A Review of Experimental Design Procedures for Regression Model Discrimination

Peter D. H. Hill
Mathematics Department
University of Waikato
Hamilton, New Zealand

We review existing methods of experimental design for discriminating between rival regression models. Box and Hill's algorithm and its extensions, and the work of Atkinson and Fedorov and Atkinson and Cox are particularly considered. The relationship between the various methods is pointed out. Several unrelated and less formal algorithms are also mentioned. The need for a multipurpose design strategy, which takes into account the need to design for estimation of the parameters of the correct model, is stressed.

KEY WORDS
Regression Models
Experimental Design
Model Discrimination

I. THE ORIGINAL PROCEDURES

Perhaps the earliest work on discrimination between specified models was that of Cox [14], [15]. He supposed that two separate probability density functions had been proposed to describe the data and using a modification of the Neyman-Pearson maximum likelihood ratio he derived a statistic to test the null hypothesis that the first distribution is appropriate. In both these papers the emphasis was on inference rather than design for discrimination. However, Chambers and Cox [11] applied this approach to the problem of designing and analysing an experiment to distinguish between logistic and probit binary response models. Apart from this application, however, other authors have not directly taken up Cox's approach in their problems of discriminatory design, although his suggestions for future investigation have proved helpful.

The problem considered by Cox is of greater generality than that of rival regression models. Hunter and Reiner [28] were the first to suggest a design procedure for this more specific, but frequently occurring, situation.

Suppose there are two rival models \( \eta(x, \theta_1) \) and \( \eta(x, \theta_2) \) expressing the expected response at a point \( x \), where \( x \) is a vector of independent or controllable variables and \( \theta_1 \) and \( \theta_2 \) are vectors of unknown parameters. Suppose also that \( n \) observations have already been taken with errors which are independently, normally distributed with zero mean and constant variance; their sequential criterion was that \( x_{n+1} \) should be chosen to maximize the function

\[
J(x_{n+1}) = \left( \eta(x_{n+1}) - \eta(x_{n}) \right)^2
\]

where \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) are the maximum likelihood estimates based on the \( n \) observations. In other words, at stage \( n \) of the experiment the two models are fitted to the data and a new observation is taken at the point where they are furthest apart. Hunter and Reiner applied this sequential procedure to a linear and a nonlinear problem and in both cases the discriminatory design converged onto a small number of points.

Their method is clearly only applicable when there are two rival models. However, Roth [32] extended the criterion for the situation where there are more than two rivals.

Independently, Fedorov and Pazman [19], with the same motivation but with the additional assumption that the variance of an observation at the point \( x \) is a known function, \( \sigma^2(x) \), derived a more complicated criterion than that of Hunter and Reiner, again for the case of two rival models. Their suggestion was that if model \( k \) were true, \( k = 1, 2 \), the point \( x_{n+1} \) should be chosen to maximize the function

\[
\left( \frac{\eta_j(x, \theta_j) - \eta_k(x, \theta_k)}{\sigma_j^2(x) + \sigma_k^2(x)} \right)^2
\]

where \( j \neq k \) and \( \sigma^2(x) \) is the variance of \( \eta_j(x, \theta_j) \). In the absence of knowledge as to which is the correct model they hoped that the two criteria would suggest new points which would be similar. If an adequate compromise is not possible they suggested maximizing the smaller of the two criterion values or some weighted combination of the two.
This approach of Hunter and Reiner, modified by Fedorov and Pazman, is both simple and intuitively appealing; it is not surprising that it has served as the basic concept to be extended and investigated in later work.

2. BOX AND HILL'S PROCEDURE

2.1 The Basic Strategy

Box and Hill [9] noted that Hunter and Reiner's [28] criterion failed to take account of the variances of the estimated responses \( \eta_j(x, \hat{\theta}_j) \). They therefore developed a more sophisticated approach which does not have this weakness, (nor indeed does the criterion of Fedorov and Pazman [19] mentioned above), and which treats the case of more than two rival models.

