Biometrika (1973), **60**, 3, *p.* 565
With 1 *text-figure*
Printed in Great Britain

Bayes estimation of a convex quadratic

BY A. O'HAGAN
University College London

SUMMARY

The estimation of a quadratic regression curve when the quadratic coefficient is known to be positive is treated by a Bayesian method. The method extends easily to deal with a general linear model under parameter constraints.

Some key words: Linear regression; Bayes; Constrained parameters; Truncation; Multivariate t distribution; Multivariate normal distribution.

1. INTRODUCTION

One of the simplest and most naturally occurring examples of a constrained parameter in linear regression is that of a quadratic which is known to be convex. That is, the model is

$$E(y_i|x_i, \alpha, \beta, \gamma) = \alpha + \beta x_i + \gamma x_i^2 \quad (i = 1, \ldots, n), \tag{1}$$

with $\gamma \geqslant 0$. In § 2 a Bayesian method is proposed to estimate α, β and γ. The accepted classical, i.e. non-Bayesian, answer is that provided by the method of maximum likelihood. In the case of normal distributions this method leads to estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ which are the usual least squares estimates provided they satisfy the constraint $\hat{\gamma} \geqslant 0$; otherwise it sets $\hat{\gamma} = 0$ with $\hat{\alpha}$ and $\hat{\beta}$ equal to the least squares estimates for a straight line regression. The corresponding estimates when the constraint is that γ is strictly positive will not always exist.

The maximum likelihood method generalizes easily to any multiple linear regression model with constrained parameters, and when the constraints take the form of linear inequalities the solution may be obtained by quadratic programming, as described by Judge & Takayama (1966) and by Hudson (1969). At the end of § 2, we indicate the obvious generalization of the Bayesian method. Section 3 discusses the choice of prior distribution, and in § 4 an example is given.

2. ESTIMATION

We assume that the vector of observations $\mathbf{y}$ given the vector $\mathbf{x}$ has a multivariate normal distribution with mean given by (1) and covariance matrix $\sigma^2\mathbf{W}$, where $\mathbf{W}$ is known. Although $\mathbf{x}$ may itself be a random variable, we assume that it is distributed independently of α, β, γ and σ^2, so that the likelihood function is

$$L(\boldsymbol{\theta}, \sigma^2|\mathbf{x}, \mathbf{y}) \propto \sigma^{-n} \exp\{-\tfrac{1}{2}\sigma^{-2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'\mathbf{W}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\}, \tag{2}$$

where $\boldsymbol{\theta}' = (\alpha, \beta, \gamma)$ and $\mathbf{X}$ is the $n \times 3$ design matrix whose ith row is $(1, x_i, x_i^2)$.

In order to use the Bayesian method we now require a prior distribution for $\boldsymbol{\theta}$ and σ^2, of such a form as to represent our beliefs about these parameters before $\mathbf{x}$ and $\mathbf{y}$ are observed. Part of our prior information is that $\gamma \geqslant 0$, so that the prior distribution will be confined to the space

$$T = (\boldsymbol{\theta}, \sigma^2\colon -\infty < \alpha < \infty,\ -\infty < \beta < \infty,\ 0 \leqslant \gamma < \infty,\ 0 \leqslant \sigma < \infty),$$

and be zero elsewhere. A convenient prior distribution has the density

$$\pi_1(\boldsymbol{\theta}, \sigma^2) \propto \sigma^{-3} \exp\{-\tfrac{1}{2}\sigma^{-2}(\boldsymbol{\theta}-\mathbf{m}_1)' \mathbf{V}_1^{-1}(\boldsymbol{\theta}-\mathbf{m}_1)\} \sigma^{-\nu_1-2} \exp(-\tfrac{1}{2}\sigma^{-2}d_1), \tag{3}$$

for $(\boldsymbol{\theta}, \sigma^2) \in T$, where $\mathbf{V}_1$ is positive definite and symmetric, and ν_1 and d_1 are positive. This is a truncated 'normal inverted gamma' distribution. In its untruncated form, this distribution has been extensively used by Bayesians. Although it has some disadvantages, it is usually sufficiently flexible to describe prior beliefs quite accurately.

