1

Why Prediction?

> … any model … is merely a human attempt to describe or explain reality … models are to be assessed in terms of their success at this task. It is misguided … to believe in Nature as obeying some theory … Even if we can find a completely successful theory, this does not mean we have identified Nature's true model – some other, distinct theory might be just as successful … In this view, theories can only be distinguished by means of their predictions about observables …
>
> Dawid (1992)

For centuries, perhaps millennia, people have tried all sorts of divination methods, from yarrow sticks to tarot cards, from the innards of animals to the positions of planets. With sufficiently skilled interpreters, these methods probably work a little: a tarot card reader may use the cards to evoke the frame of mind of the subject. To the extent that the future is implicit in a subject's frame of mind, the predictions may therefore be accurate. After all, if you know some one, even a little, you can predict some of their behavior. This is second nature for good salespeople, politicians, and others whose career success depends on detecting people's preferences. Arguably, this sort of procedure might even help with economic predictions that include market psychology. Note, however, that divination methods are rarely used to predict such outcomes as how much product a given chemical reaction will produce, or other outcomes that have essentially no element of human choice.

Here, by contrast, the goal is to make predictions by rules in such a way that evaluating how well the rules work will be unambiguous. The fortunate case occurs when the rules accurately reflect something about the mechanism used by the data generator (DG) to generate outcomes. This is the main goal of much of conventional science. However, there are vast classes of data where it is implausible to model the DG. As a slightly facetious example, one can treat *MacBeth* as a sequence of letters and try to predict the $(n+1)$th letter using the first n letters. A variant on this is predicting the $(n+1)$th nucleotide – or any finite sequence of nucleotides – on a chromosome, given the first n nucleotides in the chromosome. In both cases, the DG is so complex that detailed modeling for the purposes of prediction would be premature, to say the least. Indeed, if we want to make predictions, it's unlikely modeling will help much.

Worse, many DGs might not function by rules at all. The easiest way to think of this is that the outcome y_1 at a given time step is from one distribution, say Q_1, but the outcome y_2 from the next time step is from another distribution, Q_2, chosen by some agent who may not even know Q_1 or y_1 and chooses Q_2 using a hidden mechanism or no mechanism at all.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

Doing this repeatedly means there is nothing stable enough to model, so we cannot, even conceptually, use models to generate prediction rules. Another way to think of this is to ask whether more data can be generated – at least in principle – that would be informationally equivalent to the data we already have. This is not the same as asking whether an experiment is repeatable in practice – many aren't, as for example in econometrics – it only asks whether in principle we could generate further data sets of the same general form. Clearly, if the answer is no (think of *MacBeth*) then there may be no rule that the DG follows and hence it will be impossible to formulate a prediction rule that matches the DG. However, even when we admit that the DG does not follow rules, we may still want a well-defined prediction rule so that we can evaluate how good it is. Crazy as it sounds, this is not entirely impossible, as we shall see.

The stance of this book, stated concisely and unabashedly, is that predictive statistics proposes an alternative formulation of the paradigm statistical problem. The central feature of predictive statistics, as opposed to other schools of thought in statistics, is to use the data to predict ahead rather than try to find out what underlies the data generator, then try to model it, and finally use the model to make predictions. In either case, predictive or not, predictions must be compared with new data for validation. The question is when this comparison is made – is it before or after 'modelling' has been attempted? In predictive statistics, modeling does not start until good prediction has been achieved. This is the reverse of the conventional approach.

One of the key arguments for predictive statistics at this time of writing, and for the foreseeable future, is that so much of the statistical world has changed. Volumes of data have massively increased, preprocessing techniques for raw data (e.g., in the 'omics world) have increased and become more diverse, multitype data is more prevalent than before (and often very hard to model), and the complexity of data streams that confront the Statistician is nearly overwhelming. Together, these features make modeling difficult, if not infeasible. Indeed, often only a small fraction of the available data can be used in an analysis. Taking a predictive approach, and hence achieving good prediction, is likely to be better in the long run than direct modeling for understanding a data generator. Even where modeling is infeasible, good prediction is the sine qua non of a good theory. The reason is that modeling requires dealing with model uncertainty and misspecification and these can be extraordinarily difficult.

1.1 Motivating the Predictive Stance

It may seem strange to ask, but it's important to answer the question 'Why is prediction so important?' First, one obvious answer is that sometimes the goal really is to know what the next outcome is likely to be: prediction may be the goal of the statistical analysis. For instance, it might be helpful to predict who will get post-traumatic stress disorder (PTSD). That way, to prevent PTSD, or minimize its effects, a physician would want to know who it is most important to treat prophylactically. Sometimes the goal is prediction even when it's not phrased predictively. For instance, one may estimate a probability of recurrence of cancer (with a standard error), but it would be more informative to give a point predictor for when a patient will get a recurrence along with an assessment of the variability of that prediction. Aside from being the information that a patient or physician wants, a prediction

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

interval is less abstract and more intelligible than a probability, let alone a confidence or credible interval for a probability. Of course, people might want to predict the weather, the economy, the response to a new treatment, and so forth. Who hasn't wanted to know the future for some purpose, base or laudable?

A second answer (that is less obvious) is that most other goals of statistical analysis can be subsumed within prediction. What, after all, are the main goals of statistical analyses? Any list would have to include (1) model identification, (2) decision making, and (3) answering a question about a population – even if there is some overlap among these goals as stated. For instance, identifying a model may amount to making a decision: in classification, model identification amounts to identifying a classifier, and using a classifier amounts to deciding the rule by which one will assign a future subject to a class. In general, it's hard to find a statistical problem that doesn't have a direct connection with prediction.

Let's start with *model identification* – which is essentially estimation in one guise or another. This includes, among other possibilities, parametric estimation, classification, regression, nonparametric estimation, and model selection. It also includes some hypothesis testing. A simple versus simple test such as $\mathcal{H}_0: P = P_1$ versus $\mathcal{H}_1: P = P_2$, where P_1 and P_2 are the only two candidate probabilities for P, is an obvious example. In general, one can use a series of goodness-of-fit tests to determine models that can't be rejected. Moreover, tests such as whether a given variable should be in a regression function must also be included as part of model identification. Even though such a test does not by itself identify a model, it constitutes an effort to reduce the class of models that must be considered and is therefore a step toward model identification.

In all these cases, how can one know in reality that a model has been successfully identified without using it to generate accurate predictions? Even more, how can one know that another model with an equally good fit, possibly using different variables, can be ignored if it is not predictively discredited – for instance, on the grounds of high bias or high variance? Loosely, outside very simple problems, model identification without predictive verification is little more than conjecture. Put otherwise, whenever a model is selected, or a parameter estimated, a predictor is formed and, if it doesn't perform well, the model it came from is discredited.

Essentially this means that the search for a good model is merely a special case of the search for a good predictor. The substitution is thought worthwhile because a scientist can bring 'modeling information' into the search for a predictor. The problem is that modeling information is usually not itself predictively verified and hence is often of dubious value. Thus, taking a purely predictive view and treating modeling information as likely to be unreliable guards against the use of such suspect 'information'.

Let us now turn this around. Just as a credible model, when it exists, can often be used to generate predictions, a predictor can sometimes be used to identify a model. In the simplest case, it is assumed that there is a parametric family $\mathcal{P} = \{p(\cdot|\theta)|\ \theta \in \Omega\}$, where $\Omega \subset \mathbf{R}^d$ for some integer $d \geq 0$, equipped with a prior on Ω having a density with respect to μ, say. Then, the predictive distribution for a random variable Y_{n+1} with outcomes y_{n+1} given $Y^n = (Y_1, \ldots, Y_n) = (y_1, \ldots, y_n) = y^n$ is

$$m(y_{n+1}|y^n) = \int p(y_{n+1}|\theta) w(\theta|y^n) \mu(\mathrm{d}\theta),$$

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

where $w(\theta|y^n)$ is the posterior density. It is easy to verify that $m(y_{n+1}|y^n)$ is optimal in a relative entropy sense (see Clarke *et al.* (2014)) and that when $Y_i \sim P_{\theta_0}$, with density p_0, is independent and identically distributed (IID) for all i that

$$m(y_{n+1}|y^n) \rightarrow p_{\theta_0} \tag{1.1}$$

pointwise in y_{n+1} in distribution, as $n \rightarrow \infty$. So, one could fix a distance d on densities for Y_{n+1} and choose a model based on $\theta^* = \theta^*(y^n)$ satisfying

$$p_{\theta^*}(\cdot) = \arg\min_{\theta} d(m(\cdot|y^n), p_\theta(\cdot)).$$

Not all predictors can be so obviously converted into models, just some of the good ones. Moreover, (1.1) only holds under highly restricted conditions.

Decision making can also be subsumed into prediction. Suppose that there is a prior, a parametric family, data, and a loss function and that the task is to seek a decision rule minimizing the Bayes risk. In these cases, the challenge is to verify that the decision rule gives a good performance; at root this depends on whether some element of the parametric family matches the DG. For instance, if the decision regards which parameter value to choose then there is a model that can be used for prediction. If the decision regards which stock to buy on a given day – i.e., an action – then the gain or loss afterward gives an assessment of how good the decision is; this evaluation is disjoint from the procedure that generated the decision or action in the first place. Another common decision problem is to decide which treatment is best for a given patient or for a given patient population. Again, one is selecting an action. One can make a decision based on data, but it is only when the predictions following from that decision are tested that one can be sure the decision was the best possibile. That is, the goal in decision making is, at root, to predict the action that will be most advantageous in some sense. Otherwise stated, the merit in a given decision is determined by how well it performs, and taking empirical performance into account (since that's what's important) makes decision making merely a way to choose a predictor. Therefore, essentially, a decision problem is a one-stage prediction problem, i.e., there is one prediction, not a sequence of predictions, so there is no chance to reformulate the predictor.

As before, in some cases decision making procedures can be turned into predictors. Indeed, the decision problem may be to find a good predictor. However, even when the decision problem is not directly about prediction, it has a predictive angle. For instance, consider the frequentist test of $\mathcal{H}_0\colon F \neq G$ versus $\mathcal{H}_1\colon F = G$ using IID data y^n from F and z^n from G. This is sometimes called an equivalence test. The two-sample Kolmogorov–Smirnovtest would be one of the natural test statistics if the hypotheses were reversed. To address the test, choose a distance d, write $\mathcal{H}_0$ as $\mathcal{H}_0\colon d(F,G) > \delta$, and let $\{(F,G)|\, d(F,G) > \delta\} = \cup_k S_k$, where the S_k are sets of pairs of distribution functions, $k = 1, \ldots, K$, and the diameter of S_k is small in terms of d. Then, $\mathcal{H}_0$ vs. $\mathcal{H}_1$ is equivalent to the K tests $\mathcal{H}_{0,k} : S_k$ vs. $\mathcal{H}_1$. If the S_k are small enough, they can be approximated by their midpoints, say s_k. This gives the approximate simple-versus-simple when testing the problems $\mathcal{H}'_{0,k}\colon (F,G) = s_k$ vs. $\mathcal{H}_1\colon F = G$. Now, in principle these tests can be done, using a multiple comparisons correction, and a single approximate model (or a small collection of approximate models) can be given from which to make predictions. Note that this is not modeling, and in fact nonparametric approaches can be used in a decision problem to generate predictions. (This approach will arise in Sec. 1.2.2.)

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

Answering a question about a *population*, or, more generally, understanding the structure of a data set, is a more nebulous goal. However, it may be regarded as trying to identify some feature of a population, or of an individual within a population, that is not obviously expressible in model identification terms. As an example, imagine trying to identify which dietary supplements the residents of a city buy or determining whether two random variables are associated. The predictive angle in these cases is one of confirmation. De facto, the prediction is that residents of the city use dietary supplements from a given list. So, if a resident of the city is chosen, does the resident use one of the identified supplements or not? If predictions are made assuming the independence of two random variables, are these predictions noticeably worse than including a dependence between them? Equally important, it is relatively rare that the final end point of an analysis is the description of a data set or the answering of a question about a population. Usually, one is doing this sort of task with a greater goal in mind, such as deciding whether to offer a new supplement for sale or classifying subjects into two classes on the basis of covariates.

