On the probability of a model

Cesareo Villegas and Tim Swartz*
Department of Statistics and Actuarial Science,
Simon Fraser University, Canada.
Carmillo Martinez
Department of Mathematics and Statistics,
University College of the Fraser Valley, Canada.

Abstract

The posterior probabilities of K given models when improper priors are used depend on the proportionality constants assigned to the prior densities corresponding to each of the models. It is shown that this assignment can be done using natural geometric priors in multiple regression problems if the normal distribution of the residual errors is truncated. This truncation is a realistic modification of the regression models, and since it will be made far away from the mean, it has no other effect beyond the determination of the proportionality constants, provided that the sample size is not too large. In the case $K = 2$, the posterior odds ratio is related to the usual F statistic in "classical" statistics. Assuming zero-one losses the optimal selection of a regression model is achieved by maximizing the posterior probability of a submodel. It is shown that the geometric criterion obtained in this way is asymptotically equivalent to Schwarz's asymptotic Bayesian criterion, sometimes called the BIC criterion. An example of polynomial regression is used to provide numerical comparisons between the new geometric criterion, the BIC criterion and the Akaike information criterion.

Key Words: Bayesian testing, geometric Bayesian inference, geometric priors, model selection, probability of a model, sharp hypotheses, variable selection.
AMS subject classification: 62J20, 62A05.

1 Introduction

As was pointed out by D.R. Cox in a comment to Dempster (1971), a Bayesian analysis of two models with flat priors requires a value for the relative heights of the two priors densities. In the present paper it will be shown that such a constant value can be found if the normal distribution of the residual error is truncated. This truncation is a realistic modification

Villegas and Swartz were partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

*Correspondence to: Tim Swartz, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby B.C., Canada V5A1S6. Email:tim@stat.sfu.ca

of the model, and since it is made far away from the mean, it has no other effect beyond the determination of the constant, provided that the sample size is not too large.

In choosing a regression model one usually wants to test the null hypothesis that some of the regression coefficients are zero, or equivalently that some of the independent variables have no influence on the response. Elimination of these variables leads to a regression submodel having fewer independent variables. This process is called variable selection or subset selection. We restrict our attention to a particular set of K submodels. After the appropriate truncation of the likelihood function, the improper prior is equivalent to a truncated proper prior that assigns equal prior probabilities to these K submodels. The corresponding Bayesian analysis then provides posterior probabilities for each of the K submodels. Assuming zero-one losses the optimal subset selection is achieved by maximizing the posterior probability of a submodel. When $K = 2$, the corresponding posterior odds ratio is related to the usual F statistic. It will be shown that this criterion is asymptotically equivalent to Schwarz's (1978) asymptotic Bayesian criterion, which is sometimes called the BIC criterion. For other procedures of model selection see Chipman, George and McCulloch (2002), Berger and Pericchi (1996), O'Hagan (1995), Kass and Wasserman (1995), Gelfand and Dey (1994), Rueda (1992), Bhansali (1986), Smith and Spiegelhalter (1980), Geisser and Eddy (1979), Akaike (1973) and Jeffreys (1961), as well as the references therein.

In Section 2, we consider the truncation approach in the context of the univariate normal model. This is extended in Section 3 to a simple multivariate model. The general problem of variable selection in multiple regression is then addressed in Section 4 where comparisons are made with well known variable selection criteria.

2 Univariate normal model

Consider the problem of testing the simple null hypothesis $\mathrm{H}_0 : \mu = \mu_0$ against the composite alternative $\mathrm{H}_a : \mu \neq \mu_0$ when the standard deviation σ is known, and a sample of size n is available. Since we assume a normal distribution, the sample mean $\bar{X}$ is a sufficient statistic, and, if $\bar{X} = \bar{x}$ is

the observed value, the likelihood function is proportional to

$$\exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right\}. \tag{2.1}$$

The uniform prior on the whole real line cannot be used to find the posterior probability of the point null hypothesis H_0: $\mu = \mu_0$ because it assigns zero prior probability to it, and therefore the corresponding posterior probability is zero regardless of the data. Bartlett (1957) considered a proper prior that puts probability π_0 at $\mu = \mu_0$ and distributes the rest, $1 - \pi_0$ uniformly over

$$\mu_0 - h\sigma < \mu < \mu_0 + h\sigma$$

for some fixed h. He found that when $h \to \infty$, the posterior probability of H_0 converges to 1, and so this method fails to identify a limiting prior that represents ignorance concerning μ. Indeed, to obtain such a limiting prior, π_0 should be a function of h that converges to zero when $h \to \infty$.

Limiting priors that represent ignorance are usually improper priors that assign an infinite measure to the whole parameter space. The value assigned to any subset is not a prior probability because it is not necessarily bounded by 1, and it may be called a *probability index*. In particular, the probability index of the null hypothesis H_0 may be any positive constant $C > 0$. Consider then the improper prior which assigns to $\mu = \mu_0$ the probability index $C > 0$ and is otherwise uniform on the whole real line, with index density equal to 1.

The *posterior index* is the product of this improper prior and the likelihood function (2.1). The posterior index of H_a is the integral of (2.1), namely

$$\mathrm{I}\left(\mathrm{H}_a|\bar{x}\right) = \sigma\sqrt{2\pi/n}$$

and the posterior index of H_0 is

$$\mathrm{I}\left(\mathrm{H}_0|\bar{x}\right) = C\exp\left\{-z^2/2\right\}$$

where $z = \sqrt{n}(\bar{x} - \mu_0)/\sigma$. The posterior odds ratio for H_0 is then

$$\mathrm{PO} = \frac{\mathrm{I}\left(\mathrm{H}_0|\bar{x}\right)}{\mathrm{I}\left(\mathrm{H}_a|\bar{x}\right)} = \frac{C\sqrt{n}}{\sqrt{2\pi}\,\sigma}\exp\left\{-z^2/2\right\}$$

and depends on the prior index C of the null hypothesis H_0, as would be expected. Therefore, in order to have a prior that represents ignorance, we need to find a value C that is determined by the model. It will be shown that this requires a slight modification of the model.

It should be recognized that in the real world, sampling distributions have compact support. In the usual case when the sample space is a finite dimensional vector space, this means that the sampling distribution assigns probability 1 to a bounded set. For example, we usually say that the height of a person of a given sex, age and race has a normal distribution, and a normal distribution does not have compact support. But the normality assumption is only a convenient approximation because we very well know that (i) there cannot be persons with negative heights, and (ii) there cannot be persons with arbitrarily large heights (say greater than 3 metres). In the same book in which Gauss derived the normal distribution for the first time, he recognized that normal distributions are really approximations:

> The function just found cannot, it is true, express rigorously the probabilities of the errors: for since the possible errors are in all cases confined within certain limits, the probability of errors exceeding these limits ought always to be zero, while our formula always gives some value (Gauss 1809).