The posterior density is proportional to the product of (2) and (3), and after some simplification may be written

$$\pi_2(\boldsymbol{\theta}, \sigma^2) \propto \sigma^{-3} \exp\{-\tfrac{1}{2}\sigma^{-2}(\boldsymbol{\theta}-\mathbf{m}_2)' \mathbf{V}_2^{-1}(\boldsymbol{\theta}-\mathbf{m}_2)\} \sigma^{-\nu_2-2} \exp(-\tfrac{1}{2}\sigma^{-2}d_2), \tag{4}$$

for $(\boldsymbol{\theta}, \sigma^2) \in T$, where

$$\begin{aligned} &\mathbf{V}_2 = (\mathbf{V}_1^{-1}+\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}, \quad \mathbf{m}_2 = \mathbf{V}_2(\mathbf{V}_1^{-1}\mathbf{m}_1+\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}), \\ &\nu_2 = \nu_1+n, \quad d_2 = d_1+\mathbf{m}_1'\mathbf{V}_1^{-1}\mathbf{m}_1+\mathbf{y}'\mathbf{W}^{-1}\mathbf{y}-\mathbf{m}_2'\mathbf{V}_2^{-1}\mathbf{m}_2. \end{aligned} \tag{5}$$

The posterior distribution is therefore also truncated 'normal inverted gamma'. Some of its properties are listed below, labelled 1–4. With the exception of one result which is proved in the Appendix, all the properties are very easily obtained. The subscripts on $\mathbf{m}$, $\mathbf{V}$, ν and d have been dropped so that the results may apply to either the prior distribution (3) or the posterior distribution (4). We first partition the parameters by

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}^* \\ \gamma \end{bmatrix}, \quad \mathbf{m} = \begin{bmatrix} \mathbf{m}^* \\ m_* \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}^* & \mathbf{u} \\ \mathbf{u}' & v_* \end{bmatrix},$$

where γ, m_* and v_* are scalars, thus determining the dimensions of $\boldsymbol{\theta}^*$, $\mathbf{m}^*$, $\mathbf{V}^*$ and $\mathbf{u}$.

Property 1. The conditional distribution of $\boldsymbol{\theta}^*$, given γ and σ^2, is multivariate normal with mean

$$\mathbf{m}^* + \mathbf{u}(\gamma - m_*)/v_* \tag{6}$$

and covariance matrix $\sigma^2(\mathbf{V}^* - \mathbf{u}\mathbf{u}'/v_*)$.

Property 2. The unconditional distribution of γ is truncated t; that is,

$$z = \gamma - m_*\{\nu/(v_* d)^{\frac{1}{2}}\} \tag{7}$$

has a density proportional to that of a Student's t variable with ν degrees of freedom, over the range $[q, \infty)$, where

$$q = -m_*\{\nu/(v_* d)^{\frac{1}{2}}\}. \tag{8}$$

Property 3. The moments of z are found in the Appendix to be

$$E(z^r) = \nu^{\frac{1}{2}r}\{G(r)/G(0)\}, \tag{9}$$

where

$$G(r) = \begin{cases} B_p\{\tfrac{1}{2}(\nu-r), \tfrac{1}{2}(r+1)\} & (q \geqslant 0 \text{ or } r \text{ is odd}), \\ 2B\{\tfrac{1}{2}(\nu-r), \tfrac{1}{2}(r+1)\} - B_p\{\tfrac{1}{2}(\nu-r), \tfrac{1}{2}(r+1)\} & (q < 0 \text{ and } r \text{ is even}), \end{cases} \tag{10}$$

$$p = \nu/(\nu+q^2). \tag{11}$$

Here $B(.,.)$ and $B_p(.,.)$ are the Beta function and Incomplete Beta function, respectively. The mode of γ is m_* if this is positive, otherwise it is zero.

