An Elicitation Method for Multiple Linear Regression Models

by

Paul H. Garthwaite and James M. Dickey

University of Minnesota, School of Statistics
Technical Report No. 548
9 May 1990

Research supported in part by National Science Foundation Grant DMS-8911548.

An Elicitation Method for Multiple Linear Regression Models

ABSTRACT

This paper describes a method of quantifying subjective opinion about a normal linear regression model. Opinion about the regression coefficients and experimental error is elicited and modeled by a multivariate probability distribution (a Bayesian conjugate prior distribution). The distribution model is richly parameterized and various assessment tasks are used to estimate its parameters. These tasks include the revision of opinion in the light of hypothetical data, the assessment of credible intervals, and a task commonly performed in cue-weighting experiments. A new assessment task is also introduced. In addition, implementation of the method in an interactive computer program is described and the method is illustrated with a practical example.

KEY WORDS Probability assessment Regression Elicitation tasks Prior distributions Expert opinion probabilities

This paper addresses the task of assessing a probability distribution to quantify an expert's subjective opinion about the unknown parameters in a multiple linear regression model. The task is important as there are many situations where traditional statistical data is absent or sparse relative to the number of variables involved, and then a large proportion of available information may be contained in an expert's personal opinion. Cases include the predicted performance of multiple-ingredient manufactured products, as in the trial marketing of household soap products, or their performance in eventual use, as with tar-macadam for road surfaces, where data on longevity may require a delay of ten years or more. In other cases, experiments may be so expensive to perform that expert opinion is essential in their planning. Lubricating oils, for example, are tested by running engines for weeks in the laboratory and then dismantling them to measure rusting, wear, and various deposits. Each data point can cost thousands of pounds. In these types of situation, expert personal opinion is of great potential value and can be used more efficiently, communicated more accurately, and judged more critically if it is expressed as a probability distribution. The probability distribution can also be used as a *prior* distribution in Bayesian statistical methods, the purpose that motivated the present research. If new sample data becomes available, Bayes' theorem can be used to combine the information given by the data with the prior distribution. In principle, the resulting *posterior* distribution would contain all the available information and hence should be used for making inferences, estimates, and predictions.

A multiple linear regression model specifies that a dependent variable, Y say, is related to other independent variables, $X_1, X_2, \ldots, X_r$, by an equation

of the form :

$$E(Y \mid x_1, x_2, \ldots, x_r) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x_r \qquad (1)$$

where $\beta_1, \beta_2, \ldots, \beta_r$ (the regression coefficients) are constant and $E(Y \mid x_1, x_2, \ldots, x_r)$ is the expected (or average or true) value of Y when $X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r$. If the model should have a constant term, then X_1 is identically equal to 1 and the constant term will be β_1.

In vector notation the equation can be written

$$E(Y \mid \mathbf{x}) = \underline{\beta}'\mathbf{x} \qquad (2)$$

where $\underline{\beta} = (\beta_1, \ldots, \beta_r)'$ and $\mathbf{x} = (x_1, \ldots, x_r)'$. One purpose of the elicitation method is to determine an expert's opinion about $\underline{\beta}$. *Point estimates* of $\beta_1, \beta_2, \ldots, \beta_r$ must be determined, *variances* expressing the expert's uncertainty about the β_i's, and *covariances* expressing the strength of association in the expert's opinion about pairs β_i, β_j.

In most situations, such as experimental work or when sampling from a population, an observed value of Y will differ from $\underline{\beta}'\mathbf{x}$ because of chance variation or *random error*. Some assumptions must be made about the distribution of the random error and we shall make the common assumption that it is a normal distribution with a mean of 0 and a variance, σ^2 say, that does not depend upon the values of the X's. A second purpose of the elicitation method is to quantify the expert's opinion about σ^2.

One elicitation method for quantifying opinion about $\underline{\beta}$ and σ^2 was developed by Kadane, Dickey, Winkler, Smith and Peters (1980). The central

assessment task in their method is the assessment of fractiles of the predictive distribution of Y at specified values of $\mathbf{X}$. The expert uses the method of bisection (Huber, 1974; Ludke, Stauss and Gustafson, 1977) to successively assess the median and the 0.75, 0.875 and 0.9375 fractiles. The elicitation method developed in the present paper, similarly, requires assessments of the median, the 0.75 fractile, and the 0.25 fractile, but not the other fractiles. Assessing the median of Y at specified X-values is a task that has been widely studied in cue weighting experiments, where the X-variables are referred to as 'cues', the expected β-coefficients as 'cue-weights', and the Y-variable as the 'criterion'. This research has found that, in a wide variety of situations, assessments can be quite well represented by a linear model relating the criterion to the cues. Also, in cue–learning experiments where the true relationship between the criterion and cues is known to the experimenter, it has been found that assessors can learn to use linear cues appropriately. A good review of this research is given by Slovic and Lichtenstein (1971).

Experiments have also examined abilities at assessing 50% central credible intervals. This task requires both the 0.75 and 0.25 fractiles to be assessed and, if an assessor is "well-calibrated", the interval between the fractiles should have a long run frequency of 50% of containing the actual value of the quantity of interest. Almost all experiments have been conducted outside the framework of a regression model and assessors' abilities have varied. Meteorologists have assessed 50% credible intervals for tomorrow's maximum and minimum temperatures and have been well calibrated (Peterson, Snapper and Murphy, 1972; Murphy and Winkler, 1977). Other assessors, in both artificial and real situations, have tended to be overconfident and have

provided intervals that were too narrow, in that noticeably fewer than 50% of the intervals contained the actual value of the quantity of interest [see, for example, Lichtenstein, Fischhoff and Phillips (1982) for a review]. However, whether overconfidence should be expected in the context of regression models is uncertain; Garthwaite (1989) found subjects were *under*confident when assessing 50% credible intervals for the Y-variable in models containing one independent variable. The explanation suggested was that subjects utilized the linear relationship when giving median assessments of Y, which improved their accuracy, but subjects did not appreciate this greater accuracy and hence assessed credible intervals that were too wide.