Since this class of problems is less well defined, it is not obvious how to convert methods from it generically into a predictive interpretation beyond what has already been discussed. It is enough to be aware of the centrality of prediction among the various statistical goals subsumed under the term population description.

For the sake of completeness, recall that there are other statistical goals such as data presentation (graphics), data summarization, and the design of experiments. These too are generally in the service of some greater goal. Data presentation may be used to explain a statistical finding to non-statisticians, but these people generally have a reason why they want the analysis and a goal that they want fulfilled, which is usually predictive. Similarly, data summarization is rarely an end in itself but a subsidiary goal towards some other presumably greater goal. The design of experiments is done before data is collected, and its primary goal is to ensure that the data collected will suffice for the analytic goal – which, as has been argued, generally has a predictive perspective even if prediction is not recognized as the main explicit goal.

A third benefit of focusing on prediction is ensuring that inferences are testable and hence that any theories they represent are testable. Testability is not the same as interpretability, but a good predictor will typically permit some, perhaps limited, interpretation. For instance, given a predictor that uses explanatory variables one can often determine which of the explanatory variables are most important for good prediction. One would expect these to be the most important for modeling as well. More generally, apart from interpretability, theories for physical phenomena that arise from estimating a model and using hypothesis tests to simplify it must be validated predictively.

It is worth noting that, heuristically, there is almost an ‘uncertainty principle’ between interpretability and predictive accuracy: it’s as if the more interpretability one wants, the more predictive performance one must sacrifice, and conversely. After all, the best predictors are often uninterpretable (e.g., those for the Tour and Fires data in Sec. 1.2.1, the Bacterial NGS data in Sec. 1.2.2, and the Music data in Sec. 1.2.3). Moreover, interpretable predictors (typically based on models) are almost always predictively suboptimal: it is a mathematical fact that, for instance, Bayes model averaging[1] (which is difficult to interpret) is better than

[1] Here and elsewhere Bayesian is abbreviated to Bayes for brevity when convenient.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

using any one of the models in the average (which is usually easy to interpret), at least under squared error. Also, the adaptivity of predictors to data which has little interpretability often outperforms conventional model averages or model selection; see Wong and Clarke (2004), Clarke *et al.* (2013). Thus, interpretability does not lead to good prediction and good prediction does not require interpretability – although sometimes interpretations can be derived from predictors. Indeed, relevance vector machines (RVMs) are mathematically the best predictors in some settings (reproducing kernel Hilbert spaces), but statistically they overfit and can therefore be suboptimal because of excessive variance, meaning some terms have to be dropped for improved predictive error. This does not make RVMs more interpretable – if anything it makes them more complex and hence less interpretable – but it can make them excellent predictors.

Importantly, prediction in and of itself does not require an unseen world of abstract population quantities or measure spaces. Predictors such as ‘someone with high coronary artery calcium is likely to benefit from statin treatment’, paraphrased from Blaha *et al.* (2011), do not require anything we have not measured or cannot measure. Similarly, ‘tomorrow’s weather will be the same as today’s’ is a purely empirical statement. We may wish to invoke the mathematical rigor of measure theory to provide a theoretical evaluation of our prediction methods under various assumptions but this is a separate task from prediction per se. Indeed, in many cases the asymptotic properties of predictors, in terms of sample size or other indices, are of interest but cannot be obtained without making assumptions that bear scant relation to reality. For instance, formally a random variable is a deterministic function on an invisible and unspecified set. Is this a reasonable way to encapsulate the concept of randomness mathematically? The answer is probably no; it’s just that a better one has yet to be proposed and accepted.

A fourth reason to focus on prediction is that predictive errors automatically include the effect of uncertainty due to the data and to all the choices used for prediction. That is, when a predictor $\hat{Y}$ of Y is wrong by $|\hat{y} - y|$, the error includes not just the bias and variability of any parameters that had to be estimated to form $\hat{Y}$ but also the bias and variability due to the predictor class (or model class if models are used) itself as well as the variability in the data. This is a blessing and a curse. One of the problems with prediction is that point predictors are more variable than point estimators, so prediction intervals (PIs) are typically wider than confidence or credibility intervals (CIs). Moreover, just like CIs, model-based PIs tend to enlarge when model uncertainty is taken into account. The consequence of this is that predictive inferences tend to be weaker than parametric or other inferences about model classes. It would be natural for investigators to prefer stronger statements – even if the justification for them rests heavily on ignoring model uncertainty. However, even though inferentially weaker, point predictors and PIs have the benefit of direct testability and accurate reflection of uncertainty, which point estimators and CIs usually lack.

One of the earliest explorations of model uncertainty was by Draper (1995), who compared two ways of accounting for model uncertainty in post-model selection inference that include prediction. Draper (1995) argued that model enlargement – basically adding an extra level to a Bayesian hierarchical model – is a better solution than trying to account for the variability of model selection from criteria such as the Akaike or Bayes information criteria in terms of the sample space. He also argued that it is better to tolerate larger prediction intervals than to model uncertainty incorrectly. (As a curious note, Draper (1997) found that

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

there are cases where correctly accounting for modeling uncertainty actually reduces predictive uncertainty.) Of course, if PIs are too large to be useful then the arguments that a modeling approach is valid are more difficult to make, and any other inferences – estimates, hypothesis tests – may be called into question. However, to quote Draper (1995): 'Which is worse – widening the bands now or missing the truth later?'

There are *two criticisms* of the predictive approach that must be answered and dispensed with. First, a criticism of the predictive approach that is used to justify direct modeling approaches is that being able to predict well does not imply that the phenomenon in question is understood. The answer to this criticism is that modeling only implies understanding when the model has been extensively validated, i.e., found to be true, and this validation is primarily predictive. So, announcing a model before doing extensive validation – as is typically done – provides only the illusion of understanding. Prediction is a step toward model building, not the reverse, and predictive evaluation is therefore more honest. While the result of this kind of validation may be a predictor that is not interpretable, it is better than having an interpretable model with poor predictive performance. It may be that the traditional concept of modeling is too restrictive to be useful, especially in complex problems.

Second, it must be admitted that in practice the predictive approach is frequently harder than modeling. It's usually easier to find a not-implausible model (based on 'modeling assumptions' that boil down to the hope that they are not too far wrong), estimate a few parameters, verify that the fit is not too bad and then use the model to make statements about the population as a whole than it is to find a model that is not just plausible but actually close enough to being correct to give good predictions for new individual members of the population. Here, 'close enough' means that the errors from model misspecification or model uncertainty are small enough, compared with those from other sources of error, that they can be ignored. The problem, however, is that there are so many plausible models that finite data sets often cannot discriminate effectively amongst them. That is, as a generality, the plausibility of a model is insufficient for good prediction because one is quite likely to have found an incorrect model that the data have not been able to discredit yet. Since models that do not give sufficiently good prediction have to be disqualified, their suitability for other inferential goals must be justified by some argument other than goodness of fit. Thus, on the one hand, the task of finding a good predictor is usually harder than the task of finding a plausible model.

On the other hand, in reality a predictive approach is easier than implementing a true, accurate, modeling approach. Truly implementing a modeling approach requires that the model be correct or at least indistinguishable from correct. Given that the true model (when it exists) is rarely knowable this is an extremely difficult task. However, finding a serviceable predictor is easier, because it asks for less: giving good predictions is an easier task than uncovering a true model because bad predictions from a model invalidate the model while failure to provide good modeling inferences does not per se invalidate a good predictor. For example, if one predicts tomorrow's weather to be the same as today's weather this predictions may be reasonably accurate even though there is no underlying model from which to make inferences. Indeed, a good predictor may correspond to a dramatic simplification of a true model such that the prediction is good but the specific modeling inferences are poor.

Taking a predictive approach also requires another shift of perspective, namely that the data to be collected in the future are extremely important. This flies in the face of modeling

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

which focuses on a specific data set and what it says about a hypothetical population rather than what it says about future outcomes. It also flies in the face of standard scientific practice, which underweights confirmatory studies. As a gedanken (thought) experiment, imagine how scientific practice, funding decisions, and scientific publishing would change if the confirmation of studies (by different experimental teams) were weighted as highly as initial findings. It's not that prediction is against rapid scientific advancement; rather, it's that prediction is a check that the advancement is real (not based on errors, luck, or malfeasance) so that scarce resources don't get squandered on spurious results.

Despite the considerations so far, which are fairly well known, the main approach taken by statisticians has been to look at *model classes* and use them to generate predictors in cases where prediction was an acknowledged goal. Here, however, the key point is that much of traditional statistics has been done precisely backward: instead of modeling, or more generally choosing a model, and then predicting, one should propose a predictor class and find a member that performs well. Then, if model identification is desirable for some reason, in principle one can convert the predictor to a model within a class of models that are believed to be plausible. For instance, in some settings Bayes model averages yield good predictors. One can form a single model from a Bayes-model-average predictor by looking at the most important terms in the models that go into the average. In the special case of averaging linear models, one can regard this as a way to find coefficients on variables using a criterion different from least squares. (The difference is that Bayes model averaging combines models after determining coefficients rather than determining coefficients for a combined model.) As another example, one can use a kernelized method such as an RVM, take a Taylor expansion of the kernel in each term of the RVM, and take the leading terms as a model. In this way one might obtain a model that is interpretable and gives good predictions. If the predictions are not quite as good as those from the original predictor, at least one can see the cost paid to obtain interpretability.

Forthrightly, the point of this book is that the *paradigm problem of statistics is prediction*, not estimation or other forms of inference, and problems should be formulated in such a way that they can be answered by producing a predictor and examining its properties. The tendency that analysts and methodologists have toward model formulation and therefrom to estimation, decision making, and so forth is misguided and leads to unreproducible results. This often happens because model uncertainty is usually the biggest source of error in analyses, especially with complex data. Predictor uncertainty may be just as big a problem, but it is visible in the predictive error while model uncertainty is very hard to represent accurately.

Conventional statistical modeling recognizes the problem of model uncertainty in a variety of ways. Most recently, model uncertainty has been recognized in the desire for sparse models that satisfy the oracle property (they can correctly select the nonzero coefficients with high probability; see Sec. 10.5). Clever as all this is, it is merely a way to find a model that is not implausible, i.e., cannot obviously be ruled out. Indeed, shrinkage methods generally perform better predictively when they shrink less, i.e., are less sparse and distort the data less. Moreover, as they shrink less, shrinkage methods tend to improve and become competitive with the better model-averaging methods; see Clarke and Severinski (2010). As a generality, shrinkage methods frequently are a source of model misspecification since sparse models are rarely true outside narrow contexts. Indeed, the desire for sparsity is a variant on the desire for models (and small variance), since models are a way to summarize

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

one's purported physical understanding and sparsity is a pragmatic adaptation to settings where a purported physical understanding is an unusually remote possibility.

Indeed, many people say things like 'if the model is too complex to hold in my head there's no way I can work with it'. Leaving aside the dubious utility of modeling as a way to get sparsity (rather than just bias) and the even more dubious notion that reality should be simple enough to hold in one's head, one might consider sparsity as a desideratum for predictors. There is some merit in this – but only because good predictors have a good variance–bias tradeoff. Thus, sparsity may help reduce overall error if it reduces the variance enough to overcome any increase in bias but in fact the bias is likely to increase and, moreover, sparsity becomes less desirable as sample size increases.