Property 4. The conditional distribution of σ^{-2} given γ is g^{-1} times chi-squared with $\nu+1$ degrees of freedom, where $g = d+(\gamma-m_*)^2/v_* = d(1+\nu^{-1}z^2)$, and hence

$$E(\sigma^{2r}|\gamma) = (\tfrac{1}{2}g)^r\,\Gamma\{\tfrac{1}{2}(\nu+1)-r\}/\Gamma\{\tfrac{1}{2}(\nu+1)\}.$$

We may now estimate $\boldsymbol{\theta}$ by using the appropriate loss function. For example, if we assume quadratic loss the estimates are the posterior means. The calculation is quite straightforward, and is illustrated in § 4.

This method extends with minor modifications to a general linear model; we merely redefine $\boldsymbol{\theta}$ and $\mathbf{X}$ to be the appropriate parameter vector and design matrix. Equations (5) still hold although the dimensions of $\mathbf{m}$ and $\mathbf{V}$ may have been redefined. The set T may be any subset of the complete $(\boldsymbol{\theta}, \sigma^2)$ space. In our quadratic model the truncation of the parameter space affects only one dimension and this greatly simplifies the analysis; e.g. the moments of the truncated multivariate t distribution of $\boldsymbol{\theta}$ may be obtained quite easily from those of the truncated univariate distribution of γ which are derived in the Appendix. In more general cases, when the truncation is in several dimensions, estimates, such as the posterior means, will be much harder to find. If σ^2 were known, however, the distribution of $\boldsymbol{\theta}$ would be the truncated multivariate normal, which has been considered by Birnbaum & Meyer (1953) and by Tallis (1961, 1965). This distribution will also serve as an approximation when ν_2 is large.

3. Choice of prior distribution

When applying a Bayesian analysis the most difficult part is often choosing a distribution to express one's prior knowledge about the unknown parameters. In our case this means specifying eleven quantities: three elements of $\mathbf{m}_1$, six of $\mathbf{V}_1$, ν_1 and d_1. It may be feasible to state the prior means for $\boldsymbol{\theta}$ and σ^2, but these cannot be identified with $\mathbf{m}_1$ and $\nu_1^{-1}d_1$ because of the truncation. Nor is $\sigma^2\mathbf{V}_1$ the prior covariance matrix of $\boldsymbol{\theta}$ given σ^2. Given, for instance, four means, four variances, and the three covariances between elements of $\boldsymbol{\theta}$, we have the analytical problem of converting these into values for $\mathbf{m}_1$, $\mathbf{V}_1$, ν_1 and d_1. The problem is simplified if we can assume that α and β are *a priori* independent of γ, i.e. $\mathbf{u} = \mathbf{0}$, for then $\mathbf{m}_1^*$ and $\sigma^2\mathbf{V}_1^*$ are the means and covariance matrix of $\boldsymbol{\theta}^*$ given σ^2. In practice this independence is perhaps rather unlikely, and in general we may have to fall back on a trial-and-error approach to specifying the prior distribution.

It is often of interest in Bayesian theory to try to express prior ignorance through the prior distribution. This is because when prior ignorance is assumed the Bayesian methods usually come closest to the corresponding classical theory. In the analysis of linear models with no parameter constraints it is common to express prior ignorance by a 'normal inverted gamma' distribution with $\mathbf{V}_1^{-1} = \mathbf{0}$, $\nu_1 = 0$ and $d_1 = 0$. Reference to equations (5) will show that then $\mathbf{m}_2$ is the usual unconstrained least squares estimator of $\boldsymbol{\theta}$, and in fact in the truncated distribution the posterior mode is the maximum likelihood estimator. The posterior mean is always strictly inside T, whereas the mode need not be.

However, $\mathbf{V}_1^{-1} = \mathbf{0}$ cannot really represent prior ignorance in our case, according to most theories of what the phrase should mean. This is because it gives a prior distribution for γ which is uniform over the positive real numbers, resulting in far too much weight being given to very large values of γ. The usual recipe for prior ignorance for a parameter such as γ confined to the positive half-line is a density proportional to γ^{-1}, but in fact this goes to the other extreme, for a little algebra shows that the resulting posterior distributions for γ are always improper, cannot be normalized, and have all their mass concentrated at $\gamma = 0$. Therefore, there seems to be no practical alternative to a genuine informative prior distribution, at least for γ.

