EXPLICIT ESTIMATORS OF PARAMETRIC FUNCTIONS IN NONLINEAR REGRESSION

by

A. Ronald Gallant

Institute of Statistics
Mimeo Series No. 1108
February, 1977
Revised April, 1978

Abstract

The possibility of employing explicitly defined functions of the observations as estimators of parametric functions in nonlinear regression analysis is explored. A general theory of best average mean square error estimation leading to an explicit estimator is set forth. The estimator is shown to be a truncated Fourier series expansion of the Bayes rule which minimizes expected posterior square error loss (equivalent to average mean square error). Sufficient conditions are given for a polynomial estimator to converge in probability to the Bayes rule and for its average mean square error to converge to the minimum achievable as degree increases. In an example it is found that a linear function of the observations outperforms the maximum likelihood estimator and performs nearly as well as the Bayes rule.

Key Words

Nonlinear regression

Explicit estimators

Estimation

Bayes estimator

Fourier series

Average mean square error

1. Introduction

The fundamental difference between linear and nonlinear regression analysis is that the statistics customarily employed in linear regression analysis are explicit functions of the observations while those employed in nonlinear regression analysis are implicitly defined functions of the observations. As an instance, the maximum likelihood estimator of the parameters of a linear model with normal errors is a linear function of the observed responses but, when the model is nonlinear in the parameters, it is defined as the solution of a nonlinear quadratic programming problem and cannot be given an explicit representation. In consequence, there are two sources of error in nonlinear analysis over and above those encountered in linear analysis. The first, a statistic may be incorrectly computed due to the inherent unreliability of iterative nonlinear optimization algorithms. The second, the sampling distribution of a statistic must be approximated using an asymptotic theory which leads, for example, to errors in computing the moments of a statistic or in computing the probability statement associated with an inference. If statistics with explicit representations were available for use in nonlinear analysis some progress might be made in eliminating these sources of error.

An attempt to find such statistics in an estimation context is reported here. A general theory of best average mean square error estimation leading to explicit estimators is set forth in the next two sections. Such estimators are given a Bayesian interpretation as Fourier expansions of the estimator which minimizes expected posterior square error loss in Section 4. In the example of Section 5, a linear function of the observations performs better than the maximum likelihood estimator and nearly as well as the Bayes estimator

according to the criterion of average mean square error. A procedure for finding an exact confidence interval for a parametric function which uses the linear estimator is given in Section 6. The article concludes with some comments.

2. The Structure of the Problem

The observed data y_t, x_t $(t = 1, 2, \ldots, n)$ are assumed to have been generated according to the statistical model

$$y_t = f(x_t, \theta) + e_t \qquad (t = 1, 2, \ldots, n)$$

The form of the response function $f(x, \theta)$ is known, the unknown parameter θ is a p-vector known to be contained in the parameter space Θ, the inputs x_t are known k-vectors, the univariate responses y_t are observed, and the errors e_t are assumed to have mean zero. The distribution function of the errors is denoted as $N_e(e|\sigma)$ where $e = (e_1, e_2, \ldots, e_n)'$, an n-vector. A typical assumption is that the errors are independent normal in which case σ is univariate, but in general σ is an r-vector contained in Σ. The problem of interest is the estimation of a (possibly) nonlinear parametric function $g^*(\gamma)$ where $\gamma = (\theta', \sigma')'$.

As an example, the relative yield at time x of an intermediate substance in a chemical reaction is given as

$$f(x, \theta) = \theta_1(e^{-x\theta_2} - e^{-x\theta_1})/(\theta_1 - \theta_2)$$

under specified conditions. It may be of interest to know the time at which the maximum yield of the substance occurs. This time is given by the parametric function

$$g^*(\theta_1, \theta_2, \sigma) = (\theta_1 - \theta_2)^{-1}\ell n(\theta_1/\theta_2)$$

which depends only trivially on σ . The approach taken here shall be to estimate this parametric function directly by an estimator of the form, say,

$$\hat{g}(y) = a_0 + \Sigma_{t=1}^{n} a_t y_t$$

where y_t is the yield observed at time x_t . Note that this approach differs from the procedure of estimation of the parameters θ_1, θ_2, σ and subsequent evaluation of the parametric function $g^*(\theta_1, \theta_2, \sigma)$ at the estimated parameter values.

The estimators considered are obtained as follows. The observed y_t , denoted as the n-vector

$$y = (y_1, y_2, \cdots, y_n)' ,$$

are transformed according to

$$z = Z(y) \quad .$$

The function Z may be any vector-valued mapping of n-space to m-space but the convenient choice in applications is to take a basis for the polynomials in y of, say, second degree as the components of Z(y) . The class of estimators $\hat{g}$ of the form

$$\hat{g} = a'z$$

where a is an m-vector restricted to a linear space G are considered. Estimators of this form are explicit functions of the observations, viz. $\hat{g}(y) = a'Z(y)$.

The estimator $\hat{g}$ to be chosen from this class is that which minimizes average mean square error with respect to a weighting measure ρ defined on $\Gamma = \Theta \times \Sigma$. That is, one seeks to find $\hat{g}$ of the form $\hat{g} = a'z$ minimizing

$$\text{AMSE } (\hat{g}) = \int_{\Gamma} \mathcal{E}_{\gamma}[\hat{g} - g^*(\gamma)]^2 \, d\rho(\gamma)$$

where a is restricted to G .

The notation $\mathcal{E}_\gamma(\cdot)$ refers to expectation with respect to the distribution of y; that is, with respect to the distribution function

$$N_y[y|f(\theta), \sigma] = N_e[y - f(\theta)|\sigma]$$

where

$$f(\theta) = [f(x_1, \theta), f(x_2, \theta), \ldots, f(x_n, \theta)]' \quad .$$

It shall be assumed throughout that the components of $Z(y)$ are measurable and square integrable with respect to $N_y[y|f(\theta), \sigma]$ and that the functions $g^*(\gamma)$, σ, $\mathcal{E}_\gamma[Z_i(y)]$, and the like are measurable and square integrable with respect to ρ .

3. A Sampling Theoretic Derivation of the Estimator

Let $b_1, b_2, \ldots, b_r$ be m-vectors constituting a basis for $\mathcal{C}$ and arrange them as column vectors in the $m \times r$ matrix

$$B = [b_1 \vdots b_2 \vdots \ldots \vdots b_r] \quad .$$

The following result is derived in the Appendix: in the class of estimators of the form $a'z$ with a in $\mathcal{C}$, that choice of a which minimizes the average mean square error with respect to ρ is

$$\hat{a} = B\{B'[\int_\Gamma \mathcal{E}_\gamma(zz')\, d\rho(\gamma)]B\}^- B'[\int_\Gamma g^*(\gamma)\, \mathcal{E}_\gamma(z)\, d\rho(\gamma)] \quad .$$

The notation $\int_\Gamma \mathcal{E}_\gamma(zz')\, d\rho(\gamma)$ denotes the $m \times m$ matrix with typical element

$$\int_\Gamma \mathcal{E}_\gamma(z_i z_j)\, d\rho(\gamma) \quad .$$

Similarly $\int_\Gamma g^*(\gamma)\, \mathcal{E}_\gamma(z)\, d\rho(\gamma)$ denotes the m-vector with typical element

$$\int_\Gamma g^*(\gamma)\, \mathcal{E}_\gamma(z_i)\, d\rho(\gamma) \quad .$$

The notation A^- denotes any generalized inverse of A . If there are no restrictions on the choice of a then B may be taken as the identity matrix and the formula simplifies to

$$\hat{a} = [\int_\Gamma \mathcal{E}_\gamma(zz')\, d\rho(\gamma)]^- \int_\Gamma g^*(\gamma)\, \mathcal{E}_\gamma(z)\, d\rho(\gamma) \quad .$$

Often, in applications, the error distribution $N_e(e|\sigma)$ is taken the n-variate normal with mean zero and variance-covariance matrix $\sigma^2 I$. If the functions $Z_i(y)$ are taken as polynomials in y then the elements of $\mathcal{E}_\gamma(z)$ and $\mathcal{E}_\gamma(zz')$ may be obtained as the moments about zero of a spherical normal distribution centered at $f(\theta)$; these will be polynomials in $f(x_i, \theta)$ and σ . A natural ordering of the functions $Z_i(y)$ is such that if $m = 1 + n$ the components of $Z(y)$ are a basis for the polynomials in y of first degree, if $m = 1 + n + n(n + 1)/2$ the components of $Z(y)$ are a basis for the polynomials in y of second degree, and so on. The degree of the polynomial may be varied by varying m or by choice of B ; also, a sufficiency reduction may be accomplished by choice of columns for B . The use of B for these purposes may permit some flexibility in writing code to implement the estimator.