However, one must recognize that sparsity is rare in reality, so there is negligible merit in seeking it on esthetic grounds or insisting as doctrine that a model or predictor be sparse. Otherwise stated, most problems in the real world are complex, and if a sparse predictor or model is not so far wrong as to be discredited it is likely that gathering more data will discredit it. Aside from being an argument against standard modeling (which is generally a severe oversimplification of a physical problem), this is an argument in favor of prediction because predictors usually include terms that are useful regardless of their meaning. In short, one wants sparsity as a way to control the variance part of a variance–bias analysis. Beyond this, sparsity is of little use to prediction since bias is bad for prediction and, in the case of a model, is an indicator that the model is wrong and hence inferences from it are called into question. One must accept that the real world has little inclination to conform itself to a short, or otherwise sparse, list of quantities that we can imagine and express conveniently, model-based or not. Seeking sparsity because it pleases us or gives us the illusion of 'understanding' has a cost in terms of bias and is all too often merely a pitfall to be avoided.

The net effect of all this is that the stance of this book is predictive. That is, the fundamental task is taken to be prediction, and problems should be formulated in such a way that a predictive approach can resolve them. This book explores a large number of ways to do this and tries to elucidate some of the main properties of predictor classes, including when to use each. This means that estimators, tests, or other statistical quantities are only used when they are helpful to the predictive enterprise. For instance, in Chapter 2, the problem of estimating a mean will be recast into that of identifying point predictors for future outcomes. Instead of being concerned with standard errors or posterior variances, the focus will be on identifying prediction intervals for future outcomes. Likewise, in Parts II and III, instead of considering classes of estimators or tests, classes of predictors will be considered. The role of parameters will be mostly to form a predictor or to quantify how well a predictor performs. An overall approach to statistics can – and should – be based on prediction, and the main reason why this book was written was to demonstrate this.

1.2 Some Examples

In this section four examples of data sets will be presented where a predictive approach is either different from (and better than) a modeling approach or for which prediction is the central issue. The first two examples (in Sec. 1.2.1) are low dimensional, and we see that even in such seemingly anodyne settings a simple predictive approach is usually better than a simple modeling approach. The third example shows how a question that presents as a

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

decision problem (hypothesis testing) with complex data is actually better interpreted as a prediction problem. The fourth example shows how a classification problem with complex data may be better conceptualized as a prediction problem and how clustering techniques might be useful to formulate a predictor. These are widely divergent data types and statistical techniques. However, taken together they show that predictive approaches are ubiquitous and frequently superior to conventional model building.

1.2.1 Prediction with Ensembles rather than Models

Let us begin by looking at two low-dimensional data sets and contrasting a conventional model building approach (linear regression) with three predictive approaches (L2-BMA, RVMs, and bagged RVMs). Once the four predictors have been described, a computational comparison of their performances will be given. The predictive approaches are better in several senses.

The first data set is from the Tour de France. The Tour de France is an annual bicycle race consisting (currently) of 21 segments raced over 23 days; two days are set aside for the cyclists to rest. The race was started in 1903 but was not held from 1915–1918 or from 1940–1946 owing to the World Wars; thus up to 2012 there are 99 data points, one for each year. The exact route changes each year but the format stays the same. There are at least two time trials and segments of the course which run through the Pyrenees and the Alps, and the finish is on the Champs-Élysées. The length has ranged from 1484 miles in 1904 to 3570 miles in 1926. While many aspects of the race were recorded for each year the race was held, here it is enough to look only at the logarithm (base e) of the average speed (LSPEED) of the winner as a function of the year (YEAR) and the length (LENGTH) of the course. Here this will be called the Tour data. Through trial and error (or more formally Tukey's Ladder, see Tukey (1977)) one can find that using the logarithm of the average speed is a good choice because it leads to better fit in linear models. Also, the data for years 1919–1926 were removed as outliers because the LSPEED values for these years are unnaturally low; this reflects the huge number of young men who died during WWI who would have otherwise been potential competitors during those years. This leaves 91 data points.

Let's describe a standard model-based analysis. First, the scatterplots of LSPEED versus YEAR and LENGTH are given in Fig. 1.1. The plot on the left shows that LSPEED increases over time, possibly because of improved training, bicycle technology, and a larger field of competitors. There are no obvious outliers and the spread of LSPEED does not appear to change much with YEAR. The plot on the right shows that as LENGTH increases the winning speed tends to decrease, possibly because the cyclists are more tired by the end of the course. There are three outliers and it appears that the spread of the scatterplot increases slightly with LENGTH. There appears to be a small amount of curvature in LSPEED as a function of both YEAR and LENGTH.

Consider finding the usual least squares fit of the second-order model

$$\begin{aligned} \text{LSPEED} = {} & \beta_0 + \beta_1\text{YEAR} + \beta_2\text{YEAR}^2 + \beta_3\text{LENGTH} + \beta_4\text{LENGTH}^2 \\ & + \beta_5\text{YEAR} \times \text{LENGTH} + \epsilon, \end{aligned} \tag{1.2}$$

assuming ϵ is a mean-zero random error. Higher-order terms such as those in LENGTH^3 or YEAR^3 are so highly correlated with LENGTH^2 and YEAR^2 that they make the regression

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

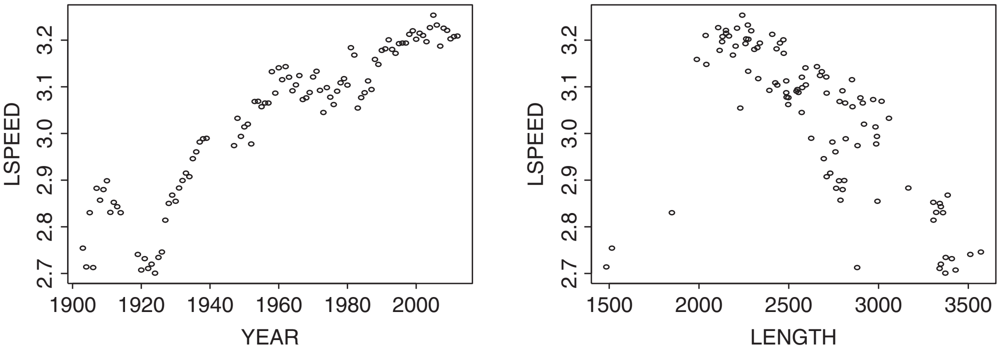

Figure 1.1 Left: Plot of LSPEED versus YEAR. Right: Plot of LSPEED versus LENGTH.

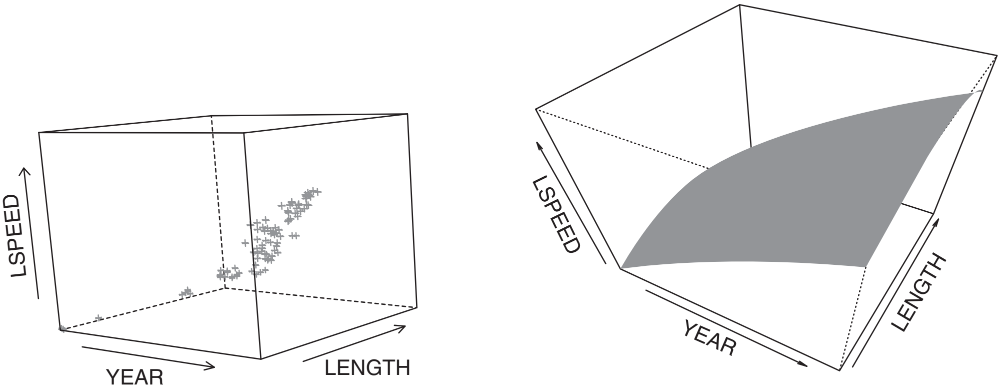

Figure 1.2 Left: Three-dimensional scatterplot of the Tour data. Right: Plot of the estimated regression function in (1.3).

worse. Moreover, a partial F-test to see whether YEAR × LENGTH is worth including, given the other included terms, shows that it can be dropped. The p-values for $\mathcal{H}_k : \beta_k = 0$ are below 0.05 except for β_2, which is 0.068. We retain the term in YEAR^2 since it is obvious that there is curvature in YEAR even if the square does not capture it well. Thus, the estimated regression function is

$$\begin{aligned} \text{LSPEED} = & -41.4 + 0.04121\text{YEAR} - 0.000009573 \times \text{YEAR}^2 \\ & + 0.0004098 \times \text{LENGTH} - 0.00000008263 \times \text{LENGTH}^2, \end{aligned} \tag{1.3}$$

and if desired one could obtain CIs, PIs, diagnostics, and so forth from standard linear model analysis. Indeed, an individual parameter estimate such as β_1 represents the change in LSPEED as a result of a unit change in YEAR and this means that parameter estimates can be converted readily into predictions.

The left-hand panel in Fig. 1.2 shows that the three-dimensional scatterplot is not very informative because the points bunch up too much for the shape to be seen in three dimensions. However, the right-hand panel shows the estimated regression function, in which the decrease of LSPEED with LENGTH and the increase in LSPEED with YEAR can be seen.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

Note that a model of the form (1.2) cannot possibly be universally true; indeed, it can only be approximately valid on a range of values for LENGTH and YEAR. Indeed, if YEAR were to increase and $\beta_2 < 0$ then, beyond a certain year, LSPEED would be negative, meaning that in the limit the SPEED was zero. On the other hand, if $\beta_2 > 0$ then the speed would be increasing and sooner or later the cyclists would be traveling faster than the speed of light. If curvature is neglected, the same problem occurs with β_1. Analogous problems occur with the coefficients β_3 and β_4 on LENGTH – but it makes some sense to treat LENGTH as bounded. These criticisms, based on bias, are different from the other criticisms that one might make of the linear model. For example, the data are not independent from year to year since often the same competitors participate, and the data are not identical from year to year because the competitors and courses differ. However, one can argue that these latter criticisms represent minor deviations from the assumptions compared with the problems with bias.

This model-based analysis can be tested predictively by comparing its prediction error to the prediction error of a predictive analysis that does not even try to model LSPEED. This requires a predictive criterion to evaluate, and there are two obvious ones to consider. First is the cumulative predictive error (CPE). That is, for a predictor sequence where the elements in the sequence are formed by the same techniques we can look at

$$\mathrm{CPE} = \sum_{i=6}^{91} \left(\hat{F}_{i-1}(Y_i, L_i) - \mathrm{LSPEED}_i \right)^2. \tag{1.4}$$

Expression (1.4) is a sum of squared errors since the estimated linear model was found using least squares estimators with all the data. In (1.4), $\hat{F}_{i-1}$ is the predictor for LSPEED$_i$ formed from the first $i - 1$ data points. In particular, the point predictor for LSPEED$_i$ formed from fitting a linear model of the form (1.2) using the first $i-1$ data points and evaluating $\hat{F}_{i-1}$ at $(Y_i, L_i) = (\mathrm{YEAR}_i, \mathrm{LENGTH}_i)$, can be used since these data points rely on data obtained before LSPEED$_i$ is measured. For the Tour data, it is natural to assume that they are ordered by YEAR, represented by $i = 1$ to 91. This means that (1.4) includes a burn-in of five data points, the minimum necessary to fit a five-term linear model. The result is that

$$\hat{F}_{90}(Y_{91}, L_{91}) \quad \text{and} \quad \hat{F}_{91}(Y_{92}, L_{92}) \tag{1.5}$$

are the predictions for the 91st and 92nd data points, but the prediction error can only be evaluated for the 91st data point.