Assessing the median and 0.75 fractile of Y at specified X-values are the only tasks common to the elicitation method developed here and the method of Kadane *et al.* In particular, we avoid eliciting assessments of the 0.875 and 0.9375 fractiles, since assessing such fractiles in the tails of a distribution is a task that subjects often perform poorly (Alpert and Raiffa, 1982; Lichtenstein *et al.*, 1982). Also, Kadane *et al.* make extensive use of hypothetical data, asking the expert to imagine that certain Y-values have been obtained and to update his or her opinions to reflect this information. Their hypothetical data set is increased in stages and at each stage the expert assesses the conditional median and the upper quartile of Y at various X-values. Research has found that subjects tend to revise their opinions insufficiently when given hypothetical data, the bias of *conservatism* (Edwards and Phillips, 1964). To restrict the effect of this bias, the elicitation method developed in this paper makes very little use of hypothetical data : only one hypothetical datum is given and only one conditional assessment elicited. Instead, the method uses

an assessment task that seems not to been suggested before. The task exploits the fact that an expert's accuracy in predicting Y will depend in part upon the values of the X-variables for which the prediction is made. The values of some of the X-variables are given to the expert who is then asked to choose values for the remaining X-variables, and to make the choice to maximize the accuracy of his or her prediction. The values chosen give information about the expert's subjective distribution. We refer to this task as choosing points of Constrained Minimum Variance (CMV) and we describe it more fully later.

The original purpose of the elicitation method was to help research chemists quantify their opinions about industrial chemical processes. In experiments they conduct, the X-variables would typically be controllable quantities, such as temperatures, pressures, amounts of reactants, etc. The elicitation method reflects this application in the phrasing of questions; the questions presuppose that experiments are to be conducted and the X-variables can be controlled. Also, terminology familiar to the industrial chemists is adopted; a particular setting of the X-variables is referred to as a *design point*, and the Y-variable as the *yield*.

The elicitation method has been implemented in an interactive computer program. To quantify his or her opinion, an expert types in answers in response to prompts from the computer and a subjective distribution is determined from these assessments. The statistical theory underlying the method is described in Garthwaite and Dickey (1988). The purpose of the present paper is to give the method, describe its implementation and illustrate its use with a practical example. A user-guide and listing of the computer program

are given in Garthwaite (1986). The program is written in Microsoft Basic and can run on a microcomputer.

THE SUBJECTIVE DISTRIBUTION

To make the elicitation problem feasible, structure must be imposed on the form of the expert's subjective probability distribution and we shall suppose that it can be modeled by a natural conjugate distribution (Raiffa and Schlaifer, 1961). This form of distribution is widely used as a prior distribution in Bayesian statistics, and hence has been studied extensively. Its primary advantage is that it is mathematically convenient : predictive distributions for future Y-values are easily determined, confidence intervals (credible intervals) for β-coefficients can be calculated analytically, hypotheses that particular β-coefficients are zero can be tested, and so on. Also, information contained in sample data can readily be combined with that contained in the natural conjugate prior distribution (DeGroot, 1970). While adopting this model we do not wish to suggest that an expert's opinion will correspond exactly to it. Rather, we hope to represent important features of the expert's opinions in a form that is usable and can be elicited.

For a regression model the conjugate distribution has four parameters, usually referred to as 'hyperparameters', denoted here by ω, n, $\mathbf{b}$ and $\mathbf{U}$, of which $\mathbf{b}$ and $\mathbf{U}$ are arrays. The hyperparameters ω and n relate primarily to opinion about σ^2, ω being a subjective estimate of σ^2 and n reflecting the amount of information on which the estimate is based. Specifically, the conjugate distribution specifies that σ^2 is distributed as ωn times the reciprocal of a chi-squared random variable with n degrees of freedom. The arrays $\mathbf{b}$

and $\mathbf{U}$ relate to $\underline{\beta}$. The conjugate distribution specifies that, given σ, the vector $\underline{\beta}$ has a multivariate normal distribution with mean $\mathbf{b}$ and variance-covariance matrix $\sigma^2\mathbf{U}/\omega$. Thus $\mathbf{b} = (b_1, \ldots, b_r)'$ is the point estimate of $(\beta_1, \ldots, \beta_r)'$, and given σ^2, we have the subjective variances and covariances

$$\begin{pmatrix} var(\beta_1) & cov(\beta_1,\beta_2) & cov(\beta_1,\beta_3) & \ldots & cov(\beta_1,\beta_r) \\ cov(\beta_1,\beta_2) & var(\beta_2) & cov(\beta_2,\beta_3) & \ldots & cov(\beta_2,\beta_r) \\ cov(\beta_1,\beta_3) & cov(\beta_2,\beta_3) & var(\beta_3) & & \\ \vdots & \vdots & & \ddots & \\ cov(\beta_1,\beta_r) & cov(\beta_2,\beta_r) & & & var(\beta_r) \end{pmatrix} = \sigma^2\mathbf{U}/\omega$$

The purpose of the elicitation method is to determine $\omega, n, \mathbf{b}$ and $\mathbf{U}$.

A new assessment task the expert performs is to specify design points as constrained minimum variance (CMV) points. For this task to yield the required information, all design points must be feasible within the region where the regression model holds and this region must be a sufficiently large set. In particular, independent variables must not be discrete-valued, since values between the discrete values could not be chosen by the expert. Hence, the variables cannot be indicators of factor levels. Also, the independent variables must not be functionally related, so for example, X_5 cannot equal X_1^2 or X_2X_4. Thus the model cannot include polynomial or interaction terms. This should seldom be a severe restriction, since cue-weighting experiments have found that subjective opinion can often be usefully represented by a simple linear model involving no second or higher order terms, provided the response is monotonically related to the independent variables (Slovic and Lichtenstein, 1971).

