University of Victoria

Department of Economics

Econometrics Working Paper EWP1004

ISSN 1485-6441

Bayesian Estimation of a Possibly Mis-Specified Linear Regression Model

David E. Giles*

Department of Economics, University of Victoria
Victoria, B.C., Canada V8W 2Y2

December, 2010

Abstract

We consider Bayesian estimation of the coefficients in a linear regression model, using a conjugate prior, when certain additional exact restrictions are placed on these coefficients. The bias and matrix mean squared errors of the Bayes and restricted Bayes estimators are compared when these restrictions are both true and false. These results are then used to determine the consequences of model mis-specification in terms of over-fitting or under-fitting the model. Our results can also be applied directly to determine the properties of the "ridge" regression estimator when the model may be mis-specified, and other such applications are also suggested.

Keywords: Bayes estimator, regression model, linear restrictions, model mis-specification, bias, matrix mean squared error

Mathematics Subject Classification: 62F07, 62F10, 62F15, 62J05, 62J07

Author Contact:

David E. Giles, Dept. of Economics, University of Victoria, P.O. Box 1700, STN CSC, Victoria, B.C., Canada V8W 2Y2; e-mail: dgiles@uvic.ca; Phone: (250) 721-8540; FAX: (250) 721-6214

1. Introduction

Consider the standard linear regression model,

$$y = X\beta + u \quad ; \quad u \sim N(0, \sigma^2 I) \quad , \tag{1}$$

where X is $(n \times k)$ and non-stochastic; β is $(k \times 1)$; and u and y are $(n \times 1)$. For many of our results, X need not have full rank, k. This rank condition is noted below when it is required. We will consider some of the properties of the Bayes estimator of β, under a natural-conjugate prior for the parameters, when the model is mis-specified through the inclusion (exclusion) of irrelevant (relevant) regressors. The properties of this estimator when a diffuse (uninformative) prior is used are well-known. In this case the Bayes estimator of β coincides with the maximum likelihood (and least squares) estimator (MLE). Then, over-fitting the model by including extraneous columns in X leaves the Bayes estimator unbiased and weakly consistent, but there is a loss of precision in both small and large samples. Accordingly, in finite samples, over-fitting increases the mean squared error (MSE) of the estimator of β. Conversely, under-fitting the model by wrongly excluding columns from X generally results in an estimator of β that is both biased and inconsistent. However, in this case there a gain in precision, and so in some parts of the parameter space there will be a reduction in MSE if the model is under-fitted (*e.g.*, Toro-Vizcarrondo and Wallace, 1968).

In this paper we adopt the natural-conjugate prior p.d.f. for (β, σ), and consider some of the finite-sample and (large-n) asymptotic consequences of these two types of model mis-specification for the point estimation of β under a Bayes' (minimum expected loss) rule when the loss function is either quadratic, absolute error, or zero-one in form. Initially, we take a fairly general, if slightly unorthodox, formulation of the mis-specification problem by considering various features of the posterior p.d.f. when exact linear restrictions are imposed on the elements of β. This enables us to make quite general comparisons between "unrestricted" and "restricted" Bayes estimators of β. Subsequently, we show that for the particular restrictions of primary interest here, this approach is equivalent to working with the joint and conditional posterior p.d.f.'s for β.

When the sample size, n, is finite, we maintain a quadratic loss structure and pay special attention to the matrix MSE (MMSE) of this Bayes estimator of β under these two types of model mis-

specification. We show that MMSE is unambiguously worsened if the model is over-fitted, and we derive a condition under which the MMSE is actually improved when the model is under-fitted. The fact that we focus on the sampling properties of Bayes estimators may be seen by some as somewhat unorthodox – bias and MSE are, after all, frequentist notions. However, this type of analysis has precedents. Bester and Hansen (2005) note that there are interesting connections between certain commonly used bias-correction techniques for MLEs and Bayesian prior densities. They argue "…that thinking about bias-reduction may offer a potentially useful approach to formulating Bayesian priors" (Bester and Hansen, 2005, p.2). There are a number of specific studies of the frequentist properties of Bayes estimators. For example, Giles and Rayner (1979) compare the natural-conjugate Bayes regression estimator with the least squares estimator in a similar manner. Howlader and Weiss (1988) compare the biases and MSEs of the maximum likelihood and Bayes estimators of the parameters of the Cauchy distribution; Bolstad (2004, pp.151-153) makes similar frequentist comparisons between the MLE and conjugate Bayes estimators of a sample proportion; and Rao and Shinozaki (1978) and Rao (2009) evaluate the performances of certain empirical Bayes estimators in the same way. While a staunch Bayesian reader may question the relevance of such results, we believe in the merits of a flexible approach to such matters.