Therefore, if we want to be more realistic, we have to truncate the normal distribution in our model. If the truncation points are far away from the mean, this truncation will have almost no effect on the common statistical inferences, except, as we are going to see now, on the determination of the constant C needed to find the posterior probability of the null hypothesis H_0. Suppose then that we truncate the normal distribution at the points $\mu \pm h\sigma$. Consider the case $n = 1$. The truncated normal density is

$$f(x) = \begin{cases} c_h \sigma^{-1} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} & \text{if} \quad \mu - h\sigma \leq x \leq \mu + h\sigma \\ 0 & \text{otherwise} \end{cases}$$

where $c_h^{-1} = \int_{-h}^{h} exp\left\{-t^2/2\right\} dt$. Under H_a , the likelihood function is

$$c_h \sigma^{-1} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad \text{if} \quad x - h\sigma \leq \mu \leq x + h\sigma$$

and 0 otherwise. At this point in the analysis, we know that under H_a the unknown mean belongs to the interval

$$x - h\sigma \leq \mu \leq x + h\sigma \tag{2.2}$$

with probability 1. Therefore it is possible to truncate the parameter space to the interval (2.2). With this restricted model, inferences are focused on the sampling distribution of a hypothetical response Y, independent of the observed $X = x$. See Villegas and Martinez (1999). The prior is then truncated to the interval (2.2) and the prior index under H_a is the length of the interval $I_x(\mathrm{H}_a) = 2h\sigma$. We note that $I_x(\mathrm{H}_a) = 2h\sigma$ does not depend on x, and hence, we refer to it as the *effective prior index* $I^*(\mathrm{H}_a)$. Under the assumption of prior ignorance, the prior index of the null hypothesis $I(\mathrm{H}_0) = C$ is then set equal to $I^*(\mathrm{H}_a)$. This implies $C = 2h\sigma$, and we have therefore completed the construction of the prior when $n = 1$.

Of course the prior should be the same for any sample size n. Suppose that we have a sample $x = (x_1, ..., x_n)$. Then the likelihood function under H_a is

$$c_h^n \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \quad \text{if } \; x_{(n)} - h\sigma \leq \mu \leq x_{(1)} + h\sigma$$

and 0 otherwise where $x_{(1)} = \min_i\{x_i\}$ and $x_{(n)} = \max_i\{x_i\}$. The posterior index of H_a is

$$I\left(H_a|x\right) = c_h^n \sigma^{-n} \int_{x_{(n)}-h\sigma}^{x_{(1)}+h\sigma} \exp\left\{-\tfrac{1}{2\sigma^2}\sum\left(x_i - \mu\right)^2\right\}\}d\mu$$

and the posterior index of H_0 is

$$I\left(H_0|x\right) = c_h^n 2h\sigma^{\,-n+1} \exp\left\{-\tfrac{1}{2\sigma^2}\sum\left(x_i - \mu_0\right)^2\right\}. \tag{2.3}$$

Assumption 2.1. We assume that h is large enough so that the integral

$$\int_{x_{(n)}-h\sigma}^{x_{(1)}+h\sigma} \exp\left\{-\tfrac{1}{2\sigma^2}\sum\left(x_i - \mu\right)^2\right\}d\mu$$

can be replaced by the integral from $-\infty$ to $+\infty$ with negligible error.

Disregarding this error, the integral is equal to

$$\frac{\sqrt{2\pi}\ \sigma}{\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\sum (x_i-\bar{x})^2\right\}$$

and substituting $\sum (x_i-\mu_0)^2 = n(\bar{x}-\mu_0)^2 + \sum (x_i-\bar{x})^2$ in (2.3), it follows that the posterior odds ratio is

$$\mathrm{PO} = \frac{I(\mathrm{H}_0|x)}{I(\mathrm{H}_a|x)} = \sqrt{\frac{2n}{\pi}} h\ \exp\left\{-z^2/2\right\} \tag{2.4}$$

where $z = \sqrt{n}(\bar{x}-\mu_0)/\sigma$. This is maximized by $z=0$, and the maximum is

$$\mathrm{MPO} = h\sqrt{2n/\pi}.$$

The relative posterior odds ratio RPO is therefore

$$\mathrm{RPO} = \mathrm{PO}/\mathrm{MPO} = \exp\left\{-z^2/2\right\}$$

and we note that the relative posterior odds ratio does not depend on the constant h.

The acceptance of one of the two hypotheses may be considered as a decision problem. Under H_0 the dimension p of the parameter space is 0 and under H_a, the dimension is 1. Let the decisions be $\hat{p}$ =0 (accept H_0) and $\hat{p}$ =1 (accept H_a) where $\hat{p}$ is an estimate of p. There is underestimation if $\hat{p}$ =0 when $p=1$ and overestimation if $\hat{p}$ =1 when $p=0$. If c_u is the cost of underestimation and c_o is the cost of overestimation, then the optimal decision (i.e. Bayes rule) is to reject H_0 if

$$\mathrm{PO} \leq c_u/c_o.$$

The ratio c_u/c_o is therefore the critical value of the posterior odds ratio PO. Under the usual zero-one losses $c_u = c_o = 1$, H_0 is rejected if $\mathrm{PO} \leq 1$. The formula (2.4) relates the critical value of PO to the critical value of z. Solving for z^2 we have

$$z^2 = 2\log\left(\sqrt{\tfrac{2n}{\pi}}\ \tfrac{h}{PO}\right). \tag{2.5}$$

Equation (2.5) shows that for fixed PO, the critical value of z increases as n increases, a qualitative rule that is now widely accepted. Bickel and Doksum (1977, page 175) state:

> This problem arises particularly in goodness of fit tests when we test the hypothesis that a very large sample comes from a particular distribution. Such hypotheses are often rejected even though for practical purposes "the fit is good enough". The reason is that n is so large that unimportant small discrepancies are picked up.

Table 1 shows critical values of z based on (2.5) for PO = 1 and selected values of h and n. The value ϵ is the joint probability of the truncated tails in the normal distribution. Logarithms of the probabilities of one tail were taken from Pearson and Hartley (1954). Note that values of h smaller than $h = 5$ have not been included because in these cases Assumption 2.1 may not be valid. Consider then comparisons with usual significance levels based on $n = 1$. When $h = 5$, the critical value $z = 1.66$ corresponds to a 0.10 significance level. Also when $h = 8.5$, by interpolation from the table, we get the critical value $z = 1.96$, corresponding to the significance level 0.05. Finally, when $h = 19$, the critical value of z is 2.33, corresponding to the significance level 0.01.

