Chapter 9
Model Selection and Model Averaging

In Chaps. 4 and 5 we explored Maximum Likelihood estimation and Bayesian statistics and, given a particular model, used our data to estimate the unknown parameter θ. The validity of the model itself was not questioned, except for a brief detour into the Bayes factor. In this chapter we will delve a little deeper into this idea and explore common routines for selecting the most appropriate model for the data. Before we start, let us make the goal of model selection a little more explicit. We defined with

$$l(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i; \theta)$$

the log-likelihood. This likelihood of the data clearly depends on the probability model $f(.; \theta)$, which we thus far have simply assumed to be true and have not specified in the notation. However, the specification and selection of the probability model is already an important step in statistical modelling and should absolutely be explicitly notated. In principle, the likelihood should then be denoted with $l(\theta; y_1, \ldots, y_n | f)$. Hence, a likelihood can only be derived if we condition upon the model, that is, we assume that $f(.; \theta)$ holds. The phrase "model" is loosely defined here and simply states that we condition on a class of probability models by assuming

$$Y_i \sim f(y; \theta) \quad i.i.d. \tag{9.0.1}$$

This "specification" step (9.0.1) is what we will focus on for the moment. Note that all derivations of Chap. 4 for the Maximum Likelihood estimate were based on the validity of (9.0.1). But what happens if the data come from a different distribution? For instance, let

$$Y_i \sim g(y) \quad i.i.d., \tag{9.0.2}$$

where $g(.)$ is the unknown true distribution.

If we do not know $g(.)$, the question arises whether it is possible to learn anything about it at all. More importantly, if we have falsely taken $f(.;\theta)$ as the true distribution, what is the meaning and nature of the Maximum Likelihood estimate $\hat{\theta}$ under this misspecification? That is, what is the use of (9.0.1) if $f(.;\theta) \neq g(.)$? Should we be warned? Is it wrong to use the Maximum Likelihood estimate when the model is incorrect? After all, as data in general are complex and their distributions usually unknown, we have no a priori insight into their true distribution $g(.)$. This clearly implies that most (if not all) models are wrong. But are they still useful?

At the very least, we want to guarantee that we are not acting like the proverbial drunkard, looking for his keys under a lamppost. A policeman arrives, helps him with his search and, after a while, asks the man where exactly he dropped the keys. The drunkard replies that he dropped them on the other side of the street, but it was dark over there, so he decided to instead search under the lamppost where there was light. That is, we need to know if, when we search for the parameters of a model, we are searching under the lamppost or where we lost the key, i.e. where it is convenient or from where the data actually stem. We will fortunately see that the Maximum Likelihood approach makes sense, even if the true model is not within the assumed model class (9.0.1) but some unknown model (9.0.2). Hence, the likelihood approach sheds light on the true model $g(.)$, even if we do not know its structure.

Sometimes it happens that we have two models that describe the data equally well. This begs the question, why we should explicitly select one of the two, instead of using both models. This perspective will lead us to model averaging, where we fit multiple models and, instead of discarding all but the best model, we apply them together as a group. But first we will begin with Akaike Information Criterion—an essential component of model selection.

9.1 Akaike Information Criterion

9.1.1 Maximum Likelihood in Misspecified Models

A key component of both model selection and model averaging is the Akaike Information Criterion (AIC), which we already briefly introduced in Chap. 7. We will explore the theory behind the AIC in detail in the following section. Our exploration begins with the Kullback–Leibler (KL) divergence, defined in Sect. 3.2.5. Recall that for two densities $f(.)$ and $g(.)$, the KL divergence is defined as

$$\begin{aligned} KL(g, f) &= \int \log\left(\frac{g(y)}{f(y)}\right) g(y)dy \\ &= \int \log(g(y))g(y)dy - \int \log(f(y))g(y)dy. \end{aligned} \tag{9.1.1}$$

We now consider $g(.)$ to be the true, but unknown, distribution, while $f(y) = f(y;\theta)$ is our distributional model. If $g(.)$ belongs to the model class, i.e. there exists a true parameter value θ_0 such that $g(y) = f(y;\theta_0)$ for all y, then the Kullback–Leibler divergence takes the value zero for $\theta = \theta_0$. Generally, however, the Kullback–Leibler divergence is positive, meaning that $f(y;\theta) \neq g(y)$, for some y, no matter what value θ takes. A sensible strategy would be to choose a θ, so as to minimise the Kullback–Leibler divergence between $f(y;\theta)$ and the unknown true distribution $g(.)$. We denote this parameter value with θ_0 such that

$$\theta_0 = \underset{\theta\in\Theta}{\arg\min}\, KL(g, f). \tag{9.1.2}$$

Looking at (9.1.1), it is clear that the first component does not depend on $f(.;\theta)$, but only on the true but unknown distribution $g(.)$. From here on, this component can therefore be ignored, as it is effectively a constant of no particular interest. So equivalently, we can instead maximise

$$\int \log(f(y;\theta))g(y)dy$$

with respect to θ. Taking the derivative gives $\int s(\theta; y)g(y)dy$, where $s(\theta; y) = \partial \log f(y;\theta)/\partial\theta$ is the score function. The best parameter θ_0 apparently needs to fulfil

$$\int s(\theta_0; y)g(y)dy = E_g(s(\theta_0; Y)) = 0, \tag{9.1.3}$$

where the expectation is calculated using the true but unknown distribution $g(.)$. In other words, the best parameter θ_0 gives a vanishing expectation of the score function. This property holds regardless of the true model.

Given this result and observed data $y_1, \ldots, y_n$ drawn from (9.0.2), let us now attempt to estimate θ_0. This is done by setting the empirical score to zero, i.e.

$$\sum_{i=1}^{n} \frac{\partial \log f(y_i;\hat{\theta})}{\partial\theta} = \sum_{i=1}^{n} s_i(\hat{\theta}; y_i) = 0. \tag{9.1.4}$$

Clearly, this is identical to Maximum Likelihood estimation, i.e. the Maximum Likelihood approach provides an estimate for θ_0. This finding serves an important purpose, in that it allows us to interpret the Maximum Likelihood estimate from a different perspective. Setting θ to θ_0 minimises the difference between $g(.)$ and $f(y;\theta)$, and hence $f(.,\theta_0)$ is the closest density to $g(.)$ as measured by the Kullback–Leibler divergence.

We can go even further and derive, as in Chap. 4, asymptotic properties of $\hat{\theta}$.

Property 9.1 For data $Y_1, \ldots, Y_n$ drawn *i.i.d.* from $g(.)$, the Maximum Likelihood estimate $\hat{\theta}$ for model $f(.;\theta)$ fulfils

$$\hat{\theta} - \theta_0 \overset{a}{\sim} N(0, I^{-1}(\theta_0)V(\theta_0)I^{-1}(\theta_0)),$$

where θ_0 is defined as per (9.1.3), $I(\theta_0)$ is the Fisher matrix and $V(\theta_0)$ is the variance of the score function.

