

An Introduction to Model Selection

Walter Zucchini

University of Göttingen

This paper is an introduction to model selection intended for nonspecialists who have knowledge of the statistical concepts covered in a typical first (occasionally second) statistics course. The intention is to explain the ideas that generate frequentist methodology for model selection, for example the Akaike information criterion, bootstrap criteria, and cross-validation criteria. Bayesian methods, including the Bayesian information criterion, are also mentioned in the context of the framework outlined in the paper. The ideas are illustrated using an example in which observations are available for the entire population of interest. This enables us to examine and to measure effects that are usually invisible, because in practical applications only a sample from the population is observed. The problem of selection bias, a hazard of which one needs to be aware in the context of model selection, is also discussed. © 2000 Academic Press

INTRODUCTION

The objective of this paper is to explain the issues involved in model selection to nonspecialists. This is an introduction to the subject, not a review of recent developments and methodology. Apart from the final section on selection bias the material presented here is an outline of the first four chapters of Linhart and Zucchini (1986) with the technical details reduced to a minimum. The focus is on frequentist methods although Bayesian methods are also mentioned to interpret them from the point of view adopted in this paper.

The examples used to illustrate the ideas are intended to provide simple, even exaggerated, concrete images to make the (unproved) general statements about model selection plausible; they are not offered as polished statistical analyses. In particular the data set on which the examples are based is not typical because observations are available for the entire population of interest. In practice these are available only for a sample from the population. This device has the advantage of enabling us to examine and to measure effects that are usually invisible in practice because we can, whenever we wish, remove the screen of uncertainty and assess precisely how well each model or selection method is performing.

I thank Michael Browne and two anonymous referees for their helpful comments on an earlier version of this paper. Correspondence and reprint requests should be addressed to Institut für Statistik and Ökonometrie, Universität Göttingen, Platz der Göttinger Sieben 5, D-37073 Göttingen, Germany.



The next section introduces the terminology and covers the basic issues that need to be considered when selecting a statistical model. I then outline how these ideas lead to model selection criteria, such as the Akaike information criterion (AIC), the bootstrap criterion, the cross-validation criterion and the Bayes information criterion (BIC). The final section explains the problem of selection bias, a serious hazard associated with model selection and one of which it is necessary to be aware.

THE BASIC IDEAS

A probability model is a useful concept for making sense of observations by regarding them as realizations of random variables, but the model that we can think of as having given rise to the observations is usually too complex to be described in every detail from the information available. The following example will be used to illustrate this and other model selection concepts. It concerns the ages and the number of visits to a General practitioner (GP) during 1995 by a well-defined group of 23,607 inhabitants of the Sydney suburb Ryde¹ which we will regard as the population of interest.

Figure 1 shows a plot of the number of visits against age for a random sample of 200 inhabitants from this population.² I will focus, for the moment, on the age distribution.

Being a *simple random sample* we can regard the observations as independent and identical realizations of a random variable having some non-negative-valued probability density function (pdf), specifically that shown in Fig. 2, which is based on the entire population. This probability distribution is the model we can regard as having given rise to the observations, the underlying model, or, as we will call it here, the *operating model*, in this case a particular pdf, $f(x)$. In practice the operating model is unknown because only a sample from the population is observed. The sample values are insufficient to faithfully reconstruct every detail of $f(x)$ but they can be used to estimate $f(x)$.

To estimate $f(x)$ we need to specify some *approximating family of models*, such as the following two. The first is the family of histograms (normalized so that the sum of the areas of the rectangles is equal to one) with $I=10$ equally spaced intervals. This family of models has nine parameters $\theta=(\theta_1, \theta_2, \dots, \theta_9)$, because to specify a particular *approximating model* in the family it is sufficient to give the heights of nine of the rectangles; the tenth is determined by the area constraint. Denote the pdf of these models by $g_{\theta}^{(10)}(x)$. The second is the family of histograms with $I=50$ equally spaced intervals, having 49 parameters $\theta=(\theta_1, \theta_2, \dots, \theta_{49})$ and pdf $g_{\theta}^{(50)}(x)$.

Before we can compare the performance of competing models we must decide what measure we intend to use to assess the fit or lack of fit. We will call a measure

¹ A description of these data is given in Heller (1997). Note, however, that the data used here are not identical to hers because they were extracted 12 months apart, during which interval the database was updated. Furthermore, I have truncated the number of visits to a maximum of 100. I thank Dr. Heller for helping me to obtain these data and Günter Kratz for his help in producing the figures.

² For convenience of explanation the sample was drawn with replacement but, as it turned out, no individual was drawn twice.

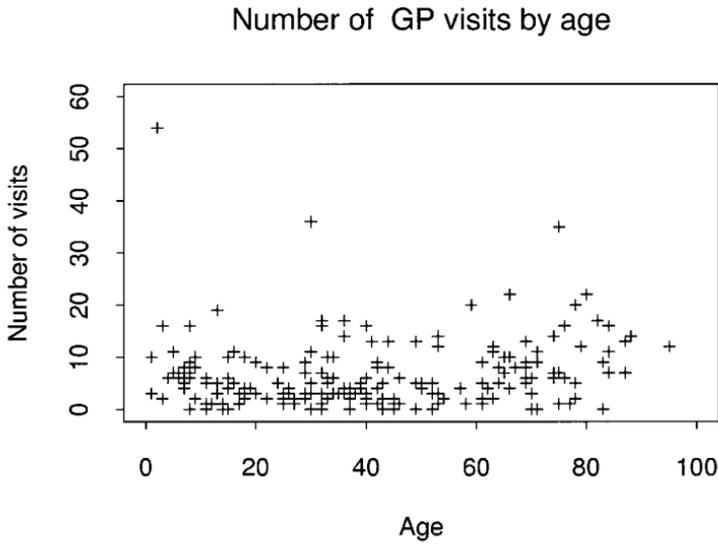


FIG. 1. The number of visits to a GP (in 1995) plotted against age for a simple random sample of 200 residents in Ryde.

of lack of fit a *discrepancy* and denote it by $\Delta(f, g_\theta)$. Some general-purpose discrepancies will be listed later but we are free to choose whatever discrepancy best suits the objectives of the envisaged statistical analysis. A possible discrepancy in our example is

$$\Delta(f, g_\theta^{(I)}(x)) = \int_0^{100} (f(x) - g_\theta^{(I)}(x))^2 dx ,$$

where $g_\theta^{(I)}(x)$ is the pdf of a histogram with I equally spaced intervals. As we happen to know $f(x)$ we can determine which approximating model within each of the



FIG. 2. The operating model, $f(x)$, for the age distribution of the Ryde population. (The ages in the database were recorded to the nearest year.)