4. Bayesian Interpretation of the Estimator

The problem of estimating $g^*(\gamma)$ may be recast in a Bayesian setting by regarding ρ as the prior distribution on Γ and $N_y(y|f(\theta), \sigma)$ as the conditional distribution of y given γ with density $n[y|f(\theta), \sigma]$. Take for the Bayes rule the mean of $g^*(\gamma)$ with respect to the posterior density, denoted by $\tilde{g}(y)$. This choice minimizes expected posterior loss for a square error loss function. Under the previous integrability assumptions, the expected posterior square error loss of an estimator is numerically the same as average mean square error; the different terminology corresponds to the nomenclature of Bayesian or sampling-theoretic inference, respectively. Thus, $\tilde{g}(y)$ is the optimal (unrestricted) choice under either criterion.

For a variety of purposes it may be convenient to approximate the Bayes' rule by truncating a series expansion of $\tilde{g}(y)$. Consider a Fourier series expansion in terms of a set of basis functions $\{Z_i(y)\}_{i=1}^{\infty}$. The most natural choice of a weight function, in the present context, is the marginal distribution of the observations

$$m(y) = \int n[y|f(\theta), \sigma] \, d\rho(\gamma)$$

Let $\{\varphi_i(y)\}_{i=1}^{\infty}$ be a sequence of orthonormal basis functions with respect to $p(y)$ generated from the sequence $\{Z_i(y)\}_{i=1}^{\infty}$ by the Gram-Schmidt process (with linear dependencies in the Z_i accomodated by deletion). The Fourier series expansion of the Bayes rule is

$$\tilde{g}(y) = \Sigma_{i=1}^{\infty} c_i \varphi_i(y)$$

with Fourier coefficients

$$c_i = \int_{\mathcal{Y}} \tilde{g}(y) \varphi_i(y)\ p(y)\ dy \quad .$$

Corollary 1 of the Appendix states that if this expansion is truncated at m then[1]

$$\hat{a}'Z(y) = \Sigma_{i=1}^{m} c_i \varphi_i(y)$$

provided $\mathcal{A} = R^m$.

This result suggests that the choice of a class of functions to which attention is to be restricted in the construction of an explicit estimator may be regarded as a choice of basis functions for a Fourier series approximation of the Bayes rule. A common choice of basis functions for the approximation of a non-periodic function is the collection of polynomials; eg., the Hermite polynomials. Corollary 2 of the Appendix states that if the predictive density function possesses a moment generating function and if $\{Z_i(y)\}_{i=1}^{\infty}$ is a basis for the polynomials then

$$\lim_{m \to \infty} \int_{\mathcal{Y}} [\hat{g}(y) - \tilde{g}(y)]^2\ p(y)\ dy = 0 \quad .$$

Thus, if the explicit estimator is a polynomial and the predictive density possesses a moment generating function then the expected posterior square error loss (average mean square error) of the explicit estimator may be made as nea the achievable minimum as desired by taking the degree of the polynomial suitabl

large. This would seem to serve as further justification for the choice of low order polynomials for the construction of the explicit estimator.

A sampling theorist is interested in the extent to which the explicit estimator constructed from a polynomial inherits the sampling properties of the Bayes rule. Theorems 4 and 5 of the Appendix show that if the errors e_t are independent normal with common variance σ^2 and if $f(\theta)$ is continuous then

$$\lim_{m \to \infty} \int_{\mathcal{Y}} [\hat{g}(y) - \tilde{g}(y)]^2 \, p(y) \, dy = 0$$

implies that $\hat{g}(y)$ converges in probability to $\tilde{g}(y)$ for any $\gamma_0 = (\theta_0', \sigma_0)'$ for which $\rho(\Gamma_0) > 0$ for all open sets Γ_0 containing γ_0 ; convergence in probability for such γ_0 implies convergence in distribution. In this sense, the effect upon the sampling distribution of the explicit estimator corresponding to some choice of a weighing measure ρ may be anticipated from the sampling behavior of the Bayes estimator corresponding to that choice of ρ .

5. Example

Consider the statistical model

$$y_t = f(x_t, \theta) + e_t$$

with response function

$$f(x, \theta) = \begin{cases} \theta_1(e^{-x\theta_2} - e^{-x\theta_1})/(\theta_1 - \theta_2) & \theta_1 \neq \theta_2 \ , \\ \theta_1 x \, e^{-x\theta_1} & \theta_1 = \theta_2 \ , \end{cases}$$

parameter space

$$\Theta = \{(\theta_1, \theta_2): \theta_1 \geq \theta_2\} \ ,$$

and design

$$\{x_t\} = \{.25, .5, 1, 1.5, 2, 4, .25, .5, 1, 1.5, 2, 4\} \ .$$

This model has appeared several times in the nonlinear regression literature (Box and Lucas 1959; Guttman and Meeter 1964; Gallant 1976) and its use here will facilitate comparison with standard methods of nonlinear regression analysis.

As noted earlier, the response function $f(x, \theta)$ describes the relative yield at time x of an intermediate substance in a chemical reaction; consider estimation of the time at which the maximum yield occurs:

$$g^*(\theta) = (\theta_1 - \theta_2)^{-1} \ell n(\theta_1/\theta_2) \quad .$$

Some additional notation required is: let $F(\theta)$ be the $n \times p$ matrix with typical element $(\partial/\partial\theta_j) f(x_t, \theta)$, where t is the row index and j is the column index; $C = [F'(\theta)F(\theta)]^{-1}$.

To obtain the explicit estimator of the time of maximum yield, the weighting measure ρ is taken as the product measure comprised of a uniform distribution on the ellipsoid $(\theta - \tau)'C^{-1}(\theta - \tau) \leq r^2$ with C evaluated at τ and the inverted gamma distribution (Zellner 1971, p. 371) with parameters v and s^2 . Or, if ρ is viewed as a prior measure then θ has the uniform distribution on the ellipsoid $\{\theta: (\theta - \tau)'C^{-1}(\theta - \tau) \leq r^2\}$ and σ is independently distributed as the inverted gamma distribution. Appropriate transformations yield an expression which is convenient for use with quadrature formulae:

$$\int g(\theta, \sigma)\, d\rho(\gamma)$$

$$= [\pi r^2 \Gamma(\alpha)]^{-1} \int_0^\infty u^{\alpha-1} \{ \int_{y'y \leq r^2} g[R'y + \tau\ ,\ 1/\sqrt{(\gamma u)}]\, dy \}\, e^{-u}\, du \ .$$

where $\alpha = v/2$, $\gamma = 2/(vs^2)$, and R is obtained by factoring C as $C = R'R$. The integral within braces may be evaluated by means of a 12-point quadrature formula (Stroud 1971, p. 281)[2/] which integrates polynomials in y up to degree

seven exactly; the outer integral may be evaluated using a 4-point quadrature formula (IBM 1968, DQL4) which integrates polynomials in u up to degree seven exactly. In terms of the arguments of $g(\theta, \sigma)$, if g is a polynomial of degree less than or equal to seven in θ and of degree less than or equal to v-2 in σ it will be integrated exactly.

The parametric choices employed here are $\tau = (1.4, .4)$, $s = .025$, $v = 10$, and $r^2 = 8 s^2$. These choices were motivated by the study of the sampling distribution of the least squares estimator reported in (Gallant 1976). With these choices, the numerical integration procedure above is equivalent to the substitution of a discrete prior measure $\hat{\rho}$,

$$\int g(\theta, \sigma)\, d\hat{\rho}(\gamma) = \Sigma_i\, p_i g(\theta_i, \sigma_i) \quad ,$$

for the measure ρ . This discrete measure is displayed in Table 1. In fact, it may be best to regard $\hat{\rho}$ as the prior in the sequel.

Table 1 here

As a referee points out, the central idea behind the choice of parameter points for Table 1 (and later Table 2) becomes lost in the details. The idea is that a numerical quadrature formula approximates an integral by a weighted average; viz.

$$\int_\Gamma g(\gamma)\, d\rho(\gamma) \doteq \Sigma_{i=1}^{N}\, p_i\, g(\gamma_i) \quad .$$

The choice of points $\gamma_i = (\theta_i, \sigma_i)$ and weights p_i are determined according to numerical analytic considerations so as to achieve an accurate approximation. Table 1 displays the choice of γ_i and p_i implied by the use the two quadrature formulas cited above. It seems useful later in the discussion (Table 2) to display the bias and mean square error surfaces of the Bayes estimator, the

explicit estimator, and the maximum likelihood estimator. Since the moments of the Bayes estimator and the maximum likelihood estimator are somewhat costly to compute and since they must be computed at the 49 points γ_i displayed in Table 1 to obtain average mean square error.it then becomes convenient to use these 49 points of the quadrature formula for the purpose of display. This is the reason for the choice of the 49 points $\gamma_i = (\theta_i, \sigma_i)$ used in Tables 1 and 2.

