2

Akaike’s information criterion

Data can often be modelled in different ways. There might be simple approaches and more advanced ones that perhaps have more parameters. When many covariates are measured we could attempt to use them all to model their influence on a response, or only a subset of them, which would make it easier to interpret and communicate the results. For selecting a model among a list of candidates, Akaike’s information criterion (AIC) is among the most popular and versatile strategies. Its essence is a penalised version of the attained maximum log-likelihood, for each model. In this chapter we shall see AIC at work in a range of applications, in addition to unravelling its basic construction and properties. Attention is also given to natural generalisations and modifications of AIC that in various situations aim at performing more accurately.

2.1 Information criteria for balancing fit with complexity

In Chapter 1 various problems were discussed where the task of selecting a suitable statistical model, from a list of candidates, was an important ingredient. By necessity there are different model selection strategies, corresponding to different aims and uses associated with the selected model. Most (but not all) selection methods are defined in terms of an appropriate *information criterion*, a mechanism that uses data to give each candidate model a certain score; this then leads to a fully ranked list of candidate models, from the ostensibly best to the worst.

The aim of the present chapter is to introduce, discuss and illustrate one of the more important of these information criteria, namely AIC (Akaike’s information criterion). Its general formula is

$$\mathrm{AIC}(M) = 2\,\text{log-likelihood}_{\max}(M) - 2\ \dim(M), \tag{2.1}$$

for each candidate model M, where $\dim(M)$ is the length of its parameter vector. Thus AIC acts as a penalised log-likelihood criterion, affording a balance between good fit (high value of log-likelihood) and complexity (complex models are penalised more than simple ones). The model with the highest AIC score is then selected.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Directly comparing the values of the attained log-likelihood maxima for different models is not good enough for model comparison. Including more parameters in a model always gives rise to an increased value of the maximised log-likelihood (for an illustration, see Table 2.1). Hence, without a penalty, such as that used in (2.1), searching for the model with maximal log-likelihood would simply lead to the model with the most parameters. The penalty punishes the models for being too complex in the sense of containing many parameters. Akaike's method aims at finding models that in a sense have few parameters but nevertheless fit the data well.

The AIC strategy is admirably general in spirit, and works in principle for any situation where parametric models are compared. The method applies in particular to traditional models for i.i.d. data and regression models, in addition to time series and spatial models, and parametric hazard rate models for survival and event history analysis. Most software packages, when dealing with the most frequently used parametric regression models, have AIC values as a built-in option.

Of course there are other ways of penalising for complexity than in (2.1), and there are other ways of measuring fit of data to a model than via the maximal log-likelihood; variations will indeed be discussed later. But as we shall see in this chapter, there are precise mathematical reasons behind the AIC version. These are related to behaviour of the maximum likelihood estimators and their relation to the Kullback–Leibler distance function, as we discuss in the following section.

2.2 Maximum likelihood and the Kullback–Leibler distance

As explained above, a study of the proper comparison of candidate models, when each of these are estimated using likelihood methods, necessitates an initial discussion of maximum likelihood estimators and their behaviour, and, specifically, their relation to a certain way of measuring the statistical distance from one probability density to another, namely the Kullback–Leibler distance. This is the aim of the present section. When these matters are sorted out we proceed to a derivation of AIC as defined in (2.1).

We begin with a simple illustration of how the maximum likelihood method operates; it uses data and a given parametric model to provide an estimated model.

Example 2.1 Low birthweight data: estimation

In the data set on low birthweights (Hosmer and Lemeshow, 1999) there is a total of $n = 189$ women with newborn babies; see the introductory Section 1.5. Here we indicate how the maximum likelihood method is being used to estimate the parameters of a given model. The independent outcome variables $Y_1, \ldots, Y_n$ are binary (0–1) random variables that take the value 1 when the baby has low birthweight and 0 otherwise. Other recorded variables are $x_{2,i}$, weight; $x_{3,i}$, age of the mother; $x_{4,i}$, indicator for 'race black'; and $x_{5,i}$, indicator for 'race other'. We let $x_i = (1, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i})^{\mathrm{t}}$. The most usual model for

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

such situations is the logistic regression model, which takes the form

$$P(Y_i = 1 \mid x_i) = p_i = \frac{\exp(x_i^{\rm t}\theta)}{1 + \exp(x_i^{\rm t}\theta)} \quad \text{for } i = 1, \ldots, n,$$

with θ a five-dimensional parameter vector. The likelihood $\mathcal{L}_n(\theta)$ is a product of $p_i^{y_i}(1 - p_i)^{1-y_i}$ terms, leading to a log-likelihood of the form

$$\ell_n(\theta) = \sum_{i=1}^{n} \{y_i \log p_i + (1 - y_i) \log(1 - p_i)\} = \sum_{i=1}^{n} [y_i x_i^{\rm t}\theta - \log\{1 + \exp(x_i^{\rm t}\theta)\}].$$

A maximum likelihood estimate for θ is found by maximising $\ell_n(\theta)$ with respect to θ. This gives $\widehat{\theta} = (1.307, -0.014, -0.026, 1.004, 0.443)^{\rm t}$. ■

In general, the models that we construct for observations $Y = (Y_1, \ldots, Y_n)$ contain a number of parameters, say $\theta = (\theta_1, \ldots, \theta_p)^{\rm t}$. This translates into a joint (simultaneous) density for Y, $f_{\rm joint}(y, \theta)$. The likelihood function is then

$$\mathcal{L}_n(\theta) = f_{\rm joint}(y_{\rm obs}, \theta),$$

seen as a function of θ, with $y = y_{\rm obs}$ the observed data values. We often work with the log-likelihood function $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$ instead of the likelihood itself. The maximum likelihood estimator of θ is the maximiser of $\mathcal{L}_n(\theta)$,

$$\widehat{\theta} = \widehat{\theta}_{\rm ML} = \arg\max_{\theta}(\mathcal{L}_n) = \arg\max_{\theta}(\ell_n),$$

and is of course a function of $y_{\rm obs}$. In most of the situations we shall encounter in this book, the model will be such that the maximum likelihood estimator exists and is unique, for all data sets, with probability 1. If the data Y are independent and identically distributed, the likelihood and log-likelihood functions can be written as

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(y_i, \theta) \quad \text{and} \quad \ell_n(\theta) = \sum_{i=1}^{n} \log f(y_i, \theta),$$

in terms of the density $f(y, \theta)$ for an individual observation. It is important to make a distinction between the model $f(y, \theta)$ that we construct for the data, and the actual, true density $g(y)$ of the data, that is nearly always unknown. The density $g(\cdot)$ is often called the data-generating density.

There are several ways of measuring closeness of a parametric approximation $f(\cdot, \theta)$ to the true density g, but the distance intimately linked to the maximum likelihood method, as we shall see, is the Kullback–Leibler (KL) distance

$$\mathrm{KL}(g, f(\cdot, \theta)) = \int g(y) \log \frac{g(y)}{f(y, \theta)} \,{\rm d}y, \tag{2.2}$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

to be viewed as the distance from the true g to its approximation $f(\cdot, \theta)$. Applying the strong law of large numbers, one sees that for each value of the parameter vector θ,

$$n^{-1}\ell_n(\theta) \overset{\text{a.s.}}{\to} \int g(y) \log f(y, \theta)\,\mathrm{d}y = \mathrm{E}_g \log f(Y, \theta),$$

provided only that this integral is finite; the convergence takes place 'almost surely' (a.s.), i.e. with probability 1. The maximum likelihood estimator $\widehat{\theta}$ that maximises $\ell_n(\theta)$ will therefore, under suitable and natural conditions, tend a.s. to the minimiser θ_0 of the Kullback–Leibler distance from true model to approximating model. Thus

$$\widehat{\theta} \overset{\text{a.s.}}{\to} \theta_0 = \arg\min_{\theta}\{\mathrm{KL}(g, f(\cdot, \theta))\}. \tag{2.3}$$

The value θ_0 is called the least false, or best approximating, parameter value. Thus the maximum likelihood estimator aims at providing the best parametric approximation to the real density g inside the parametric class $f(\cdot, \theta)$. If the parametric model is actually fully correct, then $g(y) = f(y, \theta_0)$, and the minimum Kullback–Leibler distance is zero.

Regression models involve observations (x_i, Y_i) on say n individuals or objects, where Y_i is response and x_i is a covariate vector. Maximum likelihood theory for regression models is similar to that for the i.i.d. case, but somewhat more laborious. The maximum likelihood estimators aim for least false parameter values, defined as minimisers of certain weighted Kullback–Leibler distances, as we shall see now. There is a true (but, again, typically unknown) data-generating density $g(y \mid x)$ for $Y \mid x$. The parametric model uses the density $f(y \mid x, \theta)$. Under independence, the log-likelihood function is $\ell_n(\theta) = \sum_{i=1}^n \log f(y_i \mid x_i, \theta)$. Assume furthermore that there is some underlying covariate distribution C that generates the covariate vectors $x_1, \ldots, x_n$. Then averages of the form $n^{-1}\sum_{i=1}^n a(x_i)$ tend to well-defined limits $\int a(x)\,\mathrm{d}C(x)$, for any function a for which this integral exists, and the normalised log-likelihood function $n^{-1}\ell_n(\theta)$ tends for each θ to $\int\int g(y \mid x) \log f(y \mid x, \theta)\,\mathrm{d}y\,\mathrm{d}C(x)$. For given covariate vector x, consider now the Kullback–Leibler distance from the true to the approximating model, conditional on x,

$$\mathrm{KL}_x(g(\cdot \mid x), f(\cdot \mid x, \theta)) = \int g(y \mid x) \log \frac{g(y \mid x)}{f(y \mid x, \theta)}\,\mathrm{d}y.$$

An overall (weighted) Kullback–Leibler distance is obtained by integrating KL_x over x with respect to the covariate distribution,

$$\mathrm{KL}(g, f_\theta) = \int\int g(y \mid x) \log \frac{g(y \mid x)}{f(y \mid x, \theta)}\,\mathrm{d}y\,\mathrm{d}C(x). \tag{2.4}$$

Under mild conditions, which must involve both the regularity of the parametric model and the behaviour of the sequence of covariate vectors, the maximum likelihood estimator $\widehat{\theta}$ based on the n first observations tends almost surely to the least false parameter value θ_0 that minimises $\mathrm{KL}(g, f_\theta)$.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Large-sample theory for the distribution of the maximum likelihood estimator is particularly well developed for the case of data assumed to follow precisely the parametric model being used; such situations certainly exist when subject-matter knowledge is well developed, but in most applied statistics contexts such an assumption would be too bold. Importantly, the large-sample likelihood theory has also been extended to the case of the density g not belonging to the assumed parametric class. We now briefly survey a couple of key results of this nature, which will be useful for later developments.

First, for the i.i.d. situation, define

$$u(y,\theta) = \frac{\partial \log f(y,\theta)}{\partial\theta} \quad \text{and} \quad I(y,\theta) = \frac{\partial^2 \log f(y,\theta)}{\partial\theta\,\partial\theta^{\mathrm{t}}}. \tag{2.5}$$

The first expression is a p-vector function, often called the *score vector* of the model, with components $\partial \log f(y,\theta)/\partial\theta_j$ for $j = 1,\dots,p$. The second function is a $p \times p$ matrix, sometimes called the information matrix function for the model. Its components are the mixed second-order derivatives $\partial^2 \log f(y,\theta)/\partial\theta_j\partial\theta_k$ for $j, k = 1,\dots,p$. The score function and information matrix function are used both for numerically finding maximum likelihood estimates and for characterising their behaviour. Note that since the least false parameter minimises the Kullback–Leibler distance,

$$\mathrm{E}_g u(Y,\theta_0) = \int g(y)u(y,\theta_0)\,\mathrm{d}y = 0, \tag{2.6}$$

that is, the score function has zero mean at precisely the least false parameter value. We also need to define

$$J = -\mathrm{E}_g\, I(Y,\theta_0) \quad \text{and} \quad K = \mathrm{Var}_g\, u(Y,\theta_0). \tag{2.7}$$

These $p \times p$ matrices are identical when $g(y)$ is actually equal to $f(y,\theta_0)$ for all y. In such cases, the matrix

$$J(\theta_0) = \int f(y,\theta_0)u(y,\theta_0)u(y,\theta_0)^{\mathrm{t}}\,\mathrm{d}y = -\int f(y,\theta_0)I(y,\theta_0)\,\mathrm{d}y \tag{2.8}$$

is called the *Fisher information matrix* of the model.

Under various and essentially rather mild regularity conditions, one may prove that

$$\widehat{\theta} = \theta_0 + J^{-1}\bar{U}_n + o_P(n^{-1/2}), \tag{2.9}$$

where $\bar{U}_n = n^{-1}\sum_{i=1}^n u(Y_i,\theta_0)$; see e.g. Hjort and Pollard (1993). This may be considered the basic asymptotic description of the maximum likelihood estimator. The size of the remainder term is concisely captured by the o_P notation; that $Z_n = o_P(n^{-1/2})$ means that $\sqrt{n}Z_n$ is $o_P(1)$ and tends to zero in probability. From the central limit theorem there is convergence in distribution $\sqrt{n}\bar{U}_n \to_d U' \sim \mathrm{N}_p(0, K)$, which in combination with (2.9) leads to

$$\sqrt{n}(\widehat{\theta} - \theta_0) \overset{d}{\to} J^{-1}U' = \mathrm{N}_p(0, J^{-1}KJ^{-1}). \tag{2.10}$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Next, we deal with the regression case. To give these results, we need the $p \times 1$ score function and $p \times p$ information function of the model,

$$u(y \mid x, \theta) = \frac{\partial \log f(y \mid x, \theta)}{\partial \theta} \quad \text{and} \quad I(y \mid x, \theta) = \frac{\partial^2 \log f(y \mid x, \theta)}{\partial \theta \partial \theta^{\mathrm{t}}}.$$

Let $\theta_{0,n}$ be the least false parameter value associated with densities $g(y \mid x)$ when the covariate distribution is C_n, the empirical distribution of $x_1, \ldots, x_n$. Define the matrices

$$\begin{aligned} J_n &= -n^{-1} \sum_{i=1}^{n} \int g(y \mid x_i) I(y \mid x_i, \theta_{0,n}) \,\mathrm{d}y, \\ K_n &= n^{-1} \sum_{i=1}^{n} \mathrm{Var}_g u(Y \mid x_i, \theta_{0,n}); \end{aligned} \tag{2.11}$$

these are the regression model parallels of J and K of (2.7). Under natural conditions, of the Lindeberg type, there is convergence in probability of J_n and K_n to limits J and K, and $\sqrt{n}\bar{U}_n = n^{-1/2} \sum_{i=1}^{n} u(Y_i \mid x_i, \theta_{0,n})$ tends in distribution to a $U' \sim \mathrm{N}_p(0, K)$. An important representation for the maximum likelihood estimator is $\sqrt{n}(\widehat{\theta} - \theta_{0,n}) = J_n^{-1} \sqrt{n}\bar{U}_n + o_P(1)$, which also leads to a normal limit distribution, even when the assumed model is not equal to the true model,

$$\sqrt{n}(\widehat{\theta} - \theta_{0,n}) \xrightarrow{d} J^{-1} U' \sim \mathrm{N}_p(0, J^{-1} K J^{-1}). \tag{2.12}$$

This properly generalises (2.10). Estimators for J_n and K_n are

$$\begin{aligned} \widehat{J}_n &= -n^{-1} \partial^2 \ell_n(\widehat{\theta}) / \partial \theta \partial \theta^{\mathrm{t}} = -n^{-1} \sum_{i=1}^{n} I(y_i \mid x_i, \widehat{\theta}), \\ \widehat{K}_n &= n^{-1} \sum_{i=1}^{n} u(y_i \mid x_i, \widehat{\theta}) u(y_i \mid x_i, \widehat{\theta})^{\mathrm{t}}. \end{aligned} \tag{2.13}$$

We note that $J_n = K_n$ when the assumed model is equal to the true model, in which case $\widehat{J}_n$ and $\widehat{K}_n$ are estimators of the same matrix, cf. (2.8). The familiar type of maximum likelihood-based inference does assume that the model is correct or nearly correct, and utilises precisely that the distribution of $\widehat{\theta}$ is approximately that of a $\mathrm{N}_p(\theta_0, n^{-1}\widehat{J}_n^{-1})$, which follows from (2.12), leading to confidence intervals, p-values, and so on. Model-robust inference, in the sense of leading to approximately correct confidence intervals, etc., without the assumption of the parametric model being correct, uses a ‘sandwich matrix’ instead to approximate the variance matrix of $\widehat{\theta}$, namely $n^{-1}\widehat{J}_n^{-1}\widehat{K}_n\widehat{J}_n^{-1}$.

We now illustrate these general results for two well-known regression models.

Example 2.2 Normal linear regression

Assume $Y_i = x_i^{\mathrm{t}}\beta + \sigma\varepsilon_i$ for some p-dimensional vector β of regression coefficients, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. and standard normal under traditional conditions. Then the log-likelihood function is $\sum_{i=1}^{n} \{-\frac{1}{2}(y_i - x_i^{\mathrm{t}}\beta)^2/\sigma^2 - \log\sigma - \frac{1}{2}\log(2\pi)\}$. Assume that the ε_i are not necessarily standard normal, but that they have mean zero, standard deviation 1, skewness $\kappa_3 = \mathrm{E}\,\varepsilon_i^3$ and kurtosis $\kappa_4 = \mathrm{E}\,\varepsilon_i^4 - 3$. Then calculations lead to

$$J_n = \frac{1}{\sigma^2}\begin{pmatrix} \Sigma_n & 0 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad K_n = \frac{1}{\sigma^2}\begin{pmatrix} \Sigma_n & \kappa_3 \bar{x}_n \\ \kappa_3 \bar{x}_n^{\mathrm{t}} & 2+\kappa_4 \end{pmatrix},$$

in terms of $\Sigma_n = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}}$. ■

Example 2.3 Poisson regression

Consider a Poisson regression model for independent count data $Y_1, \ldots, Y_n$ in terms of p-dimensional covariate vectors $x_1, \ldots, x_n$, which takes Y_i to be Poisson with parameter $\xi_i = \exp(x_i^{\mathrm{t}}\beta)$. The general method outlined above leads to two matrices J_n and K_n with estimates

$$\widehat{J}_n = n^{-1}\sum_{i=1}^n \widehat{\xi}_i x_i x_i^{\mathrm{t}} \quad \text{and} \quad \widehat{K}_n = n^{-1}\sum_{i=1}^n (Y_i - \widehat{\xi}_i)^2 x_i x_i^{\mathrm{t}},$$

where $\widehat{\xi}_i = \exp(x_i^{\mathrm{t}}\widehat{\beta})$. When the assumed model is equal to the true model these matrices estimate the same quantity, but if there is over-dispersion, for example, then $n^{-1}\widehat{J}_n^{-1}\widehat{K}_n\widehat{J}_n^{-1}$ reflects the sampling variance of $\widehat{\beta} - \beta$ better than $n^{-1}\widehat{J}_n^{-1}$. See in this connection also Section 2.5. ■

2.3 AIC and the Kullback–Leibler distance

As we have seen, a parametric model M for data gives rise to a log-likelihood function $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$. Its maximiser is the maximum likelihood estimator $\widehat{\theta}$. The value of Akaike's information criterion (Akaike, 1973) for the model is defined as in (2.1), which may also be spelled out as

$$\mathrm{AIC}(M) = 2\ell_n(\widehat{\theta}) - 2\,\mathrm{length}(\theta) = 2\ell_{n,\max} - 2\,\mathrm{length}(\theta), \tag{2.14}$$

with length(θ) denoting the number of estimated parameters. To use the AIC with a collection of candidate models one computes each model's AIC value and compares these. A good model has a large value of AIC, relative to the others; cf. the general remarks made in Section 2.1. We first illustrate AIC on a data example before explaining its connection to the Kullback–Leibler distance.