4. An example

This example concerns some data collected to test the efficiency of a pump. The dependent variable is the efficiency measured in per cent, and the independent variable is the water flow rate, expressed as a percentage of rated flow. In the region of 100 % of rated flow the efficiency is known to respond approximately quadratically and must be concave. The

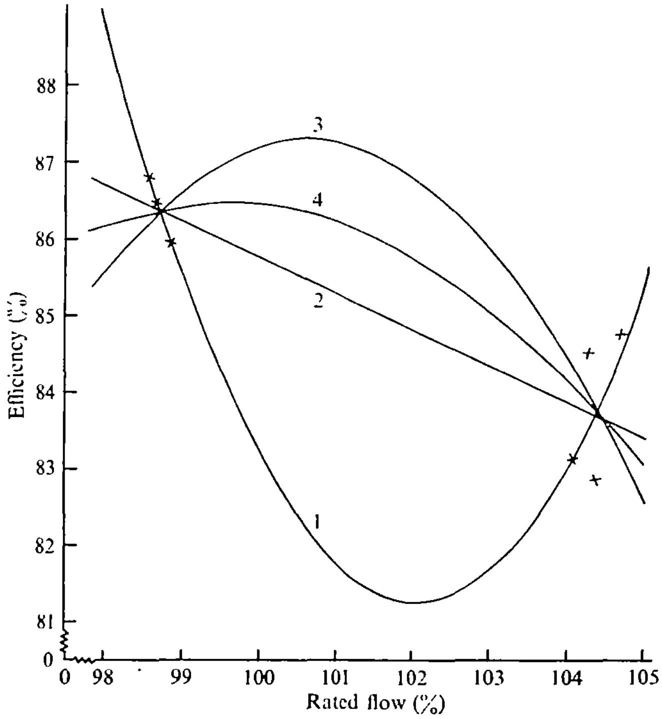

Fig. 1. Pump efficiency curves. (1) Least squares quadratic. (2) Least squares straight line. (3) Bayes estimate under squared-error loss and uniform prior distribution. (4) Bayes estimate under squared-error loss and a proper prior distribution for γ and σ^2. Data points: $\times$.

model, then, is our quadratic model, except that now γ is constrained to be negative. Alternatively, redefining $\mathbf{X}$ so that the third column contains elements $-x_i^2$ we can use the theory for $\gamma \geqslant 0$. It is assumed that the observations have homoscedastic and independent errors; thus $\mathbf{W}$ is the identity matrix.

Only seven observations could be made, and these are marked as crosses in Fig. 1. As they are roughly clustered around two separate flow rates it is clear that they provide relatively little information about γ. In fact the unrestricted least squares solution, line 1 in Fig. 1, has $\hat{\gamma} < 0$. The restricted solution is the straight line 2, but this was also unacceptable to the experimenters as one of their chief interests was in the efficiency at 100 % of rated flow. Clearly lines 1 and 2 must both underestimate this efficiency, and so the Bayes approach was tried. We shall illustrate the calculations with an 'ignorance' prior distribution having parameters $\mathbf{V}_1^{-1} = \mathbf{0}$, $\nu_1 = 0$ and $d_1 = 0$. As explained above this is expected to considerably overestimate γ, but it has the merit of simplifying the calculations for this example.