An idea about realistic values of h can be obtained in the case of human heights from the Guinness Book of Records (1996), according to which the world all-time record belongs to Robert Pershing Wadlow, born in 1918 in Alton, Illinois. When he died in 1940, his height was 8 feet 11.1 inches, or 272 cm. In Macdonell (1901) it is reported that the average height of 25,878 U.S. recruits was 170.94 cm. and the standard deviation was 6.56 cm. However, the mean height for men in the U.S. has increased during the last century. Recent data for human heights in the U.S. may be found in the NCHS Growth Curves for Children (1977). These data are based on a national sample survey designed by the U.S. Bureau of the Census. On page 44, we find that for males of age between 24 and 25 years the mean height is 178.0 cm. and the standard derivation is 7.0 cm. Therefore, the world all-time record is 13.4 standard deviations above the mean. According to the same Guinness Book of Records (1996), the world's shortest living adult is believed to be Gul Mohammad of New Delhi, India, with a height of 22.5 in. or 17.3 standard deviations below the mean. When choosing h to accommodate outliers such as the world's tallest and shortest people, it is in keeping with common statistical practice of using error distributions with long tails. The Guinness Book of Records gives many other records for plants and animals that suggest that it may not be unreasonable to

					n				
h	1	2	4	8	10	50	200	500	ϵ
5	1.66	1.86	2.04	2.20	2.25	2.58	2.84	3.00	$.6 \times 10^{-6}$
6	1.77	1.96	2.13	2.29	2.33	2.65	2.90	3.06	$.2 \times 10^{-8}$
7	1.85	2.03	2.20	2.35	2.40	2.71	2.96	3.11	$.3 \times 10^{-11}$
8	1.93	2.10	2.26	2.41	2.45	2.76	3.00	3.15	$.1 \times 10^{-14}$
9	1.99	2.15	2.31	2.45	2.50	2.80	3.04	3.19	$.2 \times 10^{-18}$
10	2.04	2.20	2.35	2.50	2.54	2.84	3.07	3.22	$.8 \times 10^{-23}$
11	2.08	2.24	2.39	2.53	2.58	2.87	3.11	3.25	$.4 \times 10^{-27}$
12	2.13	2.28	2.43	2.57	2.61	2.90	3.13	3.28	$.4 \times 10^{-32}$
13	2.16	2.32	2.46	2.60	2.64	2.93	3.16	3.30	$.1 \times 10^{-37}$
14	2.20	2.35	2.49	2.63	2.67	2.96	3.18	3.32	$.2 \times 10^{-43}$
15	2.23	2.38	2.52	2.65	2.70	2.98	3.20	3.34	$.7 \times 10^{-50}$
16	2.26	2.41	2.55	2.68	2.72	3.00	3.22	3.36	$.1 \times 10^{-56}$
17	2.28	2.43	2.57	2.70	2.74	3.02	3.24	3.38	$.8 \times 10^{-64}$
18	2.31	2.45	2.59	2.72	2.76	3.04	3.26	3.40	$.2 \times 10^{-71}$
19	2.33	2.48	2.61	2.74	2.78	3.06	3.28	3.41	$.2 \times 10^{-79}$
20	2.35	2.50	2.63	2.76	2.80	3.07	3.29	3.43	$.6 \times 10^{-88}$

Table 1: *Critical values of z based on (2.5) for* PO $= 1$ *and selected values of h and n. The value ϵ is the joint probability of the truncated tails in the normal distribution.*

choose similar values of h whenever the response is of a biological nature.

It is worth remembering that $C = 2h\sigma$ and hence one could instead choose C. The advantage of choosing h over C is that h has a nice interpretation as the number of standard deviations beyond the mean for which sampling probability is negligible. When choosing h based on past data as suggested above, one is really advocating an empirical Bayes approach where the prior depends on the data. However, referring again to Table 1, one may instead consider an objective approach by choosing a large value, say $h = 20$, which may be seen to be adequate for many applications, and is on the conservative side in the sense that H_0 is rejected less often. In the final analysis, it must be recognized that the truncation constant h is a component of the model, and its determination is not very different from the determination of the significance level of a "classical" test.

2.1 Unknown variance

In the previous section it was assumed that the standard deviation σ is known. When σ is unknown, the parameter space is the union of

$$\Omega_0 = \{(\mu, \sigma) : \ \mu = \mu_0\},$$

the subset corresponding to H_0, and

$$\Omega_a = \{(\mu, \sigma) : \ \mu \neq \mu_0\},$$

the subset corresponding to H_a. The likelihood function is

$$\tfrac{1}{\sigma^n} \exp\left\{ - \tfrac{n}{2\sigma^2}[\hat{\sigma}^2 + (\mu - \bar{x})^2] \right\}$$

where $\hat{\sigma}$ is the maximum likelihood estimate of the unknown σ. In the previous section, the prior had a point mass $C = 2h\sigma$ placed at $\mu = \mu_0$ and a density 1 on $\mu \neq \mu_0$. Equivalently, since an improper prior is determined only up to an arbitrary scale factor, the prior could have been chosen to have a point mass 1 placed at $\mu = \mu_0$ and a density $(2h\sigma)^{-1}$ on $\mu \neq \mu_0$. When σ is unknown the natural parameter space for σ is the multiplicative group of positive integers. The invariant measure in this group has differential $d\sigma/\sigma$ This structural prior is, in a Kleinian sense, a geometric prior for σ, and therefore is a good candidate to represent ignorance. See Villegas (1990). Thus the prior differential is

$$d(\mu, \sigma) = \begin{cases} \frac{d\mu}{2h} \frac{d\sigma}{\sigma^2} & \text{on } \Omega_a \\ \\ \frac{d\sigma}{\sigma} & \text{on } \Omega_0 \end{cases}.$$

Note that this prior is invariant under changes of scale (see Villegas 1990), and that on Ω_a it is the inner (or Jeffreys) prior $d\mu d\sigma/\sigma^2$, which was used for the first time by Edgeworth (1883). For a good understanding of the rationale behind the use of Jeffreys prior, see George and McCulloch (1993). A simple calculation shows that the posterior odds ratio for H_0 is

$$\mathrm{PO} = h\sqrt{\tfrac{2n}{\pi}} \ \left[1 + \tfrac{t^2}{n-1}\right]^{-\frac{n}{2}}$$

where $t = \sqrt{n}(\bar{x} - \mu_0)/s$ and $s^2 = \sum (x_i - \bar{x})^2/(n-1)$.

Therefore the relative posterior odds ratio is simply

$$\text{RPO} = \left[1 + \frac{t^2}{n-1}\right]^{-\frac{n}{2}}$$

and is equal to the corresponding relative density of a t-distribution with n-1 degrees of freedom.