QUANTIFYING OPINION ABOUT ERROR VARIANCE

Empirical evidence indicates that humans have only limited information processing capacity and should not be asked to simultaneously consider a great deal of information (Hogarth, 1975). This suggests that complex assessment tasks should be decomposed into simpler tasks and the information they yield recombined mathematically. For this reason, the task of eliciting the hyperparameters ω and n is separated from that of eliciting $\mathbf{b}$ and $\mathbf{U}$. This is a natural division, since ω and n relate to the error variance, σ^2, while $\mathbf{b}$ and $\mathbf{U}$ relate to the regression coefficients, $\underline{\beta}$.

Assessing n

The expert is asked first to suppose that two separate experiments are to be conducted at the *same* design point. Because of experimental error the yields that are obtained will differ and the expert is questioned about the magnitude of the difference. Let Y_1 and Y_2 denote the yields in the first and second experiments, respectively, and put $Z = Y_1 - Y_2$. [The difference Z has a t-distribution with mean 0, variance $2\omega^2 n/(n-2)$ and n degrees of freedom.] The expert's median assessment of the absolute magnitude of the difference, the median of $|Z|$, is elicited and its value, k_1 say, is the semi-interquartile range of the distribution of Z (c.f. Exhibit 1).

Exhibit 1 about here

After assessing k_1, the expert is asked to imagine that the difference between the yields in the two experiments was observed to be a value the computer specifies, z_1 say, and that two further experiments are to be conducted at a single design point (either the same point as before, or a new design point). Bearing in mind the hypothetical datum z_1, the expert is

asked to give the median assessment of the absolute magnitude of the difference between the further yields that will be obtained. Let k_2 denote this new semi-interquartile assessment, and let q_n denote the semi-interquartile range of a t-distribution on n degrees of freedom. It can be shown that

$$\frac{k_1}{k_2} = \frac{q_n}{q_{n+1}} \left[\frac{n+1}{(\alpha q_n)^2 + n} \right]^{\frac{1}{2}} \tag{3}$$

where $\alpha = z_1/k_1$.

In the implementation of the elicitation method, the computer chooses z_1 so that $\alpha = \frac{1}{2}$. This choice is made so that the hypothetical datum differs noticeably from the expert's assessed value, k_1, but is not so different that z_1 and k_1 appear contradictory. The values of k_1/k_2 for various values of n and this choice of α are tabulated in Exhibit 2. After assessment of the ratio k_1/k_2, the corresponding value of n is read from the table.

Exhibit 2 about here

As the table indicates, error in estimating k_1/k_2 has a far greater effect on the estimate of n when n is large than when it is small. Often this fault will be minor, since in many applications the value of n is most critical when n is small. [Large values of n imply that the marginal distribution of $\underline{\beta}$ is approximately normal.] Also, in assessing a value for k_2, the expert might use the strategy 'anchoring and adjustment' (Tversky, 1974). The 'anchor' would be k_1 and the expert would assess k_2 by making an adjustment to k_1. If this is the case, then k_1/k_2 should be more accurately assessed when its value is close to 1 since then the 'adjustment', $(k_1 - k_2)$, is small. Since (k_1/k_2) approaches 1 as n increases, the value of k_1/k_2 would be more accurate when n is large. This may partly offset the sensitivity that large values of n have to error in k_1/k_2.

A criticism of the method of eliciting n is that it requires subjects to revise their opinions to reflect hypothetical data. As mentioned earlier, experiments have found that this type of task is not performed well, with the size of revisions typically being conservative. Unfortunately, obtaining useful information about n is difficult and we have found no preferable alternative. The method used by Kadane *et al.* (1980) requires the assessment of fractiles in a tail of the distribution of Y, which is another assessment task which subjects tend not to perform well. Kadane *et al.* found that this task commonly lead to estimates of n that were extreme and, presumably, unlikely.

A natural modification to our method which might improve accuracy would be to present subjects with a variety of hypothetical data (e.g. choosing $\frac{1}{4}k_1, \frac{1}{2}k_1, \frac{3}{4}k_1, \frac{3}{2}k_1$ and $2k_1$ as the value of z_1). For each of the data an assessment of k_2 could be elicited and the corresponding estimate of n determined. Some form of averaging could then be used to reconcile these estimates.

Assessing ω

Once n has been estimated, the other hyperparameter that relates to error variance, ω, can be obtained from the equation

$$\omega = \frac{1}{2}\left(k_1/q_n\right)^2 \tag{4}$$

As Exhibit 2 indicates, q_n does not vary much (particularly for values of n between 5 and infinity). Consequently, the estimate of ω is largely determined by the unconditional median assessment, k_1, and is far less dependent (through n) upon the conditional assessment that is made after the hypo-

thetical datum has been presented.

REGRESSION COEFFICIENTS ASSESSMENT

The hyperparameters remaining to be elicited are $\mathbf{b}$ and $\mathbf{U}$, which relate to the regression coefficients. To determine these, the expert is asked to select design points that have specified properties and is also questioned about the subjective distribution of Y at these points.

Selection of Design Points

The predictive distribution of Y depends upon the design point, $\mathbf{x}$, at which the prediction is being made. Suppose attention is restricted to design points whose first i components take particular values. Specifically, consider design points for which these components are given by $x_1 = a_1, x_2 = a_2, \ldots, x_i = a_i$ but whose remaining $r - i$ components can take any values in the region for which the regression model holds. We refer to such design points as $\mathbf{a}_i$-constrained points, where $\mathbf{a}_i = (a_1, \ldots, a_i)'$. Given the form of the subjective distribution, among such design points there is a unique point at which the variance, $var(Y \mid \mathbf{x})$, is minimized i.e. where predictive accuracy is greatest. We call this the $\mathbf{a}_i$-constrained minimum variance point ($\mathbf{a}_i$-CMV point).

The elicitation method requires the expert to assess CMV points. This task can be made meaningful to the expert by using the following type of question. *'You must conduct an experiment, but you must first predict the yield that you will obtain. After the experiment you will be rewarded, with a greater reward for a more accurate prediction. In the experiment, certain of the X variables must take specified values (which you will be given), but*

you can choose values for the remaining X variables. What values would you choose?' The expert should choose the CMV point in order to maximize the reward, and in experiments we have conducted, experts have been able to perform this task (Garthwaite, 1983,pp.70-80 and 130-136).