Example 2.4 Low birthweight data: AIC variable selection

We continue Example 2.1. It is a priori not clear whether all variables x_i play a role in explaining low infant birthweight. Since the mother's weight is thought to be influential, we decide to include this variable x_2 in all of the possible models under investigation, as well as the intercept term ($x_1 = 1$); in other words, x_1 and x_2 are protected covariates. Let $x = (1, x_2)^{\mathrm{t}}$. Subsets of $z = (x_3, x_4, x_5)^{\mathrm{t}}$ are considered for potential inclusion. In

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Table 2.1. *AIC values for the eight logistic regression candidate models for the low birthweight data of Example 2.4.*

Extra covariates	$\ell_n(\widehat{\theta})$	length(θ)	AIC value	Preference order	ΔAIC
none	-114.345	2	-232.691		-1.616
x_3	-113.562	3	-233.123		-2.048
x_4	-112.537	3	-231.075	(1)	0.000
x_5	-114.050	3	-234.101		-3.026
x_3, x_4	-112.087	4	-232.175	(3)	-1.100
x_3, x_5	-113.339	4	-234.677		-3.602
x_4, x_5	-111.630	4	-231.259	(2)	-0.184
x_3, x_4, x_5	-111.330	5	-232.661		-1.586

this notation the logistic regression model has the formula

$$\mathrm{P}(\text{low birthweight}\,|\,x, z) = \frac{\exp(x^{\mathrm t}\beta + z^{\mathrm t}\gamma)}{1 + \exp(x^{\mathrm t}\beta + z^{\mathrm t}\gamma)},$$

with $\beta = (\beta_1, \beta_2)^{\mathrm t}$ and $\gamma = (\gamma_1, \gamma_2, \gamma_3)^{\mathrm t}$ the parameters to be estimated. For the estimators given in Example 2.1, and using the normal approximation for the maximum likelihood estimators $\widehat{\theta} = (\widehat{\beta}, \widehat{\gamma}) \approx_d \mathrm{N}_p(\theta_0, n^{-1} J_n^{-1})$, we obtain the corresponding p-values 0.222, 0.028, 0.443, 0.044, 0.218. As seen from the p-values, only γ_2 among the three γ_j is significantly different from zero at the 5% level of significance.

For this particular model it is easy to compute the maximised log-likelihood and find the required AIC values. Indeed, see Exercise 2.4,

$$\mathrm{AIC} = 2 \sum_{i=1}^{n} \{y_i \log \widehat{p}_i + (1 - y_i) \log(1 - \widehat{p}_i)\} - 2k,$$

where $\widehat{p}_i$ is the estimated probability for $Y_i = 1$ under the model and k is the number of estimated parameters. AIC selects the model including x_4 only, see Table 2.1, with estimated low birthweight probabilities

$$\widehat{\mathrm{P}}(\text{low birthweight}\,|\,x, z) = \frac{\exp(1.198 - 0.0166\,x_2 + 0.891\,x_4)}{1 + \exp(1.198 - 0.0166\,x_2 + 0.891\,x_4)}.$$

We note that AIC differences between the best ranked models are small, so we cannot claim with any degree of certainty that the AIC selected x_4 model is necessarily better than its competitors. In fact, a different recommendation will be given by the BIC method in Example 3.3.

We use this application to illustrate one more aspect of the AIC scores, namely that they are computed in a modus of comparisons across candidate models and that hence only their differences matter. For these comparisons it is often more convenient to subtract out the maximum AIC value; these are the ΔAIC scores being displayed to the right in Table 2.1. ■

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

As discussed above, the AIC method has intuitive appeal in penalising the log-likelihood maxima for complexity, but it is not clear at the outset why the penalty factor should take the particular form of (2.14). We now present the theory behind the precise form of AIC, first for the i.i.d. case, then for regression models. The key is estimating the expected value of the Kullback–Leibler distance from the unknown true data-generating density $g(\cdot)$ to the parametric model.

As we saw in Section 2.2, the maximum likelihood estimator $\widehat\theta$ aims at the least false parameter value θ_0 that minimises the Kullback–Leibler distance (2.2). To assess how well this works, compared with other parametric models, we study the actually attained Kullback–Leibler distance

$$\mathrm{KL}(g, f(\cdot,\widehat\theta)) = \int g(y)\{\log g(y) - \log f(y,\widehat\theta)\}\,\mathrm{d}y = \int g\log g\,\mathrm{d}y - R_n.$$

The first term is the same across models, so we study R_n, which is a random variable, dependent upon the data via the maximum likelihood estimator $\widehat\theta$. Its expected value is

$$Q_n = \mathrm{E}_g R_n = \mathrm{E}_g \int g(y)\log f(y,\widehat\theta)\,\mathrm{d}y. \tag{2.15}$$

The 'outer expectation' here is with respect to the maximum likelihood estimator, under the true density g for the Y_i. This is explicitly indicated in the notation by using the subscript g. The AIC strategy is in essence to estimate Q_n for each candidate model, and then to select the model with the highest estimated Q_n; this is equivalent to searching for the model with smallest estimated Kullback–Leibler distance.

To estimate Q_n from data, one possibility is to replace $g(y)\,\mathrm{d}y$ in R_n with the empirical distribution of the data, leading to

$$\widehat Q_n = n^{-1}\sum_{i=1}^n \log f(Y_i,\widehat\theta) = n^{-1}\ell_n(\widehat\theta),$$

i.e. the normalised log-likelihood maximum value. This estimator will tend to overshoot its target Q_n, as is made clear by the following key result. To state the result we need $V_n = \sqrt{n}(\widehat\theta - \theta_0)$, studied in (2.10), and involving the least false parameter θ_0; let also $\bar Z_n$ be the average of the i.i.d. zero mean variables $Z_i = \log f(Y_i,\theta_0) - Q_0$, writing $Q_0 = \int g(y)\log f(y,\theta_0)\,\mathrm{d}y$. The result is that

$$\widehat Q_n - R_n = \bar Z_n + n^{-1}V_n^{\mathrm{t}} J V_n + o_P(n^{-1}). \tag{2.16}$$

In view of (2.10) we have $V_n^{\mathrm{t}} J V_n \to_d W = (U')^{\mathrm{t}} J^{-1} U'$, where $U' \sim \mathrm{N}_q(0, K)$. Result (2.16) therefore leads to the approximation

$$\mathrm{E}(\widehat Q_n - Q_n) \approx p^*/n, \quad \text{where } p^* = \mathrm{E}\,W = \mathrm{Tr}(J^{-1}K). \tag{2.17}$$

In its turn this leads to $\widehat Q_n - p^*/n = n^{-1}\{\ell_n(\widehat\theta) - p^*\}$ as the bias-corrected version of the naive estimator $\widehat Q_n$.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

We make some remarks before turning to the proof of (2.16).

(i) If the approximating model is correct, so that $g(y) = f(y, \theta_0)$, then $J = K$, and $p^* = p = \mathrm{length}(\theta)$, the dimension of the model. Also, in that case, the overshooting quantity $n^{-1} V_n^{\mathrm{t}} J V_n$ is close to a $n^{-1}\chi_p^2$. Taking $p^* = p$, even without any check on the adequacy of the model, leads to the AIC formula (2.14).

(ii) We may call p^* of (2.17) the generalised dimension of the model. Clearly, other approximations to or estimates of p^* than the simple $p^* = p$ are possible, and this will lead to close relatives of the AIC method; see in particular Section 2.5.

(iii) There are instances where the mean of $V_n^{\mathrm{t}} J V_n$ does not tend to the mean p^* of the limit W. For the simple binomial model, for example, the mean Q_n of R_n does not even exist (since R_n is then infinite with a certain very small but positive probability); the same difficulty arises in the logistic regression model. In such cases (2.17) is formally not correct. Result (2.16) is nevertheless true and indicates that p^*/n is a sensible bias correction, with the more cautious reading 'R_n has a distribution close to that of a variable with mean equal to that of $n^{-1}\{\ell_n(\widehat{\theta}) - p^*\}$'.

Proof of (2.16). We first use a two-term Taylor expansion for R_n, using the score and information functions of the model as in (2.5), and find

$$\begin{aligned} R_n &\doteq \int g(y)\big\{\log f(y, \theta_0) + u(y, \theta_0)^{\mathrm{t}}(\widehat{\theta} - \theta_0) + \tfrac{1}{2}(\widehat{\theta} - \theta_0)^{\mathrm{t}} I(y, \theta_0)(\widehat{\theta} - \theta_0)\big\}\,\mathrm{d}y \\ &= Q_0 - \tfrac{1}{2} n^{-1} V_n^{\mathrm{t}} J V_n. \end{aligned}$$

Similarly, a two-term expansion for $\widehat{Q}_n$ leads to

$$\begin{aligned} \widehat{Q}_n &\doteq n^{-1}\sum_{i=1}^n \big\{\log f(Y_i, \theta_0) + u(Y_i, \theta_0)^{\mathrm{t}}(\widehat{\theta} - \theta_0) + \tfrac{1}{2}(\widehat{\theta} - \theta_0)^{\mathrm{t}} I(Y_i, \theta_0)(\widehat{\theta} - \theta_0)\big\} \\ &= Q_0 + \bar{Z}_n + \bar{U}_n^{\mathrm{t}}(\widehat{\theta} - \theta_0) - \tfrac{1}{2}(\widehat{\theta} - \theta_0)^{\mathrm{t}} J_n (\widehat{\theta} - \theta_0), \end{aligned}$$

where $J_n = -n^{-1}\sum_{i=1}^n I(Y_i, \theta_0) \to_p J$. This shows that $\widehat{Q}_n - R_n$ can be expressed as $\bar{Z}_n + n^{-1}\sqrt{n}\bar{U}_n^{\mathrm{t}} V_n + o_P(n^{-1})$, and in conjunction with (2.10) this yields (2.16). □

We next turn our attention to regression models of the general type discussed in Section 2.2. As we saw there, the distance measure involved when analysing maximum likelihood estimation in such models is the appropriately weighted Kullback–Leibler distance (2.4), involving also the distribution of x vectors in their space of covariates. For a given parametric model, with observed regression data $(x_1, y_1), \ldots, (x_n, y_n)$, the regression analogy to (2.15) is

$$Q_n = \mathrm{E}_g R_n = \mathrm{E}_g\, n^{-1}\sum_{i=1}^n \int g(y \mid x_i)\log f(y \mid x_i, \widehat{\theta})\,\mathrm{d}y,$$

involving the empirical distribution of the covariate vectors $x_1, \ldots, x_n$. A straightforward initial estimator of Q_n is $\widehat{Q}_n = n^{-1}\sum_{i=1}^n \log f(Y_i \mid x_i, \widehat{\theta})$, i.e. the normalised

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

log-likelihood maximum $n^{-1}\ell_{n,\max}$. Let $\theta_{0,n}$ be the least false parameter value associated with the empirical distribution of $x_1, \ldots, x_n$, i.e. the maximiser of $n^{-1}\sum_{i=1}^n \int g(y \,|\, x_i) \log f(y \,|\, x_i, \theta)\, \mathrm{d}y$. A two-term Taylor expansion leads to $R_n \doteq Q_{0,n} - \frac{1}{2} n^{-1} V_n^{\mathrm{t}} J_n V_n$, where $V_n = \sqrt{n}(\widehat{\theta} - \theta_{0,n})$ and J_n is as in (2.11); also, $Q_{0,n} = n^{-1}\sum_{i=1}^n \int g(y \,|\, x_i) \log f(y \,|\, x_i, \theta_{0,n})\, \mathrm{d}y$. Similarly, a second expansion yields

$$\begin{aligned} \widehat{Q}_n &= Q_{0,n} + \bar{Z}_n + \bar{U}_n^{\mathrm{t}}(\widehat{\theta} - \theta_{0,n}) - \tfrac{1}{2}(\widehat{\theta} - \theta_{0,n})^{\mathrm{t}} \widetilde{J}_n (\widehat{\theta} - \theta_{0,n}) \\ &= Q_{0,n} + \bar{Z}_n + \tfrac{1}{2} n^{-1} V_n^{\mathrm{t}} J_n V_n + o_P(n^{-1}), \end{aligned}$$

with $\bar{Z}_n$ being the average of the zero mean variables

$$Z_i = \log f(Y_i \,|\, x_i, \theta_{0,n}) - \int g(y \,|\, x_i) \log f(y \,|\, x_i, \theta_{0,n})\, \mathrm{d}y.$$

A clear analogy of the i.i.d. results emerges, with the help of (2.12), and with consequences parallelling those outlined above for the i.i.d. case. In particular, the AIC formula $2(\ell_{n,\max} - p)$ is valid, for the same reasons, under the same type of conditions as for i.i.d. data.

2.4 Examples and illustrations

Example 2.5 Exponential versus Weibull

For analysis of computer processes it may be important to know whether the running processes have the memory-less property or not. If they do, their failure behaviour can be described by the simple exponential model with density at failure time $= y$ equal to $\theta \exp(-\theta y)$, assuming i.i.d. data. If, on the other hand, the failure rate decreases with time (or for wear-out failures increases with time), a Weibull model may be more appropriate. Its cumulative distribution function is

$$F(y, \theta, \gamma) = 1 - \exp\{-(\theta y)^\gamma\} \quad \text{for } y > 0.$$

The density is the derivative of the cumulative distribution function, $f(y, \theta, \gamma) = \exp\{-(\theta y)^\gamma\}\theta^\gamma \gamma y^{\gamma-1}$. Note that $\gamma = 1$ corresponds to the simpler, exponential model. To select the best model, we compute

$$\begin{aligned} \mathrm{AIC(exp)} &= 2\sum_{i=1}^n (\log \widetilde{\theta} - \widetilde{\theta} y_i) - 2, \\ \mathrm{AIC(wei)} &= 2\sum_{i=1}^n \{-(\widehat{\theta} y_i)^{\widehat{\gamma}} + \widehat{\gamma} \log \widehat{\theta} + \log \widehat{\gamma} + (\widehat{\gamma} - 1) \log y_i\} - 4. \end{aligned}$$

Here $\widetilde{\theta}$ is the maximum likelihood estimator for θ in the exponential model, while $(\widehat{\theta}, \widehat{\gamma})$ are the maximum likelihood estimators in the Weibull model. The model with the biggest value of AIC is chosen as the most appropriate one for the data at hand. See Exercise 3.1. ■

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Example 2.6 Mortality in ancient Egypt

How long is a life? A unique set of lifelengths in Roman Egypt was collected by W. Spiegelberg in 1901 and analysed by Karl Pearson (1902) in the very first volume of *Biometrika*. The data set contains the age at death for 141 Egyptian mummies in the Roman period, 82 men and 59 women, dating from around year 100 B.C. The life-lengths vary from 1 to 96, and Pearson argued that these can be considered a random sample from one of the better-living classes in that society, at a time when a fairly stable and civil government was in existence. Pearson (1902) did not try any parametric models for these data, but discussed differences between the Egyptian age distribution and that of England 2000 years later. We shall use AIC to select the most successful of a small collection of candidate parametric models for mortality rate.

For each suggested model $f(t,\theta)$ we maximise the log-likelihood $\ell_n(\theta) = \sum_{i=1}^n \log f(t_i,\theta)$, writing $t_1,\ldots,t_n$ for the life-lengths, and then compute $\mathrm{AIC} = 2\ell_n(\widehat{\theta}) - 2p$, with $p = \mathrm{length}(\theta)$. We note that finding the required maximum likelihood estimates has become drastically simpler than it used to be a decade or more ago, thanks to easily available optimisation algorithms in software packages. As demonstrated in Exercise 2.3, it does not take many lines of `R` code, or minutes of work, per model, to (i) program the log-likelihood function, using the `function` mechanism; (ii) find its maximiser, via the nonlinear minimisation algorithm `nlm`; and (iii) use this to find the appropriate AIC value. We have done this for five models:

- Model 1 is the exponential, with density $b\exp(-bt)$, for which we find $\widehat{b} = 0.033$.
- Model 2 is the Gamma density $\{b^a/\Gamma(a)\}t^{a-1}\exp(-bt)$, with parameter estimates $(\widehat{a},\widehat{b}) = (1.609, 0.052)$.
- Model 3 is the log-normal, which takes a $\mathrm{N}(\mu,\sigma^2)$ for the log-life-lengths, corresponding to a density $\phi\{(\log t-\mu)/\sigma\}/(\sigma t)$; here we find parameter estimates $(\widehat{\mu},\widehat{\sigma}) = (3.082, 0.967)$.
- Model 4 is the Gompertz, which takes the hazard or mortality rate $h(t) = f(t)/F[t,\infty)$ to be of the form $a\exp(bt)$; this corresponds to the density $f(t) = \exp\{-H(t)\}h(t)$, with $H(t) = \int_0^t h(s)\,\mathrm{d}s = (a/b)\{\exp(bt)-1\}$ being the cumulative hazard rate. Parameter estimates are $(\widehat{a},\widehat{b}) = (0.019, 0.021)$.
- Finally model 5 is the Makeham extension of the Gompertz, with hazard rate $h(t) = k + a\exp(bt)$, for k such that $k + a\exp(bt_0) > 0$, where t_0 is the minimum age under consideration (for this occasion, $t_0 = 1$ year). Estimates are $(-0.012, 0.029, 0.016)$ for (k, a, b).