Proof To derive the above property, we expand (9.1.4) around θ_0. We begin with some notation. Firstly, let us explicitly allow the parameter θ to be multidimensional, giving $\theta = (\theta_1, \ldots, \theta_p)^T$. Secondly, let us define the score as

$$s(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} s_i(\theta; y_i).$$

As defined in (3.3.3), the observed Fisher information is given by

$$J(\theta; y_1, \ldots, y_n) = -\sum_{i=1}^{n} \frac{\partial s_i(\theta; y_i)}{\partial \theta^T} = -\frac{\partial s(\theta; y_1, \ldots, y_n)}{\partial \theta^T},$$

which is a $p \times p$ dimensional matrix. The corresponding Fisher matrix is defined as

$$I(\theta) = E_g\left(-\frac{\partial s(\theta; Y_1, \ldots, Y_n)}{\partial \theta^T}\right) = \frac{\partial^2}{\partial\theta\partial\theta^T}\sum_{i=1}^{n}\int \log f(y_i;\theta)g(y_i)dy_i. \tag{9.1.5}$$

Note that the expectation is carried out with respect to the true but unknown density $g(.)$, not with the model density $f(.;\theta)$ as in Chap. 4. The first order Taylor expansion of (9.1.4) gives

$$0 = s(\hat{\theta}; y_1, \ldots, y_n) \tag{9.1.6}$$

$$\approx s(\theta_0; y_1, \ldots, y_n) + J(\theta_0; y_1, \ldots, y_n)(\hat{\theta} - \theta_0) \tag{9.1.7}$$

$$\Leftrightarrow (\hat{\theta} - \theta_0) \approx J^{-1}(\theta_0; y_1, \ldots, y_n)s(\theta_0; y_1, \ldots, y_n). \tag{9.1.8}$$

Note that $J(\theta; y_1, \ldots, y_n)$ is the observed version of $I(\theta)$, which allows the simplification of (9.1.8) to

$$(\hat{\theta} - \theta_0) \approx I^{-1}(\theta_0)s(\theta_0; y_1, \ldots, y_n). \tag{9.1.9}$$

With standard arguments based on the central limit theorem (as in Chap. 4), this gives

$$(\hat{\theta} - \theta_0) \overset{a}{\sim} N(0, I^{-1}(\theta_0)V(\theta_0)I^{-1}(\theta_0)), \tag{9.1.10}$$

where

$$V(\theta_0) = Var(s(\theta_0; Y_1, \ldots, Y_n)).$$

If the true model is of the form $f(y;\theta)$, then the variance simplifies further, as $Var(s(\theta_0; Y_1, \ldots, Y_n))$ is equal to the Fisher information $I(\theta)$. For misspecified models, however, this property does not apply and we get

$$\begin{aligned} Var(s(\theta_0; Y_1, \ldots, Y_n)) &= \int s(\theta_0; y_1, \ldots, y_n)s^T(\theta_0; y_1, \ldots, y_n) \prod_{i=1}^{n} g(y_i)dy_i \\ &= \sum_{i=1}^{n} \int s_i(\theta_0; y_i)s_i^T(\theta_0; y_i)g(y_i)dy_i, \end{aligned} \tag{9.1.11}$$

where $y = (y_1, \ldots, y_n)$. This is not necessarily equal to the Fisher information. However, the empirical version of (9.1.11) can be used for estimating the variance. □

9.1.2 Derivation of AIC

We have shown so far that the Maximum Likelihood approach gives the best possible estimate $\hat{\theta}$ for θ_0, as defined in (9.1.2). The next task is to evaluate how close our best model $f(y;\hat{\theta})$ is to $g(y)$. Note that y here represents any possible value of a random variable Y, while $\hat{\theta}$ is the Maximum Likelihood estimate derived from the data $y_1, \ldots, y_n$. Instead of looking at specific values, we take the expectation over all possible values of Y and look at the Kullback–Leibler divergence between $g(.)$ and $f(.;\hat{\theta})$. This is given by

$$KL\left(g(.); f(.;\hat{\theta})\right) = \int \log\left(\frac{g(y)}{f(y;\hat{\theta})}\right) g(y)dy \tag{9.1.12}$$

$$= \int \log(g(y))g(y)dy - \int \log(f(y;\hat{\theta}))g(y)dy. \tag{9.1.13}$$

Again, the first component is a constant and we can simply focus on the second. Clearly, because $g(.)$ is unknown, the integral cannot be directly calculated. Moreover, the estimate $\hat{\theta}$, and hence, the Kullback–Leibler divergence (9.1.13), depends on the observed data $y_1, \ldots, y_n$. To obtain a general measure that does

not depend on the observed values, it seems sensible to take the expectation over the sample $Y_1, \ldots, Y_n$. This gives the expected Kullback–Leibler divergence

$$\begin{aligned} & E_{Y_1,\ldots,Y_n}\left\{KL\big(g(.), f(,;\hat{\theta}(Y_1,\ldots,Y_n))\big)\right\} \\ = \quad & \text{const} - E_{Y_1,\ldots,Y_n}\left\{\int \log f\big(y;\hat{\theta}(Y_1,\ldots,Y_n)\big)g(y)dy\right\}, \end{aligned} \tag{9.1.14}$$

where we made the dependence of our estimate $\hat{\theta}$ on the data $y_1, \ldots, y_n$ explicit. The constant term can be ignored in the following. Note that we have two integrals in the second term in (9.1.14): the inner integral over y in the Kullback–Leibler divergence and the outer integral resulting from the expectation with respect to $Y_1, \ldots, Y_n$. To be explicit, we can write the expectation in the second component of (9.1.14) as

$$\int\left[\int \log\Big(f(y;\hat{\theta}(y_1,\ldots,y_n))\Big)\, g(y)dy\right]\, g(y_1,\ldots,y_n)dy_1\ldots dy_n, \tag{9.1.15}$$

where $g(y_1, \ldots, y_n) = \prod_i g(y_i)$. This integral may look rather clumsy, but we will nevertheless attempt to approximate it. This will lead us to the famous Akaike Information Criterion.

Definition 9.1 The Akaike Information Criterion (AIC) is defined as

$$\text{AIC} = -2\sum_{i=1}^{n} \log f(y_i;\hat{\theta}) + 2p. \tag{9.1.16}$$

The multiplication with 2 above has no specific meaning and was suggested by Akaike himself. The AIC is very broadly applicable and, as we will show, is much more than just balance between goodness of fit (the first term) and complexity (the second term in (9.1.16)). The deeper meaning of the AIC becomes clear through its derivation.

Derivation We will now show how the AIC can be explicitly derived. We begin by disentangling the double integral and approximating $f(y;\hat{\theta})$ with a second order Taylor series expansion around θ_0. To simplify notation, from now on we write $\hat{\theta}(y_1, \ldots, y_n)$ as $\hat{\theta}$, but it should be kept in mind that $\hat{\theta}$ depends on $y_1, \ldots, y_n$ and not on y. This gives the approximation of (9.1.15) as

$$\begin{aligned} & \int\int\left\{\log f(y;\theta_0) + \frac{\partial \log f(y;\theta_0)}{\partial\theta^T}(\hat{\theta}-\theta_0) - \frac{1}{2}(\hat{\theta}-\theta_0)^T J(\theta_0)(\hat{\theta}-\theta_0)\right\} \\ & \qquad \times g(y)dy\, g(y_1,\ldots,y_n)dy_1\ldots dy_n \\ \approx \quad & \int \log\big(f(y;\theta_0)\big)g(y)dy - \frac{1}{2}\int (\hat{\theta}-\theta_0)^T I(\theta_0)(\hat{\theta}-\theta_0)g(y_1,\ldots,y_n)dy_1\ldots dy_n \end{aligned} \tag{9.1.17}$$

using (9.1.10) such that the second component vanishes. Let us first look at the final component in (9.1.17). Making use of (9.1.10), this can be asymptotically approximated by $tr(I^{-1}(\theta_0)V(\theta_0))$. Note that neither the integral in the first component in (9.1.17) nor $V(\theta_0)$ and $I(\theta_0)$ can be calculated explicitly, as they depend on the unknown $g(.)$. Our aim is therefore to estimate the above quantity based on the data $y_1, \ldots, y_n$. We do know that the data $y_1, \ldots, y_n$ are drawn from $g(.)$. Therefore, we replace the integral in the first component with its arithmetic mean. To do so, we look at the empirical version of (9.1.15) but leave $\hat{\theta}$ fixed for now. That is, we replace the expectation over $g(y)$ with the arithmetic mean using $y_1, \ldots, y_n$. To be specific, we calculate