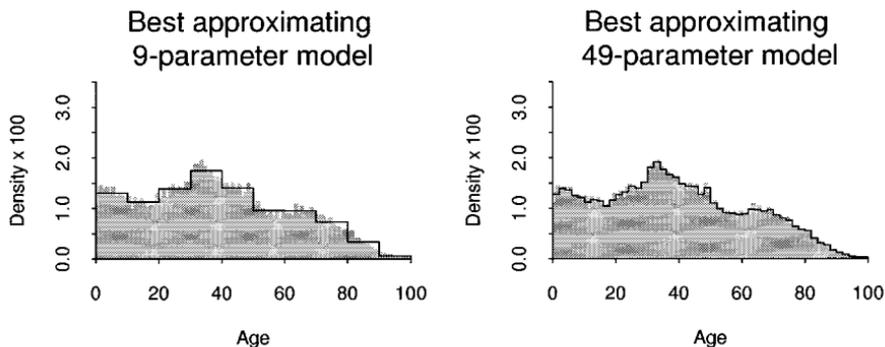


FIG. 3. The operating model for the age distribution of the Ryde population (shaded) and the best approximating models with 9 and with 49 parameters (dark lines).

two approximating families is best; that is, we can compute the parameter values, denoted by θ_0 , that minimize the discrepancy. (We simply construct the histograms using the ages for the entire population.) The models, $g_{\theta_0}^{(10)}(x)$ and $g_{\theta_0}^{(50)}(x)$, shown in Fig. 3, provide the best fits that we can ever obtain for models within each of the two contending families. Clearly the best model in the 49-parameter family provides a closer fit to $f(x)$ than does the best model in the 9-parameter family. This is to be expected because the two families are *nested*—any histogram in the latter family is also a histogram in the former family. However, even in nonnested families, the richness or flexibility of a family, manifested in our example as the variety of shapes it is able to produce, is largely determined by the number of parameters. For example the 49-parameter family of histograms considered above is more flexible than the family of normal distributions which has two parameters. The latter can only produce pdfs that are unimodal, symmetric, and bell-shaped.

We call the discrepancy between the operating model and the best approximating model the *discrepancy due to approximation*. It constitutes the lower bound for the discrepancy for models in the approximating family. In our example it is given by

$$A(f, g_{\theta_0}^{(I)}) = \int_0^{100} (f(x) - g_{\theta_0}^{(I)}(x))^2 dx = \begin{cases} 10 \times 10^5 & \text{for } I = 10 \\ 3 \times 10^5 & \text{for } I = 50. \end{cases}$$

In practice $f(x)$ is unknown and so we are not able to identify the best model in each family. The parameters have to be estimated from the observations. In our example we can use the sample relative frequencies (standardized so that the area under the histogram is equal to one); that is, $\hat{\theta}_i = n_i/n I/100$ where n_i is the number of observations that fell in the i th interval, $i = 1, 2, \dots, I$. The resulting *fitted models* for the sample of 200 Ryde residents, denoted by $g_{\hat{\theta}}^{(10)}(x)$ and $g_{\hat{\theta}}^{(50)}(x)$, are shown in Fig. 4. They differ from the best models $g_{\theta_0}^{(10)}(x)$ and $g_{\theta_0}^{(50)}(x)$.

The discrepancy between the fitted model and the best approximating model is called the *discrepancy due to estimation*. Here it is given by

$$A(g_{\hat{\theta}}^{(I)}, g_{\theta_0}^{(I)}) = \int_0^{100} (g_{\hat{\theta}}^{(I)}(x) - g_{\theta_0}^{(I)}(x))^2 dx = \begin{cases} 56 \times 10^5 & \text{for } I = 10 \\ 352 \times 10^5 & \text{for } I = 50. \end{cases}$$

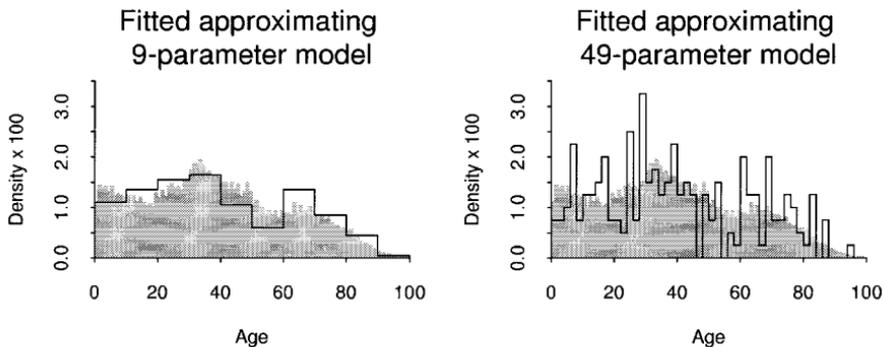


FIG. 4. The operating model (shaded) and the fitted models with 9 and with 49 parameters for the sample of 200 Ryde residents (dark lines).

Clearly the fitted 9-parameter model is much closer to the best model in its family than is the fitted 49-parameter model to the best model in its family. Note that whereas the discrepancy due to approximation did not depend on the sample values, the discrepancy due to estimation does; it would change if we used a different sample. In other words it is a random variable. The above two numerical values constitute the realizations of the discrepancy due to estimation for the sample and for the two families considered.

This example illustrates the general rule that it is necessary to take two things into account when comparing approximating families of different complexity. The best model in the more complex family is generally closer to the operating model than is the best model in the simpler family. However, the fitted model in the more complex family is likely to end up farther away from the best model than is the case in the simpler family. One can think of a complex family as having more *potential* than a simpler counterpart but that it tends to perform farther below its potential than the latter. The problem of model selection is that of finding an appropriate compromise between these two opposing properties, *potential* and *propensity to underperform*.

The *overall discrepancy*, defined as the discrepancy between the operating model and the fitted model, takes both the above factors into account. In our example it is

$$A(f, g_{\hat{\theta}}^{(I)}) = \int_0^{100} (f(x) - g_{\hat{\theta}}^{(I)}(x))^2 dx = \begin{cases} 67 \times 10^5 & \text{for } I = 10 \\ 355 \times 10^5 & \text{for } I = 50. \end{cases}$$

This overall discrepancy turns out to be the sum of its two component discrepancies, that due to approximation and that due to estimation. This is not true for all discrepancies but, even when it is not, the general rule given above remains valid. The discrepancy due to approximation favors flexible complex families, while that due to estimation favors rigid simple families. (See also Myung, 2000.)

The above values confirm what was clear from Fig. 4, namely that the nine-parameter histogram fits $f(x)$ better than does its more flexible but highly under-achieving competitor with 49 parameters. But the overall discrepancy, depending as it does on the parameter estimate $\hat{\theta}$, varies from sample to sample. It is also a

Distribution of the overall discrepancy

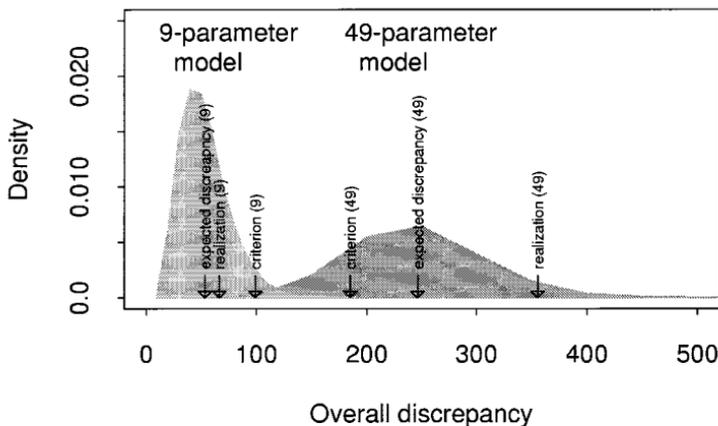


FIG. 5. The distributions of the overall discrepancies for $n = 200$ and the values of the overall discrepancy that were realized for the sample. The (entire) criterion, which is explained later in the paper, is also shown.

random variable and each of the above two numerical values constitutes only one realization from its distribution. A different sample would have led to different values. For some samples the fitted 49-parameter histogram will be closer to $f(x)$ than the 9-parameter histogram because, as we saw in Fig. 3, the 49-parameter family is certainly able to outperform the 9-parameter family.