The rest of the paper is constructed as follows. In the next section we provide a framework that enables us to incorporate exact restrictions into the natural-conjugate Bayes regression estimator. Section 3 compares the restricted and unrestricted Bayes estimators in terms of their asymptotic and finite-sample properties, and develops the condition under which the restricted estimator will dominate the unrestricted estimator in terms of matrix MMSE, when the restrictions are false. These results are then applied, in section 4, to obtain various results relating to the mis-specification of the model through the omission of relevant regressors, or the inclusion of irrelevant ones. Some connections between our analysis and the "ridge" regression estimator are noted in section 5, and section 6 concludes.

2. Bayes estimation and exact parametric restrictions

For our normal likelihood, the natural-conjugate prior p.d.f. for the parameters in (1) is:

$$p(\beta,\sigma) = p(\beta \mid \sigma)\, p(\sigma), \tag{2}$$

where

$$p(\beta \mid \sigma) \propto \sigma^{-k} \exp[(\beta - \bar{\beta})' A (\beta - \bar{\beta})/(2\sigma^2)] \quad ; \qquad |A| > 0 \tag{3}$$

$$p(\sigma) \propto \sigma^{-(\nu_0+1)} \exp[-\nu_0 c_0^2/(2\sigma^2)] \quad ; \quad \nu_0, c_0 > 0 . \tag{4}$$

Then, the marginal posterior p.d.f. for β is well-known to be multivariate Student-t:

$$p(\beta \mid y) \propto [n'c^2 + (\beta - \breve{\beta})'(A + X'X)(\beta - \breve{\beta})]^{-(n'+k)/2} , \tag{5}$$

where

$$n' = n + \nu_0 \quad ; \quad n'c^2 = \nu_0 c_0^2 + y'y + \overline{\beta}' A \overline{\beta} - \breve{\beta}'(A + X'X)\breve{\beta} \tag{6}$$

and

$$\breve{\beta} = (A + X'X)^{-1}(A\overline{\beta} + X'y) \tag{7}$$

is the mean (and mode and median) of $p(\beta \mid y)$. So, $\breve{\beta}$ is the Bayes estimator of β not only under quadratic loss, but also under an absolute error or zero-one loss function, and it is well-defined even if $rank(X) < k$. Let us now consider some properties of this estimator that have not been discussed previously.

Suppose that we have dogmatic prior information about β in the form of exact linear restrictions, $R\beta = r$, where R is non-stochastic, $(g \times k)$ and of rank g; and r is non-stochastic and $(g \times 1)$. To obtain a (quasi-) Bayes estimator of β while taking this dogmatic information into account, we can consider the conditional posterior p.d.f., $p(\beta \mid R\beta = r, y)$. However, for general R and r this p.d.f. is not very convenient to handle, and so initially we adopt an alternative approach to deriving as estimator of β that incorporates the information in $p(\beta, \sigma)$ while being consistent with the dogmatic restrictions. In section 4 we return to $p(\beta \mid R\beta = r, y)$ for specific interesting choices of R and r.