As a preliminary to the numerical work the data were subjected to a simple linear transformation; 85 % was subtracted from efficiency, and 100 % from flow rate, and both were then multiplied by 10. The posterior parameters on this new scale are $d_2 = 181{\cdot}2$, $\nu_2 = 4$ and

$$\mathbf{V}_2 = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 4{\cdot}26 & 0{\cdot}225 & 0{\cdot}00732 \\ 0{\cdot}225 & 0{\cdot}0134 & 0{\cdot}000414 \\ 0{\cdot}00732 & 0{\cdot}000414 & 0{\cdot}0000132 \end{bmatrix},$$

$$\mathbf{m}_2 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} -17{\cdot}9 \\ -1{\cdot}92 \\ -0{\cdot}0468 \end{bmatrix}.$$

The value $\nu = 4$ is explained by the fact that the term σ^{-3} is absent from the prior density (3), because $\mathbf{V}_1^{-1} = \mathbf{0}$, whereas it must be present in the posterior density (4). Thus three degrees of freedom have, in effect, been lost. These new parameters describe the posterior joint distribution of $\boldsymbol{\theta}$. In order to estimate the efficiency at 100 % of rated flow we must use this distribution to estimate α, since $x = 0$ is 100 % flow on our new scale. We assume that a squared error loss function is appropriate, and hence the estimate is the posterior mean of α. We can give the best estimate of efficiency at any given flow rate by drawing the curve $y = \hat{\alpha} + \hat{\beta}x - \hat{\gamma}x^2$, with $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ being posterior means.

We therefore calculate the basic parameters of the distribution of z. From (8) and (11),

$$q^2 = \frac{0{\cdot}0468^2 \times 4}{181{\cdot}2 \times 0{\cdot}0000132} = 3{\cdot}66,$$

$$p = (1 + 3{\cdot}66/4)^{-1} = 0{\cdot}522.$$

We may now use equations (9) and (10) to find the mean of z. For this we must evaluate the Incomplete Beta function, and Pearson's (1934) tables were used. Since $q > 0$,

$$G(0) = B_p(2, 0{\cdot}5) = 0{\cdot}128 \times 4/3, \quad G(1) = B_p(1{\cdot}5, 1) = 0{\cdot}377 \times 2/3,$$

and therefore

$$E(z) = 2 \times G(1)/G(0) = 2{\cdot}945.$$

From (7), the posterior mean of γ is

$$\begin{aligned} -0{\cdot}0468 + (181{\cdot}2 \times 0{\cdot}0000132/4)^{\frac{1}{2}} \times 2{\cdot}945 &= -0{\cdot}0468 + 0{\cdot}0245 \times 2{\cdot}945 \\ &= 0{\cdot}025, \end{aligned}$$

and from (6), the posterior mean of $\boldsymbol{\theta}^*$ is

$$\begin{bmatrix} -17{\cdot}9 \\ -1{\cdot}92 \end{bmatrix} + \frac{181{\cdot}2 \times 2{\cdot}945}{4 \times 0{\cdot}0245} \times \begin{bmatrix} 0{\cdot}00732 \\ 0{\cdot}000414 \end{bmatrix} = \begin{bmatrix} 22{\cdot}0 \\ 0{\cdot}33 \end{bmatrix}.$$

Use of these values as estimates of γ and $\boldsymbol{\theta}^*$ gives line 3 in Fig. 1. As expected this is rather too peaked, and would probably have been just as unacceptable to the experimenters as lines 1 and 2. It is interesting, however, that even with such unreliable data, as far as γ is concerned, a prior mean for γ which is effectively infinite is converted into an almost reasonable posterior mean. In fact a proper prior distribution was agreed for γ with parameters $d_1 = 50$, $\nu_1 = 4$,

$$\mathbf{m}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{V}_1^{-1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 10 \end{bmatrix}.$$

The prior density is fairly flat over the region of the parameter space for which $-\infty < \alpha < \infty$, $-\infty < \beta < \infty$, $0 \leqslant \gamma \leqslant 1$, $2 \leqslant \sigma \leqslant 5$. The posterior distribution then has mean $E(\alpha) = 14{\cdot}4$, $E(\beta) = -0{\cdot}091$ and $E(\gamma) = 0{\cdot}012$ and the resulting estimated regression line is line 4 in Fig. 1.

Note that the posterior mode of γ is at zero in both cases. We choose to use the posterior mean as a point estimate only because we are employing a squared error loss function.