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i; \hat{\theta}). \tag{9.1.18}$$

This is clearly just the likelihood function at its maximum divided by the sample size. We again expand (9.1.18) around θ_0 and obtain an approximation of (9.1.18) with

$$\frac{1}{n}\sum_{i=1}^{n} \left\{ \log f(y; \theta_0) + s_i(\theta_0; y_i)^T(\hat{\theta} - \theta_0) - \frac{1}{2}(\hat{\theta} - \theta_0)^T J_i(\theta_0; y_i)(\hat{\theta} - \theta_0) \right\}, \tag{9.1.19}$$

where $J_i(\theta_0, y_i) = -\partial^2 \log f(y_i; \theta_0)/\partial\theta\partial\theta^T$. Note that the first component in (9.1.19) can approximate the first component in (9.1.17). This means

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i; \theta_0) \quad \rightarrow \quad \int \log f(y; \theta_0) g(y) dy \tag{9.1.20}$$

for increasing n, which can in fact be proven to be a consistent estimate. Moreover, because with increasing sample size n

$$\frac{1}{n}\sum_{i} J_i(\theta_0; y) \quad \rightarrow \quad I(\theta_0), \tag{9.1.21}$$

we can argue that the integrand of the third component in (9.1.19) converges to

$$\frac{1}{2}(\hat{\theta} - \theta_0)^T I^{-1}(\theta_0)(\hat{\theta} - \theta_0).$$

Taking the expectation with respect to $Y_1, \ldots, Y_n$ and using the asymptotic distribution (9.1.10), we see that

$$E_{Y_1,\ldots,Y_n}\left(\frac{1}{2}(\hat{\theta}-\theta_0)^T I^{-1}(\theta_0)(\hat{\theta}-\theta_0)\right) = tr(I^{-1}(\theta_0)V(\theta_0)).$$

What remains is the second component in (9.1.19). If we subtract this and take (9.1.20), (9.1.21) and the approximation (9.1.17), we get

$$\frac{1}{n}\sum_{i=1}^{n}\log f(y_i;\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} s_i^T(\theta_0; y_i)(\hat{\theta}-\theta_0) \quad \rightarrow \quad \int\int \log f(y;\hat{\theta}) g(y)dy \; g(y_1,\ldots,y_n)dy_1\ldots dy_n.$$

This suggests the use of $\frac{1}{n}\sum_{i=1}^{n} s_i^T(\theta_0; y_i)(\hat{\theta}-\theta_0)$ as a bias correction, which we now try to simplify further. Using (9.1.10) gives

$$\frac{1}{n}\sum_i s_i^T(\theta_0; y_i)(\hat{\theta}-\theta_0) \approx \frac{1}{n}\sum_i s_i^T(\theta_0; y_i) I^{-1}(\theta_0)\sum_j s_j(\theta_0; y_j).$$

Bearing in mind that $E(s_i(\theta_0; y_i)) = 0$, if we take the expectation with respect to $Y_1, \ldots, Y_n$, we get

$$E_{Y_1,\ldots,Y_n}\left(\frac{1}{n}\sum s_i(\theta_0; y_i)^T(\hat{\theta}-\theta_0)\right) \approx \frac{1}{n} tr\left(I^{-1}(\theta_0)V(\theta_0)\right). \tag{9.1.22}$$

As both $I(\theta_0)$ and $V(\theta_0)$ depend on the unknown distribution $g(.)$, we cannot calculate their values. Instead, we replace the matrices with their empirical counterparts, which gives

$$\hat{I}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(y_i;\theta)}{\partial\theta\partial\theta^T} \text{ and}$$

$$\hat{V}(\theta) = \frac{1}{n}\sum_{i=1}^{n} s_i^T(\theta; y_i)s_i(\theta; y_i).$$

This approximation was proposed by Takeuchi (1979), see also Shibata (1989) or Konishi and Kitagawa (1996). It is, however, more common and much more stable for small n to approximate the term (9.1.22) even further. To do so, assume that $g(.) = f(.;\theta_0)$. Then $V(\theta_0) = I(\theta_0)$ such that

$$\frac{1}{n} tr(I^{-1}(\theta_0)V(\theta_0)) = \frac{1}{n}p,$$

where p is the dimension of the parameter. This approximation is attractively simple and was proposed by Akaike (1973). Combining the above derivations gives the famous **Akaike Information Criterion (AIC)** . □

We can conclude from the above derivation that the AIC can be interpreted more deeply than simply as a balance between goodness of fit and complexity. This deeper meaning can be drawn from the Kullback–Leibler divergence, which we write explicitly in the following important property:

Property 9.2 The AIC in (9.1.16) serves as estimate for

$$2E_{Y_1,\ldots,Y_n}\{KL(g(.), f(.;\hat{\theta}))\} - 2\int \log\big(g(y)\big)g(y)dy. \tag{9.1.23}$$

9.1.3 AIC for Model Comparison

We will now discuss how the AIC can be applied to model selection. Looking at (9.1.23), we see that the latter component is unknown, and hence, the absolute value of the AIC is not informative for us. Consequently, we are not able to explicitly estimate a value for the expected Kullback–Leibler divergence $E_{Y_1,\ldots,Y_n}\{KL(g(.), f(.;\hat{\theta}))\}$. In other words, we can never evaluate how close $f(;\theta)$ is to $g(.)$. However, the AIC is eminently useful for *relative* comparisons. Let us demonstrate the concept with two candidate models. Assume that we have the two models $f_1(.;\theta_1)$ and $f_2(.;\theta_2)$. These could be two completely different distributions or the same distribution with two different parameterisations. A common setting is $f_1(.) = f_2(.) = f(.)$, meaning they belong to the same distributional model, but θ_1 can be derived from θ_2 by setting some values of θ_2 to zero. This is also called **nested model selection**, as $f_1(.)$ is a special case of $f_2(.)$. We can now calculate the AIC values for the two models and compare them. Assume that θ_2 is p_2 dimensional and θ_1 is p_1 dimensional with $p_2 > p_1$. Then

$$AIC(1) = -2\sum_{i=1}^{n} \log\big(f(y_i;\hat{\theta}_1)\big) + 2p_1$$

$$AIC(2) = -2\sum_{i=1}^{n} \log\big(f(y_i;\hat{\theta}_2)\big) + 2p_2.$$

In this case, let us assume that AIC(1) < AIC(2). At first glance, this is a surprising result. This means that we are better able to model, as measured by our approximation of the expected Kullback–Leibler divergence, the true unknown density $g(.)$ with the smaller model $f_1(.)$ than with more complex model $f_2(.)$. In the above case, $f_1(.)$ is just a special case of $f_2(.)$ with some parameters set

to zero, and clearly we are able to get as close to $g(.)$ with $f_2(.)$ as we can with $f_1(.)$. With a higher dimensional model, we achieve more flexibility, and hence the smallest Kullback–Leibler divergence between $f_2(.)$ and $g(.)$ is always smaller than that of $f_1(.)$ and $g(.)$. However, the AIC does not measure how close we can get, but how close we are on average, after estimating the unknown parameters, i.e. we take the expectation with respect to $Y_1, \ldots, Y_n$. This takes into account the fact that additional parameters are estimated in $f_2(.)$, which are set to zero in $f_1(.)$ and this implies that we suffer from more estimation variability with $f_2(.)$ compared to $f_1(.)$, which in turn makes the expected Kullback–Leibler divergence of $f_2(.)$ larger than that of $f_1(.)$. To sum up, we emphasise that the AIC compares both how close we can get with the model to the unknown distribution $g(.)$ and the variability of parameter estimation.