They started with Shannon's concept of entropy which is defined as

\[
\sum_{j=1}^{m} \Pi_j \ln \Pi_j
\]

where \( \Pi_j \) is the probability that the \( j \)th of \( m \) models is the true one. Maximum uncertainty, corresponding to maximum entropy, occurs when

\[
\Pi_1 = \Pi_2 = \cdots = \Pi_m = 1/m
\]

They assumed that under the \( j \)th model the \( n \)th observation is normally distributed with mean \( \eta_j(x_n, \hat{\theta}_j) \) and known variance \( \sigma^2 \) (constant over all experiments); that the rival models \( \eta_j(x, \theta_j) \) are linear (or approximately linear in the neighbourhood of the parameter estimates); and that a locally uniform (non-informative) prior distribution is appropriate for the parameters in each model. With these assumptions Box and Hill showed that an upper bound \( D \) on the expected entropy change due to the \((n+1)\)th experiment about to be performed is

\[
\frac{1}{2} \sum_{j=1}^{m} \left[ \Pi_j \left( \frac{\sigma_j^2 - \hat{\sigma}_j^2}{\sigma^2 + \hat{\sigma}_j^2} \right)^2 + \left( \frac{\eta(x_{n+1}, \hat{\theta}_j) - \hat{\eta}(x_{n+1}, \hat{\theta}_j)}{\sqrt{\sigma^2 + \sigma_j^2}} \right)^2 \right] - \ln \Pi_j
\]

where \( \Pi_j \) is the probability of the \( j \)th model after \( n \) observations, \( \hat{\theta}_j \) is the maximum likelihood estimate after \( n \) observations and \( \hat{\sigma}_j^2 \) is the variance of \( \eta_j(x_{n+1}, \hat{\theta}_j) \). Their approach consisted of choosing at stage \( n + 1 \) that design point \( x_{n+1} \), which maximizes this criterion. After the observation \( y_{n+1} \) has been taken at this point the probabilities of the models are updated by the formula

\[
\Pi_{j,n+1} = \Pi_{j,n} p_j / \sum_{k=1}^{m} \Pi_{k,n} p_k
\]

where \( p_j \) is the normal density function of \( y_{n+1} \).

There are several points to notice about this criterion.

1. Maximizing an upper bound on the expected change in entropy seems a strange criterion—maximi-
data then the generalisations must be also. But Atkinson and Cox point out that if the simpler model is true, or even if it is only nearly true, the Bayesian analysis will prefer it. This may well be desirable, of course, if simplicity is more important than perfect discrimination.

An important feature of Box and Hill's procedure is that design and analysis are carried out together, stage by stage. A significance testing approach would design the discriminatory experiments, for example by the method of Hunter and Reiner [28], and then conclude that one, or some, or possibly all or none, of the models were consistent with the data. But in the Bayesian approach the models are considered to be a complete set of rivals and the analysis simply compares their relative adequacies. The informal stopping rule suggested by Box and Hill is to stop when the posterior probabilities indicate that one model is clearly superior to the rest.

However, in many applications it has become clear that this recommendation must be applied very cautiously—see, for example, Hill [22], Siddik [33], Froot and Mezaki [20] and Wentzheimer [34]. This is because the model probabilities may oscillate considerably from stage to stage of the sequential strategy. This oscillation means that a model should not be too readily accepted on the basis of a small number of discriminatory experiments.

Despite the assumptions required and the difficulties outlined above, the Box and Hill procedure has proved to be effective in practice and popular with research workers in a reasonable range of problems.

2.2 Extensions to the basic Box-Hill strategy

The Box and Hill procedure has been subject to considerable development. Hill and Hunter [23, 24] extended the criterion to the cases of multiple responses and unknown error variance. M. J. Box [10] further extended the multiresponse criterion to the case where the error covariance matrix is not constant over the range of experiments. Hsiang and Reilly [26] also adopted a Bayesian procedure; by allowing all the unknown parameters, both in the models and in the error distribution, to take only a finite set of possible values they were able to drop Box and Hill's assumptions of normality, known variance and local linearity of the models and to consider informative priors on the discrete parameter space. However, in problems involving many parameters in each model their method will require excessive computer storage space for the arrays of parameter values.