The restricted (natural-conjugate) Bayes estimator of β may be obtained by determining the modal value of $p(\beta \mid y)$, subject to the constraint(s) $R\beta = r$. To achieve this, we set up the Lagrangian

$$\Lambda = a[n'c^2 + (\beta - \breve{\beta})'(A + X'X)(\beta - \breve{\beta})]^{-(n'+k)/2} + \varphi'(R\beta - r), \tag{8}$$

where a is the normalizing constant for $p(\beta \mid y)$ in (5), and φ is a non-random $(g \times 1)$ vector of Lagrange multipliers. Taking the first derivative of (8) w.r.t. β and solving[1], we get:

$$\tilde{\beta} = \breve{\beta} + (A + X'X)^{-1} R'[R(A + X'X)^{-1} R']^{-1}(r - R\breve{\beta}) . \qquad (9)$$

The derivation of $\tilde{\beta}$ (which again does not require that X has full rank) is somewhat unorthodox, in that it mixes a dogmatic set of restrictions with the flexible prior p.d.f. used to obtain $p(\beta | y)$. However, $\tilde{\beta}$ gives us a very general estimator which, for the special case in which we are primarily interested, is easily shown to be equivalent to the mode of the appropriate conditional posterior p.d.f. for β, and is also equivalent to the appropriate marginal posterior p.d.f.'s mode when A and $\bar{\beta}$ in $p(\beta | \sigma)$ are assigned values to reflect *exact* prior information in certain dimensions. We now consider and contrast some of the properties of $\breve{\beta}$ and $\tilde{\beta}$.

3. Estimators' sampling properties

3.1 *Asymptotic properties*

The (large n asymptotic properties of $\breve{\beta}$ and $\tilde{\beta}$ are easily dealt with. The former is based on a proper prior p.d.f., so it is weakly consistent. Moreover, it is well known (*e.g.*, Zellner, 1971, pp. 31-33) that under quite general conditions $p(\beta | y)$ becomes normal, with a mode at the MLE, $b = (X'X)^{-1} X'y$, as $n \to \infty$, provided that X has full rank, k. So, $\breve{\beta}$ is also best asymptotically normal (BAN), relative to the information set. Intuitively, it is clear that $\tilde{\beta}$ should also be weakly consistent and BAN relative to its own information set, provided that the restrictions, $R\beta = r$, are true. In this case, the partially dogmatic prior information complicates the picture slightly, but the asymptotic properties of $\tilde{\beta}$ are easily established.

Theorem 1 If $R\beta = r$, and $\Sigma = \underset{n \to \infty}{Limit}(n^{-1} X'X)$ is finite and positive-definite (p.d.) then $\tilde{\beta}$ is weakly consistent for β.

Proof

$$\text{p lim}(\tilde{\beta}) = \text{p lim}(\breve{\beta}) + \underset{n \to \infty}{Limit}[n^{-1}A + n^{-1}X'X]^{-1} R'[R \underset{n \to \infty}{Limit}(n^{-1}A + n^{-1}X'X)^{-1} R']^{-1}(r - \text{plim}(\breve{\beta})) .$$

Using the consistency of $\breve{\beta}$, and the positive definiteness of Σ,

$$\begin{aligned} \text{p lim}(\tilde{\beta}) &= \beta + \Sigma^{-1} R'[R\Sigma^{-1}R']^{-1}(r - R\beta) \\ &= \beta \quad , \end{aligned}$$

if $R\beta = r$. ■

Theorem 2 $\sqrt{n}(\widetilde{\beta} - \beta) \xrightarrow{d} N[0, \sigma^2 Q]$, where $Q = [\Sigma^{-1} - \Sigma^{-1}R'(R\Sigma^{-1}R')^{-1}R\Sigma^{-1}]$, provided that X has full rank, and $R\beta = r$.

Proof

Because $\breve{\beta} \to b$ as $n \to \infty$, it follows that $\sqrt{n}(\breve{\beta} - \beta) \xrightarrow{d} N[0, \sigma^2\Sigma^{-1}]$.