Constraints are chosen by the program implementing the elicitation method and r CMV-points, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ say, are elicited. This is done sequentially as follows. The program assumes that the first term of the regression model is a constant term, so a_1 is set equal to unity. The expert then assesses his a_1-CMV point and $\mathbf{x}_1$ is defined as this point. For the second point, the computer will insist that the first component again equals unity and add the further restriction that the second component must equal the value a_2. We put $\mathbf{a}_2 = (a_1, a_2)'$ and the expert then assesses the $\mathbf{a}_2$-CMV point, which becomes $\mathbf{x}_2$. This procedure is continued. Hence, for both the $(i-1)$th and the ith design points, the first $i-1$ components are constrained to equal $a_1, a_2, \ldots, a_{i-1}$, respectively. For the ith point, additionally, the ith component is constrained to equal the value a_i, and $\mathbf{x}_i$ is the elicited $\mathbf{a}_i$-CMV point where $\mathbf{a}_i = (a_1, \ldots, a_i)'$. (Note that the vectors of constraints $\mathbf{a}_i$ increase in dimensionality as the elicitations proceed.)

The values the computer chooses for the a_i's depend upon the ranges of the independent variables over which the regression model is believed to hold. (These ranges are specified by the expert before the CMV points are elicited.) Let x_i^* denote the value of the ith component for the design point $\mathbf{x}_{i-1}$. This value is assessed by the expert, while at design point $\mathbf{x}_i$, the value of the ith component (a_i) must be chosen by the computer. The elicitation

method requires the computer to specify a value,

$$a_i \neq x_i^*, \tag{5}$$

for $i = 2, 3, \ldots, r$. Otherwise, equations for estimating the hyperparameters cannot be solved. For the stability of estimates, the absolute magnitude of $(a_i - x_i^*)$ must not be too small. Empirical research also suggests that $|a_i - x_i^*|$ should be a reasonable size. Garthwaite (1989) found that subjective distributions for regression coefficients were then more accurate, as measured by a scoring rule. On the other hand, under the model the expert's accuracy in predicting the yield at $\mathbf{x}_i$ decreases as $|a_i - x_i^*|$ increases, and we would prefer to elicit the expert's opinion at design points where knowledge is not too vague. As a compromise, the program adopts a strategy whereby $|a_i - x_i^*|$ equals between one quarter and one half the practical range of the ith variable x_i. Specifically, if x_i^* is in the upper or lower quarter of the range, then a_i is chosen to equal the midpoint of the range. Otherwise, a_i is taken at the one-quarter or three-quarters fraction of the range, depending on whether x_i^* is respectively above or below the midpoint. This strategy does not seem unreasonable, although it may prove preferable to choose a_i so that $|a_i - x_i^*|$ is even greater.

The accumulated constraints imposed on the expert's choice of each design point are more restrictive than for the preceding point. Hence the expert's accuracy in predicting the yield should steadily decrease. At each design point the expert is questioned about the *mean yield*, meaning the long-run average yield that would be obtained if a large number of experiments were conducted at the point. The upper and lower quartiles of the

distribution of the mean yield are elicited, and if $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$ denote these assessments at the ith design point, they should satisfy

$$\overline{y}_{i,.75} - \overline{y}_{i,.25} > \overline{y}_{i-1,.75} - \overline{y}_{i-1,.25} \tag{6}$$

for $i = 2, \ldots, r$. So that this requirement is satisfied, it is explained to the expert that the interquartile range is a measure of predictive accuracy. The expert can then understand why equation (6) holds, since otherwise $\mathbf{x}_i$ should have been chosen as the $\mathbf{a}_{i-1}$-CMV point (which it could have been), rather than $\mathbf{x}_{i-1}$. Re-assessments are elicited if (6) is not satisfied.

Fractile Assessments

To determine the hyperparameter $\mathbf{b}$, the expert assesses his or her median of the mean yield at each design point. Let $\overline{y}_{i,.50}$ denote the assessed median at $\mathbf{x}_i$. Under the model, this is also the expert's subjective median and subjective mean for the yield of a *single* experiment at the design point. This assessment task is the one usually performed in cue-weighting experiments. To improve the internal consistency between an expert's assessments, he or she is also questioned at the second and subsequent design points about the *change* in mean yield between consecutive design points. Let d_i denote the change in mean yield from $\mathbf{x}_{i-1}$ to $\mathbf{x}_i$. For $i = 2, 3, \ldots, r$, the median of the subjective distribution for d_i is elicited, $d_{i,.50}$ say, and if the expert's assessments are consistent, then

$$d_{i,.50} = \overline{y}_{i,.50} - \overline{y}_{i-1,.50} \tag{7}$$

The implementation of the elicitation procedure (described in the next section) helps an expert give values for $\overline{y}_{1,.50}, \ldots, \overline{y}_{r,.50}$ and $d_{2,.50}, \ldots, d_{r,.50}$ that satisfy equation (7) and represent opinion to the expert's satisfaction.