We see from the AIC values for models 1–5, listed in Table 2.2, that the two-parameter Gompertz model (model 4) is deemed the most successful. Figure 2.1 displays the mortality rate $\widehat{a}\exp(\widehat{b}t)$ for the Egyptian data, along with a simple nonparametric estimate. The nonparametric estimate in Figure 2.1 is of the type ‘parametric start times nonparametric correction’, with a bandwidth increasing with decreasing risk set; see Hjort (1992b). It indicates in this case that the Gompertz models are perhaps acceptable approximations, but that there are other fluctuations at work not quite captured by the parametric models, as for example the extra mortality at age around 25. The AIC analysis shows otherwise

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Table 2.2. *Mortality in ancient Egypt: parameter estimates, the maximised log-likelihoods and the AIC scores, for the nine models. The Gompertz models are better than the others.*

Parameters	Parameter estimates				$\ell_n(\widehat{\theta})$	AIC	
model 1, b:	0.033				−623.777	−1249.553	
model 2, a, b:	1.609	0.052			−615.386	−1234.772	
model 3, μ, σ:	3.082	0.967			−629.937	−1263.874	
model 4, a, b:	0.019	0.021			−611.353	−1226.706	
model 5, k, a, b:	−0.012	0.029	0.016		−611.319	−1228.637	
model 6, a, b:	0.019	0.021			−611.353	−1226.706	
model 7, a, b_1, b_2:	0.019	0.018	0.026		−610.076	−1226.151	(3)
model 8, a_1, b, a_2:	0.016	0.024	0.022		−608.520	−1223.040	(1)
model 9, a_1, b_1, a_2, b_2:	0.016	0.024	0.022	0.020	−608.520	−1225.040	(2)

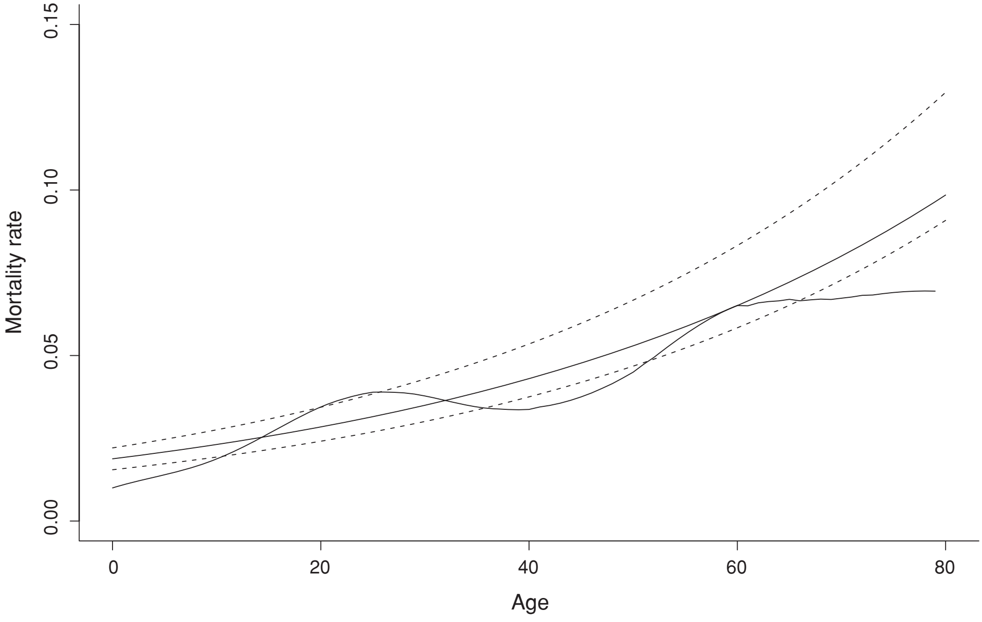

Fig. 2.1. How long is a life? For the 141 lifetimes from ancient Egypt we show the Gompertz-fitted hazard rates, for the full population (solid line), for women (dotted line, above) and for men (dotted line, below). The wiggly curve is a nonparametric hazard rate estimate.

that it does not really pay to include the extra Makeham parameter k, for example; the max log-likelihood increases merely from −611.353 to −611.319, which is not enough, as judged by the AIC value. Inclusion of one more parameter in a model is only worthwhile if the max log-likelihood is increased by at least 1.

Using the Gompertz model we attempt to separate men's and women's mortality in ancient Egypt, in spite of Pearson (1902) writing 'in dealing with [these data] I have not ventured to separate the men and women mortality, the numbers are far too insignificant'. We try out four models, corresponding to (model 6) using the same parameters (a, b) for both men and women (this is the same as model 4, of course); (model 7) using (a, b_1) and (a, b_2) for men and women (that is with the same a parameter); (model 8) using (a_1, b) and (a_2, b) for men and women (that is with the same b parameter); and (model 9) using (a_1, b_1) and (a_2, b_2) without common parameters for the two groups.

From the results listed in Table 2.2 we see that model 8 is the best one, estimating men's mortality rate as $0.016 \exp(0.022\, t)$ and women's as $0.024 \exp(0.022\, t)$; see Figure 2.1. ■

Example 2.7 Linear regression: AIC selection of covariates

The traditional linear regression model for response data y_i in relation to covariate vectors $x_i = (x_{i,1}, \ldots, x_{i,p})^{\rm t}$ for individuals $i = 1, \ldots, n$ is to take

$$Y_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \varepsilon_i = x_i^{\rm t}\beta + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with $\varepsilon_1, \ldots, \varepsilon_n$ independently drawn from $\mathrm{N}(0, \sigma^2)$ and $\beta = (\beta_1, \ldots, \beta_p)^{\rm t}$ a vector of regression coefficients. Typically one of the $x_{i,j}$, say the first, is equal to the constant 1, so that β_1 is the intercept parameter. The model is more compactly written in matrix form as $Y = X\beta + \varepsilon$, where $Y = (Y_1, \ldots, Y_n)^{\rm t}$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\rm t}$, and X is the $n \times p$ matrix having $x_i^{\rm t}$ as its ith row.

The log-likelihood function is

$$\ell_n(\beta, \sigma) = \sum_{i=1}^n \big\{ -\log\sigma - \tfrac{1}{2}(y_i - x_i^{\rm t}\beta)^2/\sigma^2 - \tfrac{1}{2}\log(2\pi) \big\}.$$

Maximisation with respect to β is equivalent to

$$\text{minimisation of } \mathrm{SSE}(\beta) = \sum_{i=1}^n (y_i - x_i^{\rm t}\beta)^2 = \|Y - X\beta\|^2,$$

which is also the definition of the least squares estimator. The solution can be written

$$\widehat\beta = (X^{\rm t}X)^{-1}X^{\rm t}Y = \Sigma_n^{-1} n^{-1} \sum_{i=1}^n x_i Y_i,$$

where $\Sigma_n = n^{-1}X^{\rm t}X = n^{-1}\sum_{i=1}^n x_i x_i^{\rm t}$, assuming that X has full rank p, making $X^{\rm t}X$ an invertible $p \times p$ matrix. The maximum likelihood estimator of σ is the maximiser of $\ell_n(\widehat\beta, \sigma)$, and is the square root of

$$\widehat\sigma^2 = n^{-1}\mathrm{SSE}(\widehat\beta) = n^{-1}\sum_{i=1}^n \mathrm{res}_i^2 = n^{-1}\|\mathrm{res}\|^2,$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

involving the residuals $\mathrm{res}_i = Y_i - x_i^{\mathrm{t}}\widehat{\beta}$. Plugging in $\widehat{\sigma}$ in $\ell_n(\widehat{\beta}, \sigma)$ gives $\ell_{n,\max} = -n \log \widehat{\sigma} - \frac{1}{2}n - \frac{1}{2}\log(2\pi)$ and

$$\mathrm{AIC} = -2n \log \widehat{\sigma} - 2(p+1) - n - n \log(2\pi). \tag{2.18}$$

Thus the best subset of covariates to use, according to the AIC method, is determined by minimising $n \log \widehat{\sigma} + p$, across all candidate models. In Section 2.6 we obtain a different estimator of the expected Kullback–Leibler distance between the estimated normal regression model and the unknown true model, which leads to a stricter complexity penalty. ■

Example 2.8 Predicting football match results

To what extent can one predict results of football matches via statistical modelling? We use the data on scores of 254 football matches; see Section 1.6 for more details regarding these data.

Denote by y and y' the number of goals scored by the teams in question. A natural type of model for outcomes (y, y') is to take these independent and Poisson distributed, with parameters (λ, λ'), with different possible specialisations for how λ and λ' should depend on the FIFA ranking scores of the two teams, say fifa and fifa$'$. A simple possibility is

$$\lambda = \lambda_0(\mathrm{fifa}/\mathrm{fifa}')^{\beta} \quad \text{and} \quad \lambda' = \lambda_0(\mathrm{fifa}'/\mathrm{fifa})^{\beta},$$

where λ_0 and β are unknown parameters that have to be estimated. The Norwegian Computing Centre (see `vm.nr.no`), which produces predictions before and during these championships, uses models similar in spirit to the model above. This is a log-linear Poisson regression model in $x = \log(\mathrm{fifa}/\mathrm{fifa}')$, with $\lambda = \exp(\alpha + \beta x)$.

We shall in fact discuss four different candidate models here. The most general is model M_3, which takes

$$\lambda(x) = \begin{cases} \exp\{a + c(x - x_0)\} & \text{for } x \le x_0, \\ \exp\{a + b(x - x_0)\} & \text{for } x \ge x_0, \end{cases} \tag{2.19}$$

where x_0 is a threshold value on the x scale of logarithmic ratios of FIFA ranking scores. In our illustration we are using the fixed value $x_0 = -0.21$. This value was found via separate profile likelihood analysis of other data, and affects matches where the ratio of the weaker FIFA score to the stronger FIFA score is less than $\exp(x_0) = 0.811$. Model M_3 has three free parameters. Model M_2 is the hockey-stick model where $c = 0$, and gives a constant rate for $x \le x_0$. Model M_1 takes $b = c$, corresponding to the traditional log-linear Poisson rate model with $\exp\{a + b(x - x_0)\}$ across all x values. Finally M_0 is the simplest one, with $b = c = 0$, leaving us with a constant $\lambda = \exp(a)$ for all matches. The point of the truncated M_2 is that the log-linear model M_1 may lead to too small goal scoring rates for weaker teams meeting stronger teams.

Models M_0 and M_1 are easily handled using Poisson regression routines, like the `glm(y ~ x, family = poisson)` algorithm in `R`, as they correspond directly to

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Table 2.3. *AIC and BIC scores for the four football match models of Example 2.8. The two criteria agree that model M_2 is best.*

Model	$\widehat{a}$	$\widehat{b}$	$\widehat{c}$	AIC	BIC
M_0	0.211	0.000	0.000	−1487.442	−1491.672
M_1	−0.174	1.690	1.690	−1453.062	−1461.523
M_2	−0.208	1.811	0.000	−1451.223	−1459.684
M_3	−0.235	1.893	−1.486	−1452.488	−1465.180

constant and log-linear modelling in x. To show how also model M_3 can be dealt with, write the indicator function $I(x) = I\{x \le x_0\}$. Then

$$\begin{aligned}\log\lambda(x) &= a + cI(x)(x - x_0) + b\{1 - I(x)\}(x - x_0)\\ &= a + b(x - x_0) + (c - b)I(x)(x - x_0),\end{aligned}$$

which means that this is a log-linear Poisson model in the two covariates $x - x_0$ and $I(x)(x - x_0)$. Model M_2 can be handled similarly.

Table 2.3 gives the result of the AIC analysis, indicating in particular that model M_2 is judged the best one. (We have also included the BIC scores, see Chapter 3; these agree with AIC that model M_2 is best.) The reason why the hockey-stick model M_2 is better than, for example, the more traditional model M_1 is that even when teams with weak FIFA score tend to lose against teams with stronger FIFA score, they still manage, sufficiently often, to score say one goal. This ‘underdog effect’ is also seen in Figure 2.2, which along with the fitted intensity for models M_1 and M_2 displays a nonparametric estimate of the $\lambda(x)$. Such an estimator is constructed locally and without any global parametric model specification. For this reason it is often used to make a comparison with parametric estimators, such as the ones obtained from models M_1 and M_2. A good parametric estimator will roughly follow the same trend as the nonparametric estimator. The latter one is defined as follows. The estimator $\widehat{\lambda}(x) = \exp(\widehat{a}_x)$, where $(\widehat{a}_x, \widehat{b}_x)$ are the parameter values maximising the kernel-smoothed log-likelihood function

$$\sum_{i=1}^{n} K_h(x_i - x)\{y_i(a + bx_i) - \exp(a + bx_i) - \log(y_i!)\},$$

where $K_h(u) = h^{-1}K(h^{-1}u)$ is a scaled version of a kernel function K. In this illustration we took K equal to the standard normal density function and selected h via a cross-validation argument. For general material on such local log-likelihood smoothing of parametric families, see Fan and Gijbels (1996), Hastie and Tibshirani (1990), and Hjort and Jones (1996).

We return to the football prediction problem in Example 3.4 (using BIC) and in Section 6.6.4 (using FIC). Interestingly, while both AIC and BIC agree that model M_2

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

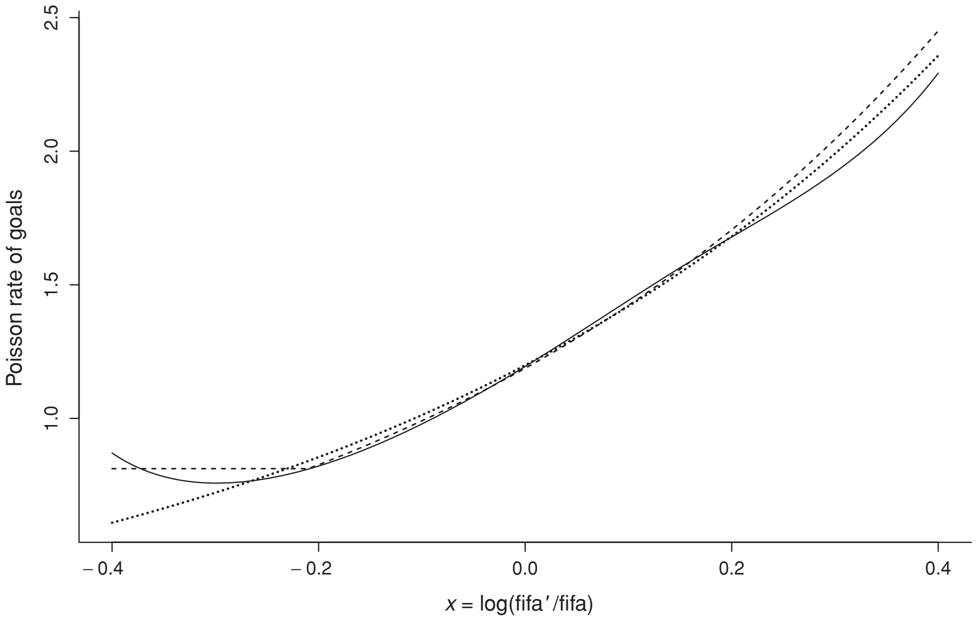

Fig. 2.2. The figure shows the fitted Poisson intensity rate $\widehat{\lambda}(x)$ of goals scored per match, as a function of $x = \log(\text{fifa}/\text{fifa}')$, where fifa and fifa$'$ are the FIFA ranking scores for the team and its opponent, for two different parametric models. These are the log-linear model $\lambda = \exp\{a + b(x - x_0)\}$ (dotted line) and the hockey-stick model (dashed line) where $\lambda(x)$ is $\exp(a)$ for $x \le x_0$ and $\exp\{a + b(x - x_0)\}$ for $x \ge x_0$, and $x_0 = -0.21$. Also shown is a nonparametric kernel-smoothed log-likelihood-based estimate (solid line).

is best, we find using the FIC methods of Chapter 6 that model M_1 may sometimes be best for estimating the probability of the event 'team 1 defeats team 2'. ■

Example 2.9 Density estimation via AIC

Suppose independent data $X_1, \ldots, X_n$ come from an unknown density f. There is a multitude of nonparametric methods for estimating f, chief among them methods using kernel smoothing. Parametric methods in combination with a model selection method are an easy-to-use alternative. We use AIC to select the right degree of complexity in the description of the density, starting out from

$$f_m(x) = f_0(x)\exp\Big\{\sum_{j=1}^{m} a_j\psi_j(x)\Big\}\Big/ c_m(a),$$

where f_0 is some specified density and the normalising constant is defined as $c_m(a) = \int f_0 \exp(\sum_{j=1}^m a_j\psi_j)\,\mathrm{d}x$. The basis functions $\psi_1, \psi_2, \ldots$ are orthogonal with respect to f_0, in the sense that $\int f_0\psi_j\psi_k\,\mathrm{d}x = \delta_{j,k} = I\{j = k\}$. We may for example take $\psi_j(x) = \sqrt{2}\cos(j\pi F_0(x))$, where F_0 is the cumulative distribution with f_0 as density. Within this

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

family f_m the maximum likelihood estimators $\widehat{a} = (\widehat{a}_1, \ldots, \widehat{a}_m)$ are those that maximise

$$\ell_n(a) = n\left\{\sum_{j=1}^{m} a_j\bar{\psi}_j - \log c_m(a)\right\},$$

where $\bar{\psi}_j = n^{-1}\sum_{i=1}^{n}\psi_j(X_i)$ and where we disregard terms not depending on a. This function is concave and has a unique maximiser as long as $n > m$. AIC selects its optimal order $\widehat{m}$ to maximise $\mathrm{AIC}(m) = 2\,\ell_n(\widehat{a}) - 2m$, perhaps among all $m \le m_{\max}$ for a reasonable upper bound of complexity. The end result is a semiparametric density estimator $\widehat{f}(x) = f_{\widehat{m}}(x)$, which may well do better than full-fledged nonparametric estimators in cases where a low-order sum captures the main aspects of an underlying density curve. See also Example 3.5 for an extension of the present method, Exercise 7.6 for further analysis, and Chapter 8 for more on order selection in combination with hypothesis testing. ■

Example 2.10 Autoregressive order selection

When a variable is observed over time, the correlation between observations needs to be carefully modelled. For a stationary time series, that is a time series for which the statistical properties such as mean, autocorrelation and variance do not depend on time, an autoregressive (AR) model is often suitable; see e.g. Brockwell and Davis (1991). In such a model, the observation at time t is written in the form

$$X_t = a_1X_{t-1} + a_2X_{t-2} + \cdots + a_kX_{t-k} + \varepsilon_t,$$

where the error terms ε_t are independent and identically distributed as $\mathrm{N}(0, \sigma^2)$ and the variables X_t have been centred around their mean. We make the assumption that the coefficients a_j are such that the complex polynomial $1 - a_1z - \cdots - a_kz^k$ is different from zero for $|z| \le 1$; this ensures that the time series is stationary. The value k is called the order of the autoregressive model, and the model is denoted by AR(k). If the value of k is small, only observations in the nearby past influence the current value X_t. If k is large, long-term effects in the past will still influence the present observation X_t. Knowing the order of the autoregressive structure is especially important for making predictions about the future, that is, predicting values of $X_{t+1}, X_{t+2}, \ldots$ when we observe the series up to and including time t; AIC can be used to select an appropriate order k. For a number of candidate orders $k = 1, 2, \ldots$ we construct AIC(k) by taking twice the value of the maximised log-likelihood for that AR(k) model, penalised with twice the number of estimated parameters, which is equal to $k + 1$ (adding one for the estimated standard deviation σ). Leaving out constants not depending on k, AIC takes a similar formula as for linear regresssion models, see (2.18). Specifically, for selecting the order k in AR(k) time series models, AIC boils down to computing

$$\mathrm{AIC}(k) = -2n\log\widehat{\sigma}_k - 2(k+1),$$

where $\widehat{\sigma}_k$ is the maximum likelihood standard deviation estimator in the model with order k. The value of the autoregressive order k which corresponds to the largest AIC(k)

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

identifies the best model. Being 'best' here, according to the AIC method, should be interpreted as the model that has the smallest estimated expected Kullback–Leiber distance from the true data-generating model. Order selection for autoregressive moving average (ARMA) models is treated extensively in Choi (1992). ■

Example 2.11 The exponential decay of beer froth*

Three German beers are investigated for their frothiness: Erdinger Weißbier, Augustinerbräu München, and Budweiser Budvar. For a given beer make, the amount of froth $V_j^0(t_i)$ is measured, in centimetres, after time t_i seconds has passed since filling. The experiment is repeated $j = 1, \ldots, m$ times with the same make (where m was 7, 4, 4 for the three brands). Observation time points $t_0 = 0, t_1, \ldots, t_n$ spanned 6 minutes (with spacing first 15 seconds, later 30 and 60 seconds). Since focus here is on the decay, let $V_j(t_i) = V_j^0(t_i)/V_j^0(t_0)$; these ratios start at 1 and decay towards zero. Leike (2002) was interested in the exponential decay hypothesis, which he formulated as

$$\mu(t) = \mathrm{E}\, V_j(t) = \exp(-t/\tau) \quad \text{for } t \geq 0,\ j = 1, \ldots, m.$$

His main claims were that (i) data supported the exponential decay hypothesis; (ii) precise estimates of the decay parameter τ can be obtained by a minimum χ^2 type procedure; and (iii) that different brands of beers have decay parameters that differ significantly from each other. Leike's 2002 paper was published in the *European Journal of Physics*, and landed the author the Ig Nobel Prize for Physics that year.