Now,

$$\widetilde{\beta} = [I - (A + X'X)^{-1}R'\{R(A + X'X)^{-1}R'\}^{-1}R]\breve{\beta} + (A + X'X)^{-1}R'[R(A + X'X)^{-1}R']^{-1}r,$$

So, using Theorem 1, $\sqrt{n}(\widetilde{\beta} - \beta) \xrightarrow{d} N[0, \sigma^2 J^*\Sigma^{-1}\Omega']$, where $J^* = [I - \Sigma^{-1}R'(R\Sigma^{-1}R')^{-1}R]$. Finally, note that

$$J^*\Sigma^{-1}J^* = (\Sigma^{-1} - \Sigma^{-1}R'(R\Sigma^{-1}R')^{-1}R\Sigma^{-1} = Q,$$

as required. ■

It is clear, then, that asymptotically $\widetilde{\beta}$ is equivalent to the restricted MLE,

$$b_R = b + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}), \tag{10}$$

provided that $rank(X) = k$. Accordingly, it is not surprising that there is a gain in asymptotic precision when using $\widetilde{\beta}$ in favour of $\breve{\beta}$, even if $R\beta \neq r$. Inspection of the asymptotic covariance matrices (a.c.m.'s) of these two estimators reveals that their difference, $\Delta = \{a.c.m.[\sqrt{n}(\breve{\beta} - \beta)] - a.c.m.[\sqrt{n}(\widetilde{\beta} - \beta)]\}$ is at least positive semi-definite (p.s.d.).

3.2 *Finite sample properties*

Giles and Rayner (1979) note that $Bias(\breve{\beta}) = \overline{W}A\beta^*$, and $V(\breve{\beta}) = \sigma^2\overline{W}X'X\overline{W}$, where $\overline{W} = (A + X'X)^{-1}$, and $\beta^* = (\overline{\beta} - \beta)$. Turning to the restricted Bayes estimator, we have:

Theorem 3 (i) $Bias(\widetilde{\beta}) = J\,Bias(\breve{\beta}) + \overline{W}R'(R\overline{W}R')^{-1}(r - R\beta)$

(ii) $V(\widetilde{\beta}) = \sigma^2 J\overline{W}X'X\overline{W}J'$,

where[2] $J = [I - \overline{W}R'(R\overline{W}R')^{-1}R]$. (11)

Proof

(i) Let $E_y(.)$ denote expectation over the sample space. Then, from (9),

$$\begin{aligned} E_y(\tilde{\beta} - \beta) &= E_y[\breve{\beta} - \beta + \overline{W}R'(R\overline{W}R')^{-1}(r - R\breve{\beta})] \\ &= Bias(\breve{\beta}) + E[\overline{W}R'(R\overline{W}R')^{-1}(r - R\beta + R\beta - R\breve{\beta})] \\ &= Bias(\breve{\beta}) + \overline{W}R'(R\overline{W}R')^{-1}(r - R\beta) + \overline{W}R'(R\overline{W}R')^{-1}R(\beta - E(\breve{\beta})) \\ &= J\,Bias(\breve{\beta}) + \overline{W}R'(R\overline{W}R')^{-1}(r - R\beta) \end{aligned}$$

■

(ii) Note that $\tilde{\beta} = J\breve{\beta} + \overline{W}R'(R\overline{W}R')^{-1}r$. So,

$$V(\tilde{\beta}) = JV(\breve{\beta})J' = \sigma^2 J\overline{W}X'X\overline{W}J'$$ ■

It follows that imposing the restrictions, $R\beta = r$, improves the small-sample precision of the Bayes estimator of β, even if these restrictions are false:

Corollary 1 The matrix $D = [V(\breve{\beta}) - V(\tilde{\beta})]$ is at least p.s.d., whether or not $R\beta = r$ holds.

Proof

Expanding the expression for $V(\tilde{\beta})$ in Theorem 3 and comparing it with $V(\breve{\beta})$, we can form the matrix

$$[R'(R\overline{W}R')^{-1}R]D[R'(R\overline{W}R')^{-1}R] = \sigma^2[R'(R\overline{W}R')^{-1}R\overline{W}X'X\overline{W}R'(R\overline{W}R')^{-1}R],$$

which is at least p.s.d., So, *D* itself is at least p.s.d. , as $[R'(R\overline{W}R')^{-1}R]$ has full rank.