When there is a natural order to the data as with the Tour data, expressions (1.4) and (1.5) make good sense. However, if one wants to minimize the effect of non-IID-ness, or use the same idea as the cumulative predictive error (CPE) on data that has no natural ordering, one might be led to permutations of the data. Indeed, owing to the variability of the predictive errors (the summands in (1.4)) it is natural to assess the cumulative prediction error by taking the average of expressions like (1.4) for several, say K, permutations of the data and making a prediction for the 91st data point by averaging the predictions of the K predictors found from the K permutations. That is, randomly draw permutations $\sigma_1, \ldots, \sigma_K$ of $\{1, \ldots, 91\}$ and find the average,

$$\mathrm{CPE}(\sigma) = \frac{1}{K} \sum_{j=1}^{K} \sum_{i=6}^{91} \left(\hat{F}_{\sigma_j(i-1)}(Y_{\sigma_j(i)}, L_{\sigma_j(i)}) - \mathrm{LSPEED}_{\sigma_j(i)} \right)^2, \tag{1.6}$$

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

where the five data points used for the burn-in are assumed to be $\sigma_j(1), \ldots, \sigma_j(5)$ for $j = 1, \ldots, K$. Note that different orderings of the data will give different values for the K summands; it might be informative to observe how terms in the model become more or less important, or even to find a standard error for the predictions, but that is not the point here. The predictions for the 91st and 92nd data points are

$$\frac{1}{K}\sum_{j=1}^{K} \hat{F}_{\sigma_j(90)}(Y_{91}, L_{91}) \quad \text{and} \quad \frac{1}{K}\sum_{j=1}^{K} \hat{F}_{\sigma_j(91)}(Y_{92}, L_{92}). \tag{1.7}$$

but again the predictive error can be evaluated only for the 91st data point.

Obviously, for updating the linear model, all the models used for the final prediction (at time 91) will be nearly the same (as given in (1.3)), since the coefficients are estimated by different sets of data points of size 90 and there are only 91 data points. So, the K predictions averaged in (1.7) may be similar even though along the way, e.g., for i between 26 and 90, the differences in the predictions may be larger. (For instance, this happens if a run of data points leads the estimate of a coefficient in one direction while a later run leads the estimate of the coefficient in a different direction, and then both runs may have larger than expected errors.) More generally, model reselection may be allowed at various stages of data accumulation, and this can give different models depending on the ordering of the data. This is an added level of complexity, to be discussed in Chapter 9.

While the cumulative predictive error is informative in the sense of how one arrives at a predictor, what is of more direct interest is the predictor found at the end of the process and evaluating how good it is. Thus, the final predictive error (FPE) is often a better criterion than the CPE. The FPE is the last predictive error that the data permits one to evaluate. Thus, the FPE is given by

$$\text{FPE} = \left(\hat{F}_{90}(Y_{91}, L_{91}) - \text{LSPEED}_{91}\right)^2 \tag{1.8}$$

or

$$\text{FPE}(\sigma) = \frac{1}{K}\sum_{j=1}^{K}\left(\hat{F}_{\sigma_j(90)}(Y_{\sigma_j(91)}, L_{\sigma_j(91)}) - \text{LSPEED}_{\sigma_j(91)}\right)^2, \tag{1.9}$$

if there is, or is not, a natural order to the data, respectively. One benefit of using the FPE instead of the CPE is that one usually needn't correct for a burn-in. After all, correcting for burn-in can be difficult; when comparing two predictors, if one requires a much higher burn-in it will be likely to have more variability and hence a higher FPE unless the burn-in greatly improves the accuracy of predictions.

Next, consider a different predictor using the same explanatory variables as in (1.2) but based on the Bayes model average (BMA) for prediction rather than linear regression with least squares estimates. Formally, the BMA for prediction (under L^2 error) is

$$p(\text{LSPEED}_i|\mathcal{D}_{i-1}) = \sum_{k=1}^{15} p(\text{LSPEED}_i|\mathcal{D}_{i-1}, \mathcal{M}_k)w(\mathcal{M}_k|\mathcal{D}_{i-1}), \tag{1.10}$$

in which the $\mathcal{M}_k = \{p(\cdot|\beta(k))\colon k = 1, \ldots, 15\}$ are the $2^4 - 1 = 15$ nontrivial submodels of (1.3), the vector of β-coefficients in the kth model being denoted by $\beta(k)$, $\mathcal{D}_{i-1}$ is the

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

data set available for forming a predictor for the ith outcome, and $w(\cdot|\mathcal{D}_{i-1})$ is the posterior distribution across the models. To keep this simple, assume a uniform prior $w(\mathcal{M}_k)$ over the 15 models, independent $N(0,1)$ priors $\phi_{(0,1)}$ for the coordinates of $\beta(k)$, i.e., $w(\beta(k)|\mathcal{M}_k) \sim \prod_{u=1}^{\dim(\beta(k))} \phi_{(0,1)}$, and that the β-coefficients are estimated by their posterior means. Obviously, (1.10) can be adapted to larger models than (1.3).

Now, the marginal likelihood of LSPEED is

$$p(\text{LSPEED}|\mathcal{M}_k) = \int p(\text{LSPEED}|\beta(k)\mathcal{M}_k)p(\beta(k)|\mathcal{M}_k)d\beta(k),$$

and the posterior model probabilities are

$$p(\mathcal{M}_k|\text{LSPEED}) = \frac{p(\text{LSPEED}|\beta(k),\mathcal{M}_k)p(\beta(k)|\mathcal{M}_k)}{\sum_k p(\text{LSPEED}|\beta(k),\mathcal{M}_k)p(\beta(k)|\mathcal{M}_k)}.$$

For (1.10) we now have

$$p(\text{LSPEED}_i|\mathcal{D}_{i-1},\mathcal{M}_k) = \int p(\text{LSPEED}_i|\beta(k),\mathcal{M}_k)p(\beta(k)|\mathcal{M}_k,\mathcal{D}_{i-1})d\beta(k)$$

for fixed k, where

$$p(\beta(k)|\mathcal{M}_k,\mathcal{D}_{i-1}) = \frac{p(\beta(k)|\mathcal{M}_k)p(\mathcal{D}_{i-1}|\beta(k))}{\int p(\beta(k)|\mathcal{M}_k)p(\mathcal{D}_{i-1}|\beta(k))d\beta_k}.$$

The BMA leads to an average of point predictors on taking expectations – a procedure which is optimal in an L^2 error sense. So, applying $E(\cdot|\mathcal{D}_{i-1})$, we get the predictor L2-BMA

$$\begin{aligned}\widehat{\text{LSPEED}}_i &= E(\text{LSPEED}_i|\mathcal{D}_{i-1})\\ &= \sum_k E(\text{LSPEED}_i|\mathcal{M}_k,\mathcal{D}_{i-1})p(\mathcal{M}_k|\mathcal{D}_{i-1}). \end{aligned}\tag{1.11}$$

In practice, it is easiest to use the approximation

$$p(\mathcal{M}_k|\mathcal{D}_{i-1}) \approx \frac{\exp(-0.5BIC_k)}{\sum_{j\in\mathcal{M}}\exp(-0.5BIC_j)},$$

where BIC_k is the Bayes information criterion value of model k (so that $\text{Var}(\epsilon)$ is estimated by its usual empirical value) and to set

$$E(\text{LSPEED}_i|\mathcal{M}_k,\mathcal{D}_{i-1}) \approx f_k(x_i|\hat{\beta}(k)),$$

where $\hat{\beta}(k)$ is an estimator for the regression parameter $\beta(k)$ of model $\mathcal{M}_k$. Here, this is the posterior mean under the prior for the parameter.

Now, choosing K sets of $i-1$ data points using K randomly chosen permutations σ_j, $j = 1,\ldots,K$, and then using (1.11) on each set of $i-1$ data points we can form the $\hat{F}_{\sigma_j(i-1)}$, which can be used to give an FPE(σ) at stage $i = 91$ for L2-BMA, as in (1.9). Using permutations means that the natural ordering on the data is being ignored – as is reasonable since the IID assumption probably holds. Again, it is natural to use $i = 91$ since the FPE represents the predictive accuracy at the last (here 91st) data point. Note that this can be done with the same explanatory variables as in (1.3) and either method, linear models or L2-BMA, can be employed with any choice of explanatory variables that the sample size will permit.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

If one really wanted a CPE, one would have to make the L2-BMA and linear models comparable by using the same burn-in with L2-BMA as for a five-term linear model, i.e., a burn-in of five data points, and take a CPE over the remaining 86. Henceforth, however, for simplicity we will focus on the FPE and neglect any burn-in for the predictors and results of this chapter.

Let us give one more technique for sequential prediction that we might use with the Tour data. There are two forms of the relevance vector machine technique that we will use here. One involves RVMs in their pure, optimal form and the other involves will be 'bagged' RVMs. Bagging – bootstrap aggregation – is a non-Bayes model-averaging technique commonly used to stabilize good but usually unstable prediction methods; see Breiman (1994, 1996b).

First, to define an RVM let $\mathcal{H}$ be a Hilbert space – basically a generalization of Euclidean space, as follows. Assume that the elements of the Hilbert space are real-valued functions on a domain $\mathcal{X}$. Then, suppose $\mathcal{H}$ has a 'reproducing kernel' k, namely, a real-valued function on $\mathcal{X} \times \mathcal{X}$ such that for any function $f \in \mathcal{H}$ we have $\langle k(\cdot, x), f(\cdot)\rangle = f(x)$, where $\langle \cdot, \cdot \rangle$ is the inner product of $\mathcal{H}$, i.e., k 'reproduces' f. We assume that k is a symmetric and positive semidefinite function on its domain, assumed compact.

Given an RKHS – a reproducing kernel Hilbert space – one can set up a minimization problem. Pick a function, say $L\colon (\mathcal{X} \times \mathbb{R}^2)^{i-1} \to \mathbb{R}$, and a convex penalty function, say $\Omega\colon [0, \infty) \to \mathbb{R}$. Let us minimize a functional J on $\mathcal{H}$ defined by

$$J(f) = L(x_1, y_1, f(x_1), \ldots, x_{i-1}, y_{i-1}, f(x_{i-1})) + \Omega(\|f\|^2_{\mathcal{H}}),$$

where $x_1, \ldots, x_{i-1} \in \mathcal{X}$, $y_1, \ldots, y_{i-1}, f(x_1), \ldots, f(x_{i-1}) \in \mathbb{R}$, and $\| \cdot \|_{\mathcal{H}}$ is the norm from the inner product on $\mathcal{H}$. The representer theorem states that there is an $f \in \mathcal{H}$ such that

$$f(x) = \arg\min_{f \in \mathcal{H}} F(f) = \sum_{j=1}^{i-1} \alpha_i k(x_j, x), \tag{1.12}$$

for some choice of $\alpha_1, \ldots, \alpha_{i-1} \in \mathbb{R}$ and, conversely, if Ω is increasing then each minimizer of $J(f)$ can be expressed in the form (1.12) . It is seen that (1.12) gives a predictor as a function of the input data, the kernel k, and $i - 1$ parameters. Loosely, an RVM is a Bayesian analysis of (1.12) treating the α_j as parameters. There are several ways to estimate the α_j. Tipping (2001) used a Bayesian approach, as did Chakraborty *et al.* (2012) and Chakraborty (2012), while Scholkopf and Smola (2002) discussed frequentist approaches.

If the α_j are estimated then (1.12) gives a third predictor, using RVMs analogous to (1.11) for L2-BMA and the predictor that one would get from finding (1.3) using $i - 1$ data points, i.e., linear models. Thus, one can use (1.8) to evaluate the final predictive error (FPE) or (1.9), by using K randomly chosen permutations σ_j, to form the $\hat{F}_{\sigma_j(i-1)}$. Setting $i = 91$ would give us the FPE for the Tour data. There is usually a tuning parameter controlling the width of the kernel function, and the value of this parameter typically has a greater effect than the shape of the kernel. This tuning parameter is usually estimated initially for the computing package being used.