Two explicit estimators are considered, a linear function[3/] in y

$$\hat{g}_1(y) = a_0 + \Sigma_{i=1}^{12} a_i y_i$$

and a quadratic function in y

$$\hat{g}_2(y) = a_0 + \Sigma_{i=1}^{12} a_i y_i + \Sigma_{i\leq j}^{12} a_{ij} y_i y_j \quad .$$

The performance of these estimators relative to the Bayes estimator and the maximum likelihood estimator is shown in Table 2. The table gives the value of the parametric function, the bias and the root mean square error for each estimator at each parameter value of the quadratic formula together with the root average mean square error for each estimator.

- -

Table 2 here

- -

These conclusions obtain from Table 2 according to the criterion of average mean square error. The slight improvement in the performance of the explicit quadratic estimator $\hat{g}_2$ relative to the explicit linear estimator $\hat{g}_1$ is not worth the added complexity entailed. The explicit linear estimator $\hat{g}_1$ is a serious competitor of the Bayes' estimator in view of its simplicity. The explicit linear estimator is improvement over the maximum likelihood estimator.

The details of the computations are as follows. The moments of the explicit estimators were obtained exactly using the usual formulae for the moments of linear and quadratic functions of normally distributed random variables. The moments of the Bayes' estimator were obtained from 2000 Monte-Carlo trials using the control variate method of variance reduction (Hammersly and Handscomb 1964, p. 59-60). The random number generator employed was GGNOF (IMSL 1975); the explicit linear estimator was used as the control variate. The estimated standard error of $\sqrt{AMSE}$ for the Bayes estimator was .000034695. The moments of the maximum likelihood estimator were approximated using the moments of the asymptotic distribution. As a partial check on accuracy the moments of $g(\hat{\theta})$ were estimated for $\gamma = (1.4, .4, .025)$ from an existing set of 4000 simulated values of the least squares estimator $\hat{\theta}$ (Gallant 1976); the results were Est $\{\mathcal{E}[g(\hat{\theta})]\} = 1.25283$ and Est$\{\text{Var}[g(\hat{\theta})]\} = .00135907$ yielding Est $\{\sqrt{MSE}[g(\hat{\theta})]\} = .03687$ which compares favorably with .03670 , the value reported in Table 2. All code was written in double precision, save the use of GGNOF, using an IBM 370/165 .

6. Confidence Intervals

Any procedure for finding an exact confidence region for γ may be used to find an exact confidence region for $g^*(\gamma)$. Let $\hat{\Gamma}$ be a random set function depending on y such that

$$P_\gamma(\{y: \gamma \text{ in } \hat{\Gamma}\}) \geq 1 - \alpha$$

Let $\hat{G} = \{g^*(\gamma): \gamma \text{ in } \hat{\Gamma}\}$; since

$$\{y: g^*(\gamma) \text{ in } \hat{G}\} \supset \{y: \gamma \text{ in } \hat{\Gamma}\}$$

it follows that

$$P_\gamma(\{y: g^*(\gamma) \text{ in } \hat{G}\}) \geq 1 - \alpha \quad .$$

Thus $\hat{G}$ is an exact confidence interval for $g^*(\gamma)$ - exact in the sense that the probability statement is correct, not approximate.

When g^* does not depend on σ , as in the example, consider

$$U = a_0 + A'y$$

where a_0 is a q-vector of known constants and A is an n by q matrix of known constants. Then, following Hartley (1964),

$$\hat{\theta} = \{\theta : [U - a_0 - A'f(\theta)]'(A' A)^{-1}[U - a_0 - A'f(\theta)]/[q\, s^2(\theta)] \leq F_\alpha\}$$

where

$$s^2(\theta) = [y - f(\theta)]'[I - A(A'A)^{-1}A'][y - f(\theta)]/(n-q)$$

is an exact confidence region for θ; F_α denotes the upper α percentage point of the F distribution with q degrees freedom for the numerator and n-q degrees freedom for the denominator. An exact confidence interval for $g^*(\theta)$ is

$$\hat{G} = \{g^*(\theta): \theta \text{ in } \hat{\theta}\} \quad .$$

To illustrate with the example, consider the sample

$$y = \begin{bmatrix} 0.31753 \\ 0.42208 \\ 0.62973 \\ 0.56630 \\ 0.54830 \\ 0.26603 \\ 0.29474 \\ 0.49830 \\ 0.58632 \\ 0.63670 \\ 0.56983 \\ 0.26299 \end{bmatrix}$$

generated according to the model with $\theta = (1.4, .4)$ and $\sigma = .025$ (corresponding

to i = 49 in Tables 1 and 2). For a_0 and A_1 take:[4]

$$a_0 = \begin{bmatrix} 1.44126 \\ 0.19376 \end{bmatrix} ,$$

$$A = \begin{bmatrix} -0.27126 & 0.45343 \\ -0.33201 & 0.58293 \\ -0.18570 & 0.42717 \\ 0.035484 & 0.14643 \\ 0.21061 & -0.077778 \\ 0.38468 & -0.30496 \\ -0.27126 & 0.45343 \\ -0.33201 & 0.58293 \\ -0.18570 & 0.42717 \\ 0.035484 & 0.14653 \\ 0.21061 & -0.077778 \\ 0.38468 & -0.30496 \end{bmatrix}$$

With these choices, the first element of $U = a_0 + A'y$ is the best linear average mean square error estimator of $g^*(\theta)$ and the second element of U is the best linear average mean square error of θ_1 .

The estimate of $g^*(\theta)$ is the value 1.2255 observed for U_1 and the exact confidence interval is

$$\hat{G} = (1.0730 , 1.3138) .$$

By way of comparison, the maximum likelihood estimate of $g^*(\theta)$ is 1.2077 and a confidence interval obtained from the asymptotic theory of the least squares estimator is

$$(1.1206 , 1.2948)$$

which is approximate, not exact. The true value of $g^*(\theta)$ is 1.2528 .

This procedure for finding a confidence interval is entirely analagous to Scheffé's method for finding multiple confidence intervals which are simultaneously valid. The parametric confidence statement "θ in $\hat{\Theta}$" may be used for several different functions of θ yielding a system of confidence intervals which are simultaneously valid. As with Scheffé's method, the procedure is conservative.

In this instance, the interval was 38% longer than that found using the asymptotic theory of the maximum likelihood estimator of $g^*(\theta)$.

The first attempt to find an exact confidence interval for this sample used the statistic $[U_1 - a_{01} - A'_1 f(\theta)]'(A'_1 A_1)^{-1}[U_1 - a_{01} - A'_1 f(\theta)]/s_1^2(\theta)$ to find the preliminary region $\hat{\Theta}$. However, the region was constrained on the left only by the parametric constraint $\theta_2 \leq \theta_1$ and appeared to continue indefinitely to the right. The addition of the best linear average mean square error estimator of θ_1 as a second componant of U remedied the problem and produced a small elliptically shaped region for $\hat{\Theta}$.

7. Discussion

It is seen that linear estimators for use in nonlinear regression analysis may be found by truncating a Fourier series expansion of a Bayes rule. And that the linearly of these estimators may be exploited to find exact confidence intervals for a parametric function of interest. Subject to the regularity conditions,[5] it is seen that the average mean square error of a polynomial explicit estimator may be made as near the minimum achievable as is desired by taking the degree of the polynomial suitably large. And that the polynomial explicit estimator converges in probability to the Bayes rule as degree increases.

The most serious deterrent to use of the explicit estimator in applications is the requirement that the user produce a prior. However, in nonlinear regression analysis one must usually have some knowledge of the situation in order to find starting values for computing the least squares estimator. But the fact that different investigators will obtain different results depending on the subjective choice of a prior when all else remains the same is disturbing . On the other hand, one can envisage an industrial application involving repetitive estimations for the purpose of calibration where the simplicity of a linear estimator would be an overwhelming advantage. The estimator may find use in such contexts.

It is possible to develop a formal theory of unbiased estimation by borrowing the ideas of estimability from linear regression and proceeding along lines roughly analogous to those used here. In examples, there was the problem that unbiased estimators of the form $a'z$ usually did not exist and another that, when they did, they depended on only one or two of the n observations. This did not appear to be a fruitful line of inquiry and was abandoned.