From the design points, a square matrix $\mathbf{T}$ is calculated as

$$\mathbf{T} = (\mathbf{x}_1, \mathbf{x}_2 - \mathbf{x}_1, \mathbf{x}_3 - \mathbf{x}_2, \ldots, \mathbf{x}_r - \mathbf{x}_{r-1})'.$$

This is an upper-triangular matrix; the elements below the main diagonal are zero. Then $\mathbf{b}$ is obtained from

$$\mathbf{b} = \mathbf{T}^{-1}(\overline{y}_{1,.50}, \overline{y}_{2,.50} - \overline{y}_{1,.50}, \overline{y}_{3,.50} - \overline{y}_{2,.50}, \ldots, \overline{y}_{r,.50} - \overline{y}_{r-1,.50})' \qquad (8)$$

To determine the hyperparameters $\mathbf{U}$, the expert assesses the lower and upper quartiles of both the mean yield at each design point and also the difference in mean yield between consecutive design points. These assessments are denoted by $\overline{y}_{i,.25}$, $\overline{y}_{i,.75}$, $d_{i,.25}$, and $d_{i,.75}$, respectively. The expert can make these assessments using equal-odds judgements, since $\overline{y}_{i,.50}$ and $d_{i,.50}$ will have been elicited earlier. For example, $\overline{y}_{i,.25}$ is the value for which the subjective probabilities $P(\overline{y}_{i,.25} < \overline{Y}_i < \overline{y}_{i,.50})$ and $P(\overline{Y}_i < \overline{y}_{i,.25})$ are equal, where $\overline{Y}_i$ denotes the unknown mean yield at $\mathbf{x}_i$. To be internally consistent, the quartile assessments must satisfy

$$(\overline{y}_{i,.75} - \overline{y}_{i,.25})^2 = (\overline{y}_{i-1,.75} - \overline{y}_{i-1,.25})^2 + (d_{i,.75} - d_{i,.25})^2 \qquad (9)$$

for $i = 2, 3, \ldots, r$. As with the median elicitations, the implementation of the elicitation procedure (described below) helps the expert to assess values

that satisfy this requirement.

To obtain **U**, the *spreads* $S(\overline{y}_1)$,$S(\overline{y}_2)$,..., $S(\overline{y}_r)$ are first calculated, where these are defined by

$$S(\overline{y}_i) = [(\overline{y}_{i,.75} - \overline{y}_{i,.25})/2q_n]^2$$

and q_n, as before, is the semi-interquartile range of a t-distribution on n degrees of freedom. The following diagonal matrix **D** is then formed :

$$\mathbf{D} = \begin{pmatrix} S(\overline{y}_1) & & & \\ & S(\overline{y}_2) - S(\overline{y}_1) & & 0 \\ & & S(\overline{y}_3) - S(\overline{y}_2) & \\ 0 & & & \ddots \\ & & & & S(\overline{y}_r) - S(\overline{y}_{r-1}) \end{pmatrix}$$

This is a squared-scale matrix for the (orthogonal) yield differences, $\overline{y}_1$, $\overline{y}_2 - \overline{y}_1, \ldots, \overline{y}_r - \overline{y}_{r-1}$. Finally **U**, the scale matrix for uncertainty concerning $\underline{\beta}$, is obtained from the equation

$$\mathbf{U} = \mathbf{T}^{-1}\mathbf{D}(\mathbf{T}')^{-1}. \tag{10}$$

[Theory underlying equations (8)–(10) is given in Garthwaite and Dickey (1988).]

IMPLEMENTATION

At design points $\mathbf{x}_i$ $(i = 2, 3, \ldots, r)$ the following procedure is used to help the expert assess values for $\overline{y}_{i,.50}$ and $d_{i,.50}$ that are internally consistent,

satisfying equation (7). The value of $\overline{y}_{i-1,.50}$ will already have been assessed at $\mathbf{x}_{i-1}$.

1. $\overline{y}_{i,.50}$ is elicited.

2. $d_{i,.50}$ is calculated from equation (7) and the expert is asked whether the calculated value represents his or her opinion.

3. If it does, then median assessments at this point are complete and the method goes on to elicit the upper and lower quartiles of $\overline{y}_i$ and d_i (see below). Otherwise, the expert assesses $d_{i.50}$ by direct elicitation.

4. $\overline{y}_{i,.50}$ is calculated from (7) and the expert is asked whether it represents opinion.

5. If it does, then acceptable values for $\overline{y}_{i,.50}$ and $d_{i,.50}$ have been obtained. Otherwise the method returns to Step 1.

Usually, the expert's opinion as expressed in $d_{i,.50}$ could be represented well by any one of a range of values. Similarly for $\overline{y}_{i,.50}$. Hence it is reasonable for an expert to accept that a calculated value of $d_{i,.50}$ or $\overline{y}_{i,.50}$ represents opinion even though the value differs somewhat from the initially expressed opinion. Another consequence is that, in practice, the above steps seldom need to be repeated at a design point even though, in theory, they could be repeated indefinitely with the expert's opinion of $\overline{y}_{i,.50}$ always conflicting with that of $d_{i,.50}$. [If serious conflict does arise, medians for the previous design point can be re-assessed.]

A broadly similar procedure is used to elicit the lower and upper quartiles of $\overline{y}_i$ and d_i. However, the method is necessarily more complicated because

more quantities are involved and two equations, (6) and (9), must be satisfied, rather than one. We first outline the procedure and then comment on it.

1. Values of $d_{i,.75}$ and $d_{i,.25}$ are elicited.

2. The computer tries to suggest values of $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$ that both represent the expert's opinion and also satisfy equation (9). [Detail below.]

3. If such values are found, the assessments at the present design point are complete and the program moves on to the next point (or stops if that was the last point). Otherwise, values of $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$ are elicited directly.

4. The computer tries to suggest values of $d_{i,.75}$ and $d_{i,.25}$ that both represent the expert's opinion and satisfy equation (9).

5. If such values are found then assessments at the present design point are complete. Otherwise the method returns to Step (1).

Under the assumptions of our model, the subjective predictive distribution for $\overline{y}_i$ is a t-distribution and hence symmetric. However, the program does not constrain elicitations of $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$ to be equidistant from $\overline{y}_{i,.50}$, for in practice, the model will seldom correspond *precisely* to an expert's beliefs. Thus our interest is in the difference between $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$, which determines $S(\overline{y}_i)$, rather than the individual values of the quartiles. Of course, if an expert's quartile assessments display extreme asymmetry, our model for the expert's beliefs is inappropriate and the elicitation method described here should not be used. At the same time, it should be mentioned that modeling asymmetric quartile assessments by a symmetric distribution is not necessarily disadvantageous. In one experiment, both symmetric and asymmetric

distributions were fitted to subjects' median and quartile assessments and although the elicited quartiles were sometimes markedly skew, a scoring rule judged the assessed symmetric distributions to be slightly the more accurate overall (Garthwaite, 1989).