It also helps to simply understand the AIC as a balance between the fit and complexity of a model, which is best explained in a regression context. To demonstrate, let

$$Y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i,$$

where x_{i1} and x_{i2} are known covariates. For the residuals, we assume normality $\varepsilon_i \sim N(0, \sigma^2)$ and independence. The log-likelihood is given by

$$l(\beta, \sigma) = \sum_{i=1}^{n} -\log(\sigma) - \frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2},$$

where $x_i = (1, x_{i1}, x_{i2})$ and $\beta = (\beta_0, \beta_1, \beta_2)^T$. The Maximum Likelihood estimate is then

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2/n,$$

where $\boldsymbol{X}$ is the design matrix and $\boldsymbol{y}$ is the vector of observations as defined in (7.1.5). The AIC is then given by ($p = 2$ in this example)

$$AIC = n\log(\hat{\sigma}^2) + 2(p+2).$$

Note that if we include more covariates, the estimated variance $\hat{\sigma}^2$ becomes smaller, but the number of parameters grows. The term $\log(\hat{\sigma}^2) = \log\left(\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2/n\right)$ measures the goodness of fit, i.e. how close the predicted value $x_i\hat{\beta}$ is to the observed value y_i. The term $2(p+2)$ measures the complexity of the model, in this case how many parameters have been fitted. These two aspects, goodness of fit and the number of parameters, are balanced by the AIC.

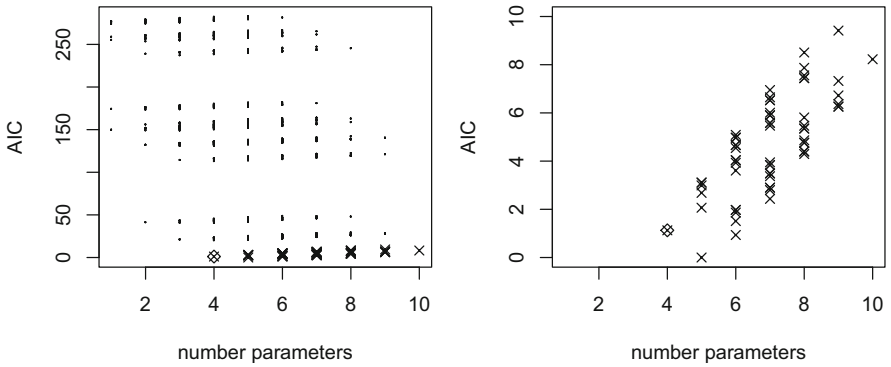

Fig. 9.1 AIC values for all 1024 possible models. Models including the true 4 covariates are indicated as crosses. The true model is shown as diamond. The right-hand plot is zoomed in on the true model, just showing the models with the smallest AIC

Example 41 Let us demonstrate the AIC with a simulated example. We generate 10 independent covariates $x_{ij} \sim N(0, 1)$, where $i = 1, \ldots, 200$ and $j = 1, \ldots, 10$, and simulate

$$Y_i \sim N(0.25x_{i1} + 0.25x_{i2} + 0.1x_{i3} + 0.1x_{i4}, 1).$$

Hence, the response variable depends on the first 4 covariates, while the remaining 6 are spurious. We fit $2^{10} = 1024$ possible regression models, with each covariate either absent or present in the model. Given the AIC value of the 1024 possible models, we want to select the best model. The results are visualised in Fig. 9.1, where we plot the number of parameters included in the model against the resulting AIC value. We indicate models that correctly contain the first 4 covariates with crosses and the true model also with a diamond. It appears that models that include the true covariates clearly reduce the AIC, but the best model selected by the AIC contains 5 parameters. This means that, in this particular example, we falsely select one of the spurious variables, which by chance explains some of the residual error, but correctly include the four relevant variables in the model. This reflects a known property of the AIC: that it tends to select overly complex models, i.e. models with too many parameters.

▷

9.1.4 Extensions and Modifications

Bias-Corrected AIC

Akaike originally proposed (9.1.16) as an approximation of the expected Kullback–Leibler divergence (9.1.15). Despite the potential violation of this elegant interpretation, alternatives to the multiplicative factor 2 of the second term have nevertheless been proposed. The most prominent suggestion that still conforms to Akaike's framework is a bias-corrected version of the AIC, proposed by Hurvich and Tsai (1989)

$$AIC_C = -2l(\hat{\theta}) + 2p\left(\frac{n}{n-p-1}\right).$$

As n increases, the correction term $n/(n-p-1)$ converges to 1, but for small samples the bias-corrected version is generally advised. A rule of thumb given by Burnham and Anderson (2002) states that the corrected AIC is preferred if n/p is less than 40.

The Bayesian Information Criterion

Very similar to the AIC is the **Bayesian Information Criterion (BIC)**. To motivate this, let us consider model selection from a Bayesian perspective. Assume we have a set of models $M_1, \ldots, M_K$, each of which corresponds to a distributional model, such that

$$M_k \Leftrightarrow Y_i \sim f_k(y; \theta_k) \quad i.i.d.$$

Model selection can now be interpreted as a decision problem and a plausible strategy is to select the most likely model or, in Bayesian terminology, the model with the highest posterior probability. To this end, we need to calculate the posterior probability of each model. For model M_k, this is given by

$$P(M_k|y) = \frac{f(y|M_k)P(M_k)}{f(y)} = \frac{\int f_k(y;\vartheta_k) f_{\theta_k}(\vartheta_k) d\vartheta_k}{f(y)} P(M_k), \tag{9.1.24}$$

where $P(M_k)$ is the prior belief in model M_k. The denominator $f(y)$ is calculated by summing over all models

$$f(y) = \sum_k \int f_k(y;\vartheta_k) f_{\theta_k}(\vartheta_k) d\vartheta_k.$$

Although (9.1.24) allows us to quantify the posterior model probability, it is often complicated, or even infeasible, to calculate. As we already discussed in Chap. 5,

this model needs to be approximated, possibly with MCMC or other alternatives. However, this problem is even more complex here, as we need to approximate the integral in each of K models separately. We therefore pursue a simplifying approach and apply a Laplace approximation to the above integral using a slightly modified version of our original approach in Sect. 5.3.2. To begin, let

$$l_k(\theta_k) = \sum_{i=1}^{n} \log f_k(y_i; \theta_k)$$

be the log-likelihood for model M_k and define with $\hat{\theta}_k$ the corresponding Maximum Likelihood estimate. With the Laplace approximation, the integral component in (9.1.24) is given by

$$\int \exp\left(l_k(\vartheta_k)\right) f_{\theta_k}(\vartheta_k) d\vartheta_k \approx \left(\frac{2\pi}{n}\right)^{\frac{p_k}{2}} \exp\left(l_k(\hat{\theta}_k)\right) f_{\theta_k}(\hat{\theta}_k) \left|\frac{1}{n} I_k(\hat{\theta}_k)\right|^{-\frac{1}{2}},$$

where p_k is the dimension of θ_k. Taking the log of the right-hand side gives

$$\frac{p_k}{2}\log(2\pi) - \frac{p_k}{2}\log(n) + l_k(\hat{\theta}_k) - \frac{1}{2}\log\left|\frac{1}{n} I_k(\hat{\theta}_k)\right| + \log f_{\theta_k}(\hat{\theta}_k).$$

If we collect all the components that grow with order n, i.e. the second and third, and multiply them by -2, we obtain the **Bayesian Information Criterion** (BIC)

$$BIC_k = -2l_k(\hat{\theta}_k) + \log(n) p_k.$$

Maximising the posterior probability of a model therefore corresponds, at least approximately, to minimising the BIC.