It is of interest to a referee, and perhaps of general interest, to obtain the explicit estimator for a model which is linear in the parameters. The classical problem is the estimation of $\lambda'\theta$ for a model of the form $y = X\theta + e$ where attention is restricted to estimators of the form $\hat{\lambda'\theta} = a_0 + a_1'y$. Rather than find the best linear unbiased estimator, consider the linear explicit estimation which is obtained by solving the system of equations

$$\int_\Gamma \mathcal{E}_\gamma \begin{pmatrix} 1 & y' \\ y & yy' \end{pmatrix} d\rho \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \int_\Gamma \lambda'\theta \, \mathcal{E}_\gamma \begin{pmatrix} 1 \\ y \end{pmatrix} d\rho \quad .$$

Upon evaluation of the integrals these equations obtain

$$\begin{pmatrix} 1 & \bar{\theta}'X' \\ X\bar{\theta} & X\bar{V}_\theta X' + X\bar{\theta}\bar{\theta}'X' + \bar{\sigma}^2 I \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \bar{\theta}' \\ X\bar{V}_\theta + X\bar{\theta}\bar{\theta}' \end{pmatrix} \lambda$$

where $\bar{\theta} = \int \theta \, d\rho$, $\bar{\sigma}^2 = \int \sigma^2 d\rho$, and $\bar{V}_\theta = \int (\theta - \bar{\theta})(\theta - \bar{\theta})' d\rho$. This system is equivalent to the system

$$\begin{pmatrix} 1 & \bar{\theta}'X' \\ 0 & X\bar{V}_\theta X' + \bar{\sigma}^2 I \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \bar{\theta}' \\ X\bar{V}_\theta \end{pmatrix} \lambda$$

which has solution

$$\hat{a}_0 = \bar{\theta}'\lambda - \bar{\theta}X'\hat{a}_1$$

$$\hat{a}_1 = X(X'X + \bar{\sigma}^2 \bar{V}_\theta^{-1})^{-1} \lambda \quad .$$

The estimator becomes

$$\hat{\lambda'\theta} = \lambda'(X'X + \bar{\sigma}^2 \bar{V}_\theta^{-1})^{-1}(X'y + \bar{\sigma}^2 \bar{V}_\theta^{-1} \bar{\theta}) \quad .$$

An estimator of θ itself may be inferred from this expression.

By virtue of the discussion in Section 4, this estimator is coincident with the Bayes rule whenever the Bayes rule is linear in the observations. As an example, let the likelihood be the n-variate normal with mean $X\theta$ and variance-covariance matrix $\sigma^2 I$; let the prior be the natural conjugate prior. The natural conjugate prior is comprised of a conditional distribution of θ given σ^2 which is a p-variate normal with mean $\bar{\theta}$ and variance-covariance matrix $\sigma^2\Sigma$ and a marginal distribution for σ which is the inverted gamma with parameters s and $v > 2$. Then $\int\theta d\rho = \bar{\theta}$, $\int\sigma^2 d\rho = vs^2/(v-2)$, and $\int(\theta-\bar{\theta})(\theta-\bar{\theta})' d\rho = [vs^2/(v-2)]\,\Sigma$. Substituting into the formula for $\hat{\lambda'\theta}$ one obtains

$$\hat{\lambda'\theta} = \lambda'(X'X + \Sigma^{-1})^{-1}(X'y + \Sigma^{-1}\bar{\theta})$$

which is immediately recognized as the Bayes rule.

The class of estimators of the form

$$\hat{\lambda'\theta} = \lambda'(X'X + \bar{\sigma^2 V}^{-1})^{-1}(X'y + \bar{\sigma^2 V}_{\theta}^{-1}\bar{\theta})$$

is quite rich. It includes the Bayes rule with normal likelihood and natural conjugate prior and any Bayes rule which is linear in the observations. It includes both the least squares estimator and the ridge regression estimator.

i	Parameter Value			Weight
	θ_{1i}	θ_{2i}	σ_i	p_i
1	1.52972	0.40788	0.01824	0.01297
2	1.27028	0.39212	0.01824	0.01297
3	1.40000	0.43339	0.01824	0.01297
4	1.40000	0.36661	0.01824	0.01297
5	1.44837	0.41539	0.01824	0.02157
6	1.35163	0.40951	0.01824	0.02157
7	1.44837	0.39049	0.01824	0.02157
8	1.35163	0.38461	0.01824	0.02157
9	1.49649	0.43070	0.01824	0.00923
10	1.30351	0.41897	0.01824	0.00923
11	1.49649	0.38103	0.01824	0.00923
12	1.30351	0.36930	0.01824	0.00923
13	1.52972	0.40788	0.02625	0.05084
14	1.27028	0.39212	0.02625	0.05084
15	1.40000	0.43339	0.02625	0.05084
16	1.40000	0.36661	0.02625	0.05084
17	1.44837	0.41539	0.02625	0.08456
18	1.35163	0.40951	0.02625	0.08456
19	1.44837	0.39049	0.02625	0.08456
20	1.35163	0.38461	0.02625	0.08456
21	1.49649	0.43070	0.02625	0.03618
22	1.30351	0.41897	0.02625	0.03618
23	1.49649	0.38103	0.02625	0.03618
24	1.30351	0.36930	0.02625	0.03618
25	1.52972	0.40788	0.04231	0.01025
26	1.27028	0.39212	0.04231	0.01025
27	1.40000	0.43339	0.04231	0.01025
28	1.40000	0.36661	0.04231	0.01025
29	1.44837	0.41539	0.04231	0.01704
30	1.35163	0.40951	0.04231	0.01704
31	1.44837	0.39049	0.04231	0.01704
32	1.35163	0.38461	0.04231	0.01704
33	1.49649	0.43070	0.04231	0.00729
34	1.30351	0.41897	0.04231	0.00729
35	1.49649	0.38103	0.04231	0.00729
36	1.30351	0.36930	0.04231	0.00729
37	1.52972	0.40788	0.09843	0.00002
38	1.27028	0.39212	0.09843	0.00002
39	1.40000	0.43339	0.09843	0.00002
40	1.40000	0.36661	0.09843	0.00002
41	1.44837	0.41539	0.09843	0.00003
42	1.35163	0.40951	0.09843	0.00003
43	1.44837	0.39049	0.09843	0.00003
44	1.35163	0.38461	0.09843	0.00003
45	1.49649	0.43070	0.09843	0.00001
46	1.30351	0.41897	0.09843	0.00001
47	1.49649	0.38103	0.09843	0.00001
48	1.30351	0.36930	0.09843	0.00001
49	1.40000	0.40000	0.02500	0.0

2. Bias and Root Mean Square Error of the Explicit Estimator, Bayes' Estimator, and the Maximum Likelihood Estimator of the Time of Maximum Yield