Hill, Hunter and Wichern [25] turned attention to the very appealing aim of designing for simultaneous model discrimination and parameter estimation. The standard approach is to design experiments to discriminate between rival models and then, when a model has been identified to design more experiments to estimate its parameters. But these two stages of experimentation would be better intertwined and the transition between them needs to be more smooth. With the disjointed approach we may take an unnecessarily large number of experiments before settling for a model and proceeding to estimation. Hill, Hunter and Wichern attempted this more integrated approach. If the jth model were definitely correct we would design to estimate its parameters as well as possible by seeking at stage n the new point x_{n+1} to maximize some criterion $E_j$, such as, for example, the inverse of Wilks' generalized variance of the parameter estimates. However, if the correct model were unidentified we would seek $x_{n+1}$ to maximize Box and Hill's discriminatory criterion $D$. Hill, Hunter and Wichern therefore suggest choosing $x_{n+1}$ to maximize

$$W_n = \frac{D_n}{D_{\text{max}}} + (1 - W_n) \sum_{j=1}^{m} \Pi_{j,n} \frac{E_j}{E_{j,\text{max}}}$$

where $D_{\text{max}}$ and $E_{j,\text{max}}$ are the maximum obtainable values of $D$ and $E_j$ over the design region. The weight $W_n$ attached to discrimination could be, for example, of the form

$$W_n = (m(1 - \Pi_{b,n})/(m - 1))^\lambda$$

where $\Pi_{b,n}$ is the largest among the $m$ posterior model probabilities at stage $n$. The positive power $\lambda$ gives the experimenter a flexible control over the rate at which interest switches from discrimination among the models to estimation within the well-discriminated model. Cochran [13] commented that design points which are best for discrimination are unfortunately not usually best for estimation. It is interesting to note that in the example treated by Hill, Hunter and Wichern their criterion does not lead to compromise points different from those for discrimination and for estimation. Rather, the effect of the criterion is to introduce an occasional "estimation point" early on and then as discrimination proceeds to introduce more "estimation points" and fewer "discrimination points."

This joint interest in discrimination and estimation is also treated by Borth [7] using the idea of total entropy which measures both the uncertainty as to the correct model and the uncertainty as to the parameters of each model. It may be that his criterion would require considerable computing time since he followed Reilly [31] in evaluating the expected decrease in total entropy directly, by means of a Gaussian-Hermite quadrature, whereas Box and Hill in their situation simply approximated the expected entropy decrease by its upper bound.

Fedorov [16, Chapter 7] adopted an approach similar to that of Borth. Läuter [29] and Atkinson [3]
also took up this problem of joint interest in discrimination and estimation.

2. FEDOROV'S PROCEDURE

Fedorov and his co-workers have advocated an approach which is essentially a formalization of the intuitive idea of Hunter and Reiner [28] that experiments should be conducted at points of maximum divergence between the models. Consider first the problem of just two rival models, $\eta(x, \theta_1)$ and $\eta(x, \theta_2)$. Suppose an experiment is to be conducted consisting of $N$ normally distributed observations with $x_j$, $j = 1, \ldots, n$, at each of $n$ different design points $x_j$, $j = 1, \ldots, n$. Let $w_j = r_j/N$ and $y_j$ be the average of the observations at the point $x_j$. Suppose also that the first model is true and that its parameter values are known. Fedorov's approach was to design the experiment to yield a maximum value for the sum of squares for lack of fit of the second model, which is equivalent to maximizing

$$\Delta(\xi_n, 1, \theta_1) = \sum_{j=1}^n w_j |\eta(x_j, \theta_1) - \eta(x_j, \hat{\theta}_2)|^2$$

where $\hat{\theta}_2$ is the least squares estimate of $\theta_2$ in the absence of experimental error. $\Delta(\xi_n, 1, \theta_1)$ symbolises the discrete design $\{x_1, \ldots, x_n; w_1, \ldots, w_n\}$ and the terms 1 and $\theta_1$ remind us that the criterion is specific for model 1 and its parameter value. We have assumed here that the error variance is constant but Fedorov and Malyutov [18] allowed a known non-constant error variance $\sigma^2(x)$. Atkinson and Fedorov [5] considered the case where there may be constraints on the parameters of the models.