To investigate how things turn out for different samples I generated 5000 random samples of size 200 from the population and, for each of them, computed the overall discrepancy for the resulting 9- and 49-parameter histograms. The resulting distributions of overall discrepancies, one for each family, are shown in Fig. 5. We see that it was no accident that, for the original sample, the 9-parameter histogram fitted $f(x)$ better than that with 49-parameters because the same is true for the vast majority of samples of size 200 from this population.

The above computations were only possible because we happened to know the operating model, which is not available in practice. (If it were then we would not bother to take a sample to estimate it.) In practice we are not in a position to compute any of the above discrepancies. They exist of course and will behave as outlined above but we cannot compute them.

Accepting that we cannot compute the overall discrepancy for our particular sample it would be of some help in deciding which family to select if we could compute its average value for samples of the given size, that is the *expected (overall) discrepancy*, $E\Delta(f, g_{\hat{\theta}})$. (See Fig. 5.) Unfortunately we cannot compute that either without knowing the operating model, *but we can estimate it*. An estimator of the expected discrepancy is called a (model selection) *criterion*.

In our example the expected discrepancy is given by (Linhart and Zucchini, 1986, p. 13):

$$E\Delta(f, g_{\hat{\theta}}^{(I)}) = \int_0^{100} f(x)^2 dx + \frac{1}{100n} \left(1 - (n+1) \sum_{i=1}^I \pi_i^2 \right),$$

where $\pi_i = \int_{100(i-1)/I}^{100i/I} f(x) dx$, $i = 1, 2, \dots, I$. The first term is the same for both approximating families and therefore can be ignored for the purposes of comparing the two families; the second term is the essential one.

An unbiased estimator of the second term (which is also referred to as a criterion even though it does not estimate the entire expected discrepancy, only its essential part) is given by

$$\text{Criterion} = \frac{I}{100n} \left[1 - \frac{n+1}{n-1} \left(\sum_{i=1}^I \frac{n_i^2}{n} - 1 \right) \right] = \begin{cases} -1149 \times 10^5 & \text{for } I=10 \\ -1063 \times 10^5 & \text{for } I=50. \end{cases}$$

Note that the criteria shown in Fig. 5 *do* include the (otherwise inessential) first term so that we can see where these criteria landed relative to the quantities they are attempting to estimate, namely their respective expected discrepancies. The criterion for the 9-parameter family is smaller and would have led us to select (what we know to be) the better model. Unfortunately this is not always the case. The expected discrepancy is a complex quantity that depends on several things: the operating model, the approximating family, the method used to estimate the parameters of the fitted model, and the sample size. It is therefore not surprising that in many situations even unbiased criteria are rather imprecise and do not always identify the best family.

To illustrate the importance of the sample size in model selection we now investigate how the mean number of GP visits, $\mu(a)$, varied with age, a . The operating model now is the bivariate distribution of the two quantities, age and number of visits, for all individuals in the Ryde population. To estimate the operating mean, $\mu(a)$, I will use a p -parameter polynomial as approximating mean,

$$v_{\theta}^{(p)}(a) = \theta_1 + \theta_2 a + \theta_3 a^2 + \dots + \theta_p a^{p-1},$$

and the discrepancy

$$A(f, g_{\theta}) = \sum_{a=1}^{90} (\mu(a) - v_{\theta}^{(p)}(a))^2.$$

Note that this discrepancy depends only on the means of the operating and approximating models. I have restricted the discrepancy to the age range 1 to 90 for convenience of illustration.³ Figure 6 shows the operating mean and the best approximating polynomials for $p = 2, 3, 4, 8$ parameters. Again, as the number of parameters increases, so the fit improves and the discrepancy due to approximation decreases.

Figure 7 shows the fitted means, $v_{\theta}^{(p)}(a)$, for the original sample and for 20 additional random samples of size 200 from the population. The method of ordinary least squares was used to fit the polynomials.

³ This makes the graphs clearer. I also took the liberty of rejuvenating the single individual in the sample aged over 90 from 95 to 90.

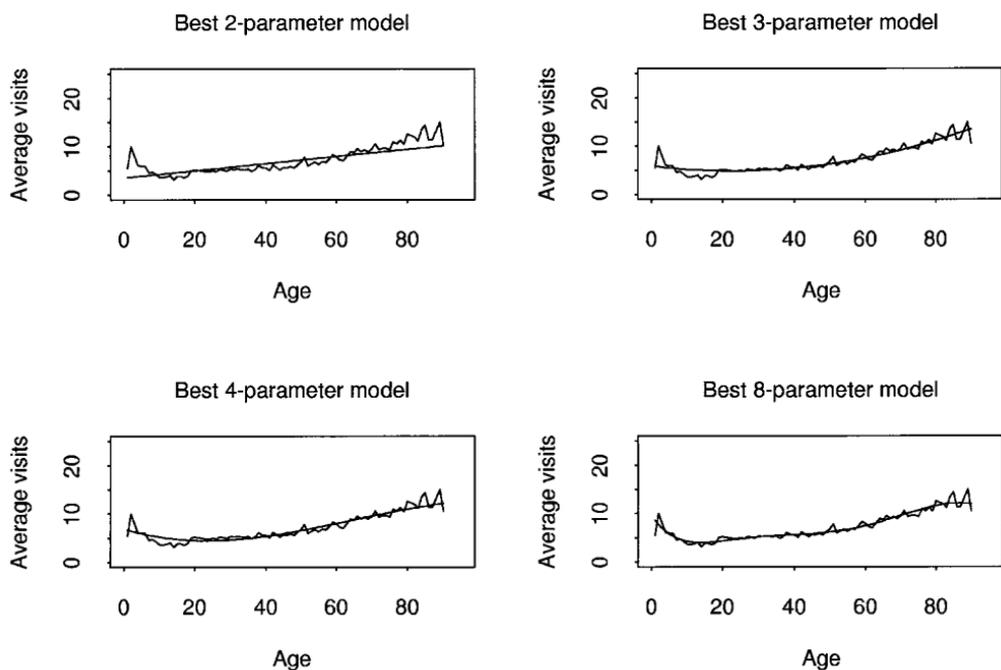


FIG. 6. Operating mean number of GP visits by age (irregular line) and the best approximating polynomial with p parameters (smooth lines).