		Explicit Estimators							
		Linear		Quadratic		Bayes' Estimator		MLE	
i	$g^*(\theta_i)$	Bias	√MSE	Bias	√MSE	Bias	√MSE	Bias	√MSE
1	1.17830	0.02820	0.03270	0.02903	0.03300	0.02640	0.02496	0.0	0.02480
2	1.33851	-0.03340	0.03727	-0.03247	0.03686	-0.02239	0.02501	0.0	0.02916
3	1.21310	0.01704	0.02376	0.01656	0.02329	0.01151	0.02161	0.0	0.02613
4	1.29663	-0.01514	0.02243	-0.01559	0.02295	-0.01144	0.02163	0.0	0.02753
5	1.20910	0.01785	0.02434	0.01749	0.02380	0.00704	0.02027	0.0	0.02575
6	1.26747	-0.00395	0.01702	-0.00448	0.01738	-0.00194	0.02289	0.0	0.02735
7	1.23907	0.00678	0.01789	0.00632	0.01744	0.00263	0.02127	0.0	0.02625
8	1.29970	-0.01686	0.02363	-0.01729	0.02419	-0.00725	0.02199	0.0	0.02789
9	1.16859	0.03204	0.03607	0.03250	0.03614	0.02449	0.02741	0.0	0.02491
10	1.28317	-0.01000	0.01934	-0.00994	0.01977	-0.00710	0.02002	0.0	0.02795
11	1.22640	0.01153	0.02017	0.01166	0.01981	0.00898	0.01908	0.0	0.02579
12	1.35002	-0.03789	0.04135	-0.03735	0.04115	-0.02753	0.03112	0.0	0.02909
13	1.17830	0.02820	0.03692	0.02911	0.03664	0.02494	0.03309	0.0	0.03569
14	1.33851	-0.03340	0.04102	-0.03239	0.04099	-0.03007	0.03695	0.0	0.04196
15	1.21310	0.01704	0.02929	0.01664	0.02885	0.01520	0.02845	0.0	0.03761
16	1.29663	-0.01514	0.02822	-0.01552	0.02861	-0.01471	0.02802	0.0	0.03962
17	1.20910	0.01785	0.02977	0.01756	0.02913	0.01259	0.02906	0.0	0.03705
18	1.26747	-0.00395	0.02415	-0.00440	0.02457	-0.00321	0.02926	0.0	0.03936
19	1.23907	0.00678	0.02477	0.00640	0.02426	0.00469	0.02790	0.0	0.03779
20	1.29970	-0.01686	0.02918	-0.01721	0.02982	-0.01211	0.02969	0.0	0.04013
21	1.16859	0.03204	0.03993	0.03258	0.03973	0.03121	0.03571	0.0	0.03571
22	1.28317	-0.01000	0.02584	-0.00986	0.02649	-0.00914	0.02649	0.0	0.04027
23	1.22640	0.01153	0.02647	0.01174	0.02586	0.01165	0.02563	0.0	0.03712
24	1.35002	-0.03789	0.04476	-0.03727	0.04486	-0.03476	0.04018	0.0	0.04187
25	1.17830	0.02820	0.04765	0.02934	0.04677	0.04050	0.04787	0.0	0.05754
26	1.33851	-0.03340	0.05089	-0.03216	0.05172	-0.04443	0.05237	0.0	0.06754
27	1.21310	0.01704	0.04201	0.01667	0.04159	0.02285	0.03675	0.0	0.06062
28	1.29663	-0.01514	0.04128	-0.01528	0.04166	-0.02189	0.03600	0.0	0.06389
29	1.20910	0.01785	0.04235	0.01750	0.04148	0.02273	0.03912	0.0	0.05975
30	1.26747	-0.00395	0.03860	-0.00417	0.03920	-0.00561	0.03250	0.0	0.06346
31	1.23907	0.00678	0.03899	0.00663	0.03831	0.00851	0.03273	0.0	0.06092
32	1.29970	-0.01686	0.04194	-0.01697	0.04279	-0.02180	0.03849	0.0	0.06469
33	1.16859	0.03204	0.05001	0.03281	0.04921	0.04696	0.05255	0.0	0.05756
34	1.28317	-0.01000	0.03968	-0.00963	0.04060	-0.01437	0.03233	0.0	0.06491
35	1.22640	0.01153	0.04010	0.01197	0.03904	0.01640	0.03301	0.0	0.05984
36	1.35002	-0.03789	0.05395	-0.03704	0.05470	-0.05235	0.05754	0.0	0.06749
37	1.17830	0.02820	0.09368	0.03103	0.09044	0.06311	0.06893	0.0	0.13385
38	1.33851	-0.03340	0.09538	-0.03047	0.09924	-0.06961	0.07975	0.0	0.15736
39	1.21310	0.01704	0.09095	0.01956	0.09057	0.03653	0.04546	0.0	0.14104
40	1.29663	-0.01514	0.09061	-0.01360	0.09139	-0.03638	0.04126	0.0	0.14860
41	1.20910	0.01785	0.09110	0.01948	0.08953	0.03933	0.04596	0.0	0.13900
42	1.26747	-0.00395	0.08943	-0.00248	0.09092	-0.00660	0.02791	0.0	0.14761
43	1.23907	0.00678	0.08959	0.00831	0.08838	0.01261	0.02662	0.0	0.14172
44	1.29970	-0.01686	0.09091	-0.01529	0.09252	-0.03781	0.04393	0.0	0.15051
45	1.16859	0.03204	0.09491	0.03449	0.09222	0.07476	0.07675	0.0	0.13392
46	1.28317	-0.01000	0.08990	-0.00794	0.09278	-0.02217	0.02464	0.0	0.15101
47	1.22640	0.01153	0.09008	0.01365	0.08773	0.02529	0.03212	0.0	0.13921
48	1.35002	-0.03789	0.09704	-0.03535	0.10029	-0.05423	0.05419	0.0	0.15702
49	1.25276	0.00143	0.02275	0.00100	0.02267	0.00072	0.02915	0.0	0.03479
√AMSE:			0.03237		0.03234		0.03095		0.04122

Appendix: Theorems and Proofs

A.1. A Sampling Theoretic Derivation of the Estimator

Let the columns b_i of B be a basis for $\mathcal{G}$ and define

$$g_{b_i}(\gamma) = e_\gamma(b_i' z) \qquad (i = 1, 2, \ldots, r) \ .$$

Let H be the r x r matrix with typical element

$$h_{ij} = \int_\Gamma g_{b_i}(\gamma) g_{b_j}(\gamma)\, d\rho(\gamma) \qquad (i, j = 1, 2, \ldots, r)$$

and let h be the r-vector with typical element

$$h_i = \int_\Gamma g_{b_i}(\gamma) g^*(\gamma)\, d\rho(\gamma) \qquad (i = 1, 2, \ldots, r) \ .$$

The matrix H is positive semi-definite since, for an arbitrary r-vector α ,

$$\alpha' H \alpha = \int_\Gamma [\Sigma_{i=1}^r \alpha_i g_{b_i}(\gamma)]^2 \, d\rho(\gamma) \geq 0 \ .$$

Moreover, the equations $H\alpha = h$ are consistent with a solution

$$\bar{\alpha} = H^- h$$

where H^- denotes any generalized inverse of H . Set

$$\bar{g}(\gamma) = \Sigma_{i=1}^r \bar{\alpha}_i \, g_{b_i}(\gamma) \ .$$

Lemma 1. Let α be an arbitrary r-vector. Then

$$\int_\Gamma [g^*(\gamma) - \Sigma_{i=1}^r \alpha_i g_{b_i}(\gamma)]^2 \, d\rho(\gamma)$$

$$= (\alpha - \bar{\alpha})' H (\alpha - \bar{\alpha}) + \int_\Gamma [g^*(\gamma) - \bar{g}(\gamma)]^2 \, d\rho(\gamma) \ .$$

Proof: Add and subtract $\bar{g}$ and expand the integrand to obtain

$$\int (g^* - \Sigma_i \alpha_i g_{b_i})^2 \, d\rho$$

$$= \int (g^* - \bar{g})^2 \, d\rho + \int (\bar{g} - \Sigma_i \alpha_i g_{b_i})^2 \, d\rho$$

$$+ 2 \int (g^* - \bar{g}) (\bar{g} - \Sigma_i \alpha_i g_{b_i}) \, d\rho \quad .$$

By substituting $\bar{g} = \Sigma_i \bar{\alpha}_i g_{b_i}$ in the last two terms one obtains

$$= \int (g^* - \bar{g})^2 \, d\rho + \int [\Sigma_i (\bar{\alpha}_i - \alpha_i) \, g_{b_i}]^2 \, d\rho$$

$$+ 2 \int \Sigma_i (\bar{\alpha}_i - \alpha_i) \, g_{b_i} (g^* - \Sigma_j \bar{\alpha}_j g_{b_j}) \, d\rho$$

Substitution, using the equations defining H and h , yields

$$= \int (g^* - \bar{g})^2 \, d\rho + (\bar{\alpha} - \alpha)' H (\bar{\alpha} - \alpha) + 2(\bar{\alpha} - \alpha) \, (h - H\bar{\alpha}) \quad .$$

The cross product term is zero because $\bar{\alpha}$ is a solution of $H\alpha = h$. □

Let $V(\gamma)$ be the variance-covariance matrix of z with typical element $v_{ij}(\gamma)$. Let $\bar{V}$ be the m x m matrix with typical element

$$\bar{v}_{ij} = \int_\Gamma v_{ij}(\gamma) \, d\rho(\gamma) \quad .$$

Note that $\bar{V}$ is positive semi-definite since, for an arbitrary m-vector a,

$$a'\bar{V}a = \int_\Gamma a'V(\gamma) \, a \, d\rho(\gamma) \geq 0$$

Lemma 2. Let a be an arbitrary choice of an m-vector from $\mathcal{C}$ and let the r-vector α satisfy $a = B\alpha$. Then

$$\text{AMSE}(a'z) = a'\bar{V}a + (\alpha - \bar{\alpha})'H(\alpha - \bar{\alpha}) + \int_\Gamma [g^*(\gamma) - \bar{g}(\gamma)]^2 \, d\rho(\gamma) \, .$$