In step 2 of the quartile assessment procedure, the program calculates values for $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$ that satisfy equation (9) and are either (a) equidistant from $\overline{y}_{i,.50}$ or (b) close to the expert's previous assessments of $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$, if such assessments have previously been given. The latter can occur when the expert has found calculated quartile values unrepresentative of his or her opinion. In this case (b), the program would mimic any asymmetry in the expert's assessments by so choosing $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$ that the ratio $(\overline{y}_{i,.75} - \overline{y}_{i,.50})/(\overline{y}_{i,.50} - \overline{y}_{i,.25})$ is the same as the expert gave. The expert is then asked whether the calculated values $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$ represent opinion. If not, $\overline{y}_{i,.75}$ is elicited directly and the program calculates the value of $\overline{y}_{i,.25}$ for which $\overline{y}_{i,.75} - \overline{y}_{i,.25}$ satisfies equation (9). The expert is then asked whether this represents opinion. If it does not, then in step 3 the program has failed to find values of $\overline{y}_{i,.75}$ and $\overline{y}_{i,.25}$ that both represent the expert's opinions and satisfy (9).

Step 4 is essentially the same as step 2 but with d_i replacing $\overline{y}_i$. However, one problem that can occur only in step 4 is that the calculated estimate of $(d_{i,.75} - d_{i,.25})^2$ may be non-positive. This happens if the requirement specified in equation (6) is not satisfied. That is, if $(\overline{y}_{i,.75} - \overline{y}_{i,.25}) \leq (\overline{y}_{i-1,.75} - \overline{y}_{i-1,.25})$. Should this occur, then some reassessment is necessary. The program explains to the expert that the interquartile range of $\overline{y}$ should be smaller at $\mathbf{x}_{i-1}$ than at $\mathbf{x}_i$, otherwise the expert should have chosen $\mathbf{x}_i$ (and not $\mathbf{x}_{i-1}$)

as the $(i-1)$th design point. The co-ordinates of each design point and the quartile assessments of $\overline{y}$ at these points are displayed to the expert, who decides which assessments do not adequately represent opinion. The expert then repeats those assessments and all assessments made subsequently.

AN EXAMPLE

To illustrate the use of the elicitation method, we describe in this section a practical example in which subjective opinion was quantified. The 'expert' was an industrial chemist seeking a viable way to manufacture a certain fluoride compound. To produce this compound, two gases are mixed in a diluent and passed through a long tube containing a catalyst. The success of the process is measured by the 'percentage yield'. The amount of the fluoride compound that is produced has a theoretical maximum and percentage yield is defined to equal *(100 × actual yield)/(maximum theoretical yield).* Critical factors which affect this yield are the temperature of the gas (*temp*), the time that the gas is in contact with the catalyst (*time*), and the quantity of each gas (*gas*1 and *gas*2) per unit volume of diluent. The chemist thought that the effect of these factors on the percentage yield would be linear for the range of values he wished to consider. Hence a linear regression model was used for this application,

$$Yield = \beta_1 + \beta_2(temp) + \beta_3(time) + \beta_4(gas1) + \beta_5(gas2) + error.$$

Before having his opinion elicited, the chemist was forewarned of the elicitation questions he would be asked and some advice was given on how he might tackle the questions. He had used a preliminary version of the

method, so this took little time. The interactive computer program that implements the method was then run, and in response to prompts from the computer, the chemist typed in answers to express his opinions.

His first set of answers determined the names and ranges of the independent variables. These were:

temp (°C) :	350–450	*time* (secs.) :	1–10
*gas*1 (%) :	1–5	*gas*2 (%) :	1–4

He was next questioned about experimental error. His assessments of k_1 and k_2 were 5 and 4.5, respectively. Using Exhibit 2, n was estimated at 5 and, from equation (4), ω was estimated at 23.7.

The chemist then assessed the position of constrained points of minimum variance and quartiles of the corresponding $\overline{y}$ and d. The co-ordinates of the selected points are given in Exhibit 3. Values with an asterisk against them were chosen by the computer and the remainder were chosen by the chemist. At point 3, for example, the computer specified the values of 425 and 7.75 for *temp* and *time*, so $\mathbf{a}_3 = (1, 425, 7.75)'$. The expert then assessed values for *gas*1 and *gas*2 (2 in each case) to form the $\mathbf{a}_3$-CMV point. From the CMV design points, the triangular matrix $\mathbf{T}$ was calculated and took the value:

Exhibit 3 about here

$$\begin{bmatrix} 1 & 400 & 7 & 1 & 1 \\ 0 & 25 & -2 & 1 & 1 \\ 0 & 0 & 2.75 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & -1.25 \end{bmatrix}$$

The quartile assessments of $\overline{y}$ and d at the design points are given in

Exhibit 4. Values with an asterisk against them were suggested by the computer and accepted by the expert as representative of his opinions. The value suggested for $d_{.50}$ was always accepted by the chemist but the suggested quartiles were sometimes changed. For $\mathbf{x}_2$, the chemist's initial assessments of $d_{.25}$ and $d_{.75}$ were 11 and 18. These suggested values of 35.97 and 44.0 for $\overline{y}_{.25}$ and $\overline{y}_{.75}$, but the chemist found these unacceptable and assessed them as 36 (essentially the same as suggested) and 43. The computer then suggested the values of 11.9 and 17.6 for $d_{.25}$ and $d_{.75}$, and the chemist felt these represented his opinions satisfactorally. For $\mathbf{x}_5$, the computer initially suggested values of 52.5 and 65.5 for $\overline{y}_{.25}$ and $\overline{y}_{.75}$, and the chemist revised the latter to $\overline{y}_{.75} = 64$. The computer then suggested 51.0 for $\overline{y}_{.25}$ [so that $\overline{y}_{.75} - \overline{y}_{.25}$ still satisfied equation (9)], and this value was accepted by the chemist.