■

Now consider the MMSE's of $\breve{\beta}$ and $\tilde{\beta}$. The latter is

$$MMSE(\tilde{\beta}) = V(\tilde{\beta}) + Bias(\tilde{\beta})Bias(\tilde{\beta})' = \tilde{M},$$ (12)

say. Define $MMSE(\breve{\beta}) = \breve{M}$ analogously. Under quadratic loss, we may say that $\tilde{\beta}$ is "preferred" to $\breve{\beta}$ if $(\breve{M} - \tilde{M})$ is at least p.s.d.. This strong MSE criterion has been used by a numerous

authors, including Toro-Vizcarrondo and Wallace (1968), Swindel (1976), Giles and Rayner (1979), Pliskin (1987) and Trenkler (1988). Let us now compare $\breve{M}$ and $\widetilde{M}$, first when $R\beta = r$ holds, and second when $R\beta \neq r$.

Theorem 4 If $R\beta = r$, then $D^* = (\breve{M} - \widetilde{M})$ is at least p.s.d..

Proof

$$\breve{M} = \sigma^2 \overline{W} X' X \overline{W} + \overline{W} A \beta^* \beta^{*\prime} A \overline{W}$$

$$\widetilde{M} = \sigma^2 J \overline{W} X' X \overline{W} J + J \overline{W} A \beta^* \beta^{*\prime} A \overline{W} J',$$

provided that $R\beta = r$. So, in this case $\widetilde{M} = J\breve{M}J'$, and from (11) it is readily shown that $RD^*R' = R\breve{M}R'$, which is at least p.s.d.. As *R* has full rank, it follows that D^* itself is also at least p.s.d..

■

So, imposing valid restrictions on the elements of β never "worsens" the MMSE of the Bayes estimator of β. This result is directly analogous to the situation when a diffuse prior p.d.f. is adopted (*i.e.*, the case of MLE) – then, imposing valid restrictions leaves the estimator unbiased, but never "worsens" its precision or its MMSE. That is, when $R\beta = r$ is true, b_R is "preferred" to b, and $\widetilde{\beta}$ is "preferred" to $\breve{\beta}$, in terms of MMSE.

Now consider the case when false restrictions are imposed on the elements of β.

Theorem 5 When $R\beta \neq r$, $\widetilde{\beta}$ is "preferred" to $\breve{\beta}$, in terms of MMSE iff

$$\lambda = (r - R\beta)'(R\breve{M}R')^{-1}(r - R\beta) \leq 1.$$

Proof

$$\breve{M} = \sigma^2 \overline{W} X' X \overline{W} + \overline{W} A \beta^* \beta^{*\prime} A \overline{W},$$

and if $R\beta \neq r$, necessarily, then from Theorem 3,

$$\widetilde{M} = J\breve{M}J' + \overline{W}R'(R\overline{W}R')^{-1}\delta\delta'(R\overline{W}R')^{-1}R\overline{W} + J\overline{W}A\beta^*\delta'(R\overline{W}R')^{-1}R\overline{W} + \overline{W}R(R\overline{W}R')\delta\beta^* A\overline{W}J',$$

where $\delta = (r - R\beta)$.

In terms of MMSE, $\widetilde{\beta}$ is "preferred" to $\breve{\beta}$ iff $(\breve{M}-\widetilde{M})$ is at least p.s.d.. Now,

$$\begin{aligned}\Delta^{*} &= (\breve{M}-\widetilde{M}) \\ &= \breve{M}-J\breve{M}J'-\overline{W}R'(R\overline{W}R')^{-1}\delta\delta'(R\overline{W}R')^{-1}R\overline{W}-J\overline{W}A\beta^{*}\delta'(R\overline{W}R')^{-1}R\overline{W}-\overline{W}R'(R\overline{W}R')^{-1}\delta\beta^{*\prime}A\overline{W}J'\end{aligned}$$

and so,

$$\begin{aligned}R\Delta^{*}R' &= R(\breve{M}-J\breve{M}J')R'-\delta\delta'-RJ\overline{W}A\beta^{*}\delta'-\delta\beta^{*}A\overline{W}J'R' \\ &= R\breve{M}R'-\delta\delta'\end{aligned}$$

as $RJ=0$.