Proof:

$$\text{AMSE}(a'z) = \int \text{Var}(a'z) + \text{Bias}\ (a'z)\ d\rho$$

$$= \int_\Gamma a'V(\gamma)a + [g^*(\gamma) - \alpha'\mathcal{E}_\gamma(B'z)]^2\ d\rho(\gamma)$$

By integrating the first term and applying Lemma 1 to the second, one obtains

$$= a'\bar{V}a + (\alpha - \bar{\alpha})'H(\alpha - \bar{\alpha}) + \int(g^* - \bar{g})^2\ d\rho\ .\ \square$$

Theorem 1. In the class of estimators of the form $a'z$ with a in $\mathcal{C}$, that choice of a which minimizes

$$\text{AMSE}(a'z) = \int_\Gamma [a'z - g^*(\gamma)]^2\ d\rho(\gamma)$$

is

$$\hat{a} = B(B'\bar{V}B + H)^-\ h\ .$$

An alternative form, for computations, is

$$\hat{a} = B[B'\int_\Gamma \mathcal{E}_\gamma(zz')\ d\rho(\gamma)B]^-\ B'\int_\Gamma g^*(\gamma)\ \mathcal{E}_\gamma(z)\ d\rho(\gamma)$$

Proof: Rearrange the terms of the expression in Lemma 2,

$$\text{AMSE}(\alpha'B'z) = \alpha'(B'\bar{V}B + H)\ \alpha - 2\alpha'H\bar{\alpha} + \text{const.}$$

This is a quadratic equation in α with stationary point

$$\hat{\alpha} = (B'\bar{V}B + H)^-H\bar{\alpha}$$

and positive semi-definite Hessian matrix $B'\bar{V}B + H$. Note that $H\bar{\alpha} = h$ to obtain the first form given in the theorem. The second form is obtained by verifying these equalities algebraicly:

$$H = B'\int\mathcal{E}(z)\mathcal{E}'(z)\ d\rho\ B\ ,$$

$$B'\bar{V}B = B'\int\mathcal{E}(zz') - \mathcal{E}(z)\mathcal{E}'(z)\ d\rho\ B\ ,$$

and

$$h = B'\int g^*\mathcal{E}(z)\ d\rho\ .\ \square$$

A.2. Bayesian Interpretation of the Estimator

The reference for notation, terminology, and general results is Hewitt and Stromberg (1965, Ch. IV, Sec. 16).

Define the measure μ on the Lebesgue subsets of $\mathcal{Y} \times \Gamma$ by

$$\mu(A) = \int_{\Gamma}\int_{\mathcal{Y}} I_A(y, \gamma)\, n[y|f(\theta), \sigma]\, dy\, d\rho(\gamma)$$

where $\mathcal{Y}$ is n-dimensional Euclidean space and $I_A(y, \gamma)$ denotes the indicator function of the set A . Consider the collection of measurable, square integrable functions with respect to this (probability) measure and denote this collection by $\mathcal{L}_2(\mathcal{Y} \times \Gamma, \mu)$. The inner product between g_1, g_2 in $\mathcal{L}_2(\mathcal{Y} \times \Gamma, \mu)$ is

$$\langle g_1, g_2 \rangle = \int_{\Gamma}\int_{\mathcal{Y}} g_1(y, \gamma) g_2(y, \gamma)\, d\mu(y, \gamma)$$

and the corresponding norm is

$$\|g\| = [\langle g, g \rangle]^{\frac{1}{2}} = \left[\int_{\Gamma}\int_{\mathcal{Y}} g^2(y, \gamma)\, d\mu(y, \gamma)\right]^{\frac{1}{2}} .$$

These definitions comprise a Hilbert space on the collection of measurable, square integrable functions with argument (y, γ) (Hewlitt and Stromberg 1965, Eg. 16.8).

An estimator, $\hat{g}(y)$ is a measurable, square integrable function of y only, that is a member of $\mathcal{L}_2(\mathcal{Y} \times \Gamma, \mu)$ which is constant with respect to variation in γ . The average mean square error of an estimator is

$$\text{AMSE}\,[\hat{g}(y)] = \|\hat{g} - g^*\|^2 ,$$

the square of the distance between the estimator $\hat{g}(y)$ and the parametric function $g^*(\gamma)$. The mean of $g^*(\gamma)$ with respect to the posterior density

$$\tilde{g}(y) = \int_{\Gamma} g^*(\gamma)\, n[y|f(\theta), \sigma^2]\, d\rho(\gamma)/p(y) ,$$

the Bayes rule, is the orthogonal projection of the parametric function $g^*(\gamma)$ into the collection of estimators with respect to the inner product $\langle \, , \, \rangle$. That is, the Bayes rule satisfies the Pythagorean identity

$$\|\hat{g} - g^*\|^2 = \|\hat{g} - \tilde{g}\|^2 + \|\tilde{g} - g^*\|^2 \quad .$$

Applying this identity to functions of the form $a'Z(y)$ with a in $\mathcal{C}$ the equation

$$\text{AMSE}\,[a'Z(y)] = \|a'z - \tilde{g}\|^2 + \|\tilde{g} - g^*\|^2$$

is obtained. Since the second term does not vary with a it follows that an attempt to minimize AMSE $[a'Z(y)]$ by varying a represents an attempt to minimize $\|a'z - \tilde{g}\|$.

Note that neither $a'Z(y)$ nor $\tilde{g}(y)$ depend on γ. The dimension of the latter minimization problem may, therefore, be reduced and attention may be restricted to measurable functions on $\mathcal{Y}$ which are square integrable with respect to the measure υ defined on the Lebesque subsets of $\mathcal{Y}$ by

$$\upsilon(A) = \int_{\Gamma}\int_{\mathcal{Y}} I_A(y)\, d\mu(y, \gamma)$$

That is, to minimize $\|a'z - g^*\|$ with respect to the norm on $\mathcal{L}_2(\mathcal{Y} \times \Gamma, \mu)$ it suffices to find a in $\mathcal{C}$ which minimizes $\|a'z - \tilde{g}\|$ with respect to the norm on $\mathcal{L}_2(\mathcal{Y}, \upsilon)$. We have, by Theorem 1, that the choice

$$\hat{a} = B[B'\int_{\Gamma} \mathcal{E}_{\gamma}(zz')\, d\rho(\gamma)\, B]^{-}B' \int_{\Gamma} g^*(\gamma)\mathcal{E}_{\gamma}(z)\, d\rho(\gamma)$$

minimizes AMSE $[a'Z(y)]$ whence this choice is the solution of both of these minimization problems. For computations, we have by Fubini's theorem that for integrable $g(y)$

$$\int_{\mathcal{Y}} g(y)\, d\omega(y) = \int_{\Gamma}\int_{\mathcal{Y}} g(y)\; n(y|f(\theta),\, \sigma\,)\; dy\; d\rho(\gamma)$$

$$= \int_{\mathcal{Y}} g(y) \int_{\Gamma} n(y|f(\theta),\, \sigma\,)\, d\rho(\gamma)\; dy$$

$$= \int_{\mathcal{Y}} g(y)\; p(y)\; dy \quad .$$

The foregoing discussion is summarized as Theorem 2 .

Theorem 2. Let $\tilde{g}(y)$ be the Bayes rule which minimizes expected posterior square error loss when estimating $g^{*}(\gamma)$ with prior measure ρ on Γ . Let $p(y)$ be the marginal density of the observations,

$$p(y) = \int_{\Gamma} n[y|f(\theta),\, \sigma\,]\; d\rho(\gamma) \quad .$$

The estimator $\hat{g}(y)$ of Theorem 1 is the best approximation of the Bayes' rule in the class of estimators of the form $a'Z(y)$ with a in G in the sense of minimizing the distance

$$\|a'z - \tilde{g}\| = [\int_{\mathcal{Y}} [a'Z(y) - \tilde{g}(y)]^2\; p(y)\; dy\;]^{\frac{1}{2}} \quad .$$

Consider the sequence of functions $\{\varphi_i(y)\}_{i=1}^{m-d}$ generated by the Gram-Schmidt orthonormalization process from $\{Z_i(y)\}_{i=1}^{m}$; the orthonormalization is with respect to $\mathcal{L}_2(y,\, \nu)$; m - d refers to the fact that there may have been linear dependencies a.e. among the $\{Z_i(y)\}_{i=1}^{m}$ which were eliminated by deletion prior to orthonormalization. Note that it is a property of the Gram-Schmidt process that to within a null set the collection of functions of the form $\Sigma_{i=1}^{m}\; a_i\; Z_i(y)$ is the same as the collection of functions of the form $\Sigma_{i=1}^{m-d}\; c_i\; \varphi_i(y)$ provided that $G = R^m$. By Theorem 16.16 of Hewitt and Stromberg (1965) the choice $\hat{c}_i = \int_{\mathcal{Y}} \tilde{g}(y)\; \varphi_i(y)\; p(y)\; dy$ minimizes $\|\, \tilde{g} - \Sigma_{i=1}^{m-d}\; c_i\; \varphi_i \|$ regarded as a function of $(c_1,\, c_2,\, \ldots,\, c_{m-d})$ and that this choice is unique. But $\Sigma_{i=1}^{m}\; \hat{a}_i\; Z_i(y)$

is of the form $\sum_{i=1}^{m-d} c_i \varphi_i(y)$ to within a null set and minimizes $\| \tilde{g} - \sum_{i=1}^{m-d} c_i \varphi_i \|$ by Theorem 2 whence it follows that $\sum_{i=1}^{m-d} \hat{c}_i \varphi_i(y) = \sum_{i=1}^{m} \hat{a}_i Z_i(y)$ a.e. Note that the choice of $(c_1, c_2, \ldots, c_{m-d})$ is unique but the choice of $\hat{a}$ may not be.