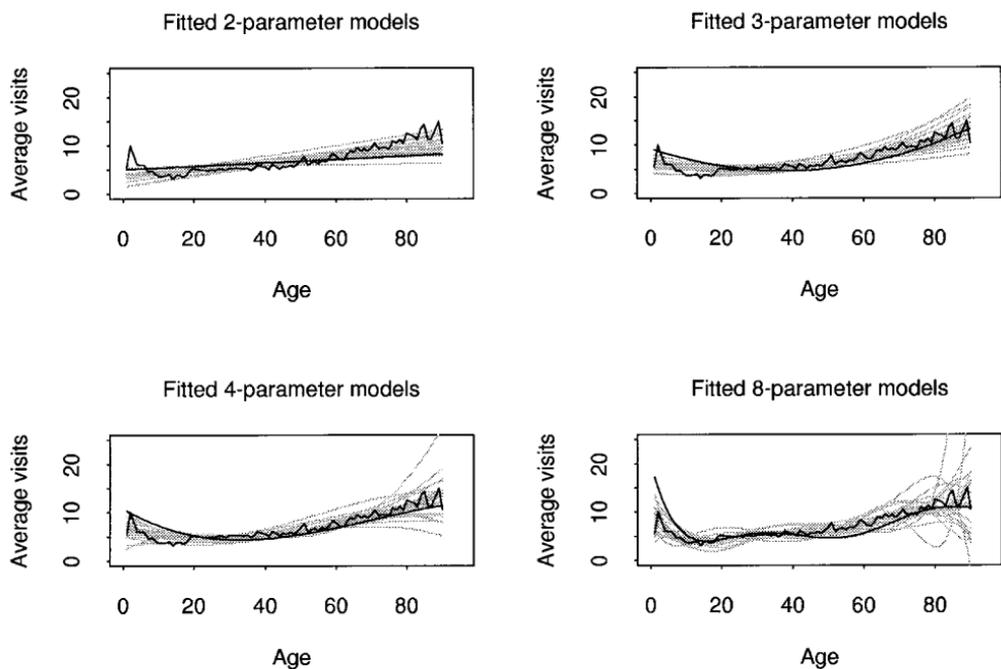


FIG. 7. Operating mean number of GP visits by age (irregular line), the mean fitted to the original sample (dark smooth line) and to 20 additional random samples (faint lines) using approximating polynomials with p parameters.

It can be seen that the estimates become less stable as p increases. That is because the discrepancy due to estimation increases. The degree of instability depends on the sample size. This is illustrated in Fig. 8 which summarizes the situation for samples of size $n = 200, 500, 1000,$ and 5000 . The expected overall discrepancies (and expected discrepancies due to estimation) in that figure were approximated using 100 samples of size n .

Figures 6, 7, and 8 illustrate a number of general points. First, the degree of model complexity that is appropriate depends substantially on the sample size. For $n = 200$ the best fit is expected with $p = 3$ parameters but for $n = 5000$ it is expected for $p = 7$. In general only simple families are stable when the sample size is small. As the sample size increases it becomes feasible to use more complex approximating families which are able to reflect smaller details of the operating model. (See Fig. 6.) Such details can be meaningfully estimated when the sample size is large because the model parameters can be estimated accurately.

The second point is that the family that is best on average (has the smallest expected discrepancy) is not necessarily much better than the second best, third best, and so on. This is illustrated in Fig. 8, especially for $n = 1000$, a case in which

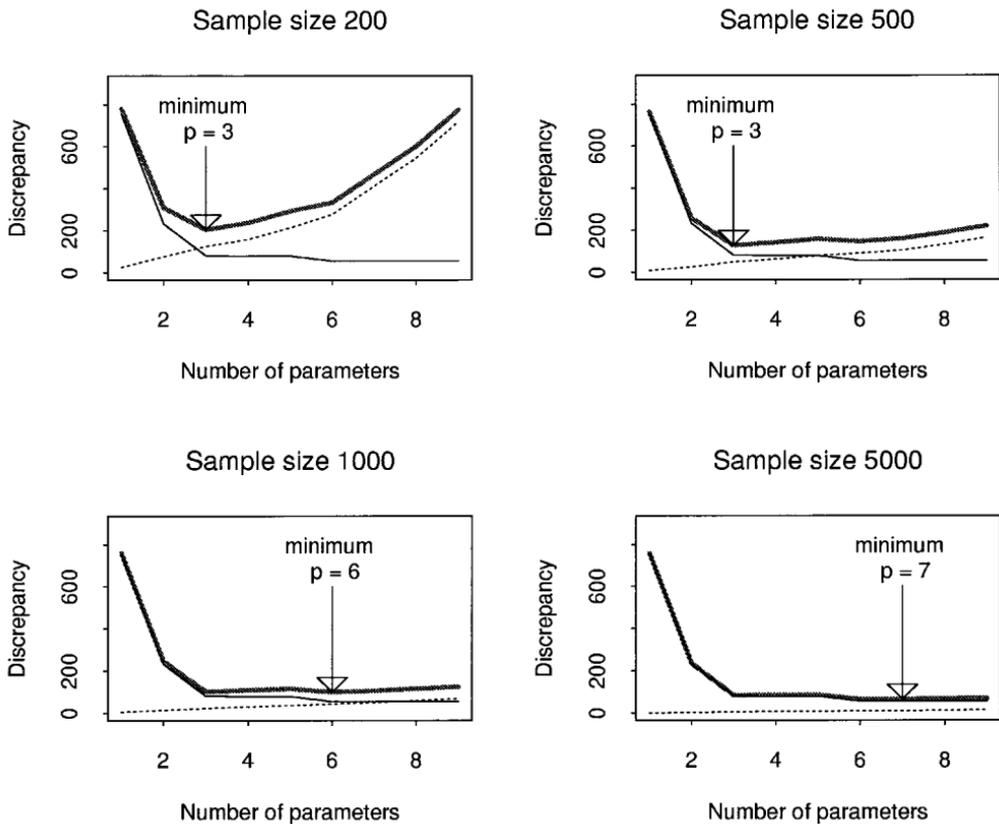


FIG. 8. The discrepancy due to approximation (solid thin line) does not depend on the sample size and so it is the same in all four graphs. The average (over 100 replications) discrepancy due to estimation (broken line) does depend on sample size and, for fixed p , decreases as the sample size increases. The expected overall discrepancy (thick line) decreases at first and then increases as a function of p .

the expected discrepancy is almost constant for $3 \leq p \leq 9$. In other words the corresponding polynomials are effectively equally accurate, or inaccurate, on average.

Third, we must consider that the expected discrepancy is not available in practice; it can only be *estimated* using a criterion. Since the approximating family that minimizes the criterion is not necessarily that which minimizes the expected discrepancy, it makes sense to examine the behavior of the criterion over the whole range of contending families and not just to automatically select the family that minimizes it.