3 A simple multivariate model

Let $y^{(1)}, \ldots, y^{(n)}$ be a sample of size n from a truncated multivariate normal distribution in $\mathcal{R}^q$ with unknown mean $\mu = (\mu_1, \ldots, \mu_q)'$ and known covariance matrix $\sigma^2 \mathrm{I}_q$. The distribution is truncated to a q dimensional coordinate cube $C_q(\mu)$ centered at μ with sides of known length $2h\sigma$. Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$ be the observed mean vector and let $\hat{\sigma}^2 = \frac{1}{nq}\sum_{i=1}^{n} \left\|y^{(i)} - \bar{y}\right\|^2$ where $||y||$ denotes the norm of the vector y.

Let $\{A_k : k = 1, \ldots, K\}$ be a family of different subsets of the finite set $A = \{1, 2, \ldots, q\}$, and let q_k denote the cardinality of A_k . Without loss of generality, we assume that the sets A_k are ordered so that the sequence $\{q_k : k = 1, \ldots, K\}$ is monotone increasing. We assume that $A_1 = \emptyset$ such that $q_1 = 0$. We then consider the hypotheses

$$\mathrm{H}_k : \quad \mu_j = 0 \quad \text{if } j \notin A_k$$

and let Ω_k be the set of all $\mu \in \mathcal{R}^q$ that satisfy the condition H_k. Note that Ω_1 is the origin in $\mathcal{R}^q$.

The parameter space Ω is the set of all pairs $\theta = (k, \mu)$ where $k \in \{1, \ldots,$ K $\}$ and $\mu \in \Omega_k$. The likelihood is a function of θ given by

$$c_h^{nq} \sigma^{-nq} \exp\left\{-\frac{n}{2\sigma^2}\left[q\hat{\sigma}^2(k) + \sum_{j \in A_k} (\mu_j - \bar{y}_j)^2\right]\right\} \tag{3.1}$$

where

$$\hat{\sigma}^2(k) = \hat{\sigma}^2 + q^{-1} \sum_{j \notin A_k} \bar{y}_j^2$$

and $\bar{y}_j$ is the jth coordinate of the mean vector $\bar{y}$. Note that when $k = 1$ the summation in (3.1) is over an empty set and can be dropped. Since we are assuming that σ is known, the factor σ^{-nq} may be dropped from (3.1). However we are not dropping it because the results obtained in this way are useful in the more realistic case in which σ is unknown.

We now construct a prior that represents ignorance. Since the prior should not depend on the sample size n, suppose for a moment that $n = 1$. The set of all pairs $\theta = (k, \mu)$ with a fixed k constitute a *copy* of $\mathcal{R}^q$ that we denote $\mathcal{R}^q_k$. The parameter space Ω can therefore be viewed as a finite set corresponding to the collection of all copies $\mathcal{R}^q_k$, $k \in \{1, \ldots, K\}$. When the parameter space is a finite set, uniform priors have been frequently used to represent ignorance. As was pointed out by J.M. Keynes (1921), this use, suggested by James Bernoulli, gained wide acceptance under the name of the principle of non-sufficient reason, later called the principle of insufficient reason.

Accordingly, a prior Π that represents ignorance should have a uniform restriction Π_k on the copy $\mathcal{R}^q_k$, with a constant density that may be different on each copy. To determine these constant densities, consider what happens when the observation y becomes known. Under H_k, and assuming $k > 1$, the mean μ belongs to the cube $C_k(y) \in \Omega_k$ with sides of length $2h\sigma$ and volume $(2h\sigma)^{q_k}$. Suppose that the parameter space Ω is truncated to a subspace $\Omega(y)$ which is the set of all pairs $\theta = (k, \mu)$ with $\mu \in C_k(y)$. Then the prior Π should also be truncated to $\Omega(y)$. This truncated prior is now a proper prior and to represent ignorance it should assign equal probability K^{-1} to all K hypotheses H_k. Therefore the original prior Π assigns mass K^{-1} to R^q_1, and for $k > 1$, the prior differential corresponding to R^q_k is

$$K^{-1}(2h\sigma)^{-q_k} \prod_{j \in A_k} d\mu_j. \tag{3.2}$$

Obviously this is also the prior that should be used for any sample size n. The posterior index of H_1 is simply

$$I(\mathrm{H}_1|Y) = K^{-1} c_h^{nq} \sigma^{-nq} \exp\left\{-\frac{1}{2\sigma^2}\|Y\|^2\right\}$$

where Y is the $n \times q$ matrix whose ith row transposed is $y^{(i)}$ and the norm $||Y||$ is defined by $\|Y\|^2 = \sum \left\|y^{(i)}\right\|^2$. If $k > 1$, the posterior index of H_k is

the integral over Ω_k of the product of (3.1) and (3.2), namely

$$\mathrm{I}\left(\mathrm{H}_k|Y\right) = K^{-1}c_h^{nq}\sigma^{-nq}\left[\frac{1}{h}\sqrt{\frac{\pi}{2m}}\right]^{q_k}\exp\left\{-\frac{nq\hat{\sigma}^2\left(k\right)}{2\sigma^2}\right\}. \tag{3.3}$$

This formula also includes the case $k = 1$ because $\hat{\sigma}^2\left(1\right) = \frac{1}{nq}\sum_{i=1}^{n}\left\|y^{(i)}\right\|^2$.

Now assume a common loss for choosing a wrong model. Then to minimize the expected loss we should choose the hypothesis that maximizes the posterior index $\mathrm{I}\left(\mathrm{H}_k|Y\right)$. Equivalently we may minimize

$$\frac{nq\hat{\sigma}^2\left(k\right)}{\sigma^2} + \ q_k \log\frac{2nh^2}{\pi}.$$

It is also natural to use a coordinate free truncation under which the support of the distribution of the random variables $y^{(i)}$, $i \in \{1,\ldots,n\}$ is a q dimensional ball (or solid sphere) centered at μ with volume $(2h\sigma)^q$ equal to the volume of the cube in the coordinatewise case. Since the truncation is done far away from the mean of the sampling distribution the difference between the likelihood functions produced by the two types of truncation is negligible. A similar argument shows that the prior has differential (3.2) when $k > 1$, and mass K^{-1} when $k = 1$. The posterior index is again given by (3.3). Therefore, our testing criteria is the same using either a coordinatewise or a coordinate free truncation.