This approach is reasonable from several points of view as outlined in Fedorov [16, Chapter 6] and in Fedorov and Malyutov. When the models are linear in the parameters (the case considered by Fedorov and Malyutov) $\Delta(\xi_n, 1, \theta_1)$ is proportional to the non-centrality parameter of the $\chi^2$ distribution of

$$\sum_{j=1}^n r_j (y_j - \eta(x_j, \hat{\theta}_2))^2$$

where $\hat{\theta}_2$ is the least squares estimate of $\theta_2$. Since the power of the $F$ test for departures from model 2 increases with $\Delta(\xi_n, 1, \theta_1)$ it is appropriate to maximize this criterion by the choice of design. When the models are nonlinear the maximization can be justified by a likelihood ratio test. Atkinson and Fedorov turned from the problem of an optimal discrete $N$ point design to the continuous or approximate theory in which the discreteness of the weights $w_j$ is abandoned and replaced by a normed measure $\xi$ defined on the design region $D$. The design measure $\xi^*$ is said to be $T$-optimal if

$$\Delta(\xi^*, 1, \theta_1) = \sup_{\xi} \Delta(\xi, 1, \theta_1)$$

Such a design has also been called locally optimal but that terminology is not exclusive to this situation. An equivalence theorem then states that $\xi^*$ is $T$-optimal if and only if

$$|\eta(x, \theta_1) - \eta(x, \hat{\theta}_2)|^2 \leq \Delta(\xi^*, 1, \theta_1)$$

where $\hat{\theta}_2$ is the least squares estimate of $\theta_2$ based on $\xi^*$ in the absence of error. The proof is given by Fedorov [17], although there is a difficulty in that the $T$-optimal design may not allow a unique estimate for $\theta_2$. The theorem in itself does not assist in obtaining $T$-optimal designs in practice but it is useful in eliciting their structure in certain simple problems and in verifying the optimality of a suggested design.

If the true model and its parameter values were known the maximization of the noncentrality parameter would be a useful approach for discriminatory design. But in fact this $\Delta(\xi, 1, \theta_1)$ is not a practical criterion at all since the identification of the true model and its parameters is our whole intention. So we are forced to consider methods of eliminating these unknown factors from our criterion. Atkinson and Fedorov suggested a Bayesian approach, multiplying the criterion by prior distributions for the models and the parameters and integrating and summing out the unknowns. Alternatively they suggested a maximin approach transforming the criterion to be maximized to

$$\inf\inf \int_\nu |\eta(x, \theta_1) - \eta(x, \theta_2)|^2\xi(dx)$$

Their final and most practical suggestion was to adopt a sequential design procedure which converges to the $T$-optimal design. Their strategy was in fact exactly that of Hunter and Reiner [28] outlined above in Section 1, except that Atkinson and Fedorov allowed constraints on the parameters and Fedorov and Malyutov allowed non-constant variance. Unlike Fedorov and Pazman [19] and Box and Hill [9] these authors took no account of the variances of the estimated responses. Asymptotically, of course, this will not be important, but it is more the small sample performance of these procedures which is the important area of comparison.

Fedorov [17] proved a theorem stating that if this sequential procedure converges to a design non-singular with respect to both models, the design is almost surely $T$-optimal. However, the proof depends upon the consistency of the estimates and this is a property not readily verified in the sequential situation.

Atkinson and Fedorov [6] turned attention to the problem of $m > 2$ rival models. Suppose again that the first model is known to be true and its parameter values are known. There are now $m - 1$ noncentrality parameters $\Delta(\xi, 1, \theta_1), j = 2, \ldots, m$, of the previous form and the criterion suggested is the maximization of the noncentrality parameter for the one or more
models closest to model 1. That is, a design \( \xi^* \) is T-optimal if and only if

\[
\Delta(\xi^*, 1, \theta_1) = \sup_{\xi} \min \Delta(\xi, 1, \theta_1)
\]

In the case when there is only one closest model this is a simple generalization of the two model criterion. But it may be that maximizing this minimum non-centrality parameter leads to a design for which there are two or more closest models with equal non-centrality parameters. This situation, occurring not infrequently in practice, leads to complexities in both the statement and the proof of the appropriate equivalence theorem and to a more complicated sequential procedure.