Exhibit 4 about here

In our experience using the elicitation method, the above pattern of reassessment is typical. Suggested values of $d_{.50}$ are rarely changed by the expert but changes to lower and upper quartiles are more common. Another typical feature is that assessments of quartiles are often asymmetric while symmetric values suggested by the computer are usually accepted. Presumably an expert's opinion is somewhat imprecise and could be adequately represented by any of a range of values.

From the median assessments, equation (8) gave the following estimate of $\mathbf{b}$.

$$\mathbf{b} = (-60.05, 0.155, 1.09, 8.70, 5.60).$$

The semi-interquartile range of a standard t-distribution with 5 degrees of freedom is 0.727. The interquartile-range assessments thus gave the following estimates for the non-zero elements of the diagonal matrix, $\mathbf{D}$: 7.57, 15.61,

1.89, 47.30, 7.57 (top-left to bottom-right). Finally the estimates of **U** were obtained from equation (10):

$$\mathbf{U} = \begin{bmatrix} 7317 & -18.59 & -9.76 & 159.2 & 36.33 \\ -18.59 & 0.0474 & 0.020 & -0.425 & -0.097 \\ -9.76 & 0.020 & 0.250 & 0 & 0 \\ 159.2 & -0.425 & 0 & 13.04 & -2.42 \\ 36.33 & -0.097 & 0 & -2.42 & 4.84 \end{bmatrix}$$

The estimated regression coefficients predict that the maximum and minimum yields within the design space will be 87% and 10%, respectively. These are, respectively, at the two boundary points: $temp = 450$, $time = 10$, $gas1 = 5$, $gas2 = 4$; and $temp = 350$, $time = 1$, $gas1 = 1$, $gas2 = 1$. This range of yields seems reasonable in that the chemist believed that a yield in excess of 75% should be achievable at some design points while a very low yield would be obtained at others. The subjective estimate of the accuracy of predicted responses can be determined from the elicited distribution. At the point of maximum predicted yield, the standard error of prediction is 14.7%, while at the minimum it is 15.7%. These are implausibly large as they imply a fair chance of obtaining a yield of more than 100% at the point of maximum yield and a negative yield at the point of minimum yield. Of course, an unbounded distribution model cannot hold exactly for a proportional yield variable; the linear model may not hold throughout the design space; and the assessed predictive variance may be too large in parts of the space. If the last of these possibilities holds, then the assessed distribution overestimates the subjective uncertainty concerning the unknown regression

coefficients. Those of a cautious disposition would regard this error, which corresponds to *under*confidence, as being in the preferable direction.

After the interactive elicitation interview, the chemist was given some explanation of the implications of the hyperparameter values that defined his assessed distribution. He thought the regression coefficient estimates represented his opinion quite well but was not really able to comment on the other hyperparameters, because of their less concrete meaning. His comments on the elicitation procedure itself were favourable. He felt he had been able to give meaningful answers to the questions asked by the computer and had found formulating his answers a stimulating task. He also liked the idea of quantifying his opinion in a form that could be utilized in the design of his experiments and their subsequent analysis.

DISCUSSION

In an assessment method, it seems sensible to elicit more values than the minimum number necessary for calculating the hyperparameters. Some form of averaging can then be used to determine the hyperparameter estimates from the elicitations. This can make the prior distribution a better representation of the expert's opinion. [Tasks of reconciling inconsistencies in an expert's expressed opinions have been discussed by Lindley, Tversky and Brown (1979) and Dickey (1980).] One criticism, then, of the elicitation method presented here is that no form of averaging is used to determine the hyperparameters $\mathbf{b}$ and $\mathbf{U}$. The r-dimensional vector $\mathbf{b}$ and the $r \times r$ matrix $\mathbf{D}$ are estimated from assessments at only r design points. However, this potential fault can be remedied by using the elicitation procedure more

than once, varying the order of the independent variables and/or the values of the constants $a_1, a_2, \ldots, a_r$, so that the design points are not identical on the separate uses of the procedure. This will yield two or more prior distributions, and from them a single 'consensus' distribution can be obtained. French (1985), Genest and Zidek (1986) and Hogarth (1975) review methods for forming consensus distributions.

ACKNOWLEDGEMENTS

The authors are grateful to staff of Imperial Chemical Industries at Runcorn who helped to test the elicitation method. This research was supported by the U.K. Science and Engineering Research Council, by Imperial Chemical Industries and by the U.S. National Science Foundation, Grant DMS-8911548.

REFERENCES

Alpert, M. and Raiffa, H. 'A Progress Report on the Training of Probability Assessors'. In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgement Under Uncertainty : Heuristics and Biases.* Cambridge : Cambridge University Press, 1982.

DeGroot, M. H. *Optimal Statistical Decisions.* London : McGraw-Hill, 1970.

Dickey, J. M. 'Beliefs about Beliefs, a Theory of Stochastic Assessments of Subjective Probabilities.' In J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (Eds.), *Bayesian Statistics*, Valencia, Spain : University Press. *Trabajos de Estadistica*, Vol. 31 (1980), 469–487.

Edwards, W. and Phillips, L. D. 'Man as Transducer for Probabilities in Bayesian Command and Control Systems.' In G. L. Bryan and M. W. Shelley (Eds.), *Human Judgements and Optimality.* New York : Wiley, 1964.

French, S. 'Group Consensus Probability Distributions : A Critical Survey.' In J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (Eds.), *Bayesian Statistics 2*, Amsterdam : North Holland, 1985.

Garthwaite, P. H. *Assessment of Prior Distributions for Normal Linear Models.* Ph.D Thesis, Department of Statistics, University College of Wales, Aberystwyth, 1983.

Garthwaite, P. H. 'Computer Program and User Guide for Quantifying Expert Opinion about Linear Regression Models', *Technical Report 9*, Department of Mathematical Sciences, University of Aberdeen, 1986.