4 Multiple regression

4.1 Basic results

Consider the multiple regression model

$$y = X\beta + \sigma u \tag{4.1}$$

where the $n \times q$ design matrix X is of full rank q, the random vector u has a standard multivariate normal distribution, σ is an unknown positive number and β is the unknown regression vector. The maximum likelihood estimate of β is the vector $\hat{\beta}$ given by

$$\hat{\beta} = \left(X'X\right)^{-1}X'y$$

and the regression equation (4.1) may be written in matrix notation as

$$\begin{bmatrix} I & 0 \\ \beta & \sigma \end{bmatrix} \begin{bmatrix} X' \\ u' \end{bmatrix} = \begin{bmatrix} X' \\ y' \end{bmatrix}. \tag{4.2}$$

The group G of matrices of the form

$$\theta = \begin{bmatrix} I & 0 \\ \beta & \sigma \end{bmatrix} \tag{4.3}$$

is called the regression group. If we set

$$v = \begin{bmatrix} X' \\ u' \end{bmatrix}, \quad z = \begin{bmatrix} X' \\ y' \end{bmatrix},$$

the regression equation (4.2) becomes

$$\theta v = z. \tag{4.4}$$

Let g be a fixed matrix of the regression group and consider the change of variables $y \to y^*$ given by

$$z^* = gz, \quad z^* = \begin{bmatrix} X & y^* \end{bmatrix}'.$$

Then the new regression equation is

$$\theta^* v = z^* \tag{4.5}$$

where the new parameter θ^* is

$$\theta^* = g\theta. \tag{4.6}$$

The new regression model (4.5) with parameter θ^* is identical to the old regression model (4.4) with parameter θ, and therefore there is no reason why the prior π^* that represents ignorance concerning θ^* should be different from the prior π that represents ignorance concerning θ. Therefore $\pi^* = \pi$ should be a left invariant measure in the regression group G (see Villegas 1981). This is, in a Kleinian sense, the *natural geometric prior* on the group G. It is well known that it is also the Jeffreys prior for the group model (see George and McCulloch 1993).

Theorem 4.1. *The natural geometric prior for the multiple regression model (4.1) has differential*

$$d\beta d\sigma / \sigma^{q+1}. \tag{4.7}$$

Proof. Set

$$g = \begin{bmatrix} I & 0 \\ \gamma' & \tau \end{bmatrix}. \tag{4.8}$$

Substitution of (4.3) and (4.8) in (4.6) gives the values for the new parameters β^* and σ^*:

$$\beta^* = \tau\beta + \gamma, \quad \sigma^* = \tau\sigma.$$

Let $\pi^*(\beta^*, \sigma^*)d\beta^* d\sigma^*$ be the prior differential for the new parameters. We substitute the old variables β and σ in two stages. In the first stage we substitute σ for σ^* and obtain the intermediate prior differential

$$\tau\pi^* (\beta^*, \tau\sigma)\, d\beta^* d\sigma.$$

In the second stage we substitute β for β^* and obtain

$$\tau^{q+1}\pi^* (\tau\beta + \gamma, \tau\sigma)\, d\beta d\sigma. \tag{4.9}$$

Let $\pi(\beta, \sigma)d\beta d\sigma$ be the prior differential for the old variables. Hence from (4.9) we have

$$\pi (\beta, \sigma) = \tau^{q+1}\pi^* (\tau\beta + \gamma, \tau\sigma). \tag{4.10}$$

Since $\pi^* = \pi$, substitution in (4.10) gives the functional equation for π,

$$\pi (\beta, \sigma) = \tau^{q+1}\pi (\tau\beta + \gamma, \tau\sigma) \tag{4.11}$$

which must hold for arbitrary values β, σ, τ and γ. Given β and σ, choose τ and γ such that $\tau\beta + \gamma = 0$ and $\tau\sigma = 1$. Substitution in (4.11) gives

$$\pi (\beta, \sigma) = \pi (0, 1) / \sigma^{q+1}.$$

But $\pi(0,1)$ is an arbitrary value which can be chosen equal to 1, and the conclusion follows immediately.□

Following Villegas (1981) the natural geometric prior (4.7) is called the *inner prior*, and the corresponding posterior is called the *inner posterior*.

Let H be the $q \times q$ lower triangular matrix with positive diagonal elements defined by the Cholesky decomposition

$$X'X = HH',$$

and let γ and $\hat{\gamma}$ be defined by

$$\gamma = H'\beta, \quad \hat{\gamma} = H'\hat{\beta}.$$

Then $\hat{\gamma}$ has a multivariate normal distribution with mean γ and covariance matrix $\sigma^2 I_q$. The maximum likelihood estimate of σ is $\hat{\sigma}$ defined by

$$\hat{\sigma}^2 = \frac{1}{n} \|y - \hat{y}\|^2$$

where $\hat{y} = X\hat{\beta}$. The random variables $\hat{\beta}$ and $\hat{\sigma}$ (i.e. $\hat{\gamma}$ and $\hat{\sigma}$) are independent sufficient statistics and their sampling distributions are given by the equations

$$\sigma^{-1} H' \left(\hat{\beta} - \beta\right) = u, \tag{4.12}$$

$$\frac{\hat{\sigma}}{\sigma} = \frac{\chi}{\sqrt{n}} \tag{4.13}$$

where χ has the chi distribution with $n - q$ degrees of freedom and is independent of u.

The likelihood function is the function of β and σ (i.e. γ and σ),

$$\frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\left[n\hat{\sigma}^2 + \|\gamma - \hat{\gamma}\|^2\right]\right\} \tag{4.14}$$

and the inner posterior differential is therefore the product of (4.14) and (4.7). It can be shown that the inner posterior distribution is also given by equations (4.12) and (4.13) if $\hat{\beta}$ and $\hat{\sigma}$ are fixed at the observed values and χ has a chi distribution with $\nu = n$ degrees of freedom.

Elimination of σ between (4.12) and (4.13) gives

$$\hat{\sigma}^{-1} H' \left(\hat{\beta} - \beta\right) = t \tag{4.15}$$

where $t = \sqrt{n}u/\chi$ has a Student multivariate t-distribution with $\nu = n$ degrees of freedom. The equation (4.15) gives the marginal posterior distribution of β. Since the posterior differential of χ is proportional to

$$\chi^n \exp\left\{-\chi^2/2\right\} \frac{d\chi}{\chi},$$

it follows that the marginal posterior differential of σ is proportional to

$$\sigma^{-n} \exp\left\{-\frac{n\hat{\sigma}^2}{2\sigma^2}\right\} \frac{d\sigma}{\sigma}.$$

4.2 Variable selection

In the present section we consider the linear model (4.1) with replications because we are interested in asymptotic results when the number of replications tends to infinity. Consider then the multiple regression model with m replications for each covariate setting

$$y^{(i)} = X\beta + \sigma u^{(i)}, \quad \{i = 1, \ldots, m\} \tag{4.16}$$

where σ is known.