Initially Atkinson and Fedorov advocated a procedure which at each stage involves ranking the \( m \) residual sums of squares and then taking the next observation at the point where the estimates of the two best fitting models are furthest apart. This sequential strategy converges to the T-optimal design if there is only one closest model to the true model. But with an artificial example the authors showed that in the contrary case it can lead to a non-optimal design. Consequently they suggested an alternative procedure. However, in Atkinson and Fedorov's examples this modified sequential procedure produced a negligible improvement over the simpler procedure. Moreover, since it is considerably more complicated in practice it is difficult to know when the modified algorithm should be preferred.

The work of Fedorov and his colleagues thus seems to amount to a justification of the method of Hunter and Reiner. The various modifications are not fundamental but the formulation of the concept of T-optimality and the proof that the sequential procedure converges to a T-optimal design if it is nonsingular have shown that Hunter and Reiner's algorithm is not only intuitively but also theoretically appealing. The extension of the concept to the case of more than two rivals is also most important and the sequential procedure (in either form) should prove helpful in most problems.

### 4. Atkinson's Procedure

The work of Atkinson and Cox [4] arose out of earlier work by Atkinson [2] on the adequacy of a single model. He supposed initially that the model was linear

\[
\eta_i(x, \theta_1) = \theta_i'f_i(x)
\]

and that interest centered on detecting a specific departure from this model, namely,

\[
\eta(x, \theta) = \theta_i'f_i(x) + \theta_0'f_0(x).
\]

The common question of whether a linear regression should be extended to a quadratic is a good example of such a situation. Suppose that a design measure \( \xi \) on the design region \( D \) is to be chosen for this purpose. Then the power of the \( F \) test (assuming normally distributed observations) for detecting departures from model 1 increases with the noncentrality parameter

\[
0 < A_j\theta_1 = \theta_0'M_{2j}M_{11}^{-1}M_{21}\theta_1
\]

where

\[
M_{jk} = \int f_j(x)f_k'(x)\xi(dx).
\]

Atkinson therefore suggested that an appropriate design criterion would be to maximize this noncentrality parameter. If \( \theta_i \) is a scalar it plays no part in the design criterion which simply estimates \( \theta_i \) with minimum variance. However, if \( \theta_i \) is a vector the noncentrality parameter depends on its unknown true value. In this situation Atkinson advocated choosing the design to maximize \( \text{det}(A_j) \). In other words, to detect departures from model 1 the design should minimize the generalized variance of the estimate of \( \theta_j \).

This approach seems sensible for the adequacy of a linear model. But when the model is nonlinear it is not usually clear what extended model is appropriate, although Atkinson made several useful suggestions.

Atkinson and Cox applied this approach to the problem of discriminating between rival models. The extended model in this situation contains all the linearly independent terms in the \( m \) individual models. To detect a departure from model \( j \) in the direction of the others they designed to estimate the parameters in the combined model which are not in model \( j \). The design criterion suggested is \( D_j \)-optimality, so for each of the \( n \) models there is a term \( \text{det}(A_j) \) to be maximized by choice of design, where \( A_j \) is of the above form. Their "equal interest" discriminatory criterion then consisted of choosing \( \xi \) to maximize

\[
\prod_{j=1}^m (\text{det}(A_j))^{1/s_j},
\]

where \( s_j \) is the number of parameters in the combined model which are not in model \( j \) and is included to adjust the determinant for the number of parameters.