Definition 9.2 The Bayesian Information Criterion (BIC) is defined as

$$BIC = -2l(\hat{\theta}) + \log(n) p.$$

The BIC clearly appears very similar to the AIC. The main difference is that the multiplication by 2 in the AIC is replaced by a $\log(n)$ in the BIC. As $\log(n) > 2$ for $n > 7$, for any reasonably sized sample, the BIC prefers models with fewer parameters, as compared to the AIC. It should also be noted that the BIC essentially shows the diminishing influence of the prior distribution with increasing n, which asymptotically goes to 0. This holds for both priors: the prior on the parameter $f_k(\theta_k)$ and the prior on the model $P(M_k)$.

Example 42 We continue with the previous example, but now calculate the BIC for each of the 1024 possible models, which can be seen in Fig. 9.2. The models containing the true parameter are shown as crosses and the true model as a diamond. The right-hand plot is an enlarged version of the left-hand plot, focusing on the area

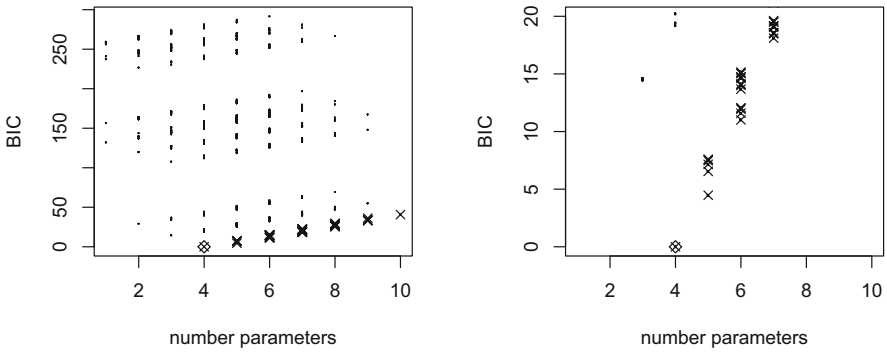

Fig. 9.2 BIC values for all 1024 possible models. Models including the true 4 covariates are indicated with crosses and the true model with a diamond. The right-hand plot is zoomed in on the true model

around the true model. Comparing the plot with Fig. 9.1, we see that the BIC tends to prefer simpler models with fewer parameters, with the true model having the lowest BIC score in this case.

▷

Deviance Information Criterion

The **Deviance Information Criterion** (DIC) was proposed by Spiegelhalter et al. (2002) and can be seen as a Bayesian version of the AIC. We do not give a detailed discussion here but only motivate its construction. The term deviance usually refers to how much the maximised likelihood for a particular model differs from that of the full model, which is sometimes called the saturated model. For this section, it is sufficient to define the deviance $D(y;\theta)$ as the log-likelihood multiplied by negative 2,

$$D(y;\theta) := -2l(\theta).$$

If θ_0 is the true parameter, then the difference in the deviance is

$$\begin{aligned}\Delta D(y;\theta_0,\hat{\theta}) &:= D(y;\theta_0) - D(y;\hat{\theta}) \\ &= 2\big\{l(\hat{\theta}) - l(\theta_0)\big\}.\end{aligned}$$

As $\hat{\theta}$ is the Maximum Likelihood estimate, we can clearly see that $\Delta D(y;\theta,\hat{\theta}) \geq 0$. In Sect. 4.4, we investigated the likelihood-ratio and derived the asymptotic distribution (4.4.3)

$$2\big\{l(\hat{\theta}) - l(\theta)\big\} \overset{a}{\sim} \mathcal{X}^2_p,$$

and in particular we get

$$E\left(2\{l(\hat{\theta}) - l(\theta)\}\right) \approx p$$

with p as the dimension of θ. Taking a Bayesian view, we can consider θ as random and $\hat{\theta}$ as fixed, conditional on the data. We continue with this line of reasoning and replace $\hat{\theta}$ with the posterior mean estimate $\hat{\theta}_{postmean}$, defined in (3.2.2). This leads us to the deviance difference

$$\Delta D(y; \theta, \hat{\theta}_{postmean}) = 2\left\{l(\hat{\theta}_{postmean}) - l(\theta)\right\}.$$

Spiegelhalter et al. (2002) use this quantity to derive a deviance-based approximation for the parameter dimension

$$p_D := E(\Delta D(y; \theta, \hat{\theta}_{postmean})|y) = \int \Delta D(y; \vartheta, \theta_{postmean}) f_\theta(\vartheta|y) d\vartheta$$

$$= \int D(y, \vartheta) f(\vartheta|y) d\vartheta - D(y, \hat{\theta}_{postmean}).$$

The first integral can be approximated using MCMC as in Chap. 5. The **Deviance Information Criterion** (DIC) is now defined by

$$DIC = D(y, \theta_{postmean}) + 2p_D = \int D(y, \vartheta) f_\theta(\vartheta|y) d\vartheta + p_D.$$

The DIC had a wide-ranging influence on model selection in Bayesian statistics, even though, when strictly following the Bayesian paradigm, the selection of a single model is of little interest. Instead, the calculation of its posterior probability is of central importance, which we will explore later in this chapter.

Cross Validation

Recall that the AIC aims to estimate the Kullback–Leibler divergence

$$E_{Y_1,\ldots,Y_n}(KL(g(.); f(,;\hat{\theta})) = E_{Y_1,\ldots,Y_n}\left\{E_Y\left(\log \frac{g(Y)}{f(Y|\hat{\theta}(Y_1,\ldots,Y_n))}\right)\right\}$$

as motivated in (9.1.16). If we now replace the Kullback–Leibler divergence with a squared distance, we can motivate the AIC as a theoretical counterpart to cross validation, as discussed in Sect. 8.5. Assume that we are interested in the mean of Y, which we denote with

$$\mu = E(Y) = \int y g(y) dy.$$

To predict the mean, we use the model $f(.;\theta)$, such that the prediction is given by

$$\hat{\mu} = \int y f(y;\hat{\theta}) dy,$$

where $\hat{\theta}$ depends on the data $y_1, \ldots, y_n$. The mean squared prediction error is then given by

$$E_Y\left((Y-\hat{\mu})^2\right) = \int \left\{y - \hat{\mu}(y_1,\ldots,y_n)\right\}^2 g(y) dy.$$

This replaces the Kullback–Leibler divergence in (9.1.14). We intend to select a model that minimises the mean squared error of prediction (MSEP), i.e.

$$E_{Y_1,\ldots,Y_n}\left\{E_Y\left[Y-\hat{\mu}(Y_1,\ldots,Y_n)\right]^2\right\} = \int \{\int \{y-\hat{\mu}(y_1,\ldots,y_n)\}^2 g(y) dy\} g(y_1,\ldots,y_n) dy_1 \ldots dy_n. \tag{9.1.25}$$

Note that (9.1.25) also shows a double integral, namely over the observations and over a new value Y. The major difference is that the Kullback–Leibler divergence is replaced with a squared distance.

9.2 AIC/BIC Model Averaging

So far we have used the AIC and some alternatives to select a single best model. Let us change our focus a little and look at multiple different models simultaneously, weighted by their relative suitability in explaining the data. In order to do so, we need to derive weights for the different models that represent their suitability or validity. The AIC values of different models provide one such measure. To begin with, let $M_1, \ldots, M_K$ be a set of models and define with

$$\Delta AIC_k = AIC_k - \min_k AIC_k,$$

where AIC_k is the Akaike Information Criterion calculated for model k and $\min_k AIC_k$ is the minimum of all AIC values. Clearly, if $\Delta AIC_k = 0$, model k is the best model (possibly one of many) and would be chosen by AIC model selection. Moreover, if ΔAIC_k is large, then model k appears to be unsuitable for modelling the data. Burnham and Anderson (2002) propose that models with $\Delta AIC < 2$ are eminently suitable for the data, while models with $\Delta AIC > 10$ are essentially unsuitable.