Here we shall compare three models for the data, and reach somewhat sharper conclusions than Leike's. Model M_1 is the one indirectly used in Leike (2002), that observations are independent with

$$V_j(t_i) \sim \mathrm{N}(\mu_i(\tau), \sigma_i^2) \quad \text{where } \mu_i(\tau) = \exp(-t_i/\tau)$$

for time points t_i and repetitions $j = 1, \ldots, m$. This is an example of a nonlinear normal regression model. The log-likelihood function is

$$\begin{aligned}\ell_n &= \sum_{i=1}^n \sum_{j=1}^m \Big[-\tfrac{1}{2}\frac{\{V_j(t_i) - \mu_i(\tau)\}^2}{\sigma_i^2} - \log\sigma_i - \tfrac{1}{2}\log(2\pi)\Big] \\ &= m\sum_{i=1}^n \Big[-\tfrac{1}{2}\frac{\widetilde{\sigma}_i^2 + \{\bar{V}(t_i) - \mu_i(\tau)\}^2}{\sigma_i^2} - \log\sigma_i - \tfrac{1}{2}\log(2\pi)\Big],\end{aligned}$$

where $\widetilde{\sigma}_i^2 = m^{-1}\sum_{j=1}^m \{V_j(t_i) - \bar{V}(t_i)\}^2$. To find the maximum likelihood estimates $(\widehat{\tau}, \widehat{\sigma}_1, \ldots, \widehat{\sigma}_n)$ we first maximise for fixed τ, and find

$$\widehat{\sigma}_i(\tau)^2 = \widetilde{\sigma}_i^2 + \{\bar{V}(t_i) - \mu_i(\tau)\}^2 \quad \text{for } i = 1, \ldots, n,$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Table 2.4. *Decay parameter estimates and AIC scores for three beer froth decay models, for the three German beer brands. Model 1 has 15 parameters per beer whereas models 2 and 3 have two parameters per beer. Model 3 is judged the best.*

	Model 1		Model 2		Model 3	
	τ	AIC	τ	AIC	τ	AIC
beer 1	275.75	235.42	277.35	254.18	298.81	389.10
beer 2	126.50	78.11	136.72	92.62	112.33	66.59
beer 3	167.62	130.21	166.07	150.80	129.98	128.60
Total:		443.75		497.59		584.28

leading to a log-likelihood profile of the form $-\frac{1}{2}m$, $-m\sum_{i=1}^{n}\log\widehat{\sigma}_i(\tau) - \frac{1}{2}mn\log(2\pi)$. This is then maximised over τ, yielding maximum likelihood estimates along with

$$\ell_{\max,1} = -\tfrac{1}{2}mn - m\sum_{i=1}^{n}\log\widehat{\sigma}_i - \tfrac{1}{2}mn\log(2\pi)$$

and $\mathrm{AIC}(M_1) = 2(\ell_{\max,1} - n - 1)$. This produces decay parameter estimates shown in Table 2.4, for the three brands. Leike used a quite related estimation method, and found similar values. (More specifically, his method corresponds to the minimum χ^2 type method that is optimal provided the 14 σ_i parameters were known, with inserted estimates for these.)

Model M_1 employs different standard deviation parameters for each time point, and has accordingly a rather high number of parameters. Model M_2 is the simplification where the σ_i are set equal across time. The log-likelihood function is then

$$\ell_n = m\sum_{i=1}^{n}\Big[-\tfrac{1}{2}\frac{\widetilde{\sigma}_i^2 + \{\bar{V}(t_i) - \mu_i(\tau)\}^2}{\sigma^2} - \log\sigma - \tfrac{1}{2}\log(2\pi)\Big].$$

Maximising for fixed τ gives the equation

$$\widehat{\sigma}(\tau)^2 = n^{-1}\sum_{i=1}^{n}[\widetilde{\sigma}_i^2 + \{\bar{V}(t_i) - \mu_i(\tau)\}^2]$$

and a corresponding log-profile function in τ. Maximising with respect to τ gives

$$\ell_{\max,2} = -\tfrac{1}{2}mn - mn\log\widehat{\sigma}(\widehat{\tau}) - \tfrac{1}{2}mn\log(2\pi),$$

and $\mathrm{AIC}(M_2) = 2(\ell_{\max,2} - 2)$.

Models M_1 and M_2 both use an assumption of independence from one time observation point to the next. This is unreasonable in view of the decay character of the data, even when the time difference between observations is 15 seconds and more. A more direct probability model M_3, that takes the physical nature of the decay process into account, uses $V_j(t) = \exp\{-Z_j(t)\}$, where Z_j is taken to have independent, non-negative

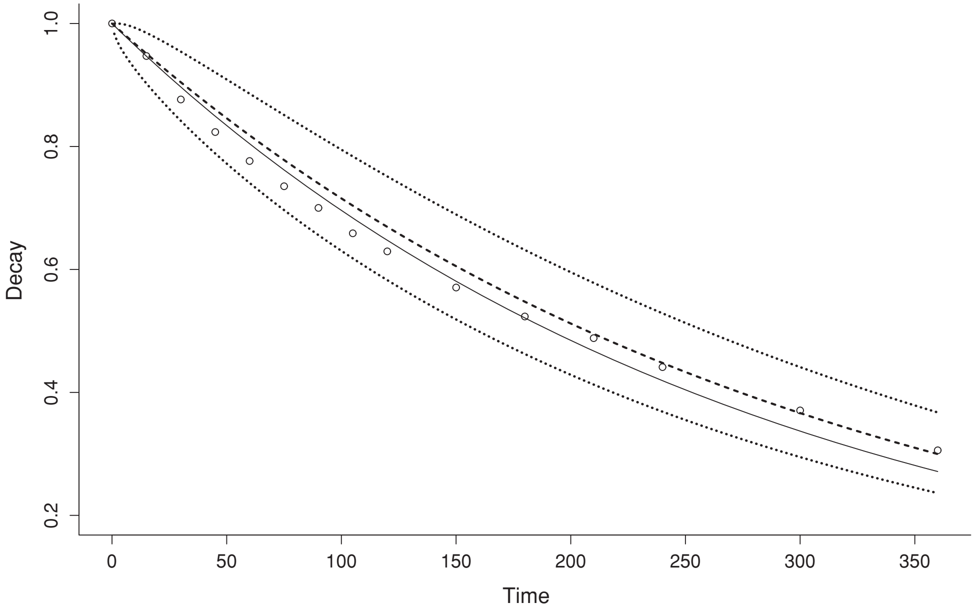

Fig. 2.3. Exponential decay of beer froth: the small circles indicate $\bar{V}(t)$, the average observed decay for seven mugs of Erdinger Weißbier, over 6 minutes. The solid line and broken line indicate respectively Leike's estimate of $\exp(-t/\tau)$ and our estimate of $\exp\{-at\log(1+1/b)\}$. The lower and upper curves form a pointwise 90% prediction band for the decay of the next mug of beer.

increments. Thus $Z_j(t) = -\log V_j(t)$ is a Lévy process, for which probability theory guarantees that

$$\mathrm{E}\exp\{-\theta Z(t)\} = \exp\{-K(t,\theta)\} \quad \text{for } t \geq 0 \text{ and } \theta \geq 0,$$

where $K(t,\theta)$ can be represented as $\int_0^\infty \{1-\exp(-\theta s)\}\,\mathrm{d}N_t(s)$. See, for example, Hjort (1990) for such Lévy theory. Exponential decay corresponds to the Lévy measures N_t being time-homogeneous, i.e. of the form $\mathrm{d}N_t(s) = t\,\mathrm{d}N(s)$ for a single Lévy measure N. Then

$$\mathrm{E}\exp\{-\theta Z(t)\} = \exp\{-tK(\theta)\} \quad \text{for } t \geq 0,$$

which also implies $\tau = 1/K(1)$. This is the Khintchine formula for Lévy processes. The simplest such Lévy model is that of a Gamma process. So we take the increments $\Delta Z_j(t_i) = Z_j(t_i) - Z_j(t_{i-1})$ to be independent Gamma variables, with parameters $(a(t_i - t_{i-1}), b)$. Note that

$$\mathrm{E}\,V(t) = \mathrm{E}\exp\{-Z(t)\} = \{b/(b+1)\}^{at} = \exp\{-at\log(1+1/b)\},$$

in agreement with the exponential decay hypothesis. Maximum likelihood estimates have been found numerically, leading also to estimates of τ and to AIC scores, as presented in Table 2.4.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

The conclusion is that the Lévy approach to beer froth watching is more successful than those of nonlinear normal regression (whether homo- or heteroscedastic); see the AIC values in the table. Figure 2.3 shows Leike's original decay curve estimate, along with that associated with model 3. Also given is a pointwise 90% confidence band for the froth decay of the next mug of Erdinger Weißbier. For more details and for comparisons with yet other models for such decay data, including Brownian bridge goodness-of-fit methods for assessing the exponential decay hypothesis, see Hjort (2007c). ■

2.5 Takeuchi's model-robust information criterion

The key property underlying the AIC method, as identified in (2.16)–(2.17), is that the bias of the estimator $\widehat{Q}_n$ can be approximated by the generalised dimension p^*/n, more precisely $\mathrm{E}(\widehat{Q}_n - Q_n) = p^*/n + o(1/n)$. Different approximations to the bias of $\widehat{Q}_n$ are obtained by using different estimators $\widehat{p}^*$ of p^*, leading to the bias correction $n^{-1}(\ell_{n,\max} - \widehat{p}^*)$ for estimating Q_n.

Using AIC in its most familiar form (2.14) amounts to simply setting the p^* of (2.17) equal to the dimension of the model $p = \mathrm{length}(\theta)$. In case the model used is equal to the true model that generated the data, it indeed holds that both dimensions are equal, thus $p^* = p$, but this is not true in general. A more model-robust version can be used in case one does not want to make the assumption that the model used is the true model. Therefore we estimate p^* by plugging in estimates of the matrices J and K. Takeuchi (1976) proposed such an estimator and the corresponding criterion,

$$\mathrm{TIC} = 2\,\ell_n(\widehat{\theta}) - 2\widehat{p}^* \quad \text{with} \quad \widehat{p}^* = \mathrm{Tr}(\widehat{J}^{-1}\widehat{K}), \tag{2.20}$$

with estimators $\widehat{J}$ and $\widehat{K}$ as in (2.13). One should consider (2.20) as an attempt at making an AIC-type selection criterion that is robust against deviations from the assumption that the model used is correct. Note that this model-robustness issue is different from that of achieving robustness against outliers; the TIC as well as AIC rest on the use of maximum likelihood estimators and as such are prone to being overly influenced by outlying data values, in many models. Selection criteria made to be robust in this sense are developed in Section 2.10.

We now give an example in which the candidate model is incorrect. Suppose that the Y_i come from some g and that the $\mathrm{N}(\mu, \sigma^2)$ model is used as a candidate model. Then it can be shown that

$$J = \frac{1}{\sigma_0^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad K = \frac{1}{\sigma_0^2}\begin{pmatrix} 1 & \kappa_3 \\ \kappa_3 & 2+\kappa_4 \end{pmatrix},$$

in terms of the skewness $\kappa_3 = \mathrm{E}U^3$ and kurtosis $\kappa_4 = \mathrm{E}U^4 - 3$ of $U = (Y - \mu_0)/\sigma_0$. This leads to $p^* = 2 + \frac{1}{2}\kappa_4$ and $\widehat{p}^* = 2 + \frac{1}{2}\widehat{\kappa}_4$, with $\widehat{\kappa}_4$ an estimator for the kurtosis of the residual distribution.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

More generally we can consider the linear regression model $Y_i = x_i^{\rm t}\beta + \sigma\varepsilon_i$, involving a k-dimensional parameter $\beta = (\beta_1, \ldots, \beta_k)^{\rm t}$, where the ε_i are i.i.d. with mean zero and unit variance, but not necessarily normal. Then we find the matrices J_n and K_n as in Example 2.2. This gives the formula $p^* = k + 1 + \frac{1}{2}\kappa_4$ and $\widehat{p}^* = k + 1 + \frac{1}{2}\widehat{\kappa}_4$ for the generalised dimension of the model.

Sometimes there would be serious sampling variability for the trace of $\widehat{J}^{-1}\widehat{K}$, particularly if the dimension $p \times p$ of the two matrices is not small. In such cases one looks for less variable estimates for p^* of (2.17). Consider for illustration the problem of selecting regressors in Poisson count models, in the framework of Example 2.3. The TIC method may be used, with $\widehat{J}_n$ and $\widehat{K}_n$ given in that example. With certain over-dispersion models one might infer that $K_n = (1 + d)J_n$, for a dispersion parameter $1 + d$, and methods are available for estimating d and the resulting $p^* = \mathrm{Tr}(J_n^{-1}K_n) = p(1 + d)$ more precisely than with the (2.20) formula. See also Section 2.7.

It is also worth noting that both the AIC and TIC methods may be generalised to various other types of parametric model, with appropriate modifications of the form of $\widehat{J}$ and $\widehat{K}$ and hence $\widehat{p}^*$ above; see Example 3.10 for such an illustration, in a context of parametric hazard regression models.

2.6 Corrected AIC for linear regression and autoregressive time series

It is important to realise that AIC typically will select more and more complex models as the sample size increases. This is because the maximal log-likelihood will increase linearly with n while the penalty term for complexity is proportional to the number of parameters. We now examine the linear regression model in more detail. In particular we shall see how some exact calculations lead to sample-size modifications of the direct AIC.

In Example 2.7 we considered the general linear regression model $Y = X\beta + \varepsilon$ and found that the direct AIC could be expressed as

$$\mathrm{AIC} = -2n \log \widehat{\sigma} - 2(p + 1) - n - n \log(2\pi), \tag{2.21}$$

with $\widehat{\sigma}^2 = \|\mathrm{res}\|^2/n$ from residuals $\mathrm{res} = Y - X\widehat{\beta}$; in particular, the AIC advice is to choose the candidate model that minimises $n \log \widehat{\sigma} + p$ across candidate models. The aim of AIC is to estimate the expected Kullback–Leibler distance from the true data-generating mechanism $g(y \mid x)$ to the estimated model $f(y \mid x, \widehat{\theta})$, see Section 2.3, where in this situation $\theta = (\beta, \sigma)$. Assume here that $g(y \mid x)$ has mean $\xi(x)$ and constant standard deviation σ_0. If the assumed model is equal to the true model, then $\xi_i = \xi(x_i) = x_i^{\rm t}\beta$.

We are not required to always use the maximum likelihood estimators. Here it is natural to modify $\widehat{\sigma}^2$ above, for example, since for the case that the assumed model is equal to the true model $\mathrm{SSE} = \|\mathrm{res}\|^2 \sim \sigma^2\chi^2_{n-p}$, which means that $\|\mathrm{res}\|^2$ should be divided by $n - p$ rather than n to make it an unbiased estimator. This is commonly done when computing estimates, but is not typical practice when working with AIC.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Write in general

$$\widehat{\sigma}^2 = \frac{\|\mathrm{res}\|^2}{n-a} = \frac{1}{n-a}\sum_{i=1}^n (Y_i - x_i^{\mathrm{t}}\widehat{\beta})^2, \tag{2.22}$$

with the cases $a = 0$ and $a = p$ corresponding to maximum likelihood and unbiased estimation respectively. In the spirit of the derivation of AIC, we examine how much

$$\begin{aligned}\widehat{Q}_n &= n^{-1}\sum_{i=1}^n \log f(Y_i \mid x_i, \widehat{\beta}, \widehat{\sigma}) \\ &= n^{-1}\sum_{i=1}^n \big\{ -\log\widehat{\sigma} - \tfrac12 (Y_i - x_i^{\mathrm{t}}\widehat{\beta})^2/\widehat{\sigma}^2 - \tfrac12 \log(2\pi)\big\} \\ &= -\log\widehat{\sigma} - \tfrac12 \frac{n-a}{n} - \tfrac12\log(2\pi)\end{aligned}$$

can be expected to overestimate

$$\begin{aligned}R_n &= n^{-1}\sum_{i=1}^n \int g(y \mid x_i)\log f(y \mid x_i, \widehat{\beta}, \widehat{\sigma})\,\mathrm{d}y \\ &= -\log\widehat{\sigma} - \tfrac12 n^{-1}\sum_{i=1}^n \frac{(\xi_i - x_i^{\mathrm{t}}\widehat{\beta})^2 + \sigma_0^2}{\widehat{\sigma}^2} - \tfrac12\log(2\pi).\end{aligned}$$

We find

$$\mathrm{E}_g(\widehat{Q}_n - R_n) = -\tfrac12 \frac{n-a}{n} + \tfrac12 \mathrm{E}_g\Big[\frac{\sigma_0^2}{\widehat{\sigma}^2}\Big\{n^{-1}\sum_{i=1}^n (x_i^{\mathrm{t}}\widehat{\beta} - \xi_i)^2/\sigma_0^2 + 1\Big\}\Big].$$

Under model circumstances, where $\xi_i = x_i^{\mathrm{t}}\beta$ and $\sigma_0 = \sigma$, it is well known that $\widehat{\sigma}^2/\sigma^2$ is distributed as a $\chi^2_{n-p}/(n-a)$ and is independent of $\widehat{\beta}$. For the fitted values,

$$X\widehat{\beta} = X(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}Y = X(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}(X\beta + \varepsilon) = X\beta + H\varepsilon$$

using the ‘hat matrix’ $H = X(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}$. This shows that

$$n^{-1}\sum_{i=1}^n (x_i^{\mathrm{t}}\widehat{\beta} - x_i^{\mathrm{t}}\beta)^2 = n^{-1}\|X\widehat{\beta} - X\beta\|^2 = n^{-1}\varepsilon^{\mathrm{t}}H\varepsilon$$

has mean equal to $n^{-1}\,\mathrm{E}\,\mathrm{Tr}(H\varepsilon\varepsilon^{\mathrm{t}}) = \sigma^2\,\mathrm{Tr}(H)/n = (p/n)\sigma^2$. Thus

$$\begin{aligned}\mathrm{E}_g(\widehat{Q}_n - R_n) &= -\tfrac12\frac{n-a}{n} + \frac12\frac{n-a}{n-p-2}\frac{p+n}{n} \\ &= \tfrac12\frac{n-a}{n}\frac{2p+2}{n-p-2} = \frac{p+1}{n}\frac{n-a}{n-p-2},\end{aligned}$$

using that $\mathrm{E}(1/\chi^2_{n-p}) = 1/(n-p-2)$ for $n > p+2$.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

This leads to a couple of strategies for modifying the direct version (2.21) to obtain a more precise penalty. The first is to keep the maximum likelihood estimator $\widehat{\sigma}$, that is, using $a = 0$ above, but to penalise the maximised log-likelihood with a more precisely calibrated factor; the result is

$$\mathrm{AIC}_c = 2\ell_n(\widehat{\beta}, \widehat{\sigma}) - 2(p+1)\frac{n}{n-p-2}. \tag{2.23}$$

This is equivalent to what has been suggested by Sugiura (1978) and Hurvich and Tsai (1989). Note that this penalises complexity (the number of parameters $p + 1$) more strongly than with the standard version of AIC, with less chance of overfitting the model. The second modification is actually simpler. It consists of using $a = p + 2$ in (2.22) and keeping the usual penalty at $2(p + 1)$:

$$\mathrm{AIC}_c^* = 2\ell_n(\widehat{\beta}, \widehat{\sigma}^*) - 2(p+1), \tag{2.24}$$

where $(\widehat{\sigma}^*)^2 = \|\mathrm{res}\|^2/(n-p-2)$. This is like the ordinary AIC but with a corrected σ estimate. In particular, this corrected AIC procedure amounts to picking the model with smallest $n \log \widehat{\sigma}^* + p$. Note that the question of which AIC correction method works the best may be made precise in different ways, and there is no 'clear winner'; see Exercise 2.6.