We may summarize this discussion as Corollary 1.

Corollary 1. Let $\tilde{g}(y)$ be the Bayes rule which minimizes expected posterior square error loss when estimating $g^*(\gamma)$ with prior measure ρ on Γ. Let $p(y)$ be the marginal density of the observations and let $\{Z_i(y)\}_{i=1}^{m}$ be a subset of $\mathcal{L}_2(y, \nu)$ where ν is defined by $\nu(A) = \int_{\mathcal{Y}} I_A(y)\, p(y)\, dy$. Let $\{\varphi_i(y)\}_{i=1}^{m-d}$ be an orthonormal sequence generated from $\{Z_i(y)\}_{i=1}^{m}$ by the Gram-Schmidt process. Then the estimator $\hat{g}(y)$ of Theorem 1 with $G = R^m$ satisfies

$$\hat{g}(y) = \sum_{i=1}^{m-d} c_i \varphi_i(y) \quad \text{a.e. } \nu$$

where $c_i = \int_{\mathcal{Y}} \tilde{g}(y)\, \varphi_i(y)\, p(y)\, dy$.

It is of interest to have a sufficient condition such that the distance $\| \hat{g} - \tilde{g} \|$ between the Bayes rule $\tilde{g}$ and the explicit estimator $\hat{g}$ may be made arbitrarily small by taking $\hat{g}$ to be a polynomial of sufficiently high degree. Theorem 3 states such a condition.

Theorem 3. Let ν be a probability measure defined on $\mathcal{Y}$, the n-dimensional real numbers, and let $\{Z_i(y)\}_{i=1}^{\infty}$ be a basis for the polynomials on $\mathcal{Y}$. If the moment generating function, $\int_{\mathcal{Y}} \exp(u'y)\, d\nu(y)$, exists on an open neighborhood of the zero vector then $\{Z_i(y)\}_{i=1}^{\infty} \subset \mathcal{L}_2(\mathcal{Y}, \nu)$ and $\{Z_i(y)\}_{i=1}^{\infty}$ is complete; that is, $f \in \mathcal{L}_2(\mathcal{Y}, \nu)$ and $< f, Z_i > = 0$ for $i = 1, 2, \ldots$ implies $f = 0$ a.e. ν. In consequence, if $\{\varphi_i(y)\}_{i=1}^{\infty}$ is a sequence of orthonormal functions generated from $\{Z_i(y)\}_{i=1}^{\infty}$ by the Gram-Schmidt process, if $f \in \mathcal{L}_2(\mathcal{Y}, \nu)$, and if $f_m(y) = \sum_{i=1}^{m} c_i \varphi_i(y)$ where $c_i = < f, \varphi_i >$ then

$$\lim_{m \to \infty} \| f_m - f \| = 0 .$$

Proof: Existence of the moment generating function implies all moments of a distribution exist whence $\{Z_i(y)\}_{i=1}^{\infty} \subset \mathcal{L}_2(\mathcal{Y},\nu)$. Let $\delta_j > 0$ be such that $\int e^{u'y}\, d\nu(y) < \infty$ for $-\delta_j < u_j < \delta_j$, $j = 1, 2, \ldots, n$. Let $f \in \mathcal{L}_2(\mathcal{Y},\nu)$ be given such that $< f, Z_i > = 0$ for $i = 1, 2, \ldots$. Consider the function

$$\hat{f}(z) = \int f(y) \exp(z'y)\, d\nu(y)$$

where z is of the form $z = u + i\,v$ and $u, v \in \mathcal{Y}$. Now

$$|\exp(z'y)| = |\exp(u'y)[\cos(v'y) + i \sin(v'y)]|$$

$$\leq |\exp(u'y)| .$$

which shows that $\hat{f}(z)$ exists if $-\delta_j < u_j < \delta_j$, $(j = 1, 2, \ldots, n)$. Decompose f as $f = f^+ - f^-$ where f^+ and f^- are nonnegative. Suppose that it can be shown that $\hat{f}(iv) \equiv 0$ then it follows that

$$\int f^+(y) \exp(iv'y)\, d\nu(y) = \int f^-(y) \exp(iv'y)\, d\nu(y) .$$

The same constant multiple will normalize both f^+ and f^- so that the measure defined by $P^+(A) = c\int_A f^+(y)\, d\nu(y)$ and $P^-(A) = c\int_A f^-(y)\, d\nu(y)$ are probability measures. But the equation above implies that these two probability measures have the same characteristic function. Consequently, $f^+ = f^-$ a.e. ν whence $f = f^+ - f^- = 0$ a.e. ν . Thus, if it can be shown that $\hat{f}(iv) \equiv 0$ then $\{Z_i(y)\}_{i=1}^{\infty}$ is complete.

Now

$$c\,\hat{f}(z) = \int \exp(z'y)\, dP^+(y) - \int \exp(z'y)\, dP^-(y)$$

and by applying Theorem 9, Chapter 2, of (Lehmann, 1959, p.52) to each function on the right one can conclude that $\hat{f}(z)$ is an analytic function of the single variable $z_j = u_j + i\,v_j$ over the region $-\delta_j < u_j < \delta_j$, $-\infty < v_j < \infty$ provided the remaining variables are held fixed at points with $-\delta_j < u_j < \delta_j$

for $j' \neq j$. By Hartogs' Theorem (Bochner and Martin, 1948, p. 140) it follows that $\hat{f}(z)$ is analytic over the region $-\delta_j < u_j < \delta_j$, $-\infty < v_j < \infty$ $(j = 1, 2, \ldots, n)$. The same inductive argument used by Lehmann in the proof of Theorem 9 to show that all partial derivatives of the form $(\partial^m/\partial z_j^m)\, \hat{f}(z)$ may be computed under the integral sign may obviously be used to conclude that any partial $(\partial^m/\pi_{j=1}^m \partial z_j^{m_j})\, \hat{f}(z)$ may be computed under the integral sign. Now

$$(\partial^m/\pi_{j=1}^m \partial z_j^{m_j})\, \hat{f}(0) = \int (\partial^m/\pi_{j=1}^m \partial z_j^{m_j})\, f(y) \exp(z'y)\, d\nu(y)\Big|_{z=0}$$

$$= \int \pi_{j=1}^m y_j^{m_j}\, f(y)\, d\nu(y)$$

$$= 0$$

since $\pi_{j=1}^n y_j^{m_j}$ is some finite linear combination of the $Z_i(y)$ $(i = 1, 2, \ldots)$. Since $\hat{f}(z)$ is analytic then $\hat{f}(z) \equiv 0$ over the region $-\delta_j < u_j < \delta_j$, $-\infty < v_j < \infty$ $(j = 1, 2, \ldots, n)$ and in particular $\hat{f}(iv) \equiv 0$ for $-\infty < v_j < \infty$ $(j = 1, 2, \ldots, n)$ as was to be shown.

The claim that $\lim_{m\to\infty} \| f_m - f \| = 0$ follows directly from Theorem 16.26 of Hewitt and Stromberg (1965, p. 245). ∎

The following result follows immediately from Corollary 1 and Theorem 3.

Corollary 2. Let $\tilde{g}(y)$ be the Bayes' rule which minimizes expected posterior square error loss when estimating $g^*(\gamma)$ with prior measure ρ on Γ . Let the moment generating function of $p(y)$, the marginal density of the observations, exist in an open neighborhood of the zero vector. Let $\{Z_i(y)\}_{i=1}^{\infty}$ be a basis for the polynomials. For each m let

$$\hat{g}_m(y) = \Sigma_{i=1}^m \hat{a}_i\, Z_i(y)$$

be the estimator given by Theorem 1 with $\Omega = R^m$.