GENERAL-PURPOSE DISCREPANCIES AND CRITERIA

In what follows the distribution function of the operating model is denoted by F and its pdf (or probability function in the case of discrete distributions) by f . The corresponding functions for an approximating family with parameter vector θ are denoted by G_θ and g_θ , respectively. Finally, F_n is used to denote the *empirical distribution function*, that is

$$F_n(x) = (\text{Number of observations in the sample } \leq x) / n .$$

Discrepancies can be, indeed should be, selected to match the objectives of the analysis. If, when estimating $\mu(a)$ for the Ryde population, we wished to emphasize the fit in a particular age group, say $40 \leq a \leq 60$, we could do this by changing the discrepancy to

$$\Delta(f, g_\theta) = \sum_{a=1}^{90} w(a)(\mu(a) - v_\theta^{(p)}(a))^2,$$

with, for example, the weights $w(a) = 2$ for $40 \leq a \leq 60$ and $w(a) = 1$ for the remaining ages. It would then be consistent with the above objective to use the method of weighted (rather than ordinary) least squares, with these weights, to estimate the parameters because that would also emphasize the fit in the age range of interest.

This illustrates the point that a natural estimator to use in conjunction with a particular discrepancy is the *minimum discrepancy estimator* or, as it is usually called in the literature, the minimum distance estimator. (See Parr, 1981.) In our context this is the estimator, $\hat{\theta}$, that minimizes what we call the *empirical discrepancy* (Linhart and Zucchini, 1986, p. 12). In most cases of interest to us here this is simply the discrepancy between the approximating model and the model obtained if one regards the sample as if it were the population. In other words it is the discrepancy between the approximating model and the empirical distribution function, $F_n(x)$, namely $\Delta(F_n, G_\theta)$, or briefly $\Delta_n(\theta)$. An example of an empirical discrepancy will be given later.

The method of maximum likelihood, an important general-purpose method of estimation, is the natural partner (minimum discrepancy estimator) for:

The *Kullback–Leibler discrepancy*,

$$\Delta_{\text{K-L}}(f, g_\theta) = -E_F \log g_\theta(x) = - \int \log g_\theta(x) f(x) dx,$$

where for discrete distributions the integral is replaced by a sum. This discrepancy focuses on the expected log-likelihood when the approximating model g_θ is used; the higher the expected log-likelihood, the better the model. Roughly speaking this discrepancy deems an approximating model good if, on average, that model assigns a high probability to sample observations. A good model makes the data seem likely; a bad model makes them seem unlikely. As we will see later it is the discrepancy associated with the AIC.

The *Pearson chi-squared discrepancy*,

$$\Delta_{\text{P}}(f, g_\theta) = \sum_x (f(x) - g_\theta(x))^2 / g_\theta(x), \quad g_\theta(x) \neq 0,$$

is a useful general-purpose discrepancy for discrete data or grouped data. We note that it would be (logically) consistent to use minimum chi-squared estimation to estimate the parameter θ in conjunction with this discrepancy but this is not essential. One can, for example, use the method of maximum likelihood to estimate the parameters and the above discrepancy to assess the fit.

The *Gauss discrepancy* is given by

$$\Delta_{\text{G}}(f, g_\theta) = \sum_x (f(x) - g_\theta(x))^2,$$

where for continuous distributions one replaces the sum by an integral and the probability functions by densities, as we did when selecting a histogram for the age distribution. Additional examples of general- and specific-purpose discrepancies are discussed in Linhart and Zucchini (1986).

Having decided which approximating families will be considered, which method will be used to estimate the parameters, and which discrepancy will be used to assess the fit, the next step is to find a criterion, an estimator of the expected discrepancy that we can use to rank the contending models. The derivation of criteria is a technical issue beyond the scope of this paper. I will confine myself to some general remarks.

In some contexts, including regression, analysis of variance, and covariance, criteria are available that are unbiased for *finite samples*. The well-known C_p and S_p variable selection criteria are examples of this type; another example is the criterion used earlier to select a histogram for the age distribution. Where such criteria are not available three approaches are currently available (apart from Bayesian methods), namely asymptotic methods, bootstrap methods, and cross-validation methods.

Asymptotic Methods

One can go about obtaining a criterion by first deriving a formula for the expected discrepancy and then finding a way to estimate it. In many contexts it is not possible to obtain a useful expression for the exact expected discrepancy (which makes it rather difficult to invent some way of estimating it). Examples of this type arise when one is selecting families of models for univariate probability distributions, contingency tables, and time series. However, in many such cases it is possible to derive an expression for the asymptotic value of expected discrepancy, that is, its limiting value as the sample increases indefinitely, and also to find an (asymptotically) unbiased estimator of that value, called an *asymptotic criterion*.

Very briefly and under a long list of so-called regularity conditions (Linhart and Zucchini, 1986, Appendix A1) as the sample size, n , increases so the expected discrepancy approaches the form:

$$EA(f, g_{\hat{\theta}}) = A(\theta_0) + K/2n.$$

The first term is the discrepancy due to approximation and K , which is called the *trace term* (because it happens to be the trace of the product of two matrices) represents a rather complicated expression for which, however, it is sometimes possible to provide an (asymptotically) unbiased estimator, \hat{K} . If so, the asymptotic criterion is given by

$$C = \hat{E}A(f, g_{\hat{\theta}}) = A_n(\hat{\theta}) + \hat{K}/n.$$

The first term is the empirical discrepancy evaluated for the fitted model (and note that the 2 has disappeared).

Clearly, the practical usefulness of asymptotic criteria is determined by how well they estimate the expected discrepancy for the *finite* sample size that is available; that they are asymptotically unbiased does not make them unbiased for *finite* samples. A second point is their standard error can be quite large. Finally their performance can only be assessed using tedious Monte Carlo simulations for specific models.

At this stage the reader might be forgiven for concluding that the situation is unsatisfactory but, fortunately, a simple alternative is available for some discrepancies, including the important cases of the Kullback–Leibler and the Pearson chi-squared. It can be shown (Linhart and Zucchini, 1986, Appendix A1) that *if the operating model belongs to the approximating family* then the term K reduces to a simple function of the number of parameters, p , in the approximating family and so the criterion simplifies enormously. We then refer to it as the *simple (asymptotic) criterion*. In the case of the Kullback–Leibler discrepancy K reduces to p and so \hat{K} can be replaced by p which leads to the simple criterion

$$C_{\hat{K}-L}^* = A_n(\hat{\theta}) + p/n.$$

(The subscript K - L stands for Kullback–Leibler and the $*$ is used to indicate that this is the simple criterion.) This is a somewhat disguised, but strictly equivalent, form of the well-known Akaike information criterion (Akaike, 1973). In fact

$$C_{K-L}^* = \text{AIC}/2n, \quad \text{where} \quad \text{AIC} = -2 \log(L) + 2p$$

and where L refers to the likelihood under the fitted model.

To derive C_{K-L}^* we made an unrealistic assumption, namely that the operating model belongs to the approximating family. Consider, for example, the 9-parameter-histogram approximating family that we looked at earlier to model the age distribution in the Ryde population. Even if we did not happen to know the operating model (that is, the age of everyone in the Ryde population) the assumption that $f(x)$ is precisely such a histogram would be more than a little far-fetched. Nevertheless, it does not necessarily follow from this that the resulting criterion, however it might have been derived, is a bad one. Indeed, the simple criterion, C_{K-L}^* , often outperforms the more complex alternative, C_{K-L} , that avoids the offending assumption, so long as the discrepancy due to approximation is not excessively large, that is unless the best model in the approximating family differs grossly from the operating model.