Although lines 3 and 4 give best estimates of efficiency given their respective prior distributions, they give no indication of the accuracy of these estimates. Indeed, with data such as we have we should expect accuracy to be very low. The variance of $\hat{\alpha}+\hat{\beta}x-\hat{\gamma}x^2$ may be calculated for any x, and at $x = 0$, the 'rated' flow, we find the standard errors of estimated efficiency are $1{\cdot}9\,\%$ and $0{\cdot}77\,\%$ for the first and second prior distributions, respectively. The difference is due to the extra information in the second prior. The magnitude of these errors shows how inadequate the data are, even after the addition of some prior information.

I wish to thank the Central Electricity Generating Board for permission to use the data of §4, and the referees for helping me to clarify its exposition. This work was financially supported by the Science Research Council.

Appendix

The density of z is

$$\lambda^{-1}(1+\nu^{-1}z^2)^{-\frac{1}{2}(\nu+1)} \quad (z \geqslant q),$$

where

$$\lambda = \int_q^\infty (1+\nu^{-1}z^2)^{-\frac{1}{2}(\nu+1)}\,dz.$$

Define

$$G(r) = 2\nu^{-\frac{1}{2}(r+1)}\int_q^\infty z^r(1+\nu^{-1}z^2)^{-\frac{1}{2}(\nu+1)}\,dz,$$

so that $\lambda = \frac{1}{2}\nu^{\frac{1}{2}}G(0)$, and $E(z^r)$ is given by (9). Make the transformation

$$x = (1+\nu^{-1}z^2)^{-1}, \quad z = \nu^{\frac{1}{2}}x^{-\frac{1}{2}}(1-x)^{\frac{1}{2}},$$

which has Jacobian

$$-(dz/dx) = \tfrac{1}{2}\nu^{\frac{1}{2}}x^{-\frac{3}{2}}(1-x)^{-\frac{1}{2}}.$$

This transformation is one-to-one in $[q,\infty)$ if $q \geqslant 0$, and we first suppose this to be the case. Then

$$G(r) = \int_0^p x^{\frac{1}{2}(\nu-r)-1}(1-x)^{\frac{1}{2}(r+1)-1}\,dx = B_p\{\tfrac{1}{2}(\nu-r),\ \tfrac{1}{2}(r+1)\}, \tag{12}$$

with p as in (11). If $q < 0$, we split the range $[q,\infty)$ into $(0,\infty)$ and $[q,0]$ so that the transformation is one-to-one in each subrange, and obtain

$$G(r) = \int_0^1 f(x)\,dx + (-1)^r\int_p^1 f(x)\,dx,$$

$$= \begin{cases} B_p\{\frac{1}{2}(\nu-r),\ \frac{1}{2}(r+1)\} & (r \text{ odd}), \\ 2B\{\frac{1}{2}(\nu-r),\ \frac{1}{2}(r+1)\} - B_p\{\frac{1}{2}(\nu-r),\ \frac{1}{2}(r+1)\} & (r \text{ even}), \end{cases}$$

where $f(x)$ is the integrand of (12). Combination of this result with (12) gives equations (10). If $r \geqslant \nu$, $G(r)$ does not exist.

References

BIRNBAUM, Z. W. & MEYER, P. L. (1953). On the effect of truncation in some or all coordinates of a multi-normal population. *J. Ind. Soc. Agric. Statist.* **5**, 17–28.

HUDSON, D. J. (1969). Least-squares fitting of a polynomial constrained to be either non-negative, non-decreasing or convex. *J. R. Statist. Soc.* B **31**, 113–8.

JUDGE, G. G. & TAKAYAMA, T. (1966). Inequality restrictions in regression. *J. Am. Statist. Ass.* **61**, 166–81.

PEARSON, K. (Ed.) (1934). *Tables of the Incomplete Beta Function*. London: The Biometrika Trust.

TALLIS, G. M. (1961). The moment-generating function of the truncated multi-normal distribution. *J. R. Statist. Soc.* B **23**, 223–9.

TALLIS, G. M. (1965). Plane truncation in normal populations. *J. R. Statist. Soc.* B **27**, 301–7.

[*Received May* 1972. *Revised April* 1973]