Our fourth predictor is a bagged version of the estimated form of (1.12). The idea of bagging is to fix a stage i at which one wants to predict. Then, if $i < n$, choose a permutation of all the data available and select the first i data points after permutation. Next, use the first

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

$i-1$ of these to predict the ith by bootstrapping. Specifically, choose $0.67(i-1)$ of the $i-1$ data points at random and use them to generate an estimated form of (1.12). Do this a number of times; in the computations below, 100 times gave a good performance. Then, take the predictions for the ith stage from all 100 (say) bootstrap samples and average them to get a prediction at the ith stage. To assess the FPE of this procedure one uses the analog of (1.9), because, even if there is a natural order to the data, it is disrupted by the bootstrapping. The resulting predictor is a bootstrapped RVM (BRVM). Although the RVM and BRVM are nonlinear and look quite different from linear models or L2-BMA with linear models, the predictors they give depend on which explanatory variables are included as well as which kernel function is chosen.

Let us now look at how these four predictors – linear models, L2-BMA (with linear models), RVMs, and BRVMs – actually perform for the Tour data. First, let's use the explanatory variables YEAR, YEAR^2, LENGTH, and LENGTH^2 for LSPEED. For the sake of clarity, note that it is the whole histogram formed from FPE values as σ_j varies that is really of interest. Despite this, often it is enough to look just at the mean FPE(σ) as in (1.9). However, to compare predictors, one should look at the variance of the terms in (1.9) as well, if not the whole distribution of the FPE. In the calculations below, the five-number summary of the K FPE values,

$$\left(\hat{F}_{\sigma_j(90)}(Y_{91}, L_{91}) - \text{LSPEED}_{\sigma_j(91)}\right)^2 \tag{1.13}$$

as σ_k varies, is given for the Tour data along with the mean, i.e., (1.9). This will show that the distribution of the FPE can be skewed even for sample sizes of 90 – but this obviously depends on the predictor and the random aspect of the DG, or, as it is more typically put, on the model being fitted and the true model including the error term.

Table 1.1 gives a comparison of results from the four methods described for the distribution of the FPE as σ varies for the Tour data using four nontrivial explanatory variables. The errors are found in L^1, but the results are qualitatively similar for $L^{1.5}$ and L^2. It is seen that the distribution of the FPEs, as approximated by the bootstrapping and reselection of σ, is somewhat skewed for all four predictors since the median is closer to Q1 than it is to the mean, especially for the linear predictor. Clearly, for this very simple class of predictors, the linear model gives the smallest (mean) FPE while BRVM gives the smallest median FPE. Arguably, the skewness makes the median more relevant. However, without an assessment of variability such as a standard error of the mean FPE or the interquartile range for the median FPE it is hard to decide whether the pure linear model or BRVM is the better predictor.

One reason why the linear predictor gives the smallest mean FPE, smaller in particular than the mean FPE for the L2-BMA, is that the linear model with four explanatory variables is a better predictor than any of its submodels. So, even though the posterior probability piles up on the full model, the L2-BMA puts nonzero mass on the other 14 models and thereby loses predictive power. Combining the models in the L2-BMA by adding the coefficients on their common terms means that one can regard the L2-BMA as a linear model in which the coefficients are found by a Bayesian criterion rather than using least squares estimators. So, the underperformance of the L2-BMA occurs because there is essentially no uncertainty as to which model is best (here the biggest) but the L2-BMA still puts mass on submodels.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

Table 1.1 *Five-number summary plus mean for the* $K = 100$ *random choices of permutations* σ *for the* Tour *data. The FPEs were found using* 100 *bootstrap samples of proportion* 0.67*. For the RVM and BRVM results, the tuning parameter in the Gaussian kernel was* 4*. The bold numbers indicate the minimum in their column.*

	min	Q1	median	mean	Q3	max
linear	0.0017	0.1238	0.1849	**0.2287**	0.3225	0.9866
L2-BMA	0.0008	0.0931	0.2008	0.2420	0.3247	1.1750
RVM	0.0093	0.1002	0.1624	0.3949	0.3341	3.72
BRVM	0.0030	0.0719	**0.1517**	0.2571	0.2400	2.069

Table 1.2 *FPE results analogous to those in Table 1.1 but for seven rather than four nontrivial explanatory variables.*

	min	Q1	median	mean	Q3	max
linear	0.0058	0.1108	0.1795	0.2315	0.3007	1.006
L2-BMA	0.0081	0.1045	0.2001	**0.2263**	0.2993	1.0750
RVM	0.0030	0.0895	0.1601	0.3331	0.3545	3.0990
BRVM	0.0039	0.0545	**0.1557**	0.2402	0.2992	1.8950

The kernel methods do better in terms of the median FPE and worse in terms of the mean FPE than the two methods using (linear) models. This may be due to the fact that kernel methods are more flexible and can track location better, even when the median is a better assessment of location than the mean. Part of the reason why the RVM does poorly in the mean FPE may be high variance. After all, the RVM permits as many terms as there are data points (here, 90), before the prediction is made so one expects that if the variability is reduced there should be an improvement in prediction. Thus, as expected, bagging i.e., using BRVMs, improves RVM quite a lot, so that BRVM has the smallest median FPE and a smaller mean FPE than RVMs alone. So, in this case, one would want to create the full histogram of the individual FPE values (for each σ) and decide whether the skewness was small enough that one should use the linear model or large enough that one should use BRVMs. Even then, the relative sizes of the median FPEs or mean FPEs would have to be assessed, to decide whether the method with the smallest error really was the best once statistical variation is taken into account. Superficially, the skewness suggests that the median FPE is more appropriate, so the preferred predictor should be the BRVM.

For comparison with Table 1.1 it is worthwhile generating the analogous Table 1.2, using more explanatory variables. So, in addition to YEAR, YEAR^2, LENGTH, and LENGTH^2, we consider including YEAR × LENGTH, YEAR × LENGTH^2, and LENGTH × YEAR^2, all terms that would test out as unimportant under a partial F-test in a conventional linear models approach. Using seven explanatory variables means that the L2-BMA has $2^7 - 1 = 127$ terms and both arguments in the kernels, in the kernel methods, are seven dimensional.

Table 1.2 is different from Table 1.1 in several ways. First, the skewness is less: it is only seen for linear models and RVMs, where the median is closer to Q1 than to Q3. So, it is hard to decide whether the median or mean FPE is more appropriate. Second, under either form of FPE, the ensemble methods L2-BMA and BRVM have the lowest errors. Neither the linear

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

model nor a single RVM does well. Third, in about half the cases the addition of the extra terms lowers the FPE, whether median or mean, in contrast with the entries in Table 1.1. So, some methods (chiefly the ensemble methods) are better able to make use of the extra information than the 'pure' methods. Otherwise, Table 1.2 is generally similar to Table 1.1: in the median FPE, the kernel methods do better than the linear-model-based methods and in the mean FPE the linear-model-based methods do better than the kernel-based methods. The computations using $L^{1.5}$ and L^2 were qualitatively similar. Furthermore, the BRVM does better than the RVM. Again, without an assessment of the variability of the errors it would be hard to conclude definitively that one method was performing better than another – even though the results here are suggestive.

As a way to show that the techniques and findings here are not atypical, the computations were redone on the Fires data set, available from `http://archive.ics.uci.edu/ml/datasets/Forest+Fires`, first studied in Cortez and Morais (2007). The idea of the data set is to predict the burn area of a forest fire in terms of 12 explanatory variables. However, one variable (the rain) varied very little and so was dropped for the analysis here. Also, the spatial coordinates in the park were dropped as being deemed not useful in more general settings. Thus, the analysis predicts the logarithm of the area (LAREA) as a function of nine covariates. The linear model predictor is obvious: one writes LAREA as a sum of ten terms, one for each explanatory variable plus a constant. For parallelism, the L2-BMA should in principle have been formed from $2^9 - 1$ submodels but for convenience here it was formed from nine simple linear regression models, each having a single explanatory variable (plus a constant). The RVMs and BRVMs were based on nine explanatory variables, used a Gaussian kernel with sigma factor 15 (rather than 4). The sample size was 517 so, apart from permutations, the first 516 data points were used to predict the 517th.

The results are shown in Table 1.3. As for Table 1.2, it can be seen that the ensemble methods BRVM and L2-BMA give the lowest median and mean FPEs, respectively, and otherwise are competitive. The median FPEs are roughly at the midpoint between Q1 and Q3, suggesting that the mean FPE might be more appropriate. On the other hand the median FPE is very much smaller than the mean, suggesting there are high tail values that might be influential and so the median FPE may be more appropriate. It is easy to see that the single RVM predictor is very poor under either FPE criterion, and the linear model predictor, while competitive, is never best. So, one is led to either the BRVM or the L2-BMA. Since there are other explanatory variables that could be included, or a larger model list could have been used in the L2-BMA, this might have changed the conclusion. So, for the present, the conclusion must remain tentative given that an analysis of the statistical variability of the FPEs has not been done.

The implication from these three computations is clear: outside very simple cases, using predictions from relatively uninterpretable ensemble methods is preferable to using individual predictors whether based on interpretable models e.g., linear models, or uninterpretable individual predictors, e.g., RVMs. Indeed, the better the predictions, the less interpretable the predictors seem to be. Otherwise stated, unless there is overwhelming evidence that a model assumption is valid, modeling frequently leads to less predictive accuracy, thereby calling the modeling itself into question. So, predictors that do not rely heavily on modeling will be more suitable for good prediction and hence any inferences made from them have greater cogency than simple or interpretable predictors.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

Table 1.3 *The FPE results for the* Fires *data parallel those in Table 1.1.*

	min	Q1	median	mean	Q3	max
linear	0.0078	0.1784	0.3463	0.5334	0.7019	3.4920
L2-BMA	0.0007	0.1586	0.3558	**0.4919**	0.5730	3.3970
RVM	0.0809	1.1750	3.5640	3.9680	5.7980	14.0900
BRVM	0.0105	0.1384	**0.2849**	0.5271	0.4100	4.714

1.2.2 Hypothesis Testing as Prediction

Although hypothesis-testing problems present themselves as decision making, in the sense of deciding which of two hypotheses to (de facto) accept, hypothesis testing, like any decision problem, is really just another technique for prediction and, like any predictive technique, its predictions must be validated on future data.

To see why this view of testing makes sense, recall that we are generally more interested in the data we will see in the future than in the data we already have. Formally, this is so because inferences are made about a population, not about a sample. So, our inferences should apply to future draws from the population. For example, the result of a hypothesis test about a parameter will tell us something about the value of the parameter. The conclusion about the parameter value is relevant to the population from which the particular sample was drawn and so should be relevant to any sample drawn from this population. Therefore the hypothesis test, and its conclusion, can be regarded as a statement about future data which such data may confirm or refute.

As a statement about the future, hypothesis testing can be regarded as a 'prediction-generating' mechanism. That is, any hypothesis test will lead to a conclusion about a parameter of interest and this conclusion represents a prediction about future samples. As a simple example, suppose that we collect data $X = (x_1, x_2, \ldots, x_n)$, where $x_i \sim N(\mu, 1), i = 1, \ldots, n$, and find that $\bar{X} = 0.4$. Consider the test

$$\mathcal{H}_0\colon \mu \leq 0 \text{ vs. } \mathcal{H}_1\colon \mu > 0; \tag{1.14}$$

we might decide, on the basis of the test statistic, to reject $\mathcal{H}_0$ and conclude that $\mu > 0$. Using this, we can form a prediction about the mean of the next sample or the value of the next observation from the population. Recall that a prediction interval is an interval associated with a random variable yet to be observed, with a specified probability that the random variable will lie within the interval. Under $\mathcal{H}_1$, a new individual observation would be predicted to lie in the interval $[z_\alpha, \infty)$, where z_α is the 100αth percentile of an $N(0,1)$, with probability at least $1-\alpha$, for some $\alpha \in (0,1)$. Analogously, we would predict that the mean of a new sample of size n, $\bar{X}_{\text{new}}$, would be in $[z_\alpha/\sqrt{n}, \infty)$ with probability at least $1-\alpha$. The same sort of reasoning applies to other frequentist parametric and nonparametric testing problems: the rejection – or acceptance – of a hypothesis limits the range of values that future outcomes may assume.