The value of ΔAIC can be transformed into a weight, with Akaike (1983) proposing the simple value

$$\exp(-\frac{1}{2}\Delta AIC_k).$$

Instead of weighting, it seems more natural to construct model probabilities out of the weights

$$P(M_k|y) := \frac{\exp(-\frac{1}{2}\Delta AIC_k)}{\sum_{k=1}^{K}\exp(-\frac{1}{2}\Delta AIC_k)},$$

which is, however, not a true Bayesian approach, because prior probabilities for the different models are not incorporated. That being said, calculating probabilities from weights is nevertheless intuitive and the approach is somewhat self-explanatory. Instead of using the AIC, one can also use the BIC by simply replacing $\Delta_k AIC$ with

$$\Delta_k BIC = BIC_k - \min_k(BIC_k),$$

with obvious definitions for BIC_k and $\min_k BIC_k$. This in turn gives the model probabilities

$$P(M_k|y) := \frac{\exp(-\frac{1}{2}\Delta BIC_k)}{\sum_{k'=1}^{K}\exp(-\frac{1}{2}\Delta BIC_{k'})}.$$

Example 43 Let us return once again to the original regression example and calculate the model probabilities $P(M_k|y)$ with both the AIC and the BIC. These can now be used to determine the probability that a particular covariate is included in the model. This probability is a measure of the importance of the variable, which we define as follows. Let $\mathcal{I}_1$ be the index set of all models that contain the first covariate. Note that with 10 potential covariates, we have $|\mathcal{I}_1| = 2^9 = 512$ models that include covariate x_1. We then define the probability that covariate 1 is in the model with

$$P(\text{covariate 1 in model}\,|y) = \sum_{k\in\mathcal{I}_1} P(M_k|y).$$

Similarly, we can calculate the probabilities for all other covariates. We show the resulting probabilities for the 10 covariates in Table 9.1, for both the AIC and the BIC. We can see that the first four covariates, which are the covariates of the true model, are always included. The remaining spurious variables have rather high posterior probabilities for the AIC, with a probability of approximately 15% that the

Table 9.1 Probability in % that the covariate is included in the model

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
AIC	100.0	100.0	100.0	100.0	13.3	12.0	14.6	15.6	16.2	14.0
BIC	100.0	100.0	100.0	100.0	0.1	0.1	0.1	0.3	0.1	0.1

variables are included in the model. For the BIC, however, they have practically no chance of inclusion.

▷

9.3 Inference After Model Selection

Inference for Maximum Likelihood estimation as derived in Chap. 4 was based on the assumption that the data are drawn from $Y \sim f(.;\theta)$ $i.i.d.$ That is, we considered the model as given. We showed at the beginning of this chapter that the Maximum Likelihood approach remains reasonable for the estimation of θ_0, returning the closest approximation of model $f(y;\theta)$ as measured by the Kullback–Leibler divergence. The inference and asymptotic behaviour of the variance of $\hat{\theta}$, however, had a more complicated structure. These properties hold if we just consider a single model $f(.;\theta)$. This begs the question: what happens if we first employ model selection to obtain a model before estimating our parameters. To be specific, let us assume we have K models

$$M_k \Leftrightarrow Y_i \sim f_k(;\theta_k) \quad i.i.d.,$$

where $k = 1,\ldots,K$, from which we select one model, e.g. the model that minimises the AIC. Let $\hat{k}$ be the corresponding model index, e.g.

$$\hat{k} = \arg\min_k \mathrm{AIC}_k.$$

The corresponding estimate is denoted with $\hat{\theta}_{\hat{k}}$, whose properties we now want to determine. In Chap. 4, we explored the behaviour of $\hat{\theta}_k$ for increasing n, but now the choice of model k is also informed by the data. Clearly, $\theta_1,\ldots,\theta_K$ may differ in size, i.e. the parameters in the different models $f_k(.;\theta_k)$ may be of a different dimensionality. For ease of explanation, let us limit ourselves to nested model classes. This means that we assume an overall model,

$$Y \sim f(.;\theta),$$

where $\theta_1,\ldots,\theta_K$ is given by θ with certain components set to 0. To make this more explicit, we need a slight change in notation and subsequently denote the parameter in the k-th model with $\theta_{(k)}$, that is, we set the model index in brackets. The probability model itself remains unchanged, but we simply set components of the parameter θ to zero, i.e. $f_k(y;) = f(y;\theta_{(k)})$. If $\theta = (\theta_1,\ldots,\theta_p)$ is a p-dimensional parameter and $\mathcal{I}_k$ is the index set of parameter components set to zero in the k-th model, i.e. $\mathcal{I}_k \subset \{1,\ldots,p\}$, then $\theta_{(k)}$ can be generated from θ by setting $\theta_j = 0$ for $j \in \mathcal{I}_k$. Assume now that we are interested in drawing inference about a single

parameter $\vartheta = \theta_j$ for some fixed $j \in \{1, \ldots, p\}$. Note that ϑ may be set to zero in some models and needs to be estimated in other models. Applying model selection as proposed above gives the estimate

$$\hat{\vartheta} = \hat{\theta}_{(\hat{k})j},$$

where $\hat{\theta}_{(\hat{k})}$ denotes the Maximum Likelihood estimate in the selected model and $\hat{\theta}_{(\hat{k})j}$ the corresponding estimate of component θ_j. This can be either zero if $j \in \mathcal{I}_{\hat{k}}$ or the estimate of the component if $j \notin \mathcal{I}_{\hat{k}}$. Hence, the estimate $\hat{\vartheta}$ can be fixed to zero but can also take values in $\mathbb{R}$, depending on the model that is selected. This property already shows that inference statements as derived in Chap. 4 do not directly apply and that we need deeper insight to understand how to derive inference after model selection.

Even though this is a relevant and very practical problem, it is surprisingly not often discussed in depth in statistics. Breiman (1992) called this the "quiet scandal", complaining that inference after model selection is not treated rigorously enough in statistics. With recent work for example by Leeb and Pötscher (2005), Berk et al. (2013) and Leeb et al. (2015), new results have been proposed, see also Claekens and Hjort (2008) or Symonds and Moussalli (2011) for a more comprehensive discussion of the topic. It appears that proper inference after model selection is methodologically demanding and rather complicated, which makes it difficult to use in an applied context. We will motivate these problems and make use of the model averaging approach that was explored in the last subsection.

Assume we select model k with probability π_k, where $\sum_k \pi_k = 1$. The probability may depend on the data, but let us for the moment assume that π_k is fixed and, somewhat unrealistically, known in advance. In practice, we would replace π_k with the model weights defined in the previous section, that is

$$\pi_k = \frac{\exp(-\frac{1}{2}\Delta AIC_k)}{\sum_{k'=1}^{K} \exp(-\frac{1}{2}\Delta AIC_{k'})},$$

which clearly depends upon the data. However, to keep the mathematics simple, we assume, as said, that π_k is known for all k. Let $\hat{\theta}_{(k)}$ be the estimate if we select model k, and define with $\hat{\vartheta}_k$ the component estimate in model k. That is, $\hat{\vartheta}_k$ equals 0 if the component is not in the model, i.e. if index $j \in \mathcal{I}_k$ where $\vartheta = \theta_j$. Given a model k, we saw in Sect. 9.1.1 that $\hat{\vartheta}_k$ converges to ϑ_k, where ϑ_k is the closest parameter to the true unknown model, as measured by the Kullback–Leibler divergence between the distributions. If the component is not in the model, then $\hat{\vartheta}_k$ is set to zero and clearly $\vartheta_k = 0$. Now let

$$\bar{\vartheta} = \sum_{k=1}^{K} \pi_k \vartheta_k$$

be the average of the parameter ϑ in the different models. This gives the expectation