Let $\bar{y} = m^{-1} \sum y^{(i)}$ be the observed mean vector. The maximum likelihood estimate of β is

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\,\bar{y}.$$

There are 2^q possible submodels of the model (4.16) where each submodel assigns the value zero to some components of the regression vector β. Equivalently, each submodel excludes a particular subset of the columns of X . We restrict our attention to a particular set of K submodels. We denote by X_k the $n \times q_k$ matrix consisting of those columns of X which are included in the kth submodel. We assume that the submodels are ordered in such a way that the sequence q_k is monotone increasing. Thus when $k = 1$, the corresponding value q_1 may be zero. If $q_1 = 0$ then the first submodel assigns the value zero to all components of the regression vector β . Otherwise the kth submodel is

$$y^{(i)} = X_k\beta^{(k)} + \sigma u^{(i)}, \quad \{i = 1, \ldots, m\} \tag{4.17}$$

where $\beta^{(k)}$ is the unknown q_k dimensional regression vector. The maximum likelihood estimate of the regression vector $\beta^{(k)}$ is

$$\hat{\beta}^{(k)} = (\mathrm{X_k}'\mathrm{X_k})^{-1}\mathrm{X_k}'\,\bar{y}.$$

Let H_k be the $q_k \times q_k$ lower triangular matrix with positive diagonal elements defined by the Cholesky decomposition

$$X_k'X_k = H_kH_k',$$

and let $\gamma^{(k)}$ and $\hat{\gamma}^{(k)}$ be defined by

$$\gamma^{(k)} = H_k'\beta^{(k)}, \quad \hat{\gamma}^{(k)} = H_k'\hat{\beta}^{(k)}.$$

Then $\hat{\gamma}^{(k)}$ has a multivariate normal distribution with mean $\gamma^{(k)}$ and covariance matrix $(\sigma^2/m)\mathrm{I}_{q_k}$. The likelihood function for the kth model assuming $q_1 \neq 0$ is the function of $\gamma^{(k)}$ (i.e. $\beta^{(k)}$),

$$\frac{1}{\sigma^N} \exp\left\{-\frac{m}{2\sigma^2}\left[n\hat{\sigma}^2(k) + \left\|\gamma^{(k)} - \hat{\gamma}^{(k)}\right\|^2\right]\right\} \tag{4.18}$$

where $N = mn$ and

$$\hat{\sigma}^2(k) = \frac{1}{N}\sum_{i=1}^{m}\left\|y^{(i)} - X_k\hat{\beta}^{(k)}\right\|^2.$$

If $q_1 = 0$, the likelihood function for the first submodel is defined only for $\beta = 0$ and is given by

$$\frac{1}{\sigma^N} \exp\left\{-\frac{1}{2\sigma^2}\|Y\|^2\right\}.$$

As was pointed out in Section 2, it should be recognized that in the real world the random vectors $y^{(i)}$ have compact support. A more realistic model is therefore obtained by assuming that under the kth submodel the $y^{(i)}$ are independent multivariate normal random vectors with covariance matrix $\sigma^2 I_n$ truncated to a coordinate cube centered at the mean $\mu^{(k)} = X_k\beta^{(k)}$ with sides of length $2h\sigma$. Note that this is a coordinatewise truncation and it is not equivalent to a coordinatewise truncation of the sampling distribution of $X_k\hat{\beta}^{(k)}$.

Let $r_k(h)$ be the radius of a ball in q_k dimensional space with volume $(2h\sigma)^{q_k}$, equal to the volume of a coordinate cube with sides of length $2h\sigma$. A ball (or solid sphere) B_k on the regression subspace centered at $X_k\beta^{(k)}$ and with radius $r_k(h)$ is the orthogonal projection of $\bar{y}$ on the regression subspace of a uniquely defined cylinder C_k. The sampling distribution of $\bar{y}$ is truncated to the cylinder C_k. Since the truncation is done far away from the mean of the sampling distribution the difference between the likelihood functions corresponding to the two types of truncation (to a cube or to a cylinder) is negligible . The truncation to a cylinder is equivalent to a spherical coordinate free truncation of the sampling distribution of $X_k\hat{\beta}^{(k)}$ in the regression subspace to the ball B_k, and it is also equivalent to the truncation of the sampling distribution of $\hat{\gamma}^{(k)}$ to a ball B_k^* centered at $\gamma^{(k)}$ and with radius $r_k(h)$. With this truncation the likelihood function is

zero if $\left\|\gamma^{(k)} - \hat{\gamma}^{(k)}\right\| \geq r_k(h)$ and otherwise it can be still considered to be proportional to (4.18), because the truncation is made far away from the population mean.

We now have to construct a prior that represents ignorance. We begin by constructing a prior for the kth submodel where $q_k > 0$. Since this regression model belongs to the exponential family, the method developed in Villegas (1990) may be applied. Under the kth submodel, a prior Π_k that represents ignorance concerning the unknown parameter $\gamma^{(k)}$ should be a uniform prior on the q_k dimensional Euclidean space, having a constant density that may depend on k. To determine these constant densities, consider what happens when $m = 1$ and the observation $y^{(1)}$ becomes available. Under the kth submodel, the mean $\gamma^{(k)}$ is then known to belong to a ball $\hat{B}_k^*$ centered at $\hat{\gamma}^{(k)}$ with radius $r_k(h)$. Suppose for a moment that the parameter space for this submodel is truncated to the ball $\hat{B}_k^*$. The prior Π_k should also be truncated to $\hat{B}_k^*$. The truncated prior is now a proper prior and to represent ignorance it should assign to this submodel a probability which should be the same for all K submodels, and is therefore K^{-1}. If $q_1 = 0$, the prior should assign to the first submodel the probability K^{-1}. If $q_k > 0$, the prior should be a uniform prior on the full q_k Euclidean space with constant density $K^{-1}(2h\sigma)^{-q_k}$. Obviously, this is also the prior that should be used for any sample size m. If $q_1 = 0$, the posterior index for the first submodel is simply

$$K^{-1}\sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}\|Y\|^2\right\} \tag{4.19}$$

and otherwise, the posterior index for the kth submodel is

$$K^{-1}\sigma^{-N}\left[\frac{1}{h}\sqrt{\frac{\pi}{2m}}\right]^{q_k}\exp\left\{-\frac{N\hat{\sigma}^2(k)}{2\sigma^2}\right\}. \tag{4.20}$$

Note that if $q_1 = 0$, the formula (4.20) for $k = 1$ gives (4.19) and therefore (4.20) gives the posterior index in all cases. Now assume a common loss for choosing a wrong model. Then to minimize the expected loss we choose the submodel that maximizes the posterior index (4.20). Equivalently we minimize

$$\frac{N\hat{\sigma}^2(k)}{\sigma^2} + q_k \log\frac{2mh^2}{\pi}.$$

Note that when m is large we effectively minimize

$$N\frac{\hat{\sigma}^2(k)}{\sigma^2} + q_k \log N. \tag{4.21}$$

The asymptotic Bayesian criterion proposed by Schwarz (1978), called the BIC criterion by Akaike (1977) minimizes (4.21) and is therefore asymptotically equivalent to our new criterion.