So, $\widetilde{\beta}$ is "preferred" to $\breve{\beta}$ iff $(R\Delta^{*}R')$ is at least p.s.d.. That is, iff

$$\eta'(R\breve{M}R'-\delta\delta')\eta \geq 0 \qquad ;\ \forall\ \eta \neq 0$$

or, iff

$$\lambda^{*} = (\eta'\delta\delta'\eta)/(\eta'R\breve{M}R'\eta) \leq 1 \qquad ;\ \forall\ \eta \neq 0\,.$$

This last inequality holds, for all η, iff $\lambda = \sup_{(\eta)}(\lambda^{*}) \leq 1$. From Rao (1973, p.60), this necessary and sufficient condition is that $\lambda = \delta'(R\breve{M}R')^{-1}\delta \leq 1$.

■

Note that if $R\beta = r$, then $\lambda = 0 \leq 1$, and so then $\widetilde{\beta}$ is "preferred" to $\breve{\beta}$ for all possible sample data, and all values of the parameters, as in Theorem 4. Theorem 5 indicates that, in general, which of $\widetilde{\beta}$ or $\breve{\beta}$ is "preferred" depends on all of the quantities *R*, *r*, β, σ^{2}, *X*, *A* and $\overline{\beta}$. It does *not* depend on the parameters in the marginal prior p.d.f. for σ.

The condition $\lambda \leq 1$ is of some interest when comparing the performances of $\widetilde{\beta}$ and $\breve{\beta}$. The fact that λ is unobservable invites the question, "can we test to see if the hypothesis $H_0 : \lambda \leq 1$ is favoured against $H_A : \lambda > 1$?" In an analogous situation involving b_R and *b*, Toro-Vizcarrondo and Wallace (1978) used the fact that their counterpart to our λ is the non-centrality parameter in the distribution for the UMP test of the validity of $R\beta = r$, together with the monotone likelihood principle, to construct a classical test of (their counterpart to our) H_0. However this type of

situation does not arise here, and it is not clear how a useful classical test of H_0 could be formulated.

Giles and Rayner (1979) also developed a counterpart to our λ - criterion in a similar comparison of b and $\breve{\beta}$. They suggested several ways of "testing" H_0. Of these, one with a strong Bayesian flavour may be mentioned here. Define:

$$F_{\sigma,y}(1) = \Pr(\lambda \leq 1 \mid \sigma, y)$$
$$= \Pr[(r - R\beta)'(R\breve{M}R')^{-1}(r - R\beta) \leq 1 \mid \sigma, y] \quad .$$

A posteriori, $F_{\sigma,y}(1)$ is a probability associated with a quadratic form in the "random" vector, $\delta = (r - R\beta)$, where $p(\delta \mid \sigma, y) \equiv N[(r - R\breve{\beta}), \sigma^2 R\overline{W}X'X\overline{W}R']$.

Numerical algorithms such as those proposed by Imhof (1961) or Davies (1980) may be used to exploit this normality and compute $F_{\sigma,y}(1)$. Then, the marginal posterior probability that $\lambda \leq 1$ may be approximated numerically by univariate integration:

$$F_y(1) = \Pr(\lambda \leq 1 \mid y) = \int_0^\infty F_{\sigma,y}(1)\, p(\sigma \mid y) d\sigma \quad ,$$

where, by natural-conjugacy, and using (6):

$$p(\sigma \mid y) \propto \sigma^{-(n'+1)} \exp[-n'c^2/(2\sigma^2)] \quad .$$

Given equal prior odds, the posterior odds in favour of $H_0 : \lambda \leq 1$ relative to $H_A : \lambda > 1$ are $O_{0A} = [F_y(1)/(1 - F_y(1)]$, and for any symmetric loss function over the λ-space H_0 is favoured over H_A if $O_{0A} > 1$.

Whatever method is used to test H_0 and subsequently adopt either $\tilde{\beta}$ or $\breve{\beta}$, a preliminary-test strategy is involved and this affects the sampling properties of the final estimator of β (*e.g.*, Giles and Giles, 1993). In particular, if one of $\tilde{\beta}$ or $\breve{\beta}$ is chosen subsequent to a test of H_0, then the bias

and MMSE expressions derived above will no longer be valid. We do not pursue this point further here.