This type of thinking holds for Bayes testing as well. Assume that the same data as before is available and the task is to test the same hypotheses but from a Bayes perspective. Suppose there is a prior probability for each hypothesis, i.e., for $i = 0, 1$, there is a

$$w_i = P(\mathcal{H}_i) = P(\mu \in \mathcal{M}_i),$$

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

where $\mathcal{M}_0 = \{\mu\colon \mu \leq 0\}$ and $\mathcal{M}_1 = \{\mu\colon \mu > 0\}$. Given the prior probabilities and the likelihoods, denoted $P(x|\mathcal{H}_i)$, the posterior probability for each hypothesis can be found, namely,

$$p_i = P(\mathcal{H}_i|x) = P(\mu \in \mathcal{M}_i|x) = P(\mathcal{H}_i)P(x|\mathcal{H}_i) = w_i P(x|\mathcal{H}_i).$$

The hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ can be compared using the Bayes factor, i.e., the ratio of the posterior odds and the prior odds; see Kass and Raftery (1995). The posterior odds in favor of $\mathcal{H}_0$ relative to $\mathcal{H}_1$ are

$$p_0/p_1 = P(\mathcal{H}_0|x)/P(\mathcal{H}_1|x),$$

and the prior odds are defined analogously.

If the Bayes factor is sufficiently large, say > 3, it is common practice to decide in favor of $\mathcal{H}_0$ and, again, this would lead to a prediction about the next outcome or mean of the next sample. One could use prediction intervals as in the Frequentist prediction case, or, to be more consistent with the Bayes approach, one could use the predictive distribution to get a Bayes prediction interval. The latter would require not just probabilities for the hypotheses themselves but a prior distribution, say w, on $\mathcal{M}_0 \cup \mathcal{M}_1$. Then, if $\mathcal{H}_0$ were taken as true, the $(1-\alpha)100\%$ Bayes prediction interval for the next outcome would be $[t_\alpha, \infty)$, where t_α is the lower 100αth percentile of

$$\int_{\mathcal{M}_0} p(x|\mu) \frac{w(\mu)p(x^n|\mu)}{\int_{\mathcal{M}_0} w(\mu)p(x^n|\mu)} d\mu.$$

Analogous expressions would hold if $\mathcal{H}_1$ were taken as true, if one wanted a Bayes prediction interval for a future sample mean $\bar{X}_{\text{new}}$, or other parametric or nonparametric hypotheses were tested. As in the frequentist testing case, the rejection – or acceptance – of a hypothesis limits the range of values that future outcomes may assume.

As a more elaborate example consider the same ideas in the context of bacterial metagenomics. A DNA sample known to contain more than one strain of bacteria is called metagenomic, and a typical task is to detect which bacterial strains are present in the sample. So, the decision problem comes down to claiming the presence or absence of a given strain from a list of possible strains and hence represents a collection of hypothesis tests, one for each candidate strain. Accordingly, this will necessitate a multiple comparisons correction and will show how testing can lead to a collection of predictions.

Suppose the DNA sample has been processed via next-generation sequencing (NGS), a complicated series of biochemical reactions that parallelizes the sequencing of DNA at the cost of generating many short DNA sequences rather than fewer longer sequences. Given the ‘short reads’ from a single physical sample, the task is to do the tests

$$\mathcal{H}_{0,j}: C_j \text{ is not in the population vs. } \mathcal{H}_{1,j}: C_j \text{ is in the population,} \tag{1.15}$$

where the C_j for $j = 1, \ldots, J$ represent a collection of reference genomes for bacterial strains. That is, each C_j is a bacterial genome written as a long string of nucleotide bases. (Most bacteria have one or two circular chromosomes and a number of other smaller circular pieces of DNA called plasmids. Plasmids are the way in which bacteria exchange genetic information without sex.)

Analyzing NGS data usually begins with alignment. Since each short read is a sequence of the nucleotides A, T, C, and G that may occur in zero, one, or more of the C_j, specialized

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

software must be used to identify where on the C_j each short read matches (to within a certain tolerance because both the short reads and the reference genomes may be inaccurate). If a given read aligns then it is possible that the read came from the bacterial strain represented by C_j. There are some common regions across bacteria, so the alignment of some short reads may not discriminate among the bacteria. However, there are some short reads that may be unique to a specific C_j. Obviously, the efficacy of the alignment and subsequent hypothesis testing will depend on J, the coverage of the genomes, and the richness of the reference genome database across the bacterial taxa, as well as on the quality of the NGS data. Typically, J will be on the order of thousands while the number of reads will be in the hundreds of thousands to millions.

Given a sample of NGS data we can do the J tests (1.15). From a frequentist standpoint one compares the number of genomic reads which would be expected to align to a given reference C_j under $\mathcal{H}_{0,j}$ with the number of genomic reads from the sample that were observed as aligning to the same reference. So, let $r_k, k = 1, \ldots, K$, denote the sample genomic reads, i.e., finite sequences of A, T, C, and G, and let l_k be the length of r_k. Also, let X_j be the random variable representing the number of reads from C_j in the sample of size K.

Given $X_j = x_j$, the p-value $P(X_j > x_j|\mathcal{H}_{0,j})$ can be estimated by permutation testing. The basic idea behind a permutation test is to permute the data from which a test statistic is calculated, effectively creating many new data sets so that many values of the test statistic can be found. The values of the statistic are only valid under the null hypothesis because the permutation is over all the data assuming that the null is true. The collection of values of the test statistic is used to form a histogram which is an estimate of the sampling distribution of the test statistic under the null hypothesis. The p-value for the test is taken to be the area under the histogram to one side (for a one-sided test) of the actual value of the test statistic. Permutation tests can be very useful when the distribution of the test statistic under $\mathcal{H}_{0,j}$ is unknown; see Valdes *et al.* (2015).

The analog of these classical permutation tests for NGS data is to mutate the nucleotides in the sample reads r_k, $k = 1, \ldots, K$, at a fixed mutation rate q. Now, if a read r_k actually came from a specific C_j then it will be mutated, thereby simulating a sample read that might be found if $\mathcal{H}_{0,j}$ were true, i.e., if DNA from C_j were not in the sample. If this is done for all K reads then the result is a collection of K mutated reads, which can be aligned to each reference genome C_j. To quantify this, let $Y_j = Y_j(q)$ for $j = 1, \ldots, J$ be the number of mutated reads that align to C_j for a given q. Now, Y_j is an estimate of the number of reads that would be expected to align to C_j under $\mathcal{H}_{0,j}$. Note that if all the sample reads were mutated M times, the result would be M different values of Y_j, namely, $\{Y_{jm} : m = 1, \ldots, M\}$. By comparing the value of X_j with the values $\{Y_{jm}\}$ from mutated versions of the data, a p-value for testing $\mathcal{H}_{0,j}$ versus $\mathcal{H}_{1,j}$ can be found. This can be repeated for each $j \in \{1, \ldots, J\}$, giving a set of J raw p-values. After adjusting the p-values for multiple comparisons, e.g., using the Westfall–Young procedure (see Westfall and Young (1993)), since such a procedure is also based on permutations, they can be used to decide the presence of any C_j. As in Sec. 1.2.1, if this is done many times, a loss function can be chosen and a CPE or FPE found for any testing procedure.

Up to this point, this has just been a hypothesis-testing problem. However, it really is prediction: the set of adjusted p-values $P(X_j > x_j|\mathcal{H}_{0,j})$ for $j = 1, \ldots, J$ represents a prediction of which bacterial strains are present in any other sample from the same population. For instance, we may consider any adjusted p-value less than a given threshold as

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

'significant', reject the associated null hypothesis, and decide that the associated bacterial strain is present. Otherwise stated, this is a prediction that the strain in question will be present in a new sample from the same population. Likewise, any adjusted p-value larger than a given threshold may be considered 'not significant', so the associated null would not be rejected and the decision would be that the associated bacterial strain is not present. This is a prediction that the strain in question will not be present in a new sample from the same population. If there is a collection of samples from which to obtain a prediction, there are many ways to pool the data to get improved prediction. One way, see Vovk (2012), is to generate adjusted p-values for each sample, calculate the average p-value for each strain, and use the set of average adjusted p-values to generate predictions.

The tests (1.15) can also be done from a Bayesian perspective. Suppose the bacterial composition of the sample is modeled as an observation from a multinomial distribution, with $M+1$ categories representing M bacterial strains and one additional category for all nonbacterial strains. Let θ_j be the proportion of the population in category j, so that $\theta_1 + \theta_2 + \cdots + \theta_{M+1} = 1$ and $\theta_j \geq 0$ for $j = 1, \ldots, M+1$. Given a sample of K reads from the population with k_j reads aligning to category j, the likelihood function is

$$p(\mathcal{D}|\theta_1, \ldots, \theta_{M+1}) = \left(\frac{K}{k_1 k_2 \cdots k_{M+1}} \right) \theta_1^{n_1} \cdots \theta_{M+1}^{k_{M+1}},$$

where $\mathcal{D}$ is the data i.e., the set of reads. For convenience, consider a conjugate prior $w(\theta) \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_{M+1})$ represented as

$$w(\theta) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_{M+1}^{\alpha_{M+1} - 1},$$

where $\alpha_j > 0$ for $j = 1, \ldots, M+1$; then

$$w(\theta_1, \ldots, \theta_{M+1}|\mathcal{D}) \sim \text{Dirichlet}(\alpha_1 + k_1 - 1, \alpha_2 + k_2 - 1, \ldots, \alpha_{M+1} + k_{M+1} - 1).$$

Treating the θ_j as random variables Θ_j gives the posterior expectation for $\Theta = (\theta_1, \ldots, \theta_{M+1})^T$ as

$$\hat{\theta}_j = E(\Theta_j|\mathcal{D}) = \frac{k_j + \alpha_j}{K + \sum_{j=1}^{M+1} \alpha_j},$$

which is a combination of two sources of information about the proportional composition of the sample, namely, the prior and the data; see Clarke *et al.* (2015).

In this formulation every category has a nonzero prior probability and a nonzero posterior probability. This may not accurately reflect the population if there is strong reason to believe that many bacterial strains are not present. After all, such a belief would lead to choosing a prior assigning zero mass to a category believed to be void. If the goal were solely estimation then the approach here could be modified by using either a shrinkage prior (Johndrow 2013) or a mixture prior (Morris *et al.* 2005). These analyses separate strains into abundant and scarce classes but are beyond our present scope.

If mass is assigned to the $\theta_j = 0$ by a continuous prior w such as the Dirichlet distribution then testing (1.15) can be reinterpreted (informally) as

$$\mathcal{H}_{0,j}: w(C_j|\mathcal{D}) < \epsilon \ vs. \ \mathcal{H}_{1,j}: w(C_j|\mathcal{D}) \geq \epsilon, \tag{1.16}$$

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

where $\epsilon > 0$ is small enough that for all practical purposes the jth strain is not present. The associated Bayes factor for testing $\mathcal{H}_{0,j}$ versus $\mathcal{H}_{1,j}$ provides evidence about whether strain j is present in the population but would be subject to multiple comparisons problems. A common way around this is to choose one compound hypothesis, i.e., one event or set $\{\theta_1, \ldots, \theta_{M+1}\}$ that corresponds to the absence of any strain C_j known to be harmful to humans using the tolerance ϵ. Then, a single test can be done to determine whether anything harmful is present in the population from which the sample was drawn.

Just as in the frequentist case, the conclusion from a Bayes test is a prediction of what is expected to be present (or absent) in the next sample from the same population, i.e., Bayes testing is a method for generating a prediction.

Note that the posterior expectation for the proportions of each bacterial strain, the $\hat{\theta}_j$, can also be regarded as a prediction for the composition of the next sample from the same population. Specifically, the $\hat{\theta}_j$ are the proportions of reads from the C_j that are expected to be in future samples. An analogous interpretation holds true for any estimation problem, parametric or nonparametric. Again, estimation and model selection can be regarded as techniques for prediction generation, as was done implicitly with linear models and L2-BMA in Sec. 1.2.1. Indeed, estimating the coefficients in an RVM has an analogous interpretation.