Of the two modifications AIC_c and AIC_c^* only the first has an immediate, but ad hoc, generalisation to general parametric regression models. The suggestion is to use the penalty term obtained for normal linear regression models also for general likelihood models, leading to

$$\mathrm{AIC}_c = 2\ell_n(\widehat{\theta}) - 2\,\mathrm{length}(\theta)\frac{n}{n - \mathrm{length}(\theta) - 1}. \tag{2.25}$$

Hurvich and Tsai (1989) show that this form is appropriate when searching for model order in normal autoregressive models. For autoregressive models AR(k) (see Example 2.10) the formula of AIC_c is again the same as that for linear regression models, namely (leaving out constants not depending on the number of parameters)

$$\mathrm{AIC}_c = -n\log(\widehat{\sigma}_k^2) - \frac{n(n+k)}{n-k-2}.$$

Outside linear regression and autoregressive models the (2.25) formula should be used with care since there is no proof of this statement for general likelihood models. The more versatile bootstrapping method, described at the end of the next section, can be used instead.

2.7 AIC, corrected AIC and bootstrap-AIC for generalised linear models*

While in a traditional linear model the mean of the response $\mathrm{E}(Y \mid x) = x^{\mathrm{t}}\beta$ is a linear function, in a *generalised linear model* there is a monotone and smooth *link function* $\mathrm{g}(\cdot)$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

such that

$$g(\mathrm{E}(Y_i\,|\,x_i)) = g(\xi_i) = \eta_i = x_i^{\mathrm{t}}\beta = \sum_{j=1}^{p} x_{i,j}\beta_j$$

for $i = 1, \ldots, n$. The Y_i are independent, and the likelihood contribution for individual i has the form

$$f(y_i, \theta_i, \phi) = \exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\Big\}.$$

The functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are known. The $b(\cdot)$ plays an important role since its derivatives yield the mean and variance function. The parameter ϕ is a scale parameter, and θ_i is the main parameter of interest, as it can be written as a function of the mean $\mathrm{E}(Y_i\,|\,x)$. Since $\mathrm{E}(\partial \log f(Y_i;\theta_i,\phi)/\partial\theta_i) = 0$, a formula valid for general models, it follows that

$$\xi_i = \mathrm{E}(Y_i\,|\,x_i) = b'(\theta_i) = \partial b(\theta_i)/\partial\theta_i.$$

Also, $\mathrm{Var}(Y_i\,|\,x_i) = a(\phi)b''(\theta_i)$. Many familiar density functions can be written in this form. The class of generalised linear models includes as important examples the normal, Poisson, binomial, and Gamma distribution. For more information on generalised linear models, we refer to McCullagh and Nelder (1989) and Dobson (2002).

The log-likelihood function $\ell_n(\beta,\phi)$ is a sum of $\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i,\phi)$ contributions. Maximum likelihood estimators $(\widehat\beta,\widehat\phi)$ determine fitted values $\widehat\theta_i$ and $\widehat\xi_i$. Let as before $g(y\,|\,x)$ denote the true density of Y given x. We now work with the Kullback–Leibler related quantity

$$R_n = n^{-1}\sum_{i=1}^{n}\int g(y\,|\,x_i)\log f(y_i\,|\,x_i,\widehat\theta)\,\mathrm{d}y.$$

The definitions above give

$$\begin{aligned} R_n &= n^{-1}\sum_{i=1}^{n}\int g(y\,|\,x_i)[\{y\widehat\theta_i - b(\widehat\theta_i)\}/\widehat a + c(y,\widehat\phi)]\,\mathrm{d}y \\ &= n^{-1}\sum_{i=1}^{n}\Big[\{\xi_i^0\widehat\theta_i - b(\widehat\theta_i)\}/\widehat a + \int c(y,\widehat\phi)g(y\,|\,x_i)\,\mathrm{d}y\Big], \end{aligned}$$

where ξ_i^0 denotes the real mean of Y_i given x_i, and with $\widehat a = a(\widehat\phi)$

$$\widehat Q_n = n^{-1}\ell_n(\widehat\beta,\widehat\phi) = n^{-1}\sum_{i=1}^{n}\big[\{Y_i\widehat\theta_i - b(\widehat\theta_i)\}/\widehat a + c(Y_i,\widehat\phi)\big].$$

From this follows, upon simplification, that

$$\mathrm{E}_g(\widehat Q_n - R_n) = n^{-1}\sum_{i=1}^{n}\mathrm{E}_g\Big\{\frac{(Y_i-\xi_i^0)\widehat\theta_i}{\widehat a}\Big\}.$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Let us define the quantity

$$\alpha_n = n\,\mathrm{E}_g(\widehat{Q}_n - R_n) = \mathrm{E}_g\,A_n, \tag{2.26}$$

with $A_n = \sum_{i=1}^n (Y_i - \xi_i^0)\widehat{\theta}_i/\widehat{a}$. Thus α_n/n is the expected bias for the normalised maximised log-likelihood, as an estimator of the estimand $Q_n = \mathrm{E}_g\,R_n$, and can be estimated from data. AIC takes $\alpha_n = p$. We derive approximations to α_n for the class of generalised linear models. For simplicity of presentation we limit discussion to generalised linear models with canonical link function, which means that $\theta_i = \eta_i = x_i^{\mathrm{t}}\beta$. Let

$$\Sigma_n = n^{-1}\sum_{i=1}^n v_i x_i x_i^{\mathrm{t}} \quad \text{and} \quad \Sigma_n^* = n^{-1}\sum_{i=1}^n v_i^* x_i x_i^{\mathrm{t}},$$

where $v_i = b''(\theta_i)$ and $v_i^* = a(\phi)^{-1}\mathrm{Var}(Y_i \mid x_i)$. Then $\widehat{\beta} - \beta_0$ may up to an $O_P(n^{-1/2})$ error be represented as $\Sigma_n^{-1} n^{-1} \sum_{i=1}^n \{Y_i - b'(\theta_i)\}x_i$, as a consequence of the property exhibited in connection with (2.12). When the assumed model is equal to the true model, $v_i^* = v_i$. We find, also outside the model conditions, that

$$\begin{aligned}
\alpha_n &\doteq \mathrm{E}_g(1/\widehat{a}\,)\sum_{i=1}^n \mathrm{E}_g\big[(Y_i - \xi_i^0)x_i^{\mathrm{t}}\widehat{\beta}\big] \\
&\doteq \mathrm{E}_g(1/\widehat{a}\,)\sum_{i=1}^n \mathrm{E}_g\big[(Y_i - x_i^0)x_i^{\mathrm{t}}\Sigma_n^{-1}\{Y_i - b'(\theta_i)\}x_i\big] \\
&\doteq \mathrm{E}_g(a/\widehat{a}\,)\,n^{-1}\sum_{i=1}^n v_i^* x_i^{\mathrm{t}}\Sigma_n^{-1}x_i = \mathrm{E}_g(a/\widehat{a}\,)\,\mathrm{Tr}\big(\Sigma_n^{-1}\Sigma_n^*\big).
\end{aligned}$$

We have used the parameter orthogonality property for generalised linear models, which implies approximate independence between the linear part estimator $\widehat{\beta}$ and the scale estimator $\widehat{a} = a(\widehat{\phi})$. For the linear normal model, for example, there is exact independence, and $a/\widehat{a} = \sigma^2/\widehat{\sigma}^2$ has exact mean $m/(m-2)$ if the unbiased estimator $\widehat{\sigma}^2$ is used with degrees of freedom $m = n - p$.

These approximations suggest some corrections to the ordinary AIC version, which uses $\alpha_n = p$. Some models for dealing with overdispersion use ideas that in the present terminology amount to $\Sigma_n^* = (1+d)\Sigma_n$, with d an overdispersion parameter. For Poisson-type regression there is overdispersion when the observed variance is too big in comparison with the variance one would expect for a Poisson variable. In such a case the dispersion is $1 + d = \mathrm{Var}_g Y/\mathrm{E}_g Y > 1$. Including such an overdispersion parameter corresponds to modelling means with $\exp(x_i^{\mathrm{t}}\beta)$ but variances with $(1+d)\exp(x_i^{\mathrm{t}}\beta)$. An estimate of this d leads to $\alpha_n \doteq (1+d)p$ and to a corrected

$$\mathrm{AIC}_c = 2\ell_n(\widehat{\theta}) - 2p(1+d)$$

for use in model selection when there is overdispersion.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Table 2.5. *AIC and bootstrap AIC values for the eight candidate models of Example 2.4 on low birthweights. The model rankings agree on the top three models but not on the rest.*

Model	$\ell_{\max}$	length(β)	$\widehat{\alpha}_n$	AIC	Ranking		AIC$_{\text{boot}}$
$\emptyset$	–114.345	2	1.989	–232.691	5	4	–232.669
x_3	–113.562	3	3.021	–233.123	6	6	–233.165
x_4	–112.537	3	3.134	–231.075	1	1	–231.344
x_5	–114.050	3	3.033	–234.101	7	7	–234.167
x_3, x_4	–112.087	4	4.078	–232.175	3	3	–232.331
x_3, x_5	–113.339	4	4.096	–234.677	8	8	–234.834
x_4, x_5	–111.630	4	4.124	–231.259	2	2	–231.507
x_3, x_4, x_5	–111.330	5	5.187	–232.661	4	5	–233.034

Using the bootstrap is another alternative for estimating α_n. Simulate $Y_1^*, \ldots, Y_n^*$ from an estimated bigger model, say one that uses all available covariates, with estimated means $\widehat{\xi}_1, \ldots, \widehat{\xi}_n$, keeping $x_1, \ldots, x_n$ fixed. For this simulated data set, produce estimates $\widehat{\beta}^*$ and $\widehat{a}^* = a(\widehat{\phi}^*)$, along with $\widehat{\theta}_i^*$. Then form $A_n^* = \sum_{i=1}^n (Y_i^* - \widehat{\xi}_i) x_i^{\rm t} \widehat{\beta}^* / \widehat{a}^*$. The α_n estimate is formed by averaging over a high number of simulated A_n^* values. This procedure can be used for each candidate model, leading to the bootstrap corrected

$$\text{AIC}_{\text{boot}} = 2\ell_n(\widehat{\theta}) - 2\widehat{\alpha}_n. \tag{2.27}$$

Again, the model with highest such value would then be selected.

Example 2.12 Low birthweight data: bootstrap AIC*

As an illustration, let us return to the low birthweight data of Example 2.4. For the logistic regression model, A_n of (2.26) is $\sum_{i=1}^n (Y_i - p_i^0) x_i^{\rm t} \widehat{\beta}$, where p_i^0 is the mean of Y_i under the true (but unknown) model. For each candidate model, we simulate a large number of $A_n^* = \sum_{i=1}^n (Y_i^* - \widehat{p}_i) x_i^{\rm t} \widehat{\beta}^*$, where Y_i^* are 0–1 variables with probabilities set at $\widehat{p}_i$ values found from the biggest model, and where $\widehat{\beta}^*$ is the maximum likelihood estimator based on the simulated data set, in the given candidate model. For the situation of Example 2.4, with 25,000 simulations for each of the eight candidate models, the results are presented in Table 2.5. The $\widehat{\alpha}_n$ numbers, which may be seen as the exact penalties (modulo simulation noise) from the real perspective of AIC, agree well with the AIC default values in this particular illustration. Also, the AIC and AIC$_{\text{boot}}$ values are in essential agreement. ■

2.8 Behaviour of AIC for moderately misspecified models*

Consider a local neighbourhood framework where data stem from a density

$$f_n(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}), \tag{2.28}$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

with a p-dimensional θ and a one-dimensional γ. The null value γ_0 corresponds to the narrow model, so (2.28) describes a one-parameter extension. This framework will be extended and discussed further in Sections 5.2, 5.7 and 6.3, also for situations with $q \geq 2$ extra parameters. The AIC method for selecting one of the two models compares

$$\mathrm{AIC}_{\mathrm{narr}} = 2\,\ell_n(\widetilde{\theta}, \gamma_0) - 2p \quad \text{with} \quad \mathrm{AIC}_{\mathrm{wide}} = 2\,\ell_n(\widehat{\theta}, \widehat{\gamma}) - 2(p+1),$$

where maximum likelihood estimators are used in each model. To understand these random AIC scores better, introduce first

$$\begin{pmatrix} U(y) \\ V(y) \end{pmatrix} = \begin{pmatrix} \partial \log f(y, \theta_0, \gamma_0)/\partial\theta \\ \partial \log f(y, \theta_0, \gamma_0)/\partial\gamma \end{pmatrix},$$

along with $\bar{U}_n = n^{-1}\sum_{i=1}^n U(Y_i)$ and $\bar{V}_n = n^{-1}\sum_{i=1}^n V(Y_i)$. The $(p+1)\times(p+1)$-size information matrix of the model is

$$J_{\mathrm{wide}} = \mathrm{Var}_0\begin{pmatrix} U(Y) \\ V(Y) \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}, \text{ with } J_{\mathrm{wide}}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}.$$

Here the $p \times p$-size J_{00} is simply the information matrix of the narrow model, evaluated at θ_0, and the scalar J_{11} is the variance of $V(Y_i)$, also computed under the narrow model. We define $\kappa^2 = J^{11}$.

Coming back to AIC for the two models, one finds via result (2.9) and Taylor expansions that

$$\mathrm{AIC}_{\mathrm{narr}} \doteq_d 2\sum_{i=1}^n \log f(Y_i, \theta_0, \gamma_0) + n\bar{U}_n^{\mathrm{t}} J_{11}^{-1}\bar{U}_n - 2p,$$

$$\mathrm{AIC}_{\mathrm{wide}} \doteq_d 2\sum_{i=1}^n \log f(Y_i, \theta_0, \gamma_0) + n\begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix}^{\mathrm{t}} J_{\mathrm{wide}}^{-1}\begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix} - 2(p+1).$$

Further algebraic calculations give, with $D \sim \mathrm{N}(\delta, \kappa^2)$,

$$\begin{aligned} \mathrm{AIC}_{\mathrm{wide}} - \mathrm{AIC}_{\mathrm{narr}} &\doteq_d n\begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix}^{\mathrm{t}} J_{\mathrm{wide}}^{-1}\begin{pmatrix} \bar{U}_n \\ \bar{V}_n \end{pmatrix} - n\bar{U}_n^{\mathrm{t}} J_{00}^{-1}\bar{U}_n - 2 \\ &= n\{\bar{U}_n^{\mathrm{t}}(J^{00} - J_{00}^{-1})\bar{U}_n + 2\bar{U}_n^{\mathrm{t}} J^{01}\bar{V}_n + \bar{V}_n^2 J^{11}\} - 2 \\ &= n(\bar{V}_n - J_{10}J_{00}^{-1}\bar{U}_n)^2\kappa^2 - 2 \\ &\xrightarrow{d} D^2/\kappa^2 - 2 \sim \chi_1^2(\delta^2/\kappa^2) - 2. \end{aligned}$$

The probability that AIC prefers the narrow model over the wide model is therefore approximately $\mathrm{P}(\chi_1^2(\delta^2/\kappa^2) \leq 2)$. In particular, if the narrow model is perfect, the probability is 0.843, and if $\delta = \kappa$, the probability is 0.653.

It is also instructive to see that $\mathrm{AIC}_{\mathrm{wide}} - \mathrm{AIC}_{\mathrm{narr}}$ is asymptotically equivalent to $D_n^2/\kappa^2 - 2$, where $D_n = \sqrt{n}(\widehat{\gamma} - \gamma_0)$. This gives a connection between the Akaike criterion and a certain pre-test strategy. Pre-testing is a form of variable selection which consists of checking the coefficients in the wide model and keeping only those which are

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

significant. In this example with only two models, we check whether the extra parameter γ is significant and decide to keep the wide model when it indeed is significant. There are various ways to test for significance (t-statistics, z-statistics with different significance level, etc.). According to AIC we would use the wide model when $D_n^2/\widehat{\kappa}^2 > 2$. Generalisations of these results to AIC behaviour in situations with more candidate models than only two, as here, are derived and discussed in Section 5.7.

2.9 Cross-validation

An approach to model selection that at least at the outset is of a somewhat different spirit from that of AIC and TIC is that of *cross-validation*. Cross-validation has a long history in applied and theoretical statistics, but was first formalised as a general methodology in Stone (1974) and Geisser (1975). The idea is to split the data into two parts: the majority of data are used for model fitting and development of prediction algorithms, which are then used for estimation or prediction of the left-out observations. In the case of leave-one-out cross-validation, perhaps the most common form, only one observation is left out at a time. Candidate models are fit with all but this one observation, and are then used to predict the case which was left out.

We first explain a relation between leave-one-out cross-validation, the Kullback–Leibler distance and Takeuchi's information criterion. For concreteness let us first use the i.i.d. framework. Since the Kullback–Leibler related quantity R_n worked with in (2.15) may be expressed as $\int g(y) \log f(y, \widehat{\theta})\, \mathrm{d}y = \mathrm{E}_g \log f(Y_{\mathrm{new}}, \widehat{\theta})$, where Y_{new} is a new datum independent of $Y_1, \ldots, Y_n$, an estimator of its expected value is

$$\mathrm{xv}_n = n^{-1} \sum_{i=1}^{n} \log f(Y_i, \widehat{\theta}_{(i)}). \tag{2.29}$$

Here $\widehat{\theta}_{(i)}$ is the maximum likelihood estimator computed based on the data set where Y_i is omitted; this should well emulate the $\log f(Y_{\mathrm{new}}, \widehat{\theta})$ situation, modulo the small difference between estimators based on sample sizes n versus $n - 1$. The cross-validation model selection method, in this context, is to compute the (2.29) value for each candidate model, and select the one with highest value.

There is a connection between cross-validation and the model-robust AIC. Specifically,

$$\mathrm{xv}_n \doteq n^{-1}\big\{\ell_n(\widehat{\theta}) - \mathrm{Tr}(\widehat{J}^{-1}\widehat{K})\big\}, \tag{2.30}$$

which means that $2n\,\mathrm{xv}_n$ is close to Takeuchi's information criterion (2.20). To prove (2.30) we use properties of influence functions.

Let in general terms $T = T(G)$ be a function of some probability distribution G. The influence function of T, at position y, is defined as the limit

$$\mathrm{infl}(G, y) = \lim_{\varepsilon \to 0} \{T((1-\varepsilon)G + \varepsilon\delta(y)) - T(G)\}/\varepsilon, \tag{2.31}$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

when it exists. Here $\delta(y)$ denotes a unit point-mass at position y, which means that $(1-\varepsilon)G+\delta(y)$ is the distribution of a variable that with probability $1-\varepsilon$ is drawn from G and with probability ε is equal to y. These influence functions are useful for several purposes in probability theory and statistics, for example in the study of robustness, see e.g. Huber (1981), Hampel *et al.* (1986), Basu *et al.* (1998) and Jones *et al.* (2001).