4. Specification analysis

We now use the general results of the last section to analyze the consequences, for the sampling properties of the natural-conjugate Bayes estimator of β, of mis-specifying the model (1) by including irrelevant regressors or excluding relevant ones. Let us re-write (1) as

$$y = X\beta + u \equiv X_1\beta_1 + X_2\beta_2 + u \,, \tag{13}$$

where X_i and β_i are $(n \times k_i)$ and $(k_i \times 1)$ respectively (i = 1, 2). Now set $R = (I, \quad 0)$ and $r = 0$. So, the restrictions, $R\beta = r$, are just $\beta_1 = 0$. If these restrictions are valid, then using $\widetilde{\beta}$ amounts to estimating β under the appropriate model specification, while using $\breve{\beta}$ amounts to over-fitting the model. On the other hand, if the restrictions are false then $\widetilde{\beta}$ is associated with under-fitting the model, while $\breve{\beta}$ is then based on the appropriate model specification[3]. As we will see in the next sub-section, approaching the specification analysis *via* the results of section 3 is equivalent to restricting (or failing to restrict) the prior p.d.f. for β directly. However, it has the advantages of convenience and transparency.

4.1 *Posterior analysis*

We can now show that for the above choice of R and r, $\widetilde{\beta}_2$ is just the mean (= mode) of the conditional posterior p.d.f. for β_2, $p(\beta_2 \mid \beta_1 = 0, y)$, and of course $\widetilde{\beta}_1 = 0$ by construction. We exploit the property of a multivariate Student-t distribution that its conditional[4] (and marginal) distributions are also Student-t. We partition A conformably with $X'X$, so:

$$(A + X'X) \equiv \begin{pmatrix} A_{11} + X_1'X_1 & A_{12} + X_1'X_2 \\ A_{12}' + X_2'X_1 & A_{22} + X_2'X_2 \end{pmatrix} \equiv W \equiv \begin{pmatrix} W_{11} & W_{12} \\ W_{12}' & W_{22} \end{pmatrix}.$$

Theorem 6 Let $R = (I, \ \ 0)$ and $r = 0$. In addition let $E_\beta(.)$ denote expectation over the (β dimensions of the) parameter space. Then $E_\beta(\beta_2 \mid \beta_1 = 0, y) = \tilde{\beta}_2$.

Proof

From Zellner (1971, p.388), and $p(\beta \mid y)$ in (5),

$$\begin{aligned} E_\beta(\beta_2 \mid \beta_1 = 0, y) &= \breve{\beta}_2 - W_{22}^{-1} W_{12}'(\beta_1 - \breve{\beta}_1) \\ &= \breve{\beta}_2 + W_{22}^{-1} W_{12} \breve{\beta}_1 \quad . \end{aligned} \tag{14}$$

Also, from (9), in this case:

$$\begin{aligned} \tilde{\beta} \equiv \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} &= \begin{pmatrix} \breve{\beta}_1 \\ \breve{\beta}_2 \end{pmatrix} + \overline{W} R' (R \overline{W} R')^{-1} \left(r - R \begin{pmatrix} \breve{\beta}_1 \\ \breve{\beta}_2 \end{pmatrix} \right) \\ &= J \begin{pmatrix} \breve{\beta}_1 \\ \breve{\beta}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \breve{\beta}_2 + W_{22}^{-1} W_{12}' \breve{\beta}_1 \end{pmatrix} \quad . \end{aligned} \tag{15}$$

So, comparing (14) and (15), we see that

$$\tilde{\beta}_2 = E_\beta(\beta_2 \mid \beta_1 = 0, y) = \mathrm{mode}(\beta_2 \mid \beta_1 = 0, y) .$$

■

Thus, $\tilde{\beta}_2$ may be viewed either as a restricted variant of $\breve{\beta}_2$, or as an important feature of the appropriate conditional posterior p.d.f. for β_2.