A further benefit of the Bayes approach is the sequential updating of the posterior. Given a posterior distribution, it can be used to generate a prediction for a future sample. Then, once the future sample is observed, it can be aggregated with the data already obtained to form a new posterior from which an updated prediction for another future sample can be obtained, and so on. Thus, again, if a loss function is chosen, the sequence of errors made by the use of this predictor sequence can be used to generate a CPE or FPE.

To complete the argument that decision problems such as hypothesis testing are really just a form of prediction, it's worth considering the standard decision-theoretic setting. Bayes testing can be shown to be the Bayes action in a formal decision-theoretic sense (under generalized zero–one loss). Moreover, even though frequentist testing emerges from the Neyman–Pearson lemma, which uses a putative non-decision-theoretic optimality criterion, standard frequentist testing can also be given a decision-theoretic interpretation. So, the reasoning already given for frequentist testing exemplifies the point that decision problems are really prediction.

Classical decision theory is based on Wald's theorem that preferences can be ordered by risk or on Savage's axioms, which can be used to show that preferences can be ordered by Bayes risk. In either case, the structure has four elements: a collection of actions $a \in \mathcal{A}$, a collection of states of Nature $\theta \in \Theta$, a distance-like function L that assigns a cost $L(a, \theta)$ to taking action a when θ is true, and a collection of distributions, either $p(\mathcal{D}|\theta)$ or $p(\theta|\mathcal{D})$, representing the current states of beliefs about Nature. Clearly, in this structure, an optimal action such as a Bayes action or maximum expected utility action is a best guess as to θ given $\mathcal{D}$. So, if θ is a parameter, the result is an estimator $\hat{\theta}$ and this identifies an element from the set of possible current states of beliefs about Nature from which predictions can be made. This four-element structure of decision theory can be adapted to prediction problems more directly by regarding θ as a future outcome, in which case a is a predictor. A lucid and comprehensive examination of the way in which decision theory, predictive model selection, and predictive model assessment are interrelated in a Bayesian context can be found in Vehtari and Ojanen (2012). In a predictive context, a loss or utility function is often called

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

a scoring function because it is an assessment of how consistent a forecast probability is to its realized value. Gneiting (2011) provides an extensive treatment of scoring functions for point predictions; see also Parry *et al.* (2012) and Dawid *et al.* (2012).

A more general and operational description of the procedures presented here is briefly given in Sec. 8.6.3 from both the Bayesian and frequentist standpoints.

1.2.3 Predicting Classes

An example of a classification problem that may be seen as typical of a broad range of settings is provided by the analysis of data generated from musical scores. As a starting point, consider the foundational work of McKay (2004). In a wide-ranging study of the features that one might extract from pieces of music in order to identify the genre they represented, McKay identified 149 features that one could compute from musical scores. He broke them down into seven categories: dynamics (four features), instrumentation (20 features), melody (20 features), pitch (26 features), rhythm (35 features), texture (20 features), and chords (28 features). In fact, in his work, McKay only implemented 111 of these features to provide a formal categorization of all music.

Later, the music21 project built on McKay's work, developing a webtool-based on his feature selection. This webtool provides a set of computational tools to help musicologists answer quantitative questions about pieces of music on the basis of their scores. To date, music21 has implemented 70 of the features McKay identified: zero of the dynamics, 20 of the instrumentation, 19 of the melody, 26 of the pitch, 35 of the rhythm, 20 of the texture, and none of the chordal features. music21 calls these symbolic features (see `http://web.mit.edu/music21/doc/html/moduleFeaturesJSymbolic.html`) and adds 21 'native' features that, unlike the symbolic features, are unprocessed counts of various sorts directly obtained from the music scores (some details can be found at `http://web.mit.edu/music21/doc/html/moduleFeaturesNative.html`). music21 also provides a corpus of musical pieces that can be found at `http://web.mit.edu/music21/doc/html/referenceCorpus.html#referencecorpus`. It contains pieces of music from a variety of periods and styles. When applied to a given piece of music in the music21 corpus, the music21 software can be used along with specialized scripts to output a vector of comma-separated values of length 91. However, the values separated by commas are sometimes vectors in turn, so the real dimension of the vectors is 633, much higher than 91.

Given that this feature selection is intended to apply to all music, it's worth seeing how it performs on a binary classification problem such as determining whether a given piece of music was written by Mozart or Haydn. Mozart lived from 1756 to 1791 and Haydn lived from 1732 to 1809, so both are entirely within the period of Western music known as classical, 1730–1820. This means that, as different as Mozart's and Haydn's works are, it is reasonable to compare them and see that different classifiers give different performances.

Note that in this sort of problem a modeling approach is infeasible. It makes little sense to try to formulate a model that would accurately represent the creative process that Mozart or Haydn might have used to produce their masterpieces. Indeed, it is essentially impossible to characterize what makes a particular piece of music, or composer, great. There is just too much variety that has not been – and probably cannot be – explored. After all, if it were

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

possible to model musical brilliance precisely it would be possible to write a ‘great new music-generating algorithm’ – hardly something that will be accomplished any time soon.[2] Indeed, leaving aside the perceived quality of music, even coming up with a reliable way to categorize the various genres of music reliably is by itself a formidable task.

For the present, let’s use the 100 pieces of music composed by Mozart and 244 pieces composed by Haydn that are contained in the corpus. In this count, different movements from the same work are counted separately. This makes sense because, say, movement 1 from a piece by Haydn could have more in common with movement 1 from another piece by Haydn than it does with movement 2 from the same piece. Each of these 344 pieces of music was summarized by a set of features computed from the score. So, the data is neither independent nor identical and its degree of nonindependence and nonidenticality cannot realistically be assessed. The best that can be done is to test out various classification techniques and see how well they perform. In fact, this classification problem is almost a survey sampling model: the complete list of pieces by Haydn and Mozart is available. So, to make the prediction problem meaningful, half the pieces from each composer were randomly selected and put together as a training set. The other half were used as the predictive test set.

For the sake of illustration, there are three distinct ways to tackle this problem. First, one can adopt a model-based approach and use logistic regression, single trees, or support vector machines (SVMs) in an effort to model the data. Support vector machines do not actually provide an explicit model, but they are designed to give ‘support vectors’ which are interpreted as the boundary points between regions in a binary classification problem. Also, built into SVMs is a transformation specified by the kernel, thereby giving an implicit model. Second, one may proceed nonparametrically. In this case, the natural approach is to use k-nearest neighbors and clustering to assess whether the classes really are meaningful. Third, one can adopt a model-averaging standpoint. In this context two possible techniques are random forests and gradient boosting (the statistical formulation of Adaboost). These are purely predictive approaches since the resulting classifiers do not generally say anything directly about the DG. Since the DG in this case is in some important sense unknowable, one expects that classifiers that do not rely on finding a model for the DG ought to do better than those that do. In fact, this is more or less what is seen.

To get the data ready for analysis two tasks were performed. First, explanatory variables that did not vary enough to provide information on the response were removed, specifically, explanatory variables with a sample variance less than 0.2. This procedure removed 602 of the 633 original real variables. Second, explanatory variables that were too correlated with each other were removed. Specifically, the correlation between each pair of explanatory variables was found and if a pair had absolute correlation strictly greater than 0.9 one of the pair was removed at random and the correlations recalculated; this procedure ended up removing only six more variables, leaving a nonunique set of 25 variables.

Once this was done, the training set was found in two stages. First, since the 100 pieces by Mozart represented 29% of the total number of pieces, 344, sampling at random from the pieces by Mozart in proportion to their prevalence gave $0.5 \times 0.29 \times 344 = 49$ pieces.

[2] In fact, there are programs that can generate more music in the style of a given composer by mimicking some of its features. However, this is not creation *de novo* of a new sort of great music. It is more like a combination of existing great music.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

Likewise, sampling at random from the pieces by Haydn gave $0.5 \times 0.71 \times 344 = 122$. Taken together this gave a total sample size of 171 for training and left 173 pieces for testing how well the classifiers performed. Second, the variance and correlation constraints that were imposed on the whole data set were imposed on the sample of size 171. This removed one more variable, which was also removed from the test set. Thus there were 24 explanatory variables, 171 samples, and each sample had a response of zero or one to indicate composition by Mozart or Haydn, respectively. Henceforth this will be called the music data and treated as a classification problem.

The results of the three model based classifiers – logistic regression, single tree, and SVM – on the test set from the music data are summarized, respectively, in the following three confusion matrices:

$$\begin{pmatrix} 24 & 25 \\ 27 & 97 \end{pmatrix}, \quad \begin{pmatrix} 19 & 13 \\ 32 & 109 \end{pmatrix}, \quad \begin{pmatrix} 8 & 0 \\ 43 & 122 \end{pmatrix}. \tag{1.17}$$

The true composer is indicated by the column (left is Mozart, right is Haydn). The predicted composer is indicated by the row (top is Mozart, bottom is Haydn). Thus, in the first matrix, logistic regression used the 24 variables to identify 24 Mozart pieces correctly and 97 Haydn pieces correctly. It misclassified 52 pieces; the upper right entry means it incorrectly identified Mozart as Haydn 25 times and the lower left entry means it incorrectly identified Haydn as Mozart 27 times. Thus, the total number of errors was $25 + 27 = 52$.

It's worth looking at the key diagnostics for this logistic regression. First, Table 1.4 shows the variables that were included, along with their estimates, SEs, and uncorrected p-values. The significant p-values are low enough that a multiple testing procedure would not make them insignificant. Note that the variables are only labeled generically. This is so because the documentation from music21 only defines the variables by their order via python code so it is hard to be sure what each one means. However, this is not essential to the analysis at this stage. It is enough to note that, aside from the intercept, only five variables, V131, V145, V425, V426, and V445, appear to be useful.

A single tree gave the middle confusion matrix in (1.17). Trees are a very rich class of models – much richer than logistic regression and consequently less stable. The contributed R package rpart was used to generate single tree. It is shown in Fig. 1.3. For this tree, the number of errors was $13 + 32 = 45$.

It is seen that the classification tree and logistic regression 'agree' that V445, V145, and V131 are important.

The standard way to assess variable importance for trees is via a permutation technique. The result is shown in Table 1.5. It is seen that a fourth variable, V425, found important by logistic regression is picked up as being important by this technique too; however, V426 is not. This sort of discrepancy is not surprising since trees are unstable: they form a large class of models that provide an overall model by fitting a series of local models using less and less data per node as the tree grows; see Breiman *et al.* (1984).