$$E(\hat{\vartheta}) = \sum_{k=1}^{K} \pi_k E(\hat{\vartheta}|\text{model } k) = \sum_{k=1}^{K} \pi_k \vartheta_k = \bar{\vartheta},$$

where the conditional expectation $E(.|\text{model } k)$ denotes the expected value if we consider model k to be the true model. The variance is calculated using Property 2.3, that is

$$\begin{aligned} Var(\hat{\vartheta}) &= E_{\text{model}}(Var(\hat{\vartheta}|\text{model})) + Var_{\text{model}}(E(\hat{\vartheta}|\text{model})) \\ &= \sum_{k=1}^{K} \pi_k Var_k(\hat{\vartheta}_k) + \sum_{k=1}^{K} \pi_k (\vartheta_k - \bar{\vartheta})^2, \end{aligned} \tag{9.3.1}$$

where $Var_k(\hat{\vartheta}_k)$ denotes the variance if model k is considered true. Estimating the variance in (9.3.1) is somewhat clumsy. The ad hoc approach would be to replace the unknown quantities in (9.3.1) with their estimates. This is not a problem for the first component, i.e. we can replace $Var_k(\hat{\vartheta}_k)$ with its estimate $\widehat{Var_k}(\hat{\vartheta}_k)$. The variance is either zero, if ϑ_k is equal to zero in model k, or the inverse Fisher information of model k. To replace the second component in (9.3.1), we first need the averaged estimate

$$\hat{\bar{\vartheta}} = \sum_{k=1}^{K} \pi_k \hat{\vartheta}_k.$$

If we now replace the latter sum in (9.3.1) with

$$\sum_{k=1}^{K} \pi_k (\hat{\vartheta}_k - \hat{\bar{\vartheta}})^2, \tag{9.3.2}$$

we can see that the estimates are correlated, as they are all derived from the same data. That is, $Cor(\hat{\vartheta}_k, \hat{\vartheta}_{k'}) \neq 0$ and in fact the correlation will be high, possibly even close to 1. This induces a bias and precludes the use of (9.3.2) as an estimate for the second component in (9.3.1). Instead, Burnham and Anderson (2002) proposed the estimation of the variance (9.3.1) with

$$\widehat{Var}(\hat{\vartheta}) = \left\{ \sum_{k=1}^{K} \pi_k \sqrt{\widehat{Var}_k(\hat{\vartheta}_k) + (\hat{\vartheta}_k - \hat{\bar{\vartheta}})^2} \right\}^2.$$

Given their results, this variance estimate can also be used as an estimate for the variance of the averaged estimator $Var(\hat{\bar{\vartheta}})$. Recently, Kabaila et al. (2016) came to the conclusion that "it seems difficult to find model-averaged confidence

intervals that compete successfully with the standard confidence intervals based on the full model". In other words, it appears most useful to just fit the full model with no parameter component set to zero and take the variance estimate from this model to quantify the variability of $\hat{\vartheta}$ or $\hat{\tilde{\vartheta}}$. However, one may also follow the recommendation of Burnham and Anderson (2002) who state on Page 202 "Do not include (...) p-values when using the information theoretic approach as this inappropriately mixes different analysis paradigms". That is to say, one either uses model selection to get $\hat{\vartheta}$, which is either the Maximum Likelihood estimate if it is in the selected model or zero if not, or one tests whether $\vartheta \neq 0$, but never both. An alternative may be to follow a bootstrap strategy and run both the model selection and the estimation to obtain bootstrap confidence intervals, which is certainly recommended if sufficient computing resources are available. The bootstrap approach also allows us to accommodate the uncertainty from the estimation of the model probabilities π_k using AIC weights, whose derivation would be even more complicated. In other words, inference after model selection is complex and clumsy, and if a simple and practical approach is desired, then bootstrapping appears as suitable method, at the cost of heavy computation.

9.4 Model Selection with Lasso

A conceptually attractive alternative to model selection is the Least Absolute Shrinkage and Selection Operator (Lasso) approach. The Lasso allows us to pursue both model selection and estimation in a single step. The method was proposed by Tibshirani (1996). Here we provide an overview using a penalised likelihood approach. A comprehensive discussion of the method is provided in Hastie et al. (2015). Let $l(\theta)$ be the likelihood of a statistical model, and let $\theta = (\theta_1, \ldots, \theta_p)$ be the parameter. We assume that p is large and our aim is, as in the previous section, to select a model by setting numerous components of θ to zero. Let $\mathcal{I}$ be the index of components of θ set to zero through model selection, where $\mathcal{I} \subset \{1, \ldots, p\}$. The Lasso approach aims to find the maximum of the extended, or penalised, log-likelihood

$$l_p(\theta, \lambda) = l(\theta) - \lambda \sum_{j \in \mathcal{I}} |\theta_j| \tag{9.4.1}$$

with λ as the tuning parameter. Clearly, setting $\lambda = 0$ gives the normal likelihood, while setting $\lambda \to \infty$ implies that $\theta_j \equiv 0$ for all $j \in \mathcal{I}$. Hence, the term λ plays the role of a model selection parameter.

The penalised likelihood (9.4.1) can be solved with iterative quadratic programming. To do so, let $\hat{\theta}_{(0)}$ be a starting value, for instance, the parameter estimate with $\theta_j = 0$ for $j \in \mathcal{I}$, and set $\theta_{(t)} = \hat{\theta}_{(0)}$. We approximate the log-likelihood with

$$l(\theta) \approx l(\theta_{(t)}) + s(\theta_{(t)})(\theta_{(t)} - \theta) - \frac{1}{2}(\theta_{(t)} - \theta) I(\theta_{(t)})(\theta_{(t)} - \theta) =: Q(\theta_{(t)}, \theta).$$

The right-hand side is a quadratic function in θ. This allows us to approximate the Lasso likelihood (9.4.1) with

$$Q(\theta_{(t)}, \theta) - \lambda \sum_{j \in \mathcal{I}} |\theta_j| \to \max .$$

It can be shown that the above maximisation problem is equivalent to

$$Q(\theta_{(t)}, \theta) \to \max \text{ subject to } \sum_{j \in \mathcal{I}} |\theta_j| \le c \tag{9.4.2}$$

for some c, which is clearly related to our penalty term λ. Equation (9.4.2) gives a quadratic optimisation problem, which leads us to a new estimate $\theta_{(t)}$. We set $\theta_{(t+1)} = \hat{\theta}_{(t)}$ and this new value is in turn used to derive $Q(\theta_{(t+1)}, \theta)$, which gives the next iteration.

The Lasso approach will result in an estimate where numerous parameter values of the index set $\mathcal{I}$ are zero. This can be nicely visualised, as shown in Fig. 9.3. We show the log-likelihood for a two-parameter model with $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ as Maximum Likelihood estimate, i.e. when λ is (9.4.1) is set to zero. The grey shaded area shows the parameter values with $|\theta_1| + |\theta_2| \le 3.7$, i.e. $c = 3.7$ in (9.4.1). Clearly, the constrained likelihood in (9.4.2) is maximised if θ_1 is set to zero.