The reason why C_{K-L}^* often outperforms C_{K-L} is that although p might be a biased estimator of K , whereas \hat{K} is (asymptotically) unbiased, the latter has a high standard error, whereas p (a constant) has zero standard error. Hurvich and Tsai (1989) derived a refinement, $\text{AIC}_e = \text{AIC} + 2p(p+1)/(n-p-1)$, to reduce the small-sample bias of the AIC for regression and time series models.

The AIC is of the same general form as the Schwarz criterion (Schwarz, 1978) or, as it is also called, the Bayesian information criterion:

$$\text{BIC} = -2 \log(L) + p \log(n).$$

Note that the BIC differs from the AIC only in the second term which now depends on the sample size n . Clearly, as n increases, the BIC favors simpler approximating families (that is families with a smaller number of parameters p) than does the AIC. But despite the superficial similarity between the AIC and BIC the latter is derived in a very different way and within a Bayesian framework. The following brief description (see, e.g., Raftery (1995) or Wasserman (2000) for substantial accounts) is intended to provide an interpretation of the BIC in the framework outlined in this paper.

Suppose that we have two competing families, G_1 and G_2 . (In the Ryde example G_1 could represent the family of histograms with 10 intervals and G_2 that with 50 intervals.) Denote the sample information (the data) by D . One begins by assigning prior probabilities, $P(G_i)$, $i=1, 2$, to the event that family i is correct or, in our terminology, the event that family i contains the operating model. One also assigns prior distributions to the model parameters in each family. This enables one to compute the *integrated likelihood*, $P(D | G_i)$, which can be interpreted as the

likelihood of the observed values if family G_i is correct. The next step is to apply Bayes' theorem to compute $P(G_i | D)$, the posterior probability that family i is correct given the observed values. A measure of the extent to which the data support family G_2 over G_1 is given by the *posterior odds*:

$$\frac{P(G_2 | D)}{P(G_1 | D)} = \frac{P(D | G_2) P(G_2)}{P(D | G_1) P(G_1)}.$$

The first quotient on the right-hand side is called the *Bayes factor* and the second the *prior odds*. The Bayes factor can be interpreted as a measure of the extent to which the data support G_2 over G_1 when the prior odds are equal to one. The prior odds are equal to one if, prior to examining the data, the families G_1 and G_2 are regarded as equally plausible, that is equally likely to contain the operating model. The BIC can be shown to be a large sample approximation to the logarithm of the Bayes factor.

One of the reasons for using the BIC rather than the Bayes factor itself is that the computations needed to evaluate the latter can be enormous, especially as it is necessary to specify prior distributions for the parameters in each of the (often numerous) competing families. However, this obstacle is being overcome (e.g., DiCiccio, Kass, Raftery, and Wasserman, 1997) and there is a growing literature on interesting applications in which Bayes factors are used for model selection (e.g., Kass and Raftery, 1995; Albert and Chib, 1997).

In the terminology of this paper the computation of Bayes factors requires one to assign a prior probability to each model in each approximating family, although in the Bayesian framework these are not regarded as *approximating families* but as families that potentially contain the operating model. To each individual model one must assign a (subjective) probability that it is *strictly correct*. In the Ryde age-distribution example this amounts to assigning a prior probability to each possible histogram with 10 intervals so that it is identical to the operating model (and to those with 50 intervals). In this example, even if one did not know the operating model, it must surely be regarded hopelessly unlikely that any of these histograms might be precisely identical to the operating model. Indeed it is difficult to imagine any family that one might normally consider and to which one could assign a priori probability *other than zero* to the event that it contains an operating model such as that given in Fig. 2.

Thus if one takes the view that operating models are generally vastly more complex than any model one is likely to consider fitting to them in practice, then, strictly speaking, the Bayesian approach to model selection is not applicable. On the other hand, as we saw earlier, the AIC is also derived under the assumption that the operating model belongs to the approximating family. The main point is to recognize how these two approaches to model selection differ. The frequentist approach accepts up-front that the approximating models are not necessarily the real thing and attempts to identify the family that leads to the best fit, on average. The Bayesian approach regards every contending model as potentially constituting the real thing and then estimates, for each model, the probability of it being that.

Bootstrap Methods

Bootstrap methods, introduced by Efron (1979), provide a simple means of circumventing the type of technical difficulties mentioned above. Recall that in the age distribution example we were able to compute *all* the discrepancies of interest because we knew the operating model. For example, by drawing repeated random samples from the population we were even able to approximate the distributions of the two overall discrepancies (given in Fig. 5). The point here is that if we know the operating model we can always compute the expected discrepancy by using a formula, if this is available, or by resorting to Monte Carlo methods if it is not. Now if we regard the *sample as a population* then the operating model for this mini-population is known and so we can compute the corresponding expected discrepancy. The expected discrepancy for the mini-population is an estimator of the expected discrepancy for the real population and is called the *bootstrap criterion*. Chung, Lee, and Koo (1996) have shown that the bootstrap criterion has a downward bias and recommend a simple adjustment which, in our notation, amounts to adding of p/n .

Cross-Validation Methods

The idea here is to split the sample data into two subsamples, a *calibration sample* of size $n - m$ and a *validation sample* of size m ; the first is used to fit the model and the second to estimate the expected discrepancy. Such an estimator is called a cross-validation criterion.

There is a problem in deciding how to select m , the number of observations allocated to the validation sample. For example, if we select $m = n/2$ then only $n/2$ observations are available to fit the model and so (in the validation step) we would be judging the model by its performance for a sample size of $n/2$. As we saw in Fig. 8 the sample size is an important factor in determining which approximating family is best, on average, and our objective is to find the best family for a sample of size n (the number of observations we in fact have) not for a sample of size $n/2$. We could reduce m but that would leave fewer observations for the validation sample and thereby erode the accuracy with which we can estimate the expected discrepancy.

The following idea is used to circumvent the above problem (see, e.g., Stone, 1974): One uses a small m , even $m = 1$ (referred to as one-item-out cross-validation) but one repeats the following steps for *all possible calibration samples* of size $n - m$:

Step 1: Fit the model to the calibration sample.

Step 2: Estimate the expected discrepancy for the fitted model using the validation sample.

The cross-validation criterion is the average, over these repetitions, of the estimates obtained of step 2. For a more comprehensive account of cross-validation methods see Browne (2000).

