 150 (5 times 30) measurements for estimating the common mean vector ξ. This is the simplest model for the mean parameters, and involves four such parameters.

(ii) If we expect a linear trend over time, we can assume that at time period t the mean components $\xi_{t,j}$ are given by formulae of the form $\xi_{t,j} = \alpha_j + \beta_j \mathrm{time}(t)$, for $j = 1, 2, 3, 4$, where $\mathrm{time}(t)$ is elapsed time from the first era to era t, for $t = 1, \ldots, 5$. Estimating the intercept α_j and slope β_j is then sufficient for obtaining estimates of the mean of measurement j at all five time periods. This model has eight mean parameters.

(iii) In the situation where we do not postulate any structure for the mean vectors, we assume that the mean vectors $\xi_1, \ldots, \xi_5$ are possibly different, with no obvious formula for computing one from the other. This corresponds to five different four-dimensional normal distributions, with a total of 20 mean parameters. This is the richest or most complex model.

In this particular situation it is clear that model (i) is contained in model (ii) (which corresponds to the slope parameters β_j being equal to zero), and likewise model (ii) is contained in model (iii). This corresponds to what is called a nested sequence of models, where simpler models are contained in more complex ones. Some of the model selection strategies we shall work with in this book are specially constructed for such situations with nested candidate models, whereas other selection methods are meant to work well regardless of such constraints.

Other relevant questions related to these data include the following. Is the correlation structure between the four measurements the same over the five time periods? In other words, is the correlation between measurements x_1 and x_2, and so on, the same for all five time periods? Or can we simplify the correlation structure by taking correlations between different measurements on the same skull to be equal? Yet another question relates to the standard deviations. Can we take equal standard deviations for the measurements, across time? Such questions, if answered in the affirmative, amount to different model simplifications, and are often associated with improved inference precision since fewer model parameters need to be estimated. Each of the possible simplifications alluded to here corresponds to a statistical model formulation for the covariance matrices. In combination with the different possibilities listed above for modelling the mean vector, we arrive at a list of different models to choose from.

We come back to this data set in Section 9.1. There we assign to each model a number, or a score, corresponding to a value of an information criterion. We use two such information criteria, called the AIC (Akaike's information criterion, see Chapter 2) and BIC (the Bayesian information criterion, see Chapter 3). Once each model is assigned a score, the models are ranked and the best ranked model is selected for further analysis of the data. For a multi-sample cluster analysis of the same data we refer to Bozdogan *et al.* (1994).

1.3 Who wrote 'The Quiet Don'?

The Nobel Prize in literature 1965 was awarded to Mikhail Sholokhov (1905–1984), for the epic *And Quiet Flows the Don*, or *The Quiet Don*, about Cossack life and the birth of a new Soviet society. In Russia alone his books have been published in more than a thousand editions, selling in total more than 60 million copies. But in the autumn of 1974 an article was published in Paris, The Rapids of Quiet Don: the Enigma of the Novel by the author and critic known as 'D'. He claimed that 'The Quiet Don' was not at all Sholokhov's work, but rather that it was written by Fiodor Kriukov, an author who fought against bolshevism and died in 1920. The article was given credibility and prestige by none other than Aleksandr Solzhenitsyn (a Nobel prize winner five years after Sholokhov), who in his preface to D's book strongly supported D's conclusion (Solzhenitsyn, 1974). Are we in fact faced with one of the most flagrant cases of theft in the history of literature?

An inter-Nordic research team was formed in the course of 1975, captained by Geir Kjetsaa, a professor of Russian literature at the University of Oslo, with the aim of disentangling the Don mystery. In addition to various linguistic analyses and some doses of detective work, quantitative data were also gathered, for example relating to sentence lengths, word lengths, frequencies of certain words and phrases, grammatical characteristics, etc. These data were extracted from three corpora (in the original Russian editions): (i) Sh, from published work guaranteed to be by Sholokhov; (ii) Kr, that which with equal trustworthiness came from the hand of the alternative hypothesis Kriukov; and (iii) QD, the Nobel winning text 'The Quiet Don'. Each of the corpora has about 50,000 words.

We shall here focus on the statistical distribution of the number of words used in sentences, as a possible discriminant between writing styles. Table 1.1 summarises these data, giving the number of sentences in each corpus with lengths between 1 and 5 words, between 6 and 10 words, etc. The sentence length distributions are also portrayed in Figure 1.3, along with fitted curves that are described below. The statistical challenge is to explore whether there are any sufficiently noteworthy differences between the three empirical distributions, and, if so, whether it is the upper or lower distribution of Figure 1.3 that most resembles the one in the middle.