Consider some parameter functional $\theta = T(G)$. Data points $y_1,\ldots,y_n$ give rise to the empirical distribution function G_n and an estimate $\widehat{\theta} = T(G_n)$. The primary use of influence functions in our context is the fact that under general and weak conditions,

$$T(G_n) = T(G) + n^{-1}\sum_{i=1}^n \mathrm{infl}(G, Y_i) + o_P(n^{-1/2}). \tag{2.32}$$

This implies the fundamental large-sample result

$$\sqrt{n}\{T(G_n) - T(G)\} \xrightarrow{d} \mathrm{N}_p(0, \Omega), \tag{2.33}$$

where Ω is the variance matrix of $\mathrm{infl}(G, Y)$ when $Y \sim G$. For $T(G)$ we now take the Kullback–Leibler minimiser $\theta_0 = \theta_0(g)$ that minimises $\mathrm{KL}(g, f(\cdot,\theta))$ of (2.2), with g the density of G. Then $T(G_n)$ is the maximum likelihood estimator $\widehat{\theta}$. The influence function for maximum likelihood estimation can be shown to be

$$\mathrm{infl}(G, y) = J^{-1}u(y,\theta_0), \tag{2.34}$$

for precisely $\theta_0 = T(G)$, with J as defined in (2.7) and $u(y,\theta)$ the score function; see Exercise 2.11.

We record one more useful consequence of influence functions. For a given data point y_i, consider the leave-one-out estimator $\widehat{\theta}_{(i)}$, constructed by leaving out y_i. Note that $G_n = (1-1/n)G_{n,(i)} + n^{-1}\delta(y_i)$, with $G_{n,(i)}$ denoting the empirical distribution of the $n-1$ data points without y_i. From (2.31), therefore, with $\varepsilon = 1/n$, the approximation $T(G_n) \doteq T(G_{n,(i)}) + n^{-1}\,\mathrm{infl}(G_{n,(i)}, y_i)$ emerges, that is

$$\widehat{\theta}_{(i)} \doteq \widehat{\theta} - n^{-1}\,\mathrm{infl}(G_{n,(i)}, y_i) \doteq \widehat{\theta} - n^{-1}\,\mathrm{infl}(G_n, y_i). \tag{2.35}$$

From (2.35) and (2.34), $\widehat{\theta}_{(i)} \doteq \widehat{\theta} - n^{-1}\widehat{J}^{-1}u(y_i,\widehat{\theta})$. Using Taylor expansion, $\log f(y_i,\widehat{\theta}_{(i)})$ is close to $\log f(y_i,\widehat{\theta}) + u(y_i,\widehat{\theta})^{\mathrm{t}}(\widehat{\theta}_{(i)} - \widehat{\theta})$. Combining these observations, we reach

$$\begin{aligned}\mathrm{xv}_n &\doteq n^{-1}\sum_{i=1}^n\{\log f(y_i,\widehat{\theta}) + u(y_i,\widehat{\theta})^{\mathrm{t}}(\widehat{\theta}_{(i)} - \widehat{\theta})\}\\ &\doteq n^{-1}\ell_n(\widehat{\theta}) - n^{-1}\sum_{i=1}^n u(y_i,\widehat{\theta})^{\mathrm{t}}n^{-1}\widehat{J}^{-1}u(y_i,\widehat{\theta}),\end{aligned}$$

which leads to (2.30) and ends the proof. The approximation holds in the sense that the difference between the two sides goes to zero in probability.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

The case of regression models can be handled similarly, inside the framework of Section 2.2. We again have

$$\mathrm{xv}_n = n^{-1}\sum_{i=1}^n \log f(y_i \,|\, x_i, \widehat\theta_{(i)}) \doteq n^{-1}\ell_{n,\max} + n^{-1}\sum_{i=1}^n u(y_i \,|\, x_i, \widehat\theta)^{\mathrm t}(\widehat\theta_{(i)} - \widehat\theta),$$

where we need a parallel result to that above for the difference $\widehat\theta_{(i)} - \widehat\theta$. Such an approximation emerges from similar arguments, where the work tool is that of the influence function $\mathrm{infl}((G, C), (x, y))$ defined in a manner similar to (2.31), but now viewed as associated with the functional $T = T(G, C)$ that operates on the combination of true densities $g(y \,|\, x)$ and the covariate distribution C. For the maximum likelihood functional, which is also the minimiser of the weighted Kullback–Leibler divergence (2.4), one may prove that the influence function takes the form $J^{-1}u(y \,|\, x, \theta_0)$, with J as in (2.12). With these ingredients one may copy the earlier arguments to reach the approximation $\widehat\theta_{(i)} - \widehat\theta \approx -n^{-1}\widehat J_n^{-1}u(y_i \,|\, x_i, \widehat\theta)$. In conclusion,

$$\begin{aligned}\mathrm{xv}_n &\doteq n^{-1}\ell_{n,\max} - n^{-2}\sum_{i=1}^n u(y_i \,|\, x_i, \widehat\theta)^{\mathrm t}\widehat J_n^{-1}u(y_i \,|\, x_i, \widehat\theta)\\ &= n^{-1}\big\{\ell_{n,\max} - \mathrm{Tr}(\widehat J_n^{-1}\widehat K_n)\big\},\end{aligned}$$

essentially proving that cross-validation in regression models is first-order large-sample equivalent to the model-robust AIC method discussed in Section 2.5.

We have so far discussed cross-validation as a tool in connection with assessing the expected size of $\int\int g(y \,|\, x)\log f(y \,|\, x, \widehat\theta)\,\mathrm dy\,\mathrm dC(x)$, associated with maximum likelihood estimation and the weighted Kullback–Leibler distance (2.4). Sometimes a more immediate problem is to assess the quality of the more direct prediction task that estimates a new y_{new} with say $\widehat y_{\mathrm{new}} = \widehat\xi(x_{\mathrm{new}})$, where $\widehat\xi(x) = \xi(x, \widehat\theta)$ estimates $\xi(x, \theta) = \mathrm E_\theta(Y \,|\, x)$. This leads to the problem of estimating the mean of $(y_{\mathrm{new}} - \widehat y_{\mathrm{new}})^2$ and related quantities. Cross-validation is also a tool for such problems.

Consider in general terms $\mathrm E_g h(Y_{\mathrm{new}} \,|\, x, \widehat\theta)$, for a suitable h function, for example of the type $\{Y_{\mathrm{new}} - \xi(x, \widehat\theta)\}^2$. We may write this as

$$\pi_n = \int\int g(y \,|\, x)h(y \,|\, x, \widehat\theta)\,\mathrm dy\,\mathrm dC_n(x) = n^{-1}\sum_{i=1}^n \mathrm E_{g(y \,|\, x_i)}h(Y_{\mathrm{new},i} \,|\, x_i, \widehat\theta),$$

where $Y_{\mathrm{new},i}$ is a new observation drawn from the distribution associated with covariate vector x_i. The direct but naive estimator is $\bar\pi_n = n^{-1}\sum_{i=1}^n h(y_i \,|\, x_i, \widehat\theta)$, but we would prefer the nearly unbiased cross-validated estimator

$$\widehat\pi_n = n^{-1}\sum_{i=1}^n h(y_i \,|\, x_i, \widehat\theta_{(i)}).$$

Assume that h is smooth in θ and that the estimation method used has influence function of the type $J(G, C)^{-1}a(y \,|\, x, \theta_0)$, for a suitable matrix $J(G, C)$ defined in terms of

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

both the distribution G of $Y \mid x$ and of the covariate distribution C of x. Such influence representations hold for various minimum distance estimators, including the maximum likelihood method. Then Taylor expansion, as above, leads to

$$\widehat{\pi}_n \doteq \bar{\pi}_n - n^{-1}\mathrm{Tr}(\widetilde{J}^{-1}\widetilde{L}), \tag{2.36}$$

where $\widetilde{J} = J(G_n, C_n)$ is the empirical version of $J(G, C)$ and

$$\widetilde{L} = n^{-1}\sum_{i=1}^{n} a(y_i \mid x_i, \widehat{\theta})h'(y_i \mid x_i, \widehat{\theta})^{\mathrm{t}},$$

with $h'(y \mid x, \theta) = \partial h(y \mid x, \theta)/\partial\theta$. The cross-validation result (2.30) corresponds to the special case where the h function is $\log f(y \mid x, \theta)$ and where the estimation method is the maximum likelihood; in this situation, both $a(y \mid x, \theta)$ and $h'(y \mid x, \theta)$ are equal to $u(y \mid x, \theta)$, and $\widetilde{L}$ is identical to $\widehat{K}$. For an illustration of various cross-validation methods, see Example 5.10.

Example 2.13 Two views on selecting Poisson models

Assume as in Example 2.3 that count data Y_i are associated with covariate vectors x_i, and that a Poisson model is being considered with parameters $\xi_i = \exp(x_i^{\mathrm{t}}\beta)$. To assess the prediction quality of such a candidate model, at least two methods might be put forward, yielding two different model selectors. The first is that implied by concentrating on maximum likelihood and the Kullback–Leibler divergence (2.4), and where the above results lead to

$$\mathrm{xv}_n = n^{-1}\sum_{i=1}^{n}\log f(y_i \mid x_i, \widehat{\beta}_{(i)}) \doteq n^{-1}\ell_{n,\max} - n^{-1}\mathrm{Tr}(\widehat{J}_n^{-1}\widehat{K}_n).$$

As we know, using the right-hand side scores for candidate models corresponds to the model-robust AIC, cf. TIC in Section 2.5. The second option is that of comparing predicted $\widehat{y}_i = \exp(x_i^{\mathrm{t}}\widehat{\beta})$ with observed y_i directly. Let therefore $h(y \mid x, \beta) = \{y - \exp(x^{\mathrm{t}}\beta)\}^2$. Here $h'(y \mid x, \beta) = -2\{y - \exp(x^{\mathrm{t}}\beta)\}\exp(x^{\mathrm{t}}\beta)x$, and the techniques above yield

$$\widehat{\pi}_n = n^{-1}\sum_{i=1}^{n}\big\{y_i - \exp(x_i^{\mathrm{t}}\widehat{\beta}_{(i)})\big\}^2 \doteq n^{-1}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2 + 2n^{-1}\mathrm{Tr}(\widehat{J}_n^{-1}\widehat{M}_n),$$

where $\widehat{M}_n = n^{-1}\sum_{i=1}^{n}\widehat{\xi}_i(y_i - \widehat{\xi}_i)^2 x_i x_i^{\mathrm{t}}$. This example serves as a reminder that 'model selection' should not be an automated task, and that what makes a 'good model' must depend on the context. ■

The first-order asymptotic equivalence of cross-validation and AIC has first been explained in Stone (1977). There is a large literature on cross-validation. For more information we refer to, for example, Stone (1978), Efron (1983), Picard and Cook (1984), Efron and Tibshirani (1993) and Hjorth (1994).

Allen (1974) proposed a method of model selection in linear regression models which is based on the sum of squared differences of the leave-one-out predictions $\widehat{y}_i$ with their true observed values y_i. This gives the PRESS statistic, an abbreviation for prediction sum of squares. In a linear regression model $Y_i = x_i^{\rm t}\beta + \varepsilon_i$ with a vector β of regression coefficients, leave out observation i, then fit the model, obtain the estimated coefficients $\widehat{\beta}_{(i)}$ and form the leave-one-out prediction $\widehat{y}_{i,\rm xv} = x_i^{\rm t}\widehat{\beta}_{(i)}$. This gives the statistic $\text{PRESS} = \sum_{i=1}^n (y_i - \widehat{y}_{i,\rm xv})^2$. This score can then be computed for different subsets of regression variables. The model with the smallest value of PRESS is selected. Alternatively one may compute the more robust $\sum_{i=1}^n |y_i - \widehat{y}_{i,\rm xv}|$. There are ways of simplifying the algebra here, making the required computations easy to perform. Let $s_1, \ldots, s_n$ be the diagonal entries of the matrix $I_n - H$, where $H = X(X^{\rm t}X)^{-1}X^{\rm t}$ is the hat matrix. Then the difference $y_i - \widehat{y}_{i,\rm xv}$ is identical to $(y_i - x_i^{\rm t}\widehat{\beta})/s_i$, where $\widehat{\beta}$ is computed using all data in the model; see Exercise 2.13. Thus $\text{PRESS} = \sum_{i=1}^n (y_i - x_i^{\rm t}\widehat{\beta})^2/s_i^2$. We note that PRESS only searches for models among those that have linear mean functions and constant variance, and that sometimes alternatives with heterogeneous variability are more important; see e.g. Section 5.6.

2.10 Outlier-robust methods

We start with an example illustrating the effect that outlying observations can have on model selection.

Example 2.14 How to repair for Ervik's 1500-m fall?

In the 2004 European Championships for speedskating, held in Heerenveen, the Netherlands, the Norwegian participant Eskil Ervik unfortunately fell in the third distance, the 1500-m. This cost Norway a lost spot at the World Championships later that season. In the result list, his 2:09.20 time, although an officially registered result, is accordingly a clear outlier, cf. Figure 2.4. Including it in a statistical analysis of the results might give misleading results. We shall work with the times in seconds computed for the two distances, and will try to relate the 1500-m time via a linear regression model to the 5000-m time. We perform our model fitting using results from the 28 skaters who completed the two distances. Including Ervik's result seriously disturbs the maximum likelihood estimators here. We fit polynomial regression models, linear, quadratic, cubic and quartic, $Y = \beta_0 + \beta_1 x + \cdots + \beta_4 x^4 + \varepsilon$, where x and Y are 1500-m time and 5000-m time, respectively, and assuming the errors to come from a $\mathrm{N}(0, \sigma^2)$ distribution with unknown variance.

Table 2.6 contains the results of applying AIC to these data. When all observations are included, AIC ranks the linear model last, and prefers a quadratic model. When leaving out the outlying observation, the linear model is deemed best by AIC. For model selection purposes, the 'best model' should be considered the one which is best for non-outlying data. The estimated linear trend is very different in these two cases. Without the

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Table 2.6. *AIC for selecting a polynomial regression model for the speedskating data of the 2004 championship: First with all 28 skaters included, next leaving out the result of Ervik who fell in the 1500-m.*

Model:	linear	quadratic	cubic	quartic
AIC (all skaters):	−227.131	−213.989	−215.913	−217.911
preference order:	(4)	(1)	(2)	(3)
AIC (without Ervik):	−206.203	−207.522	−209.502	−210.593
preference order:	(1)	(2)	(3)	(4)

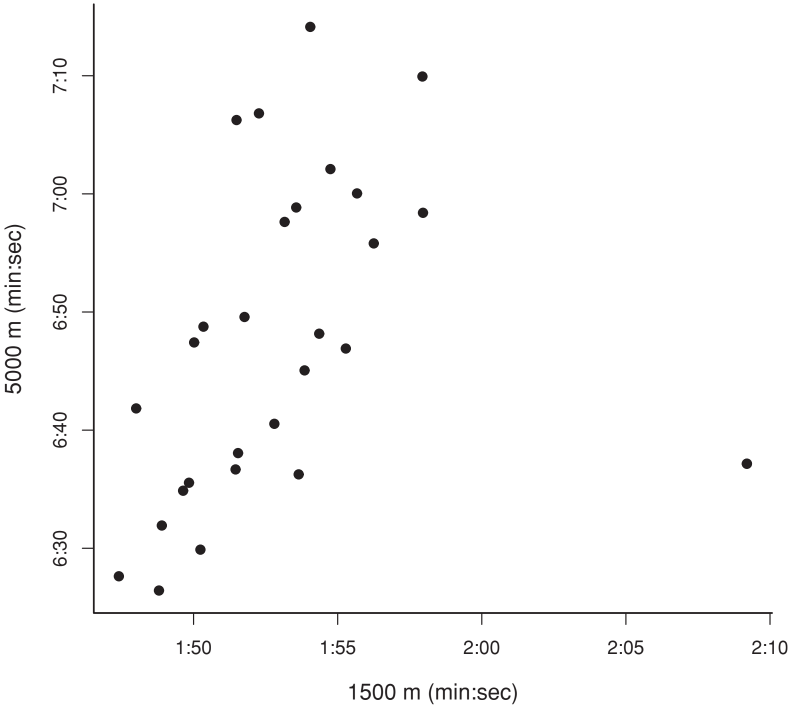

Fig. 2.4. Ervik's fall in the 2004 European Championship: 1500-m versus 5000-m results, with Ervik's outlying 1500-m time very visible to the right.

outlying observation the estimated slope is 3.215, while with Ervik included, the value is only 1.056, pushing the line downwards. Since the estimators are so much influenced by this single outlying observation, also the model selection methods using such estimators suffer. In this example it is quite obvious that Ervik's time is an outlier (not because of his recorded time per se, but because he fell), and that this observation can be removed before the analysis. In general, however, it might not always be so clear which observations are outliers; hence outlier-robust model selection methods are called for. ■

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

2.10.1 AIC from weighted penalised likelihood methods

Let us consider a linear regression model of the form $Y = X\beta + \varepsilon$, for a p-vector of regression coefficients β, where the errors ε have common variance σ^2, which is unknown. We wish to select components of β. For normal regression models, where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, maximum likelihood methods coincide with least squares estimators. In general, both maximum likelihood estimators and least squares estimators are known to be non-robust against outlying observations in many models. This means that a few observations far from the remaining part of the observations may seriously influence estimators constructed from the data. Consequently a model selected using methods based on maximum likelihood might select a model that does not fit well, or one that may be too complex for the majority of the data. Agostinelli (2002) studies weighted versions of likelihood estimators to construct weighted model selection criteria. These methods are available in the `R` package `wle`.

While maximum likelihood estimators for $\theta = (\beta, \sigma)$ are found via solving the set of score equations $\sum_{i=1}^n u(Y_i \,|\, x_i, \beta, \sigma) = 0$, where $u(y \,|\, x, \theta) = \partial \log f(y \,|\, x, \theta)/\partial\theta$, weighted maximum likelihood estimators solve a set of equations of the form

$$\sum_{i=1}^n w(y_i - x_i^{\mathrm{t}}\beta, \sigma)u(y_i \mid x_i, \beta, \sigma) = 0. \tag{2.37}$$

We now summarise details of the construction of the weights, see Agostinelli (2002). With $\widehat{e}_i$ denoting smoothed Pearson residuals (see below), the weight functions $w(\cdot)$ are defined as $w(y_i - x_i^{\mathrm{t}}\beta, \sigma) = \min\{1, \max(0, r(\widehat{e}_i) + 1)/(\widehat{e}_i + 1)\}$. The option $r(a) = a$ brings us back to classical maximum likelihood estimators, since then all weights equal one. Some robust choices for the residual adjustment function r are discussed in Agostinelli (2002). Instead of the usual residuals $e_i = y_i - x_i^{\mathrm{t}}\beta$, smoothed Pearson residuals are used to construct the weights. These are defined as $\widehat{e}_i = \widehat{f}_e(e_i)/\widehat{m}(e_i) - 1$. More precisely, for a kernel function K, a given univariate density function like the standard normal, the kernel density estimator of the residuals $e_i = y_i - x_i^{\mathrm{t}}\beta$ is defined as $\widehat{f}_e(t) = n^{-1}\sum_{i=1}^n h^{-1}K(h^{-1}(e_i - t))$, where h is a bandwidth parameter (see for example Wand and Jones, 1995). And for a normal regression model, the kernel smoothed model density equals $\widehat{m}(t) = \int h^{-1}K(h^{-1}(t - s))\phi(s, \sigma^2)\,\mathrm{d}s$; here $\phi(s, \sigma^2)$ is the $\mathrm{N}(0, \sigma^2)$ density. When K is the standard normal kernel, $\widehat{m}(t) = \phi(t, \sigma^2 + h^2)$.