Then

$$\lim_{m\to\infty} \int_{\mathcal{Y}} [\hat{g}_m(y) - \tilde{g}(y)]^2 \, p(y) \, dy = 0 \, .$$

If $\hat{g}_m$ converges in mean square to $\tilde{g}$ with respect to the predictive density $p(y)$ then $\hat{g}_m$ converges in both probability and distribution to $\tilde{g}$ with respect to the predictive density $p(y)$. From the sampling theory point of view it is of more interest to know whether or not $\hat{g}_m$ converges in probability and distribution to $\tilde{g}$ with respect to the conditional distribution $N(y|f(\theta),\sigma)$. A sufficient condition is given as Theorem 4.

Theorem 4. Let ν denote the probability measure on the measurable space $(\mathcal{Y}, \mathcal{B})$ determined according to $\nu(B) = \int_\Gamma \int_{\mathcal{Y}} I_B(y) \, n(y|f(\theta),\sigma) \, dy \, d\rho(\gamma)$ and let $\nu(\cdot|\gamma)$ denote the probability measure determined by $\nu(B|\gamma) = \int_{\mathcal{Y}} I_B(y) n[y|f(\theta),\sigma] \, dy$. Let $\hat{g}_m(y)$ converge in probability to $\tilde{g}(y)$ on $(\mathcal{Y}, \mathcal{B}, \nu)$. If $\{\nu(B|\gamma)\}_{B\in\mathcal{B}}$ is an equicontinuous family at γ_0 and if for every open set Γ_0 containing γ_0 $\rho(\Gamma_0) > 0$ then $\hat{g}_m(y)$ converges in probability to $\tilde{g}(y)$ on $[\mathcal{Y}, \mathcal{B}, \nu(\cdot|\gamma_0)]$.

Proof: Given $\delta > 0$ choose an open set Γ_0 containing γ_0 such that for all $B \in \mathcal{B}$ $|\nu(B|\gamma_0) - \nu(B|\gamma)| < \delta/2$. Given $\epsilon > 0$ choose M so that $m > M$ implies $\nu(B_m) < \rho(\Gamma_0)\delta/2$ for $B_m = \{y: |\hat{g}_m(y) - \tilde{g}(y)| < \epsilon\}$. Then

$$\begin{aligned} P(B_m|\gamma_0) &< [1/\rho(\Gamma_0)] \int_{\Gamma_0} P(B_m|\gamma) \, d\rho(\gamma) + \delta/2 \\ &\le [1/\rho(\Gamma_0)] \int_{\Gamma} P(B_m|\gamma) \, d\rho(\gamma) + \delta/2 \\ &< \nu(B_m)/\rho(\Gamma_0) + \delta/2 \\ &< \delta \, . \, \square \end{aligned}$$

A proof that normal distribution generates an equicontinuous family $\{\nu(B|\gamma)\}_{B\in\mathcal{B}}$ when $f(\theta)$ is continuous is given below. It is straightforward and could be applied to many common distributions.

Theorem 5. Let $f(\theta)$ be a continuous function on a closed set Θ, let $n[y|f(\theta), \sigma]$ be the n-variate normal density function with mean vector $f(\theta)$ and variance-covariance matrix $\sigma^2 I$, and let $\mathcal{B}$ denote the Lebesgue subsets of R^n. Then the family $\{\nu(B|\gamma)\}_{B\in\mathcal{B}}$ where $\nu(B|\gamma) = \int_B n(y|f(\theta), \sigma)\, dy$ is equicontinuous for every $\gamma = (\theta',\sigma)' \in \Theta \times (0, \infty)$.

Proof: Fix $\gamma_0 \in \Theta \times (0, \infty)$. Let S be a bounded open sphere containing θ_0, let $\bar{S}$ denote its closure and set $\bar{\Theta} = \Theta \cap \bar{S}$ whence $\bar{\Theta}$ is compact; set $\bar{\Sigma} = \{\sigma: \sigma_0^2/2 \le \sigma^2 \le 3\sigma_0^2/2\}$. Given $\epsilon > 0$ there is a radius r such that each closed sphere Y_θ with radius r and center $f(\theta)$ satisfies $\nu(Y_\theta|\gamma) > 1 - \epsilon/4$ for all $\sigma \in \bar{\Sigma}$. Then $\bar{Y} = \cup_{\bar{\Theta}} Y_\theta$ is compact since the continuous image $f[\bar{\Theta}]$ of a compact set $\bar{\Theta}$ is compact. Now the normal density function $n[y|f(\theta), \sigma]$ is continuous in (y,θ,σ) on the compact set $\bar{Y} \times \bar{\Theta} \times \bar{\Sigma}$ hence uniformly continuous. Then choose δ such that $|(y',\theta',\sigma) - (y', \theta_0', \sigma_0)| = |(\theta', \sigma) - (\theta_0',\sigma_0)| < \delta$ implies $|n[y|f(\theta),\sigma] - n[y|f(\theta_0),\sigma_0]| < \epsilon/(2 \int_{\bar{Y}} dy)$. Now

$$
\begin{aligned}
&|\nu(B|\gamma) - \nu(B|\gamma_0)| \\
&\quad = |\nu(B \cap \bar{Y}|\gamma) + \nu(B \cap \sim \bar{Y}|\gamma) - \nu(B \cap \bar{Y}|\gamma_0) - \nu(B \cap \sim \bar{Y}|\gamma)| \\
&\quad \le |\nu(B \cap \bar{Y}|\gamma) - \nu(B \cap \bar{Y}|\gamma_0)| + \epsilon/2 \\
&\quad = \left| \int_{B\cap\bar{Y}} n(y|\gamma) - n(y|\gamma_0)\, dy \right| + \epsilon/2 \\
&\quad \le \int_{B\cap\bar{Y}} \epsilon/(2\textstyle\int_{\bar{Y}} dy)\, dy + \epsilon/2 \\
&\quad \le \epsilon .
\end{aligned}
$$

Thus, $\{\nu(B|\gamma)\}_{B\in\mathcal{B}}$ is equicontinuous at γ_0 . □

Footnotes

1/ If there are deletions then the expansion is truncated at m less the number of deletions.

2/ Read $B_2 = [(551 + 41\sqrt{29})/6264]V$ to correct an error in their formula.

3/ See the next section for the values of a_i.

4/ Note that U is a function of the sufficient statistic

$$[y_1 + y_7 , y_2 + y_8 , y_3 + y_9 , y_4 + y_{10} , y_5 + y_{11} , y_6 + y_{12} , \Sigma_{i=1}^{6}(y_i - y_{i+6})^2]$$

5/ The predictive density for the example does not possess a moment generating function. Nontheless, the computations yield results one would have anticipated had it existed.

References

Bochner, Solomon and Martin, William Ted (1948) Several Complex Variables, Princeton, New Jersey: Princeton University Press.

Box, G. E. P. and Lucas, H. L. (1959) "Design of Experiments in Non-Linear Situations," Biometrika, 46, 77-90.

Gallant, A. Ronald (1976) "Confidence Regions for the Parameters of a Nonlinear Regression Model," Institute of Statistics Mimeograph Series, No. 1077, Raleigh: North Carolina State University.

Guttman, Irwin and Meeter, Duane A. (1964) "On Beale's Measures of Non-Linearity," Technometrics, 7, 623-637.

Hammersly, J. M. and Handscomb, D. C. (1964) Monte-Carlo Methods, New York: John Wiley and Sons, Inc.

Hartley, H. O. (1964) "Exact Confidence Regions for the Parameters in Non-Linear Regression Laws," Biometrika, 51, 347-353.

Hewitt, Edwin and Stromberg, Karl (1965) Real and Abstract Analysis, New York: Springer-Verlag Inc.

I. B. M. Corporation (1968) System 360 Scientific Subroutine Package, Version III, White Plains, New York: Technical Publications Department.

I.M.S.L. (1975) IMSL Library 1, Fifth Ed., Houston, Texas: International Mathematical and Statistical Libraries, Inc.

Lehmann, E. L. (1959) Testing Statistical Hypotheses, New York: John Wiley and Sons.

Stroud, A. H. (1971) Approximate Calculation of Multiple Integrals, Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Zellner, Arnold (1971) An Introduction to Bayesian Inference In Econometrics, New York: John Wiley and Sons, Inc.