When the submodels to be considered are such that no two of them have the same number of independent variables, a piecewise linear loss function

$$L(\hat{k}, k) = \begin{cases} c_u(q_k - q_{\hat{k}}) & \text{if } \hat{k} < k \\ c_o(q_{\hat{k}} - q_k) & \text{if } \hat{k} > k \end{cases}$$

may be used where $L(\hat{k}, k)$ is the loss when $\hat{k}$ is chosen and k is the true value. Here c_u is the unit cost of underestimation and c_o is the unit cost of overestimation. As is well known, the optimal decision $\hat{k}$ is simply the $c_u/(c_u + c_o)$ quantile of the posterior distribution of k.

As was stated in Section 3, the posterior distribution obtained with a coordinatewise truncation on the regression space is the same as the posterior distribution obtained with a coordinate free truncation, and therefore both cases lead to the same variable selection criterion.

4.3 Unknown variance

We now extend the multiple regression model previously considered to the more realistic case where σ is unknown.

The likelihood function (4.18) is the same as before but with σ viewed as an additional variable. The prior is the same as before but multiplied by $d\sigma/\sigma$. To find the posterior index for the kth submodel we integrate the unknown parameters σ and $\gamma^{(k)}$. This can be done in two stages. In the first stage, integration with respect to $\gamma^{(k)}$ gives (4.20) multiplied by $d\sigma/\sigma$. In the second stage, further integration with respect to σ gives, disregarding factors that do not depend on k,

$$\left[\frac{1}{h}\sqrt{\frac{\pi}{2m}}\right]^{q_k} \hat{\sigma}(k)^{-N}.$$

Therefore our variable selection criterion chooses k which minimizes

$$N \log \hat{\sigma}^2(k) + q_k \log \frac{2mh^2}{\pi}. \tag{4.22}$$

This criterion will be called the *geometric criterion*. When m is large, (4.22) is asymptotically equivalent to

$$N \log \hat{\sigma}^2 (k) + q_k \log N \tag{4.23}$$

which is the BIC criterion proposed by Schwarz (1978).

"Classical" statistics is particularly concerned with the case $K = 2$, in which we have to choose one of two regression models, M1 with q_1 independent variables and M2 with q_2 independent variables. The posterior odds ratio is given by

$$\text{PO} = \frac{\text{P}(\text{M}_1|Y)}{\text{P}(\text{M}_2|Y)} = \left[\frac{2mh^2}{\pi}\right]^{\frac{q_2-q_1}{2}} \left[\frac{\hat{\sigma}(q_2)}{\hat{\sigma}(q_1)}\right]^N.$$

If $q_1 < q_2$, we have

$$\frac{\hat{\sigma}^2(q_2)}{\hat{\sigma}^2(q_1)} = \frac{q_2 - q_1}{N - q_2} F + 1$$

where F is an F-statistic with $q_2 - q_1$ and $N - q_2$ degrees of freedom. Therefore if $q_1 < q_2$,

$$\text{PO} = \left[\frac{2mh^2}{\pi}\right]^{\frac{q_2-q_1}{2}} \left[1 + \frac{q_2 - q_1}{N - q_2} F\right]^{-\frac{N}{2}}$$

whence PO is a monotone decreasing function of the usual F statistic.

Example. Guttman (1967) considered the problem of fitting a polynomial curve to data generated by Monte Carlo simulation from the fifth degree polynomial regression model

$$y_{ij} = 5.50 + 0.07x_i + 2.64x_i^2 - 0.27x_i^3 - 1.12x_i^4 + 0.85x_i^5 + u_{ij} \tag{4.24}$$

$\{i = 1, \ldots, 9;\ j = 1, \ldots, n_i\}$. The values of the independent variable are $x_i = i - 5$ and the u_{ij}/σ are independent standard normal variates. The number of replications of x_i is $n_i = 10$ for $i \neq 5$ and $n_5 = 100$. The same model has been analyzed by Hager and Antle (1968), Halpern (1973) and Akaike (1977). Akaike (1977) also considered two modified models in which the coefficients were reduced by the factors 0.1 and 0.01.

To construct an example, we consider $m = 2$ replications of a basic polynomial model similar to (4.24) in which

(i) the coefficients are reduced by the factor 0.01

(ii) $n_i = 1$ for $i \neq 5$ and $n_5 = 10$

(iii) two different values $\sigma = 1.0$ and $\sigma = 1.5$ are used in generating data

Only submodels corresponding to polynomials of degree less than or equal to 8 are considered, because this ensures that all of the submodels are full rank.

To choose a submodel Akaike (1973) proposed the AIC criterion (Akaike information criterion) that minimizes

$$N \log \hat{\sigma}^2(k) + 2q_k. \tag{4.25}$$

The AIC criterion (4.25), the BIC criterion (4.23) and the geometric criterion (4.22) are all of the form

$$N \log \hat{\sigma}^2(k) + \lambda q_k$$

and differ only in their respective values of λ. The geometric criterion (4.22) and the BIC criterion (4.23) give the same result when

$$n = \frac{2h^2}{\pi}. \tag{4.26}$$

Similarly the AIC criterion (4.25) gives the same result as the geometric criterion (4.22) when

$$h\sqrt{\frac{2m}{\pi}} = e. \tag{4.27}$$

Since for the chosen design $m = 2$ and $n = 18$, substitution in (4.26) shows that BIC gives the same result as a geometric criterion with $h = 5.31$. Similarly, substitution in (4.27) shows that AIC gives the same result as a geometric criterion with $h = 2.41$. Note that $h = 2.41$ is small and it may not satisfy the assumption that the truncation is done far away from the mean. Therefore model probabilities computed with $h = 2.41$ may be suspect. Data sets were obtained by Monte Carlo simulation. These simulations were repeated 6 times, giving 6 independent data sets when $\sigma = 1$ and 6 independent data sets when $\sigma = 1.5$. Since the geometric

criterion depends on a given h, four different geometric criteria were used, corresponding to the values $h = 6, 8, 10, 15$.