Suggestions have been made for adjusting model selection methods based on the likelihood function with weights constructed as above, to lead to robust selection criteria. For example, Agostinelli (2002) proposes to modify the $\mathrm{AIC} = 2\,\ell_n(\widehat{\theta}) - 2(p + 1)$ of (2.14) to

$$\mathrm{wleAIC} = 2\sum_{i=1}^n w(y_i - x_i^{\mathrm{t}}\widehat{\beta}, \widehat{\sigma}) \log f(y_i \,|\, x_i, \widehat{\beta}, \widehat{\sigma}) - 2(p + 1).$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

The parameters are estimated by solving (2.37). We return to Example 2.14 and apply the wleAIC using the R function `wle.aic` with its built-in default choice for selection of the bandwidth h. This leads to the following wleAIC values for the four models (linear to quartic): -199.033, -208.756, -218.612, -228.340. These are quite sensible values in view of Table 2.6, and in particular indicate preference for the simplest linear model. The wleAIC analysis agrees also with the robust model selection method that we discuss in the following subsection.

2.10.2 Robust model selection from weighted Kullback–Leibler distances

The methods discussed in the previous subsection involve robust weighting of the log-likelihood contributions, but maximising the resulting expression is not necessarily a fruitful idea for models more general than the linear regression one. We shall in fact see below that when log-likelihood terms are weighted, then, in general, a certain additional parameter-dependent factor needs to be taken into account, in order for the procedures to have a minimum distance interpretation.

Let g be the true data-generating density and f_θ a candidate model. For any non-negative weight function $w(y)$,

$$d_w(g, f_\theta) = \int w\{g \log(g/f_\theta) - (g - f_\theta)\}\,\mathrm{d}y \tag{2.38}$$

defines a divergence, or distance from the true g to the model density f_θ; see Exercise 2.14. When w is constant, this is equivalent to the Kullback–Leibler distance. Minimising the d_w distance is the same as maximising $H(\theta) = \int w(g \log f_\theta - f_\theta)\,\mathrm{d}y$. This invites the maximum weighted likelihood estimator $\widehat{\theta}$ that maximises

$$H_n(\theta) = n^{-1} \sum_{i=1}^{n} w(y_i) \log f(y_i, \theta) - \int w f_\theta\,\mathrm{d}y. \tag{2.39}$$

This is an estimation method worked with by several authors, from different perspectives, including Hjort (1994b), Hjort and Jones (1996), Loader (1996) and Eguchi and Copas (1998).

In principle any non-negative weight function may be used, for example to bring extra estimation efforts into certain regions of the sample space or to downweight more extreme values. Robustness, in the classical first-order sense of leading to bounded influence functions, is achieved for any weight function $w(y)$ with the property that $w(y)u(y, \theta)$ is bounded at the least false value θ_0 that minimises d_w; see below. We note that simply maximising the weighted likelihood function $\sum_{i=1}^n w(y_i) \log f(y_i, \theta)$ will not work without the correction term $-\int w f_\theta\,\mathrm{d}y$, as it may easily lead to inconsistent estimators.

The argmax $\widehat{\theta}$ of H_n may be seen as a minimum distance estimator associated with the generalised Kullback–Leibler distance d_w. As such it is also an M-estimator, obtained by

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

solving $H_n^{(1)}(\theta) = 0$, where $H_n^{(1)}(\theta) = n^{-1} \sum_{i=1}^n w(y_i)u(y_i, \theta) - \xi(\theta)$ is the derivative of H_n; here $\xi(\theta) = \int w f_\theta u_\theta \,dy$ is the derivative of $\int w f_\theta \,dy$ with respect to θ. Here and below we let the '$^{(j)}$' indicate the jth partial derivative with respect to θ, for $j = 1, 2$. The estimator is consistent for the least false value θ_0 that minimises the d_w distance from truth to model density. Note that $H_n^{(1)}(\theta)$ converges in probability to $\int w(g - f_\theta)u_\theta \,dy$, so the least false θ_0 may also be characterised as the solution to $\int w g u_\theta \,dy = \int w f_\theta u_\theta \,dy$. Next consider the second derivative matrix function

$$H_n^{(2)}(\theta) = n^{-1} \sum_{i=1}^n w(y_i) I(y_i, \theta) - \int w f_\theta u_\theta u_\theta^{\rm t} \,dy - \int w f_\theta I_\theta \,dy,$$

and define $J_n = -H_n^{(2)}(\theta_0)$. Here $J_n \to_p J$ as n grows, where $J = \int w f_\theta u_\theta u_\theta^{\rm t} \,dy + \int w(f_\theta - g) I_\theta \,dy$, evaluated at $\theta = \theta_0$. Furthermore, $\sqrt{n} H_n^{(1)}(\theta_0) \to_d U' \sim \mathrm{N}_p(0, K)$, by the central limit theorem, with

$$K = \mathrm{Var}_g\{w(Y)u(Y, \theta_0) - \xi(\theta_0)\} = \int g(y) w(y)^2 u(y, \theta_0) u(y, \theta_0)^{\rm t} \,dy - \xi\xi^{\rm t},$$

and where $\xi = \int g(y)w(y)u(y, \theta_0)\,dy = \int w(y) f(y, \theta_0) u(y, \theta_0)\,dy$. With these definitions,

$$\sqrt{n}(\widehat{\theta} - \theta_0) = J^{-1}\sqrt{n} H_n^{(1)}(\theta_0) + o_P(1) \overset{d}{\to} J^{-1}U' \sim \mathrm{N}_p(0, J^{-1}KJ^{-1}). \tag{2.40}$$

The influence function for the associated $\widehat{\theta} = T(G_n)$ functional, where $T(G)$ is the argmax of $\int w(g \log f_\theta - f_\theta)\,dy$, is $\mathrm{infl}(G, y) = J^{-1}\{w(y)u(y, \theta_0) - \xi(\theta_0)\}$, see Exercise 2.14.

To judge the prediction quality of a suggested parametric model f_θ, via the weighted Kullback–Leibler distance d_w of (2.38), we need to assess the expected distance $d_w(g(\cdot), f(\cdot, \widehat{\theta}))$ in comparison with other competing models. This quantity is a constant away from $Q_n = \mathrm{E}_g R_n$, where $R_n = \int w\{g \log f(\cdot, \widehat{\theta}) - f(\cdot, \widehat{\theta})\}\,dy$, with the outer expectation in Q_n referring to the distribution of the maximum weighted likelihood estimator $\widehat{\theta}$. We have $Q_n = \mathrm{E}_g\{w(Y_{\rm new}) \log f(Y_{\rm new}, \widehat{\theta}) - \int w(y) f(y, \widehat{\theta})\,dy\}$, with $Y_{\rm new}$ denoting a new observation generated from the same distribution as the previous ones. A natural estimator is therefore the cross-validated leave-one-out statistic

$$\widehat{Q}_{n,\rm xv} = n^{-1} \sum_{i=1}^n w(Y_i) \log f(Y_i, \widehat{\theta}_{(i)}) - \int w(y) f(y, \widehat{\theta})\,dy, \tag{2.41}$$

where $\widehat{\theta}_{(i)}$ is the maximum weighted likelihood estimator found from the reduced data set that omits the ith data point.

We show now that the (2.41) estimator is close to a penalised version of the directly available statistic $\widehat{Q}_n = H_n(\widehat{\theta}) = H_{n,\max}$ associated with (2.39). Arguments similar to those used to exhibit a connection from cross-validation to AIC lead under the present

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

circumstances first to

$$\widehat\theta_{(i)} - \widehat\theta \doteq -n^{-1}\mathrm{infl}(G_n, y_i) \doteq -n^{-1}\widetilde J_n^{-1}\{w(y_i)u(y_i,\widehat\theta) - \xi(\widehat\theta)\},$$

where $\widetilde J_n = -H_n^{(2)}(\widehat\theta)$, and then to

$$\widehat Q_{n,\mathrm{xv}} \doteq H_n(\widehat\theta) + n^{-1}\sum_{i=1}^n w(y_i)u(y_i,\widehat\theta)^{\mathrm t}(\widehat\theta_{(i)} - \widehat\theta) \doteq H_n(\widehat\theta) - n^{-1}\mathrm{Tr}(\widetilde J_n^{-1}\widetilde K_n),$$

in terms of $\widetilde K_n = n^{-1}\sum_{i=1}^n w(y_i)^2 u(y_i,\widehat\theta)u(y_i,\widehat\theta)^{\mathrm t} - \xi(\widehat\theta)\xi(\widehat\theta)^{\mathrm t}$. This is a generalisation of results derived and discussed in Section 2.9.

The approximation to the leave-one-out statistic (2.41) above can be made to resemble AIC, by multiplying with $2n$, so we define the w-weighted AIC score as

$$\begin{aligned}\mathrm{wAIC} &= 2nH_{n,\max} - 2\widetilde p^* \\ &= 2\Big\{\sum_{i=1}^n w(y_i)\log f(y_i,\widehat\theta) - n\int w(y)f(y,\widehat\theta)\,\mathrm dy\Big\} - 2\widetilde p^*,\end{aligned} \tag{2.42}$$

where now $\widetilde p^* = \mathrm{Tr}(\widetilde J_n^{-1}\widetilde K_n)$. This appropriately generalises the model-robust AIC, which corresponds to $w = 1$. The wAIC selection method may in principle be put to work for any non-negative weight function w, but its primary use is in connection with weight functions that downscale the more extreme parts of the sample space, to avoid the sometimes severely non-robust aspects of ordinary maximum likelihood and AIC strategies.

There is another path leading to wAIC of (2.42), more akin to the derivation of the AIC formula given in Section 2.3. The task is to quantify the degree to which the simple direct estimator $\widehat Q_n = H_n(\widehat\theta) = H_{n,\max}$ needs to be penalised, in order to achieve approximate unbiasedness for $Q_n = \mathrm E_g R_n$. Translating arguments that led to (2.16) to the present more general case of weighted likelihood functions, one arrives after some algebra at

$$\widehat Q_n - R_n = \bar Z_n + H_n^{(1)}(\theta_0)^{\mathrm t}(\widehat\theta - \theta_0) + o_P(n^{-1}) = \bar Z_n + n^{-1}V_n^{\mathrm t}JV_n + o_P(n^{-1}),$$

with $V_n = \sqrt n(\widehat\theta - \theta_0)$, where $\bar Z_n$ is the average of the zero mean variables $Z_i = w(Y_i)\log f(Y_i,\theta_0) - \int g(y)w(y)\log f(y,\theta_0)\,\mathrm dy$, and where (2.40) is used. This result, in conjunction with $W_n = V_n^{\mathrm t}JV_n \to_d W = V^{\mathrm t}JV$, where $V = J^{-1}U'$ and $U' \sim \mathrm N_p(0,K)$. Estimating the mean of W_n with $\mathrm{Tr}(\widetilde J_n^{-1}\widetilde K_n)$ leads therefore, again, to wAIC of (2.42), as the natural generalisation of AIC.

Just as we were able in Section 2.9 to generalise easier results for the i.i.d. model to the class of regression models, we learn here, with the appropriate efforts, that two paths of arguments both lead to the robust model selection criterion

$$\mathrm{wAIC} = 2nH_n(\widehat\theta) - 2\widetilde p^* = 2nH_{n,\max} - 2\widetilde p^*. \tag{2.43}$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Here the criterion function to maximise is

$$H_n(\theta) = n^{-1}\sum_{i=1}^n w(x_i, y_i)\log f(y_i \,|\, x_i, \theta) - n^{-1}\sum_{i=1}^n \int w(x_i, y) f(y \,|\, x_i, \theta)\,\mathrm{d}y,$$

and

$$\widetilde{p}^* = \mathrm{Tr}(\widetilde{J}_n^{-1}\widetilde{K}_n), \quad \text{with } \widetilde{J}_n = -H_n^{(2)}(\widehat{\theta}) \text{ and } \widetilde{K}_n = n^{-1}\sum_{i=1}^n v_i v_i^{\mathrm{t}},$$

where we write $v_i = w(x_i, y_i)u(y_i \,|\, x_i, \widehat{\theta}) - \xi(\widehat{\theta} \,|\, x_i)$ and $\xi(\theta \,|\, x_i) = \int w(x_i, y) f(y \,|\, x_i, \theta)u(y \,|\, x_i, \theta)\,\mathrm{d}y$. The traditional AIC is again the special case of a constant weight function.

Example 2.15 (2.14 continued) How to repair for Ervik's 1500-m fall?
To put the robust weighted AIC method into action, we need to specify a weight function $w(x, y)$ for computing the maxima and penalised maxima of the $H_n(\beta, \sigma)$ criterion functions. For this particular application we have chosen

$$w(x, y) = w_1\Big(\frac{x - \mathrm{med}(x)}{\mathrm{sd}(x)}\Big) w_2\Big(\frac{z - \mathrm{med}(z)}{\mathrm{sd}(z)}\Big),$$

involving the median and standard deviations of the x_i and the z_i, where $z_i = y_i - (5/1.5)x_i$, for appropriate w_1 and w_2 functions that are equal to 1 inside standard areas but scale down towards zero for values further away. Specifically, $w_1(u)$ is 1 to the left of a threshold value λ and $(\lambda/u)^2$ to the right of λ; while $w_2(u)$ is 1 inside $[-\lambda, \lambda]$ and $(\lambda/|u|)^2$ outside. In other applications we might have let also $w_1(u)$ be symmetric around zero, but in this particular context we did not wish to lower the value of $w_1(u)$ from its standard value 1 to the left, since unusually strong performances (like national records) should enter the analysis without any down-sizing in importance, even when they might look like statistical outliers. The choice of the $1/u^2$ factor secures that the estimation methods used have bounded influence functions for all y and for all x outlying to the right, with respect to both mean and standard deviation parameters in the linear-normal models. The threshold value λ may be selected in different ways; there are, for example, instances in the robustness literature where such a value is chosen to achieve say 95% efficiency of the associated estimation method, if the parametric model is correct. For this application we have used $\lambda = 2$.

We computed robust wAIC scores via the (2.43) formula, for each of the candidate models, operating on the full data set with the $n = 28$ skaters that include the fallen Ervik. This required using numerical integration and minimisation algorithms. Values of H_n maxima and $\widetilde{p}^*$ penalties are given in Table 2.7, and lead to the satisfactory conclusion that the four models are ranked in exactly the same order as for AIC used on the reduced $n = 27$ data set that has omitted the fallen skater. In this case the present wAIC and the wleAIC analysis reported on above agree fully on the preference order for the four

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Table 2.7. *AIC and robust weighted AIC for selecting a polynomial regression model for predicting the 5000-m time from the 1500-m time, with preference order, with data from the 2004 European Championships. The four models have 3, 4, 5, 6 parameters, respectively.*

Model:	linear	quadratic	cubic	quartic
AIC (all skaters):	−227.130 (4)	−213.989 (1)	−215.913 (2)	−217.911 (3)
AIC (w/o Ervik):	−206.203 (1)	−207.522 (2)	−209.502 (3)	−210.593 (4)
$nH_{n,\max}$:	−100.213	−99.782	−99.781	−99.780
$\widetilde{p}^*$	2.312	2.883	3.087	3.624
wAIC (all skaters):	−205.049 (1)	−205.329 (2)	−205.747 (3)	−206.811 (4)

models, and also agree well on the robustly estimated selected model: the wleAIC method yields estimates 3.202, 9.779 for slope b and spread parameter σ in the linear model, while the wAIC method finds estimates 3.193, 9.895 for the same parameters. ■

2.10.3 Robust AIC using M-estimators

Solutions to the direct likelihood equations lead to non-robust estimators for many models. We here abandon the likelihood as a starting point and instead start from M-estimating equations. As in Section 2.10.1 we restrict attention to linear regression models.

An M-estimator $\widehat{\beta}$ for the regression coefficients β is the solution to an equation of the form

$$\sum_{i=1}^{n} \Psi(x_i, y_i, \beta, \sigma) = \sum_{i=1}^{n} \eta(x_i, (y_i - x_i^{\rm t}\beta)/\sigma)x_i = 0, \tag{2.44}$$

see Hampel *et al.* (1986, Section 6.3). An example of such a function η is Huber's ψ function, in which case $\eta(x_i, \varepsilon_i/\sigma) = \max\{-1.345, \min(\varepsilon_i/\sigma, 1.345)\}$, using a certain default value associated with a 0.95 efficiency level under normal conditions. With this function we define weights for each observation, $\widehat{w}_i = \eta(x_i, (y_i - x_i^{\rm t}\widehat{\beta})/\sigma)/\{(y_i - x_i^{\rm t}\widehat{\beta})/\sigma\}$. Ronchetti (1985) defines a robust version of AIC, though not via direct weighting. The idea is the following. AIC is based on a likelihood function, and maximum likelihood estimators are obtained (in most regular cases) by solving the set of score equations $\sum_{i=1}^{n} u(Y_i \mid x_i, \beta, \sigma) = 0$. M-estimators solve equations (2.44). This set of equations can (under some conditions) be seen as the derivative of a function τ, which then might take the role of the likelihood function in AIC, to obtain Ronchetti's

$$\mathrm{AICR} = 2\sum_{i=1}^{n} \tau\big((y_i - x_i^{\rm t}\widehat{\beta})/\widehat{\sigma}\big) - 2\,\mathrm{Tr}(\widehat{J}_n^{-1}\widehat{K}_n),$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

where here $\widehat{J}_n$ and $\widehat{K}_n$ are estimators of $J = \mathrm{E}(-\partial\Psi/\partial\theta)$ (with $\theta = (\beta, \sigma)$) and $K = \mathrm{E}(\Psi\Psi^{\mathrm{t}})$. Note that the type of penalty above is also used for the model-robust TIC, see (2.20), though there used with the likelihood instead of τ. For the set of equations (2.44) with Huber's ψ function, the corresponding τ function is equal to (see Ronchetti, 1985)

$$\tau\big((y_i - x_i^{\mathrm{t}}\widehat{\beta})/\widehat{\sigma}\big) = \begin{cases} \frac{1}{2}(y_i - x_i^{\mathrm{t}}\widehat{\beta})^2/\widehat{\sigma}^2 & \text{if } |y_i - x_i^{\mathrm{t}}\widehat{\beta}| < 1.345\widehat{\sigma}, \\ 1.345|y_i - x_i^{\mathrm{t}}\widehat{\beta}|/\widehat{\sigma} - 1.345^2/2 & \text{otherwise.} \end{cases}$$

For a further overview of these methods, see Ronchetti (1997).

2.10.4 The generalised information criterion

Yet another approach for robustification of AIC by changing its penalty is obtained via the generalised information criterion (GIC) of Konishi and Kitagawa (1996). This criterion is based on the use of influence functions and can not only be applied for robust models, but also works for likelihood-based methods as well as for some Bayesian procedures.