Garthwaite, P. H. 'Fractile Assessments for a Linear Regression Model : An Experimental Study', *Organizational Behavior and Human Decision Processes*, 43 (1989), 188–206.

Garthwaite, P. H. and Dickey, J. M. 'Quantifying Expert Opinion in Linear Regression Models', *Journal of the Royal Statistical Society*, B50 (1988), 462–474.

Genest, C. and Zidek, J. V. 'Combining Probability Distributions : A Critique and an Annotated Bibliography', *Statistical Science*, 1 (1986), 39–46.

Hogarth, R. M. 'Cognitive Processes and the Assessment of Subjective Probability Distributions', *Journal of the American Statistical Association*, 70 (1975), 271–294.

Huber, G. P. 'Methods for Quantifying Subjective Probabilities and Multi-attribute Utilities', *Decision Sciences*, 5 (1974), 430–458.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. 'Interactive Elicitation of Opinion for a Normal Linear Model', *Journal of the American Statistical Association*, 75 (1980), 845–854.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. 'Calibration of Probabilities : The State of the Art to 1980'. In D. Kahneman, P. Slovic and A. Tversky (Eds.), *Judgement Under Uncertainty : Heuristics and Biases*, Cambridge : Cambridge University Press, 1982.

Lindley, D. V., Tversky, A. and Brown, R. V. 'On the Reconciliation of Probability Assessments', *Journal of the Royal Statistical Society*, A142 (1979), 146–180.

Ludke, R. L., Stauss, F. F. and Gustafson, D. H. 'Comparison of Five Methods for Estimating Subjective Probability Distributions', *Organizational Behavior and Human Performance*, 19 (1977), 162–179.

Murphy, A. H. and Winkler, R. L. 'Reliability of Subjective Probability Forecasts of Precipitation and Temperature', *Applied Statistics*, 26 (1977), 41–47.

Peterson, C. R., Snapper, K. J. and Murphy, A. H. 'Credible Interval Temperature Forecasts', *Bulletin of the American Meteorological Society*, 53 (1972), 966–970.

Raiffa, H. A. and Schlaifer, R. S. *Applied Statistical Decision Making.* Boston : Graduate School of Business Administration, Harvard University, 1961.

Slovic, P. and Lichtenstein, S. C. 'Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgement', *Organizational Behavior and Human Performance*, 6 (1971), 649–744.

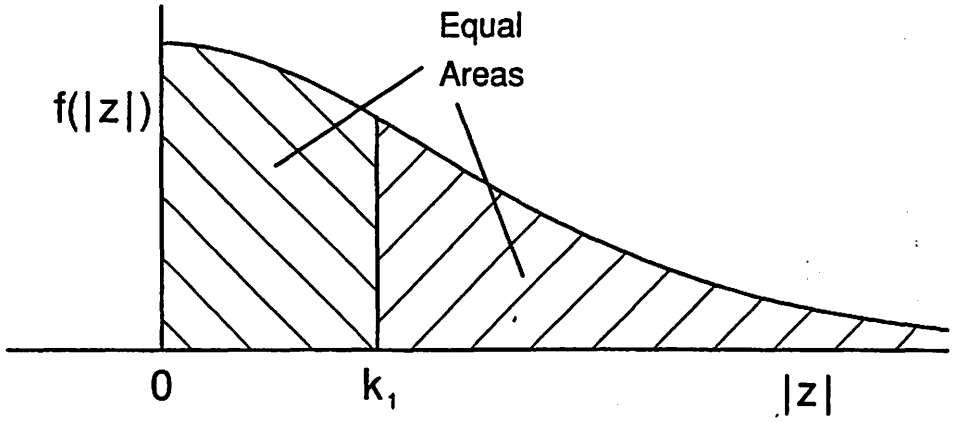
Tversky, A. 'Assessing Uncertainty', *Journal of the Royal Statistical Society*, B36 (1974), 148–159.

Authors' Biographies:

Paul H Garthwaite is a lecturer in the Department of Mathematical Sciences at the University of Aberdeen. He received his Ph.D in statistics from the University College of Wales, Aberystwyth in 1983. His research interests include Bayesian statistics and Monte Carlo methods.

James M Dickey is a professor in the School of Statistics at the University of Minnesota. He received his Ph.D in mathematics from the University of Michigan in 1965, thesis advisor Leanard J. Savage. He was professor and head of the Department of Statistics at the University College of Wales, Aberystwyth, from 1976 to 1980. His research interests include Bayesian statistics, multivariate analysis, special functions, and subjective-probability modeling.

Exhibit 1. Probability distribution function for the absolute difference between two yields at the same design point



n	q_n	k_1/k_2
1	1.000	1.550
2	0.816	1.255
3	0.765	1.164
5	0.727	1.095
7	0.711	1.067
10	0.700	1.047
15	0.691	1.030
20	0.687	1.023
∞	0.684	1

Exhibit 2. Values of the semi-interquartile range q_n of a Student-t distribution on n degrees of freedom and the ratio of elicited medians k_1/k_2 of successive absolute differences for $\alpha = z_1/k_1 = 1/2$.

Point	*Constant*	*Temp*	*Time*	*Gas 1*	*Gas 2*
x_1	1*	400	7	1	1
x_2	1*	425*	5	2	2
x_3	1*	425*	7.75*	2	2
x_4	1*	425*	7.75*	4*	3
x_5	1*	425*	7.75*	4*	1.75*

Exhibit 3. Elicited points of constrained minimun variance

Point	$\overline{y}_{.25}$	$\overline{y}_{.50}$	$\overline{y}_{.75}$	$d_{.25}$	$d_{.50}$	$d_{.75}$
x_1	22	24	26	–	–	–
x_2	36	40	43	11.9*	16*	17.6*
x_3	39.4*	43	46.6*	2	3*	4
x_4	59.8*	66	72.2*	17	23*	27
x_5	51.0*	59	64	−9	−7*	−5

Exhibit 4. Median and quartile assessments at the elicited points of constrained minimum variance