An Example

We now illustrate the above methods using the Kullback–Leibler discrepancy to select a model for the (marginal) distribution of the number of GP visits for the

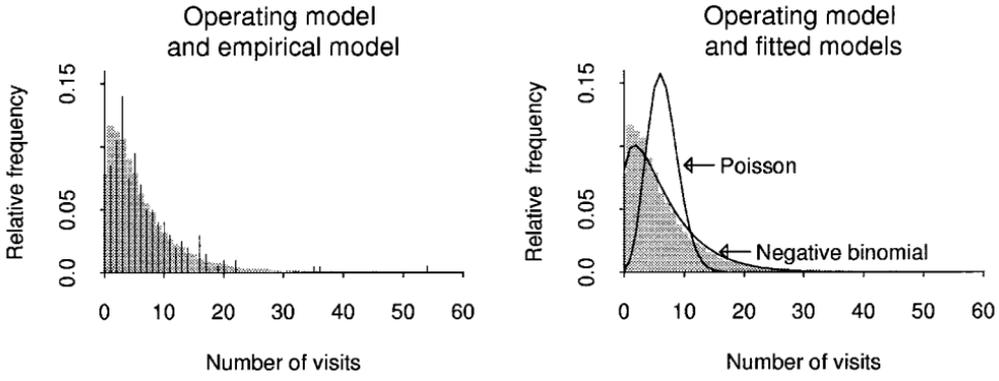


FIG. 9. The graph on the left shows the operating model (shaded) and empirical model, that is the observed relative frequencies (thin dark lines). The graph on the right shows the operating model and the fitted Poisson and negative binomial distributions.

Ryde sample shown in Fig. 1. The operating model (that is the relative frequencies of GP visits for all residents of Ryde) and the empirical model (the relative frequencies for the sample of 200 residents) are shown in Fig. 9. Also shown are two competing models that were fitted to the sample data using the method of maximum likelihood, namely the Poisson (which fits poorly) and the negative binomial (which fits better).

Again, as we are in possession of the operating model, it is possible to compute the distributions of the overall discrepancy for each of the two approximating families. I used 1000 samples of size 200 to compute those shown in Fig. 10.

The probability function for the Poisson distribution with parameter λ and that for the negative binomial distribution with parameters α and β are, respectively,

$$g_{\lambda}(x) = \lambda^x e^{-\lambda} / x!,$$

$$g_{\alpha, \beta}(x) = \Gamma(x + \beta) (1 - \alpha)^{\beta} \alpha^x / (\Gamma(\beta) \Gamma(x + 1)), \quad x = 0, 1, 2, \dots$$

Asymptotically unbiased estimators, \hat{K} , of the trace term are available for both of these distributions (Linhart and Zucchini, 1986, p. 46), but that for the negative binomial is rather complicated and will not be given here. For the Poisson distribution $\hat{K} = m_2 / m'_1$, where m_2 is the sample variance and m'_1 is the sample mean. The asymptotic criterion for the Poisson is thus given by

$$C_{K-L}(\text{Poisson}) = \Delta_n(\hat{\lambda}) + m_2 / m'_1 n, \quad \text{where} \quad \Delta_n(\hat{\lambda}) = - \sum_{i=1}^n \log g_{\hat{\lambda}}(x_i) / n,$$

and $\hat{\lambda} = m'_1$ is the maximum likelihood estimator of λ .

The simple criterion (equivalent to the AIC) simply uses $\hat{K} = 1$ for the one-parameter Poisson model and $\hat{K} = 2$ for the 2-parameter negative binomial. The latter reduces to

$$C_{K-L}^*(\text{neg. binomial}) = \Delta_n(\hat{\alpha}, \hat{\beta}) + 2/n, \quad \text{where} \quad \Delta_n(\hat{\alpha}, \hat{\beta}) = - \sum_{i=1}^n \log g_{\hat{\alpha}, \hat{\beta}}(x_i) / n,$$

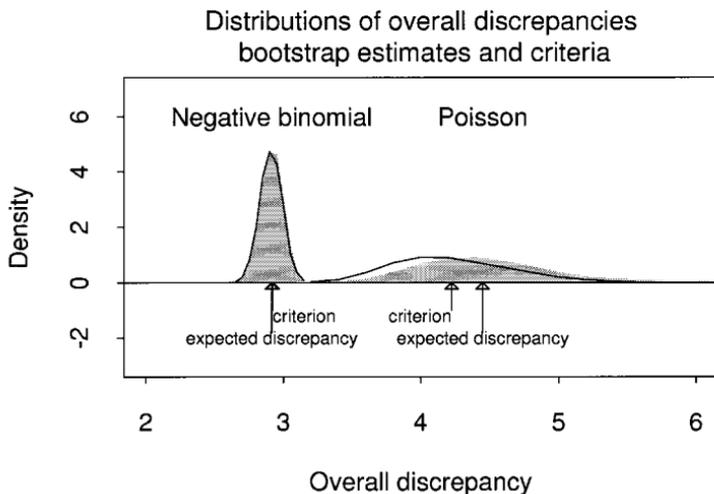


FIG. 10. The distributions of the overall discrepancies (shaded) for the two approximating families and the bootstrap estimates of these two distributions (thin lines). Also shown are the expected discrepancies and criteria. (All the criteria given in Table 1 coincide on this graph.)

where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimators of α and β . (See, e.g., Linhart and Zucchini, 1986, p. 46.)

To compute the *bootstrap criterion* we repeat the following two steps a large number of times (Fig. 10 is based on 1000 replications):

1. Take a random sample of size n with *replacement from the sample*. This is called a bootstrap sample.
2. Compute the maximum likelihood estimators ($\hat{\lambda}^*$ for the Poisson; $\hat{\alpha}^*$ and $\hat{\beta}^*$ for the negative binomial) using the values in the bootstrap sample. For the Poisson model compute $A_n(\hat{\lambda}^*)$ and for the negative binomial $A_n(\hat{\alpha}^*, \hat{\beta}^*)$.

The values of the 1000 pairs of discrepancies computed in step 2 were used to obtain the estimates of the two distributions of the overall discrepancy that are shown in Fig. 10. In this example these bootstrap distributions are quite close to the distributions they are intended to estimate. In part this is because we are fitting families with very few parameters. For families which have more parameters the agreement tends to be less precise. The average of the $A_n(\hat{\lambda}^*)$ and that of the $A_n(\hat{\alpha}^*, \hat{\beta}^*)$ values computed in step 2 are the bootstrap criteria for the respective families and are given in Table 1.

Finally, the *cross-validation discrepancy* can be computed as follows (though it can be done more efficiently). We repeat the following two steps for $i = 1, 2, \dots, n$:

1. Assemble the i th calibration sample by taking the original sample and removing the i th observation, x_i .
2. Compute the maximum likelihood estimates ($\hat{\lambda}^{(i)}$ for the Poisson; $\hat{\alpha}^{(i)}$ and $\hat{\beta}^{(i)}$ for the negative binomial) using the values in the i th calibration sample. For the Poisson model compute $CV_i(\text{Poisson}) = -\log g_{\hat{\lambda}^{(i)}}(x_i)$ and for the negative binomial model $CV_i(\text{neg. binomial}) = -\log g_{\hat{\alpha}^{(i)}, \hat{\beta}^{(i)}}(x_i)$.

TABLE 1
The Criteria and the Expected Discrepancy

Criterion	Poisson	Negative binomial
Asymptotic	4.25	2.92
Simple asymptotic (AIC/2n)	4.23	2.92
Bootstrap	4.23	2.91
Cross-validation	4.22	2.93
BIC/2n	4.23	2.94
Expected discrepancy	4.45	2.91

Note. Values of the criteria and the expected discrepancy for the Poisson and negative binomial distributions. As indicated in the text the BIC was not specifically designed to estimate the expected discrepancy. The AIC and BIC criteria have been divided by $2n$ to make them comparable to the others.