Table 2 (for $\sigma = 1$) and Table 3 (for $\sigma = 1.5$) give the frequency of selection of submodels by each of the criteria. Note that the frequencies corresponding to BIC and to the geometric criterion with $h = 6$ are indentical, as was to be expected since BIC is formally equivalent to a geometric criterion with $h = 5.31$.

	Polynomial Degree							
criterion	1	2	3	4	5	6	7	8
AIC	0	0	1	1	2	1	0	1
BIC	0	0	2	1	1	1	0	1
Geometric, $h = 6$	0	0	2	1	1	1	0	1
Geometric, $h = 8$	0	0	2	2	1	0	0	1
Geometric, $h = 10$	0	0	2	3	0	0	0	1
Geometric, $h = 15$	0	0	3	2	1	0	0	0

Table 2: *Frequency of selection of submodels by each of the criteria for* $\sigma = 1$.

	Polynomial Degree							
criterion	1	2	3	4	5	6	7	8
AIC	0	0	2	1	1	1	0	1
BIC	0	0	3	2	0	0	0	1
Geometric, $h = 6$	0	0	3	2	0	0	0	1
Geometric, $h = 8$	0	0	4	2	0	0	0	0
Geometric, $h = 10$	0	0	5	1	0	0	0	0
Geometric, $h = 15$	0	0	5	1	0	0	0	0

Table 3: *Frequency of selection of submodels by each of the criteria for* $\sigma = 1.5$.

Note that in both Tables 2 and 3, the column corresponding to a polynomial of the third degree has values that increase as the value of h increases. In general, it can be said that as the value of h increases, the criteria are more parsimonious. Note also that as we go from $\sigma = 1.5$ to $\sigma = 1$, the frequency distributions become more concentrated around the true value 5. The probabilities assigned to the 8 models using the geometric prior with $h = 6$ are given in Table 4 when $\sigma = 1.5$. The probabilities corresponding to $h = 8$ and $\sigma = 1.5$ are given in Table 5.

	Polynomial Degree							
Data Set	1	2	3	4	5	6	7	8
1	0.00	0.00	0.17	0.54	0.21	0.04	0.03	0.01
2	0.00	0.00	0.72	0.13	0.12	0.02	0.01	0.00
3	0.00	0.00	0.48	0.35	0.15	0.02	0.00	0.00
4	0.00	0.00	0.09	0.11	0.09	0.14	0.03	0.54
5	0.00	0.00	0.24	0.35	0.16	0.20	0.03	0.02
6	0.01	0.00	0.77	0.15	0.06	0.01	0.00	0.00

Table 4: *Model probabilities using the geometric prior with* $h = 6$ *and* $\sigma = 1.5$.

	Polynomial Degree							
Data Set	1	2	3	4	5	6	7	8
1	0.00	0.00	0.24	0.56	0.17	0.02	0.01	0.00
2	0.00	0.00	0.81	0.11	0.07	0.01	0.00	0.00
3	0.00	0.00	0.57	0.32	0.10	0.01	0.00	0.00
4	0.00	0.00	0.21	0.20	0.12	0.14	0.02	0.31
5	0.00	0.00	0.35	0.38	0.13	0.12	0.01	0.01
6	0.02	0.00	0.82	0.12	0.04	0.00	0.00	0.00

Table 5: *Model probabilities using the geometric prior with* $h = 8$ *and* $\sigma = 1.5$.

Acknowledgments

This paper is dedicated to the memory of our friend and lead author, Cesareo Villegas. The intellectual ideas found in the paper are almost entirely those of Professor Villegas and continue his lifelong pursuit of the development of prior distributions in statistics. We thank the Editor, W.G. Manteiga for the handling of this paper; well beyond his term of duty.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory* (B.N. Petrov and F. Csaki, eds.) Budapest: Akademiai Kiado. Reprinted in *Breakthroughs in Statistics, 1* (1992). (S. Kotz and N.L. Johnson, eds.) New York: Springer-Verlag.

Akaike, H. (1977). On entropy maximization principle. *Applications of Statistics* (P.R. Krishnaiah, ed.) Amsterdam: North-Holland, 27-42.

Bartlett, M,S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, **44**, 533-534.

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

Bhansali, R.J. (1986). Asymptotically efficient selection of the order by the criterion autoregressive transfer function. *Annals of Statistics*, **14**, 315-325.

Bickel, P.J. and Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden-Day, Inc.

Chipman, H., George, E.I. and McCulloch, R.E. (2002). The practical implementation of Bayesian model selection. To appear in the IMS monograph, *Model Selection*.

Dempster, A.P. (1971). Model searching and estimation in the logic of inference. *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott, eds.) Toronto: Holt Rinehart and Winston of Canada, 56-81.

Edgeworth, F.Y. (1883). The method of least squares. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science Series*, **16**, 360-375.

Gauss, K.F. (1809). *Theoria Motus Corporum Coelestium, Hamburg*. English translation by C.H. Davis 1963; New York: Dover.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153-160.

Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B*, **56**, 501-513.

George, E.I. and McCulloch, R. (1993). On obtaining invariant prior distributions. *Journal of Statistical Planning and Inference*, **37**, 169-179.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society Series B*, **29**, 83-100.

Hager, H. and Antle, C. (1968). The choice of the degree of a polynomial. *Journal of the Royal Statistical Society Series B*, **30**, 469-471.

Halpern, E.F. (1973). Polynomial regression from a Bayesian approach. *Journal of the American Statistical Association*, **68**, 137-143.

Jeffreys, H. (1961). *Theory of Probability, 3rd edition*. Oxford, Oxford University Press.

Kass, R.E. and Wasserman, L. (1995). A reference test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928-934.

Keynes, J.M. (1921). *A Treatise on Probability*. London: Macmillan.

Macdonell, W.R. (1901). On criminal anthropometry and the identification of criminals. *Biometrika*, **1**, 177-227.

NCHS Growth Curves for Children (1977). National Center for Health Statistics. U.S. Department of Health, Education and Welfare.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society Series B*, **57**, 99-138.

Pearson, E.S. and Hartley, H.O. (1954). *Biometrika Tables for Statisticians, vol. 1, 3rd edition*. Cambridge: University Press.

Rueda, R. (1992). A Bayesian alternative to parametric hypothesis testing. *Test*, **1**, 61-68.

Smith, A.F.M. and Spiegelhalter, D.J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B*, **42**, 768-776.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Villegas, C. (1981). Inner statistical inference II. *Annals of Statistics*, **9**, 768-776.

Villegas, C. (1990). Bayesian inference in models with Euclidean structures. *Journal of the American Statistical Association*, **85**, 1159-1164.

Villegas, C. and Martnez, C.J. (1999). On the concepts of coherence and admissibility. *Test*, **8**, 319-338.