We use notation for influence functions as in Section 2.9. Denote with G the true distribution of the data and G_n the empirical distribution. The data are modelled via a density function $f(y, \theta)$ with corresponding distribution F_θ. The GIC deals with functional estimators of the type $\widetilde{\theta} = T(G_n)$, for suitable estimator functionals $T = T(G)$. Maximum likelihood is one example, where the T is the minimiser of the Kullback–Leibler distance from the density of G to the parametric family. We assume that $T(F_\theta) = \theta$, i.e. T is consistent at the model itself, and let $\theta_{0,T} = T(G)$, the least false parameter value as implicitly defined by the $T(G_n)$ estimation procedure. To explain the GIC, we return to the derivation of AIC in Section 2.3. The arguments that led to $\mathrm{E}_g(\widehat{Q}_n - Q_n) = p^*/n + o(1/n)$ of (2.17), for the maximum likelihood estimator, can be used to work through the behaviour of $\widetilde{Q}_n = n^{-1}\ell_n(\widetilde{\theta})$ (which is now not the normalised log-likelihood maximum, since we employ a different estimator). Konishi and Kitagawa (1996) found that $\mathrm{E}_g(\widetilde{Q}_n - Q_n) = p^*_{\mathrm{infl}}/n + o(1/n)$, where

$$p^*_{\mathrm{infl}} = \mathrm{Tr}\left\{\int \mathrm{infl}(G, y)u(y, \theta_{0,T})^{\mathrm{t}}\,\mathrm{d}G(y)\right\},$$

with infl the influence function of $T(G)$, see (2.31). The GIC method estimates p^*_{infl} to arrive at

$$\mathrm{GIC} = 2\ell_n(\widetilde{\theta}) - 2\sum_{i=1}^{n} \mathrm{Tr}\{\mathrm{infl}(G_n, Y_i)^{\mathrm{t}} u(Y_i, \widetilde{\theta})\}.$$

For M-estimators $\widetilde{\theta} = T(G_n)$ which solve a set of equations $\sum_{i=1}^n \Psi(Y_i, \widetilde{\theta}) = 0$, the influence function can be shown to be

$$\mathrm{infl}(G, y) = J_T^{-1}\Psi(y, T(G)) = \left\{-\int \Psi^*(y, \theta_{0,T})\,\mathrm{d}G(y)\right\}^{-1}\Psi(y, \theta_{0,T}),$$

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

where $\Psi^*(y,\theta) = \partial\Psi(y,\theta)/\partial\theta$. This makes p^*_{infl} take the form $\text{Tr}(J_T^{-1}K_T)$, with $K_T = \int \Psi(y,\theta_{0,T})u(y,\theta_{0,T})^{\text{t}}\,\text{d}y$. Clearly the maximum likelihood method is a special case, corresponding to using $\Psi(y,\theta) = u(y,\theta)$, in which case p^*_{infl} coincides with p^* of (2.17) and GIC coincides with TIC of Section 2.5.

The above brief discussion was kept to the framework of i.i.d. models for observations $y_1,\ldots,y_n$, but the methods and results can be extended to regression models without serious obstacles. There are various M-estimation schemes that make the resulting $\widetilde{\theta}$ more robust than the maximum likelihood estimator. Some such are specifically constructed for robustification of linear regression type models, see e.g. Hampel *et al.* (1986), while others are more general in spirit and may work for any parametric models, like the minimum distance method of Section 2.10.2; see also Basu *et al.* (1998) and Jones *et al.* (2001).

Remark 2.1 One or two levels of robustness

Looking below the surface of the simple AIC formula (2.14), we have learned that AIC is inextricably linked to the Kullback–Leibler divergence in two separate ways; it aims at estimating the expected Kullback–Leibler distance from the true data generator to the parametric family, *and* it uses maximum likelihood estimators. The GIC sticks to the non-robust Kullback–Leibler as the basic discrepancy measure for measuring prediction quality, but inserts robust estimates for the parameters. Reciprocally, one might also consider the strategy of using a robustified distance measure, like that of (2.38), but inserting maximum likelihood estimators when comparing their minima. It may however appear more natural to simultaneously robustify both steps of the process, which the generalised Kullback–Leibler methods of Section 2.10.2 manage to do in a unified manner. ■

2.11 Notes on the literature

Maximum likelihood theory belongs to the core of theoretical and applied statistics; see e.g. Lehmann (1983) and Bickel and Doksum (2001) for comprehensive treatments. The standard theory associated with likelihoods relies on the assumption that the true data-generating mechanism is inside the parametric model in question. The necessary generalisations to results that do not require the parametric model to hold appeared some 50 years later than the standard theory, but are now considered standard, along with terms like least false parameter values and the sandwich matrix; see e.g. White (1982, 1994); Hjort (1986b, 1992a).

Applications of AIC have a long history. A large traditional application area is time series analysis, where one has studied how best to select the order of autoregressive and autoregressive moving average models. See for example Shibata (1976), Hurvich and Tsai (1989) or general books on time series analysis such as Brockwell and Davis (2002),

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Chatfield (2004). The book by McQuarrie and Tsai (1998) deals with model selection in both time series and regression models. For multiple regression models, see also Shibata (1981) or Nishii (1984). Burnham and Anderson (2002) deal extensively with AIC, also in the model averaging context. The form of the penalty in TIC has also been observed by Stone (1977).

Criteria related to AIC have sometimes been renamed to reflect the particular application area. The network information criterion (NIC) is such an example (Murata *et al.*, 1994), where AIC is adjusted for application with model selection in neural networks. The question of selecting the optimal number of parameters here corresponds to whether or not more neurons should be added to the network. Uchida and Yoshida (2001) construct information criteria for stochastic processes based on estimated Kullback–Leibler information for mixing processes with a continuous time parameter. As examples they include diffusion processes with jumps, mixing point processes and nonlinear time series models.

The GIC of Konishi and Kitagawa (1996) specified to M-estimators belongs to the domain of robust model selection criteria. Robust estimation is treated in depth in the books by Huber (1981) and Hampel *et al.* (1986). See also Carroll and Ruppert (1988) for an overview. Choi *et al.* (2000) use empirical tilting for the likelihood function to make maximum likelihood methods more robust. An idea of reweighted down-scaled likelihoods is explored in Windham (1995). Different classes of minimum distance estimators that contain Windham's method as a special case are developed in Basu *et al.* (1998) and Jones *et al.* (2001). Ronchetti and Staudte (1994) construct a robust version of Mallow's C_p criterion, see Chapter 4. Several of the robust model selection methods are surveyed in Ronchetti (1997). We refer to that paper for more references. Ronchetti *et al.* (1997) develop a robust version of cross-validation for model selection. The sample is split in two parts, one part is used for parameter estimation and outliers are here dealt with by using optimal bounded influence estimators. A robust criterion for prediction error deals with outliers in the validation set. Müller and Welsh (2005) use a stratified bootstrap to estimate a robust conditional expected prediction loss function, and combine this with a robust penalised criterion to arrive at a consistent model selection method. Qian and Künsch (1998) perform robust model selection via stochastic complexity. For a stochastic complexity criterion for robust linear regression, see Qian (1999).

Extensions of AIC that deal with missing data are proposed by Shimodaira (1994), Cavanaugh and Shumway (1998), Hens *et al.* (2006) and Claeskens and Consentino (2008); see also Section 10.4.

Hurvich *et al.* (1998) developed a version of the AIC_c for use in nonparametric regression to select the smoothing parameter. This is further extended for use in semiparametric and additive models by Simonoff and Tsai (1999). Hart and Yi (1998) develop one-sided cross-validation to find the smoothing parameter. Their approach only uses data at one side of x_i when predicting the value for Y_i.

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

Exercises

2.1 *The Kullback–Leibler distance:*

(a) Show that $\mathrm{KL}(g, f) = \int g \log(g/f)\,\mathrm{d}y$ is always non-negative, and is equal to zero only if $g = f$ almost everywhere. (One may e.g. use the Jensen inequality which states that for a convex function h, $\mathrm{E}\,h(X) \geq h(\mathrm{E}X)$, and strict inequality holds when h is strictly convex and X is nondegenerate.)

(b) Find the KL distance from a $\mathrm{N}(\xi_1, \sigma_1^2)$ to a $\mathrm{N}(\xi_2, \sigma_2^2)$. Generalise to the case of d-dimensional normal distributions.

(c) Find the KL distance from a $\mathrm{Bin}(n, p)$ model to a $\mathrm{Bin}(n, p')$ model.

2.2 *Subset selection in linear regression:* Show that in a linear regression model $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$, with independent normally distributed errors $\mathrm{N}(0, \sigma^2)$, AIC is given by

$$\mathrm{AIC}(p) = -n\,\log(\mathrm{SSE}_p/n) - n\{1 + \log(2\pi)\} - 2(p + 2),$$

where SSE_p is the residual sum of squares. Except for the minus sign, this is the result of an application of the function `AIC()` in `R`. The `R` function `stepAIC()`, on the other hand, uses as its AIC the value $\mathrm{AIC}_{\mathrm{step}}(p) = n\log(\mathrm{SSE}_p/n) + 2(p + 1)$. Verify that the maximiser of $\mathrm{AIC}_{\mathrm{step}}(p)$ is identical to the maximiser of $\mathrm{AIC}(p)$ over p.

2.3 *ML and AIC computations in R:* This exercise is meant to make users familiar with ways of finding maximum likelihood estimates and AIC scores for given models via algorithms in `R`.

Consider the Gompertz model first, the following is all that is required to compute estimates and AIC score. One first defines the log-likelihood function

```
logL = function(para, x) {
a = para[1]
b = para[2]
return(sum(log(a) + b * x - (a/b) * (exp(b * x) - 1)))}
```

where `x` is the data set, and then proceeds to its maximisation, using the nonlinear minimisation algorithm `nlm`, perhaps involving some trial and error for finding an effective starting point:

```
minuslogL = function(para, x){-logL(para, x)}
nlm(minuslogL, c(0.03, 0.03), x)
```

This then leads to `parahat = c(0.0188, 0.0207)`, and afterwards to `maxlogL = logL(parahat, x)` and `aic = 2 * maxlogL - 4`. Carry out analysis, as above, for the other models discussed in Example 2.6.

2.4 *Model-robust AIC for logistic regression:* Obtain TIC for the logistic regression model

$$\mathrm{P}(Y = 1 \mid x, z) = \frac{\exp(x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma)}{1 + \exp(x^{\mathrm{t}}\beta + z^{\mathrm{t}}\gamma)},$$

where $x = (1, x_2)^{\mathrm{t}}$ and $z = (x_3, x_4, x_5)^{\mathrm{t}}$. Verify that the approximation of $\mathrm{Tr}(J^{-1}K)$ by k, the number of free parameters in the model, leads to the AIC expression for logistic regression as in Example 2.4. Apply TIC to the low birthweight data, including in the model the same variables as used in Example 2.4.

2.5 *The* AIC$_c$ *for linear regression:* Verify that for the linear regression models $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$, with independent normally distributed errors $\mathrm{N}(0, \sigma^2)$, the corrected AIC is

$$\mathrm{AIC}_c = \mathrm{AIC} - 2(p+2)(p+3)/(n-p-3).$$

2.6 *Two AIC correction methods for linear regression:* Simulate data from a linear regression model with say $p = 5$ and $n = 15$, and implement in addition to the AIC score both correction methods AIC$_c$ of (2.23) and AIC$_c^*$ of (2.24). Compare values of these three scores and their ability to pick the 'right' model in a little simulation study. Try to formalise what it should mean that one of the two correction methods works better than the other.

2.7 *Stackloss:* Read the stackloss data in `R` by typing `data(stackloss)`. Information on the variables in this data set is obtained by typing `?stackloss`. We learn that there are three regression variables 'Air.Flow', 'Water.Temp' and 'Acid.Conc.'. The response variable is 'stack.loss'.

(a) Consider first a model without interactions, only main effects. With three regression variables, there are $2^3 = 8$ possible models to fit. Fit separately all eight models (no regression variables included; only one included; only two included; and all three included). For each model obtain AIC (using function `AIC()`). Make a table with these values and write down the model which is selected.

(b) Consider the model with main effects and pairwise interactions. Leave all main effects in the model, and search for the best model including interactions. This may be done using the `R` function `stepAIC` in library(MASS).

(c) Now also allow for the main effects to be left out of the final selected model and search for the best model, possibly including interactions. The function `stepAIC` may again be used for this purpose.

2.8 *Switzerland in 1888:* Perform model selection for the `R` data set called `swiss`, which gives a standardised fertility measure and socio-economic indicators for each of the 47 French-speaking provinces of Switzerland, at around 1888. The response variable is called 'fertility', and there are five regression variables. Type `?swiss` in `R` for more information about the data set.

2.9 *The Ladies Adelskalenderen:* Consider the data from the Adelskalenderen for ladies' speedskating (available from the book's website). Times on four distances 500-m, 1500-m, 3000-m and 5000-m are given. See Section 1.7 for more information, there concerning the data set for men. Construct a scatterplot of the 1500-m time versus the 500-m time, and of the 5000-m time versus the 3000-m time. Try to find a good model to estimate the 1500-m time from the 500-m time, and a second model to estimate the 5000-m time from the 3000-m time. Suggest several possible models and use AIC and AIC$_c$ to perform the model selection.

2.10 *Hofstedt's highway data:* This data set is available via `data(highway)` in the `R` library `alr3` (see also Weisberg, 2005, Section 7.2). There are 39 observations on several

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press

highway-related measurements. The response value of interest is Rate, which is the accident rate per million vehicle miles in the year 1973. Eleven other variables $(x_1, \ldots, x_{11})$ are possibly needed in a linear regression model with Rate as response.

(a) Fit the full model $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{11} x_{i,11} + \varepsilon_i$ using robust regression with M-estimators for the regression coefficients, using Huber's ψ function. The function `rlm` in the `R` library `MASS` may be used for this purpose. Obtain for each observation the value of the weight $\widehat{w}_i$ using the residuals of the fitted full model. Weights below one indicate downweighted observations. Identify these cases.

(b) Use AICR to perform variable selection.

2.11 *Influence function for maximum likelihood:* For a regular parametric family $f(y, \theta)$, consider the maximum likelihood functional $T = T(G)$ that for a given density g (e.g. the imagined truth) finds the least false value $\theta_0 = \theta_0(g)$ that minimises the Kullback–Leibler distance $\int g \log(g/f_\theta)\,\mathrm{d}y$. Two continuous derivatives are assumed to exist, as is the matrix $J = -\int g(y)\mathrm{infl}(y, \theta_0)\,\mathrm{d}y$, taken to be positive definite.

(a) Show that $T(G)$ also may be expressed as the solution θ_0 to $\int g(y)u(y, \theta_0)\,\mathrm{d}y = 0$.

(b) Let y be fixed and consider the distribution $G_\varepsilon = (1 - \varepsilon)G + \varepsilon\delta(y)$, where $\delta(y)$ denotes unit point mass at position y. Write $\theta_\varepsilon = \theta_0 + z$ for the least false parameter when f_θ is compared to G_ε. Show that this is the solution to

$$(1 - \varepsilon)\int g(y')u(y', \theta_\varepsilon)\,\mathrm{d}y' + \varepsilon u(y, \theta_\varepsilon) = 0.$$

By Taylor expansion, show that z must be equal to $J^{-1}u(y, \theta_0)\varepsilon$ plus smaller terms. Show that this implies $\mathrm{infl}(G, y) = J^{-1}u(y, \theta_0)$, as claimed in (2.34).

2.12 *GIC for M-estimation:* Start from Huber's ψ function for M-estimation: for a specified value $b > 0$, $\psi(y) = y$ if $|y| \le b$, $\psi(y) = -b$ if $y < -b$ and $\psi(y) = b$ if $y > b$, and show that the influence function $\mathrm{infl}(G, y)$ at the standard normal distribution function $G = \Phi$ equals $\psi(y)/\{2\Phi(b) - 1\}$. Next, specify GIC for M-estimation in this setting.

2.13 *Leave-one-out for linear regression:* Consider the linear regression model where y_i has mean $x_i^{\mathrm{t}}\beta$ for $i = 1, \ldots, n$, i.e. the vector y has mean $X\beta$, with β a parameter vector of length p and with X of dimension $n \times p$, assumed of full rank p. The direct predictor of y_i is $\widehat{y}_i = x_i^{\mathrm{t}}\widehat{\beta}$, while its cross-validated predictor is $\widehat{y}_{(i)} = x_i^{\mathrm{t}}\widehat{\beta}_{(i)}$, where $\widehat{\beta}$ and $\widehat{\beta}_{(i)}$ are the ordinary least squares estimates of β based on respectively the full data set of size n and the reduced data set of size $n - 1$ that omits (x_i, y_i).

(a) Let $A = X^{\mathrm{t}}X = \sum_{i=1}^n x_i x_i^{\mathrm{t}} = A_i + x_i x_i^{\mathrm{t}}$, with $A_i = \sum_{j \ne i} x_j x_j^{\mathrm{t}}$. Show that

$$A^{-1} = A_i^{-1} - \frac{A_i^{-1}x_i x_i^{\mathrm{t}} A_i^{-1}}{1 + x_i^{\mathrm{t}} A_i^{-1} x_i},$$

assuming that also A_i has full rank p.

(b) Let $s_1, \ldots, s_n$ be the diagonal elements of $I_n - H = I_n - X(X^{\mathrm{t}}X)^{-1}X$. Show that $s_i = 1 - x_i^{\mathrm{t}} A^{-1} x_i = 1/(1 + x_i^{\mathrm{t}} A_i^{-1} x_i)$.

(c) Use $\widehat{\beta} = (A_i + x_i x_i^{\mathrm{t}})^{-1}(w + x_i y_i)$ and $x_i^{\mathrm{t}}\widehat{\beta}_{(i)} = x_i^{\mathrm{t}} A_i^{-1} w$, with $w = \sum_{j \ne i} x_j y_j$, to show that $y_i - \widehat{y}_i = (y_i - \widehat{y}_{(i)})/(1 + x_i^{\mathrm{t}} A_i^{-1} x_i)$. Combine these findings to conclude that

$y_i - \widehat{y}_{(i)} = (y_i - \widehat{y}_i)/s_i$ for $i = 1, \ldots, n$. This makes it easy to carry out leave-one-out cross-validation for linear regression models.

2.14 *The robustified Kullback–Leibler divergence:* For a given non-negative weight function w defined on the sample space, consider $d_w(g, f_\theta)$ of (2.38).

(a) Show that $d_w(g, f_\theta)$ is always non-negative, and that it is equal to zero only when $g(y) = f(y, \theta_0)$ almost everywhere, for some θ_0. Show also that the case of a constant w is equivalent to the ordinary Kullback–Leibler divergence.

(b) Let $T(G) = \theta_0$ be the minimiser of $d_w(g, f_\theta)$, for a proposed parametric family $\{f_\theta\colon \theta \in \Theta\}$. With G_n the empirical distribution of data $y_1, \ldots, y_n$, show that $T(G_n)$ is the estimator that maximises $H_n(\theta)$ of (2.39). Show that the influence function of T may be expressed as $J^{-1}a(y, \theta_0)$, where $a(y, \theta) = w(y)u(y, \theta) - \xi(\theta)$ and J is the limit of $J_n = -H_n^{(2)}(\theta_0)$; also, $\xi(\theta) = \int w f_\theta u_\theta \, \mathrm{d}y$.

(c) It is sometimes desirable to let the data influence which weight function w to use in (2.38). Suppose a parametric $w(y, \alpha)$ is used for weight function, with $\hat{\alpha}$ some estimator with the property that $\sqrt{n}(\hat{\alpha} - \alpha)$ has a normal limit. Extend the theory of Section 2.10.2 to cover also these cases of an estimated weight function, and, in particular, derive an appropriate generalisation of the robust model selection criterion wAIC of (2.42).

https://doi.org/10.1017/CBO9780511790485.003 Published online by Cambridge University Press