The cross-validation criterion for the Poisson is the average of the n values $CV_i(\text{Poisson})$ computed in step 2 and that for the negative binomial model is the average of the n values $CV_i(\text{neg. binomial})$, $i = 1, 2, \dots, n$.

All the above criteria (given in Table 1) are practically equal within each of the two models and they cannot be distinguished at the resolution displayed in Fig. 10. That is because the two families considered have very few parameters and so the overall discrepancy is dominated by the discrepancy due to approximation. The criteria all indicate that the negative binomial is the better of the two distributions for these data.

SELECTION BIAS

The objectivity of formal model selection procedures and the ease with which they can be applied with increasingly powerful computers on increasingly complex problems has tended to obscure the fact that too much selection can do more harm than good. An overdose of selection manifests itself in a problem called selection bias which occurs when one uses the same data to select a model and also to carry out statistical inference, for example to compute a confidence interval on the basis of the selected model. The purpose of this section is to explain the problem; the solution is still being invented. A more comprehensive account of selection bias is given in Chatfield (1995) and the discussions that follow that paper.

The expected discrepancy associated with each approximating family depends on two things that we know (the sample size and the method used to estimate the parameters) and one thing that we do not know (the operating model) and so we cannot compute it. If we could compute it we could arrange the contending families in order of increasing expected discrepancy, that is from best to worst, on average. This ideal list, *List A*, is not available but we can compute estimates of each expected discrepancy, the criteria, and use those to compile a *List B* in which the contending families are arranged in order of increasing criterion, that is from *apparently best* to *apparently worst*.

In general the rankings in *List B* will differ from those in *List A*. For some families the criterion will be higher than the expected discrepancy and for others it will be lower, depending on the details of the particular sample that was drawn, that is depending on the *luck of the draw*.

Those families whose criteria happen to have underestimated the expected discrepancy will tend to be rated higher on *List B* than on *List A* and vice versa for families whose criteria happen to have overestimated their expected discrepancy. In Fig. 5 the expected discrepancy of the 9-parameter family is smaller than that of the 49-parameter family and so the former is ranked higher on *List A*. The criterion for the 9-parameter family landed above (on the *unlucky* side of) its expectation and that for the 49-parameter family fell below (on the *lucky* side of) its expectation, but this did not matter because the criteria did rank the families correctly. Thus, *List B* and *List A* are the same for this sample and so the top family in *List B* is indeed the best family.

However, in general, we cannot be certain that the top family in *List B* is indeed the best. It might be the second best that was moderately lucky, or it might be the 73rd best that was very lucky. The top position in *List B* is biased in favor of families that were lucky or, more precisely, families whose criteria underestimated the expected discrepancy for the available sample. This bias is especially relevant when the number of approximating families considered for selection is large, and in practice that number can get very large.

That, in turn, gives rise to a second problem, referred to as *selection bias*: *The top model in List B will, in general, appear to perform better than it really does.*

We consider how the above two problems are manifested in the context of variable selection in multiple linear regression with a large number of predictors. The usual objective here is to identify that subset of the predictors that minimizes some expected discrepancy, such as the mean squared error of prediction.

The standard *all-subsets variable selection* procedure in multiple linear regression with 15 potential predictors examines $2^{15} = 32,768$ approximating families—even more if one is also investigating transformations of the predictors, such as their logarithms, squares, or cross-products. Thus the equivalent of Fig. 5 will have 32,768, or more, distributions instead of two and furthermore many of them would overlap more substantially than the two unusually well-separated distributions shown in Fig. 5. Consequently some of the variables that appear in the selected subset (the one that is ranked first in *List B*) are there by good luck rather than by merit. (For more details and examples see Miller (1990) and the papers referenced in Chatfield (1995, Section 2.3.)) Furthermore, in the absence of additional evidence, it is not possible to determine which of the selected variables should be taken seriously, that is how to separate the wheat from the chaff.

The second problem is that the selected family will, on average, give an optimistic impression of how well it fits. In the context of multiple regression this is manifested in a substantial (or even gross) underestimate of the residual variance. The selected predictors will *appear to be* more accurate than they turn out to be when predicting *future values* of the dependent variable.

The first of the two problems is insoluble; it is not possible to identify with certainty which approximating family is the top of *List A* on the basis of *List B*. In

many situations one would be content to be confident that the top-ranking family in *List B* is, if not the best, then at least not a bad one that was lucky. The risk that a bad approximating family will make it to the top of *List B* can be reduced by restricting selection to a small number of well-considered families. The risk becomes substantial if one casts a large selection net intended to cover every effect and complication that could conceivably be relevant.

Progress has been made toward solving the problem of selection bias. It is an advantage of the Bayesian approach (see, e.g., Raftery, 1995) that the methodology exists to address this issue, even if it is computationally demanding. Briefly, one does not restrict one's attention to a single selected family; one works instead with all the models that were considered, weighting the contribution of each (via the posterior probability that is the correct model) to adjust for model uncertainty.

A frequentist approach for incorporating model uncertainty into statistical inference has been suggested by Buckland, Burnham, and Augustin (1997). They also base inference on a suitably weighted linear combination of estimates from all the contending models. Ye (1998) has introduced the concept of generalized degrees of freedom to correct for selection bias.

Note added in proof. Some important publications have appeared subsequent to the preparation of this article. In particular the book by K. P. Burnham and D. R. Anderson, "Model Selection and Inference: A Practical Information-Theoretic Approach," Springer (1998) provides an excellent account of frequentist methods.

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory*, pp. 267–281. Budapest: Akademiai Kiado.
- Albert, J., & Chib, S. (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association*, **92**, 916–925.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108–132.
- Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, **53**, 603–618.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, **158**, 419–466.
- Chung, H.-Y., Lee, K.-W., & Koo, J.-Y. (1996). A note on bootstrap model selection criterion. *Statistic & Probability Letters*, **26**, 35–41.
- DiCiccio, T. J., Kass, R. E., Raftery, A., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903–915.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.
- Heller, G. Z. (1997). Who visits the GP? Demographic patterns in a Sydney suburb. Technical report, Department of Statistics, Macquarie University.
- Hurvich, C. M., & Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
- Linhart, H., & Zucchini, W. (1986). *Model selection*. New York: Wiley.
- Miller, A. J. (1990). *Subset selection in regression*. London: Chapman and Hall.

- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190–204.
- Parr, W. C. (1981). Minimum distance estimation: a bibliography. *Communication in Statistics—Theory and Methods*, **A10**, 1205–1224.
- Raftery, A. E. (1995). Bayesian model selection in social research (with discussion by Andrew Gelman, Donald B. Rubin and Robert M. Hauser). In P. V. Marsden (Ed.), *Sociological methodology 1995*, pp. 111–196. Oxford: Blackwells.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.
- Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika*, **61**, 509–515.
- Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92–107.
- Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–131.

Received: November 4, 1997; revised August 24, 1998