The Lasso penalisation can also be taken from a Bayesian perspective. In this case, we interpret the penalty as the logarithm of a prior distribution over the coefficients θ_j, $j \in \mathcal{I}$. That is, we assume

$$f_\theta(\theta_j,\, j \in \mathcal{I}) \propto \exp(-\sum_{j \in \mathcal{I}} |\theta_j|)$$

or, due to its factorisation,

$$f_{\theta_j}(\theta_j) \propto \exp(-|\theta_j|) \quad i.i.d. \text{ for } j \in \mathcal{I}.$$

The resulting distribution is a Laplace prior (see Park and Casella 2008), i.e.

$$(\theta_j;\, j \in \mathcal{I}) \sim \prod_{j \in \mathcal{I}} \frac{1}{\sigma} \exp\left(-\frac{|\theta_j|}{\sigma}\right),$$

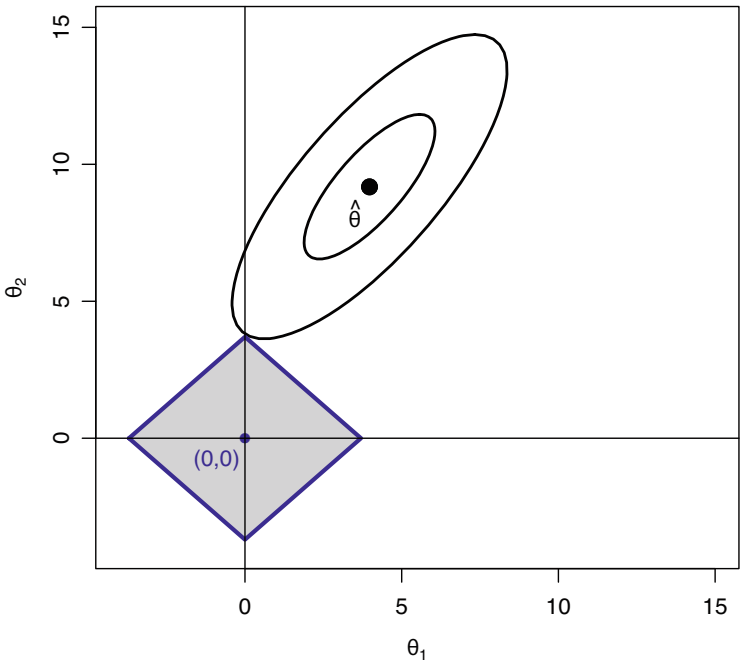

Fig. 9.3 Representation of Lasso penalty

where σ is a scaling parameter. In this case, σ plays the role of a hyperparameter, which needs to be set by the user or estimated from the data as per empirical Bayes.

Variance estimates for the fitted parameters after Lasso estimation have been derived by Lockhart et al. (2014). As motivated in the previous section, inference after model selection is clumsy, so we refer to the cited article for details, see also Hastie et al. (2015).

Example 44 Let us continue with the data from the previous examples and apply the Lasso estimation approach. The likelihood in this case results in a normal distribution, such that the Lasso estimate minimises

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \to \min_{\beta}.$$

The resulting estimates for different values of λ are shown in Fig. 9.4. The left-hand plot shows the estimates β_j plotted against $\log \lambda$. It is more convenient to show the estimates plotted against $\sum_{j=1}^{p} |\hat{\beta}_j|$, which is shown in the right-hand plot. For clarification we include the true values of the parameters as horizontal grey lines, which are 0.25 for the β_1 and β_2 and 0.1 for β_3 and β_4. We see that the Lasso

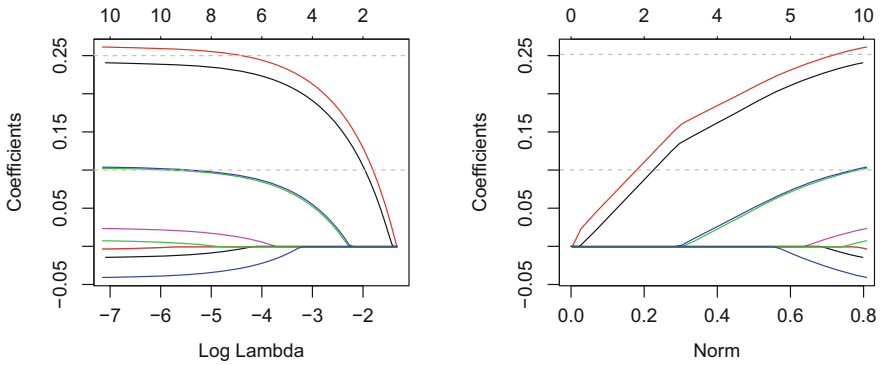

Fig. 9.4 Lasso estimates for different values of λ (left-hand side) and the norm $\sum_{j=1}^{p} |\hat{\beta}_j|$ (right-hand side). The true parameter values are shown as dashed horizontal lines

approach picks the correct values of β and sets the other coefficients to zero as long as $\lambda > \exp(-3)$.

▷

9.5 The Bayesian Model Selection

We already briefly addressed the subject of Bayesian model selection in Sect. 6.7, where we introduced the Bayes factor. We will extend this here a little and sketch the fundamentals of how the Bayesian paradigm can allow us to validate and select a model. We refer, for example, to Ando (2010) for a deeper discussion of the field. Let $M_1, \ldots, M_K$ be a set of K candidate models, each parameterised by θ_k. For each model, we define with $f_{\theta_k}(.)$ the prior distribution of the parameter, such that

$$f(y|M_k) = \int f_k(y; \theta_k) f_{\theta_k}(\theta_k) d\theta_k$$

is the distribution of the data in the k-th model, where $f_k(.)$ denotes the distribution if the k-th model holds. We further include the model prior $P(M_k)$ for $k = 1, \ldots, M$, where we may allow simpler models to have a higher prior probability. This, by extension, is equivalent to the penalisation of complex models. The posterior model probability is then

$$P(M_k|y) = \frac{f(y|M_k) P(M_k)}{f(y)},$$

where $f(y) = \sum_{k=1}^{K} f(y|M_k)P(M_k)$. Model selection is now simply choosing the model with the largest posterior probability. We may also use the posterior model probability to pursue model averaging, taking $P(M_k|y)$ as the weight. This setting can be further extended by sampling directly from the space of all possible models. Generally, the Bayesian view is quite suitable for model selection, and in fact the idea of model averaging discussed in Sect. 9.2 is very much in this vein. We do not go deeper into the field here but refer to Berger and Pericchi (2001), Cui and George (2008) or Hoeting et al. (1999) for further reading.

9.6 Exercises

Exercise 1 (Use R Statistical Software)
Let us once again consider Example 41. We generate 10 independent covariates $x_{ij} \sim N(0, 1)$, where $i = 1, \ldots, n$ and $j = 1, \ldots, 10$, and simulate

$$Y_i \sim N(10 + 0.25x_{i1} + 0.25x_{i2} + 0.1x_{i3} + 0.1x_{i4}, 1) \ .$$

Note that we include an additional intercept term $\beta_0 = 10$. Therefore, the response variable depends only on the first 4 covariates with true coefficients $\beta = (10, 0.25, 0.25, 0.1, 0.1)$, while the remaining 6 are spurious. Perform the following with different sample sizes $n = 50, 100, 150, 200, 500, 1000, 5000$ and with $S = 500$ repetitions.

1. For every of the $S = 500$ experiments, find the best model (of all possible models) using the AIC criterion. Calculate the percentage of experiments in which the selected model (1) contains exactly four correct predictors and (2) contains at least four correct predictors. Calculate (using the best model) the average value of all 10 coefficients over the $S = 500$ experiments.
2. Instead of selecting the best model, use the function `stepAIC` of the `R` package `MASS` to select a final model and repeat the experiments again $S = 500$ times. Evaluate the coverage probabilities, i.e. the percentage of cases where the confidence interval contains the true coefficient (coefficient-wise, assume a value of zero and a variance of zero for a coefficient not included in the final model). Comment on your results.
3. Evaluate how the results change, when you increase or decrease the absolute size of the coefficients or the number of coefficients which are different from zero. As an example, consider $\beta = (10, 5, 5, 1, 1)$ and $\beta = (10, 0.05, 0.05, 0.01, 0.01)$ and $\beta = (10, 1, 1, 1, 1, 1, 0.1, 0.1, 0.1, 0.1, 0.1)$.