A simple model for sentence lengths is that of the Poisson, but one sees quickly that the variance is larger than the mean (in fact, by a factor of around six). Another possibility is that of a mixed Poisson, where the parameter is not constant but varies in the space of sentences. If Y given λ is Poisson with this parameter, but λ has a Gamma (a, b) distribution, then the marginal takes the form

$$f^*(y, a, b) = \frac{b^a}{\Gamma(a)} \frac{1}{y!} \frac{\Gamma(a + y)}{(b + 1)^{a+y}} \quad \text{for } y = 0, 1, 2, \ldots,$$

Table 1.1. *The Quiet Don: number of sentences N_x in the three corpora Sh, Kr, QD of the given lengths, along with predicted numbers* pred_x *under the four-parameter model (1.1), and Pearson residuals* res_x*, for the 13 length groups. Note: The first five columns have been compiled from tables in Kjetsaa et al. (1984).*

Words		N_x			pred_x			res_x		
from	to	Sh	Kr	QD	Sh	Kr	QD	Sh	Kr	QD
1	5	799	714	684	803.4	717.6	690.1	−0.15	−0.13	−0.23
6	10	1408	1046	1212	1397.0	1038.9	1188.5	0.30	0.22	0.68
11	15	875	787	826	884.8	793.3	854.4	−0.33	−0.22	−0.97
16	20	492	528	480	461.3	504.5	418.7	1.43	1.04	3.00
21	25	285	317	244	275.9	305.2	248.1	0.55	0.67	−0.26
26	30	144	165	121	161.5	174.8	151.1	−1.38	−0.74	−2.45
31	35	78	78	75	91.3	96.1	89.7	−1.40	−1.85	−1.55
36	40	37	44	48	50.3	51.3	52.1	−1.88	−1.02	−0.56
41	45	32	28	31	27.2	26.8	29.8	0.92	0.24	0.23
46	50	13	11	16	14.5	13.7	16.8	−0.39	−0.73	−0.19
51	55	8	8	12	7.6	6.9	9.4	0.14	0.41	0.85
56	60	8	5	3	4.0	3.5	5.2	2.03	0.83	−0.96
61	65	4	5	8	2.1	1.7	2.9	1.36	2.51	3.04
Total:		4183	3736	3760						

which is the negative binomial. Its mean is $\mu = a/b$ and its variance $a/b + a/b^2 = \mu(1 + 1/b)$, indicating the level of over-dispersion. Fitting this two-parameter model to the data was also found to be too simplistic; patterns are more variegated than those dictated by a mere negative binomial. Therefore we use the following mixture of a Poisson (a degenerate negative binomial) and another negative binomial, with a modification to leave out the possibility of having zero words in a sentence:

$$f(y, p, \xi, a, b) = p\frac{\exp(-\xi)\xi^y/y!}{1 - \exp(-\xi)} + (1 - p)\frac{f^*(y, a, b)}{1 - f^*(0, a, b)} \qquad (1.1)$$

for $y = 1, 2, 3, \ldots$ It is this four-parameter family that has been fitted to the data in Figure 1.3. The model fit is judged adequate, see Table 1.1, which in addition to the observed number N_x shows the expected or predicted number pred_x of sentences of the various lengths, for length groups $x = 1, 2, 3, \ldots, 13$. Also included are Pearson residuals $(N_x - \text{pred}_x)/\text{pred}_x^{1/2}$. These residuals should essentially be on the standard normal scale if the parametric model used to produce the predicted numbers is correct; here there are no clear clashes with this hypothesis, particularly in view of the large sample sizes involved, with respectively 4183, 3736, 3760 sentences in the three corpora. The pred_x numbers in the table come from minimum chi-squared fitting for each of the three

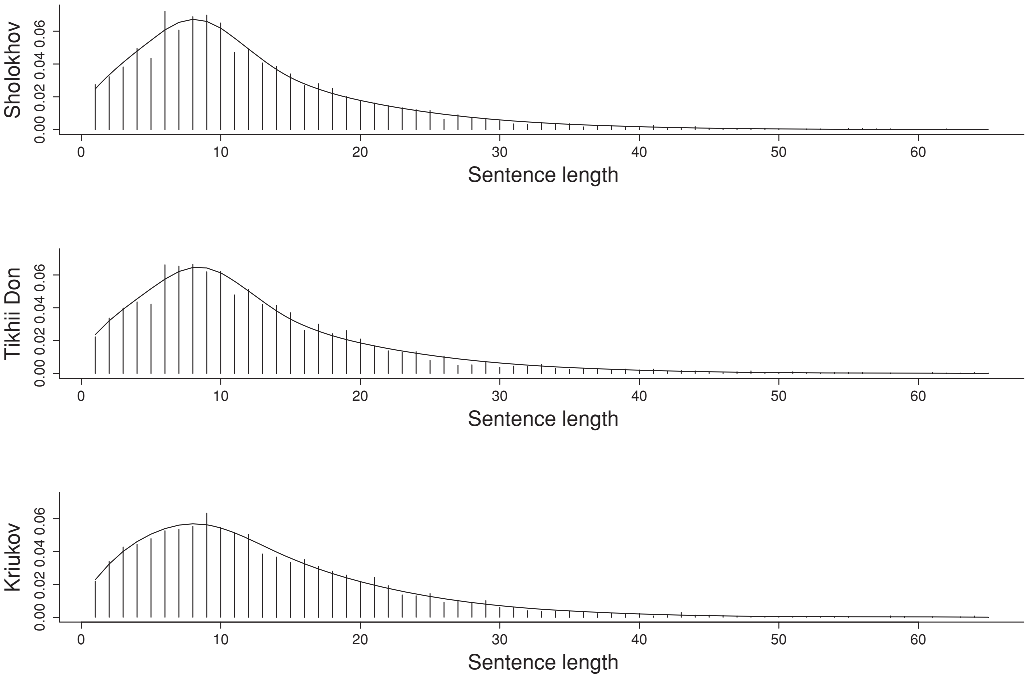

Fig. 1.3. Sentence length distributions, from 1 word to 65 words, for Sholokhov (top), Kriukov (bottom), and for ‘The Quiet Don’ (middle). Also shown, as continuous curves, are the distributions (1.1), fitted via maximum likelihood.