An SVM gave the right-most confusion matrix in (1.17). Support vector machines are the least model dependent of the three methods. Underlying an SVM there is a model, but it is defined by the use of a kernel which represents a transformation of the feature space and is usually only defined implicitly. Using the Gaussian kernel, the contributed R package kernlab gave an SVM classifier that made 43 errors on the test set. The SVM also has 121

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

Table 1.4 *This table shows the* 24 *variables (plus intercept term), with their coefficients, SEs, and two-sided* p*-values from a logistic regression for classifying the pieces by Mozart and Haydn in the test set of size* 173 *from the* music *data. Bold type indicates uncorrected* p*-values for significant variables.*

Variable	Estimate	Std error	p-value
Intercept	−50.861	15.45	**0.001**
V130	0.114	0.61	0.851
V131	−0.758	0.34	**0.026**
V135	−0.208	0.30	0.496
V145	−0.040	0.01	**0.002**
V146	−0.050	0.11	0.641
V415	−0.124	0.07	0.092
V417	0.075	0.11	0.486
V424	0.017	0.01	0.185
V425	−1.142	0.40	**0.004**
V426	0.895	0.31	**0.004**
V428	−0.285	0.67	0.669
V431	0.328	0.43	0.443
V438	−0.035	0.06	0.572
V439	−0.062	0.22	0.780
V440	0.315	0.32	0.331
V441	0.202	0.11	0.075
V442	−1.019	0.63	0.107
V443	0.031	0.11	0.783
V445	0.892	0.23	**9.86e-05**
V449	−0.105	0.08	0.193
V605	−0.130	0.08	0.107
V609	−0.004	0.01	0.732
V610	0.127	0.07	0.064
V632	−3.216	219.37	0.988

Table 1.5 *This table shows the most important variables in the* music *data as determined by the variable importance assessment (based on permuting values) from* rpart. *Bold indicates variables with significant p-values from logistic regression in Table 1.4.*

V445	V443	**V145**	V146	V610	V441	V417
17	11	11	8	8	8	8
V449	V609	V605	**V131**	V130	V439	V425
7	6	6	4	3	1	1

support vectors – meaning about one-third of the data is on the boundary between the two classes. This would be considered high. Even worse, SVM did best by predicting, approximately, that all the pieces were written by Haydn. Roughly, as the model class gets richer, i.e., the model if it exists is harder to identify, the classifier does better at prediction. Note that even the best of these classifiers, the SVM, indicates a breakdown. (A valid criticism of this approach is that the results used the built-in estimates of tuning parameters in the tree

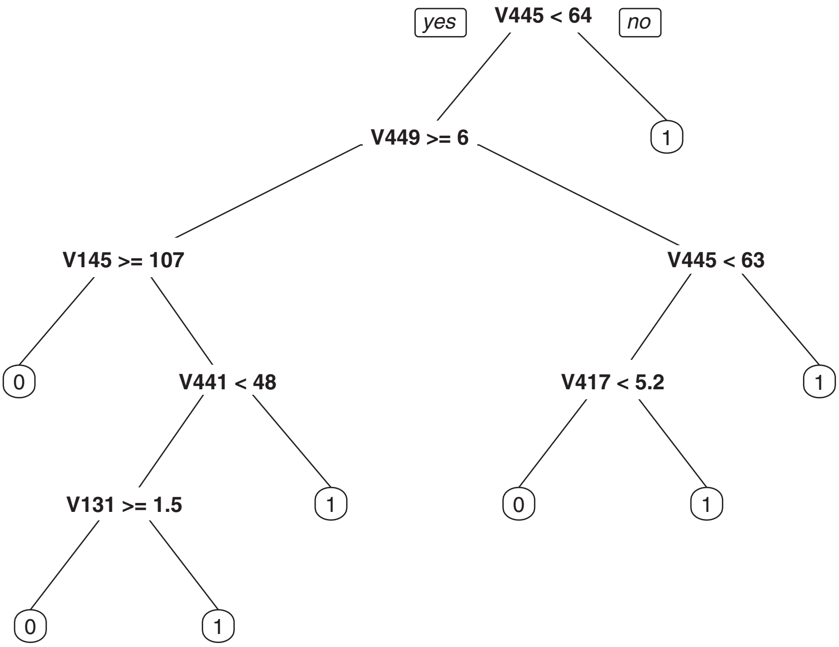

Figure 1.3 Single classification tree generated by rpart for the Mozart vs. Haydn problem and the music data. It only uses V445, V449, V417, V145, and V131.

and SVM; a more careful selection of the tuning parameters might have given better values. However, this is true of all the methods presented here. They are, in some average sense, equally disadvantaged.)

It's worth seeing whether a nonparametric approach will give better results. Using the R package RWEKA, confusion matrices for a k-nearest-neighbors classifier were generated. For the values $k = 1, 2, 3$ the confusion matrices for the music data are, respectively:

$$\begin{pmatrix} 24 & 21 \\ 27 & 101 \end{pmatrix}, \quad \begin{pmatrix} 24 & 21 \\ 27 & 101 \end{pmatrix}, \quad \begin{pmatrix} 15 & 11 \\ 36 & 111 \end{pmatrix}. \tag{1.18}$$

The first and second nearest neighbors give the same result, and the total error is 48. Including a third nearest neighbor doesn't help much – the total error is still 47. Nearest neighbor methods are known to be quite stable, so this small difference is no surprise. Overall, this method does better than logistic regression, the most restrictive model, which made 52 errors, but worse than the less restrictive tree or SVM models, which made 45 and 43 errors, respectively. It may be that the nonparametric model class is to a big for the data to overcome the model uncertainty – not to suggest that the notion of a model makes any sense in this setting. More precisely, the variability in the selection of the model from which to obtain a predictor may be hard for the data to overcome.

It is reasonable to be concerned that part of the problem with classification arises because the points in the training and test sets do not separate well. That is, the clusters of Mozart points and Haydn points in the 24-dimensional feature space are not well defined. To test this, drop the zero–one response and consider clustering on the vectors of explanatory variables. This can be done using the contributed R package pamk. For each number of clusters

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

Table 1.6 *The top row of this table shows the number of clusters in the* music *data found using* pamk. *The lower row shows the average silhouette distance for the clustering.*

2	3	4	5	6	7	8	9	10
0.414	0.320	0.400	0.391	0.375	0.341	0.343	0.348	0.279

k it will give a partitioning-around-medoids clustering that it can evaluate by the average silhouette distance. The results are given in Table 1.6. It is seen that $k = 2$ clusters is optimal (as 2 has the largest silhouette distance), but only weakly so. Four or even five clusters would be nearly as good. This means that the clusters are only weakly defined. Part of the problem may be that centroid-based clustering methods and even silhouette distances are better suited to convex clusterings than to more general clusterings and that if there are clusters in the 24-dimensional data then it is likely they are not convex. Nevertheless, Table 1.6 indicates that the problem is difficult. The concerns about weakly defined clusters would apply to logistic regression but less so to trees, since trees provide a rich class of models, and even less so to SVMs since they have a kernel transformation.

Turning at last to two techniques that are not at all model based, the confusion matrices for the random forests technique randomForest and the gradient boosting technique gbm applied to the music data are, respectively,

$$\begin{pmatrix} 17 & 6 \\ 34 & 110 \end{pmatrix}, \quad \begin{pmatrix} 21 & 11 \\ 30 & 11 \end{pmatrix}. \tag{1.19}$$

It is seen that the errors are 40 and 41, the smallest of the methods. The random forests technique is based on aggregating single classification trees like that found earlier. The aggregation is done via bootstrapping on the training set and *b*ootstrap *agg*regation is often abbreviated as ‘bagging’. Thus, random forests is just bagged trees. Basically, random forests inherits all the success of individual trees but refines them by the averaging process, thereby improving the performance of the final classifier. By contrast, gradient boosting is a model-averaging technique that is a refinement of the original boosting procedure, which was called Adaboost (Friedman *et al.* 2000). The idea behind boosting is that one averages successive classifiers that are derived by putting more and more weight on the points where misclassification occurred. The refinement was to recognize that boosting could be re-expressed as fitting an additive logistic regression model by minimizing a functional; see Friedman *et al.* (2000).

It is important to note that in this problem, for which a model is unlikely to exist, the classifiers that do best are those that do not rely on model-type assumptions. Indeed, the performance of the methods roughly corresponds inversely to the strength of the modeling assumptions. For logistic regression, the modeling assumptions are very strong and it performs worst. The nearest neighbors method implicitly makes the weakest modeling assumptions and it’s second worst. The trees method makes weaker modeling assumptions than logistic regression but not as weak as nearest neighbors, so its performance is intermediate between logistic and nearest neighbors. However, as the model becomes less important, as with SVMs, performance improves and, as modeling is abandoned completely in favor of relying on the data more and more, performance is improved further, as seen with random forests (which does best) and boosting (a very close second).

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

1.3 General Issues

The main point of the examples given in this chapter is that modeling is not as helpful as it is assumed to be, because its predictive performance can be improved rather handily by finding predictors not based on modeling. This was seen clearly in the Tour, Fires, and music data. If a worse predictor rests on modeling assumptions then, at a minimum, one is led to question the validity of those modeling assumptions. Indeed, in many realistic cases (i) models will be too complicated to formulate or (ii) simply do not exist. As noted, formulating a model for the Tour or Fires data is conceivable but practically impossible except in the most approximate sense. Also, it is unlikely that a model for the NGS data or for the music data exists at all. Even where one can imagine (with some effort) that a true model might exist, one should refer to data as coming from a data generator (DG) or source rather than a model since at best one will only be able to approximate the model, and the quality of that approximation will be hard to determine.

The better strategy for statistical analysis is to seek a predictor, not a model. The best way to find a good predictor is to look for one directly: choose a class of predictors that on the basis of experience and theory seems likely to perform well and find the best predictor in it. The optimal predictor can then be used to make predictions and to extract modeling information. For instance, one might be able to find a model that generates a predictor which approximates the optimal predictor in some sense. Or, one might be able to derive only some aspects of modeling directly from the predictor, such as identifying which variables are important even if the functional form by which they influence a response cannot be determined.

It is true that there are special, narrow, cases where modeling can be done quite accurately and the modeling gives predictions that are competitive with the best predictors. However, these cases are relatively rare in practice. When they occur, however, seeking the best predictor and seeking the best model may be equivalent, in which case it may be easier to find a predictor by obtaining it from a model. However, even though these cases exist, they are sufficiently rare that it is imprudent to build a statistical theory around the concept of a true model; better to build a statistical theory around quantities that make physical statements which can be compared with physical outcomes.

Another special case that bears mentioning is the setting where there is a true model that is extremely complex but can be well approximated in a verifiable way. This is usually the hope of subject matter specialists who seek statistical analysis. However, as much as this is a hope, it cannot be said to be realized in the absence of validation, which all too often is not even attempted. That is, it is not enough that the model 'fits' the data nor is it enough that inferences from the model be intuitively plausible. Many models will fit a given data set equally well, and if the inferences from a model are intuitively plausible that may mean only that the model has successfully encapsulated the intuition that went into it, i.e., the argument is effectively circular. If one is in the fortunate case that (i) a simple model 'fits', (ii) it gives predictions comparable with the best predictor, (iii) any inferences that one can derive from good predictors are consistent with the model, and (iv) other substantially different models that might fit well can be discarded for other reasons, then indeed one has had success. However, this is rare and mostly occurs in settings where there is so little variability that statistical analysis is not relevant.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press

As a generality, one approach that seems ill advised in many if not most problems is to seek a good model directly. This is so because of the likely deterioration in predictive performance resulting from using a model-based predictor rather than seeking a predictor from a more general class. Otherwise stated, model validation is likely to fail precisely because a good model was found instead of a good predictor. Moreover, to find a good model one must confront model uncertainty (and likely misspecification) or model nonexistence. The issue is therefore whether model uncertainty or the inability to specify a modeling approach is so severe as to preclude modeling as a useful approach to analysis. It may often be the case that the best modeling will result from finding a good predictor and trying to derive inferences about the DG from it.

The following chapters elaborate a predictive paradigm focusing on the properties of point prediction. The centerpiece is the goal of generating predictions on the same scale as the outcomes to be predicted. Decision-theoretic techniques for doing this have been proposed; however, this is not the same as constructing a theory of statistics centered on prediction. Comparing predictions with outcomes is different from asking whether an outcome materialized in a prediction interval, although the two are related: one can choose a scoring function (much the same as a loss function) that gives a cost of zero when the outcome is in the prediction interval (PI) and a cost of one when it isn't. So, at least in a limited sense, the use of PIs and scoring functions is conceptually included in the framework developed in the coming chapters. Likewise, probability forecasting, i.e., assigning probabilities to possible outcomes, is closely related to the approach taken here even though probabilities and outcomes are not on the same scale. In these settings one assesses the degree to which an outcome is representative of the probability of an event. Again, this is accommodated, at least in a limited sense, within the framework to be developed here.

https://doi.org/10.1017/9781139236003.002 Published online by Cambridge University Press