It is interesting to notice the connection between this equal interest criterion and the T-optimality of Atkinson and Fedorov [5]. Suppose there are two distinct rival linear models:

\[
\eta_i(x, \theta_1) = \theta_i'f_i(x), \quad \eta_i(x, \theta_2) = \theta_0'f_0(x)
\]

The T-optimality assuming model 1 is true with parameters known is \( \Delta(\xi, 1, \theta_1) \). It is straightforward to show that in this case

\[
\Delta(\xi, 1, \theta_1) = \theta_i'(M_{11} - M_{12}M_{22}^{-1}M_{21})\theta_1 = \theta_i'A_1\theta_1
\]

where the \( M_{jk} \) are defined as above. The dilemma at
this stage, as discussed in Section 3, is how to apply this criterion when the true model and its parameters are unknown. Bayesian, maximin and sequential approaches were suggested in Section 3. Now if model 2 were true and its parameters known the T-optimal criterion would be $\theta_2' A_2 \theta_2$ where $A_2$ is defined as above. Atkinson and Cox's equal interest criterion $\{\det (A_2)^{1/s} \}$ is thus seen to be another method of treating the uncertainty as to model and parameters in the T-optimal criterion.

In practice these equal interest designs are obtained by an iterative algorithm (analogous to Wynn's [35] algorithm for obtaining a D-optimal design) where at each stage a new point is chosen to maximize a weighted combination of the $m$ relevant variance functions. The motivation for this algorithm and the proof of its convergence derive from a generalised equivalence theorem stated by Atkinson and Cox.

A weakness of the equal interest approach is that it is non-data-dependent. Clearly as the number of rival models increases this approach will become less efficient since the lack of data does not enable extremely implausible rivals to be eliminated quickly. Consequently Atkinson and Cox replaced the equal interest weights, $1/s$, with a system whereby the weights at each stage depend on the likelihood of the models. The beneficial result of this introduction of data dependence is that as evidence against a model increases its importance in the design criterion decreases. On the other hand a defect is that the criterion converges towards that for optimal estimation of parameters in the T-optimal criterion. So here again we have returned to the important need for simultaneous consideration of discrimination and estimation.

5. SOME FURTHER PROCEDURES

Andrews [1] proposed a novel sequential procedure. At each stage the value of the response at the next design point can be predicted either from the values of previous responses at that point or by means of a general model combining all the models proposed. Given this predicted response the residual sum of squares and the significance level of the resulting $F$ test for adequacy are calculated for each model. Then the next design point chosen is the one which minimizes a certain criterion involving these predicted significance levels. Andrews' choice of the criterion function was somewhat arbitrary but in his problem involving simple polynomial models he obtained an effective design which discriminates between adequate and inadequate models.

Lütter [29] treated the problem of estimating the parameters of a linear model when it is not known which model it is among a family $M$ of linear models. She called a design measure $\xi S$-optimal if it maximizes the criterion

$$\int \ln \det \{ M_n(\xi);Q(dm)$$

where $Q$ is a normed measure over the family $M$ of models and $M_n(\xi)$ is the information matrix for the $m$th model. She proved an equivalence between the S-optimal design and the appropriately defined minimax design—this is essentially the same equivalence theorem as that of Atkinson and Cox [4]—and she developed an iterative algorithm for obtaining $S$-optimal designs.

Chernoff [12] was concerned with the wider problem of sequentially designing experiments in order to make an inference as to the true value of a parameter governing the outcome of the experiment. His game theory approach involved the use of Kullback's information theory. More recently Meeter, Pirie and Blot [30] have incorporated the problem as to the true form of a regression model into this more general framework. They compared Chernoff's strategy with that of Box and Hill [9] by a simulation study of three different problems, one of which concerned rival regression models.

Informal graphical methods, such as that suggested by Herzberg and Tukey [21] for comparing the residuals from the models at each stage, may also be useful in many problems. The development of computer graphics will presumably assist in this direction and in interpreting such techniques. An even more straightforward comparison and analysis of residuals will also often be helpful. These less formal discriminatory methods may be especially useful in communicating with research workers.

A new discriminatory design procedure [22] devised by the author will be described elsewhere.

6. ACKNOWLEDGEMENTS

This research formed part of the author's doctoral thesis at the University of Glasgow. Helpful discussions with Professor S. D. Silvey and financial support from the Commonwealth Scholarship Plan are gratefully acknowledged, as are some interesting comments from a referee.

REFERENCES