corpora, that is, finding parameter estimates to minimise

$$P_n(\theta) = \sum_x \frac{\{N_x - \mathrm{pred}_x(\theta)\}^2}{\mathrm{pred}_x(\theta)^2}$$

with respect to the four parameters, where $\mathrm{pred}_x(\theta) = np_x(\theta)$ in terms of the sample size for the corpus worked with and the inferred probability $p_x(\theta)$ of writing a sentence with length landing in group x.

The statistical problem may be approached in different ways; see Hjort (2007a) for a wider discussion. Kjetsaa’s group quite sensibly put up Sholokhov’s authorship as the null hypothesis, and D’s speculations as the alternative hypothesis, in several of their analyses. Here we shall formulate the problem in terms of selecting one of three models, inside the framework of three data sets from the four-parameter family (1.1):

M_1: Sholokhov is the rightful author, so that text corpora Sh and QD come from the same statistical distribution, while Kr represents another;

M_2: D and Solzhenitsyn were correct in denouncing Sholokhov, whose text corpus Sh is therefore not statistically compatible with Kr and QD, which are however coming from the same distribution;

M_3: Sh, Kr, QD represent three statistically disparate corpora.

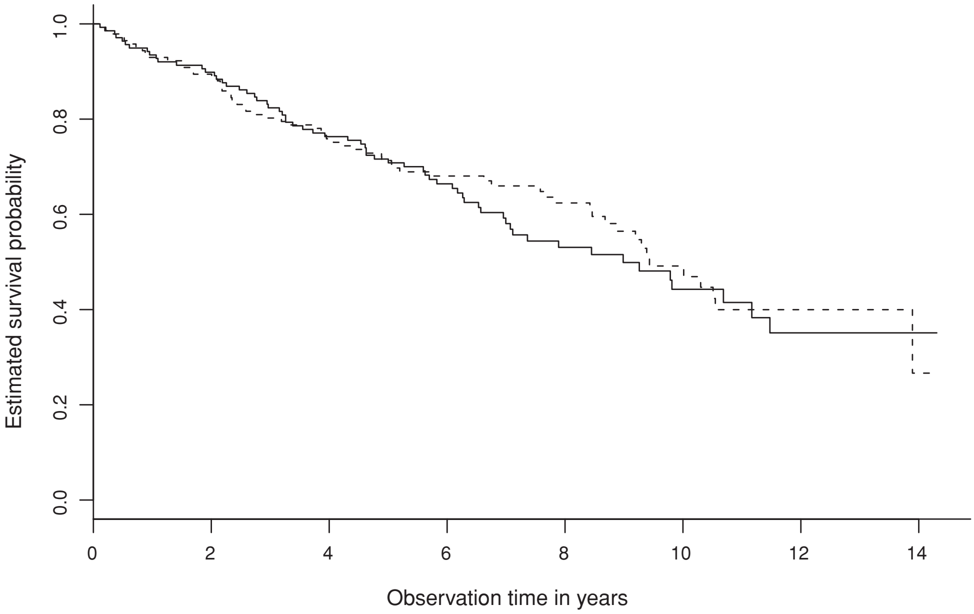

Fig. 1.4. Estimated survival probabilities (Kaplan–Meier curves) for the drug group (solid line) and placebo group (dashed line) in the study on primary biliary cirrhosis.

Selecting one of these models via statistical methodology will provide an answer to the question about who is most probably the author. (In this problem formulation we are disregarding the initial stage of model selection that is associated with using the parametric (1.1) model for the sentence distributions; the methods we shall use may be extended to encompass also this additional layer of complication, but this does not affect the conclusions we reach.) Further discussion and an analysis of this data set using a method related to the Bayesian information criterion is the topic of Section 3.3.

1.4 Survival data on primary biliary cirrhosis

PBC (primary biliary cirrhosis) is a condition which leads to progressive loss of liver function. It is commonly associated with Hepatitis C or high-volume use of alcohol, but has many other likely causes. The data set we use here for examining risk factors and treatment methods associated with PBC is the follow-up to the original PBC data set presented in appendix D of Fleming and Harrington (1991); see Murtaugh *et al.* (1994) and the data overview on page 287. This is a randomised double-blinded study where patients received either the drug D-pencillamine or placebo. Of the 280 patients for whom the information is included in this data set, 126 died before the end of the study. Figure 1.4 gives Kaplan–Meier curves, i.e. estimated survival probability curves, for the two groups. The solid line is for the drug group, the dashed line for the placebo group.