Note that unless $\breve{\beta}_1 = 0$ (which is most unlikely), $\tilde{\beta}_2 = \breve{\beta}_2$ iff $W_{12} = 0$. The latter condition is that $A_{12} = -X_1' X_2$. One way in which this condition would be satisfied, without requiring that the prior depend on the X data, is that $A_{12} = X_1' X_2 = 0$. Let us consider the interpretation of the condition $A_{12} = 0$. The conditional prior covariance between β_1 and β_2 is $\sigma^2 A^{12}$, where A^{ij} is the $(i,j)^{\text{th}}$ block of A^{-1}. So, this conditional prior covariance is

$$\sigma^2 A^{12} = -\sigma^2 (A_{11} - A_{12} A_{22}^{-1} A_{12}')^{-1} A_{12} A_{22}^{-1} ,$$

which is zero iff $A_{12}=0$, or $(A_{11}-A_{12}A_{22}^{-1}A_{12}')^{-1}=0$. The latter inverse matrix is simply A^{11}. Of course, one way to impose the (zero) restrictions on β_1 directly is by setting $\overline{\beta}_1=0$ and $A^{11}=0$ in the (conditional) prior for β, (3).

Further, the *marginal* prior p.d.f. for β is

$$p(\beta)=\int_0^{\infty} p(\beta\,|\,\sigma)\,p(\sigma)\,d\sigma$$
$$\propto[\nu_0+(\beta-\overline{\beta})'(A/c_0^2)(\beta-\overline{\beta})]^{-(k+\nu_0)/2}$$

which is multivariate Student-t, with covariance matrix $[\nu_0 c_0^2/(\nu_0-2)]A^{-1}$. So, the *marginal* prior covariance between β_1 and β_2 is

$$\sigma^2[\nu_0 c_0^2/(\nu_0-2)]A^{12}=-[\sigma^2\nu_0 c_0^2/(\nu_0-2)](A_{11}-A_{12}A_{22}^{-1}A_{12}')^{-1}A_{12}A_{22}^{-1}, \qquad (16)$$

which again is zero iff[5] $A_{12}=0$ or $A^{11}\equiv(A_{11}-A_{12}A_{22}^{-1}A_{12}')^{-1}=0$.

So, orthogonality of the two sub-sets of regressors *and* a zero (marginal or conditional) prior covariance between β_1 and β_2 are jointly sufficient for the restricted and unrestricted natural-conjugate Bayes estimators of β to coincide. Alternatively, this will be achieved if the two sub-sets of regressors are orthogonal *and* the conditional prior variance for β_1 is set to zero, which (together with assigning $\overline{\beta}_1=0$) is an obvious way of imposing the exact restrictions directly through the prior pd.f. In contrast, recall that in the case of a diffuse prior p.d.f. (or MLE) the corresponding condition for the restricted and unrestricted estimators to coincide is simply $X_1'X_2=0$.

4.2 *Asymptotic properties*

The weak consistency of $\breve{\beta}$ ensures that over-fitting the model still results in a Bayes estimator of β that is weakly consistent. Consider the situation when the model is under-fitted.

Theorem 7 The natural-conjugate Bayes estimator of β_2 in the under-fitted model is weakly consistent if $\Sigma_{12} = 0$, where $\Sigma = \underset{n\to\infty}{Limit}(n^{-1}X'X) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}$.

Proof

From the proof of Theorem 1,

$$\text{p lim}(\widetilde{\beta}) = \beta + \Sigma^{-1}R'(R\Sigma^{-1}R')^{-1}(r - R\beta) .$$

Substituting $r = 0$ and $R = (I, \quad 0)$,

$$\text{p lim}\begin{pmatrix} \widetilde{\beta}_1 \\ \widetilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} -\beta_1 \\ \Sigma_{22}^{-1}\Sigma_{12}'\beta_1 \end{pmatrix} = \begin{pmatrix} 0 \\ \beta_2 + \Sigma_{22}^{-1}\Sigma_{12}'\beta_1 \end{pmatrix}.$$

For the under-fitted model, $\beta_1 \neq 0$, so $\widetilde{\beta}_1$ is always inconsistent, and $\widetilde{\beta}_2$ is consistent iff $\Sigma_{12} \equiv \text{p lim}(n^{-1}X_1'X_2) = 0$. This last condition is precisely that which ensures the weak consistency of the Bayes estimator of β_2 based on a diffuse prior (the MLE of β_2) when the model is under-fitted. This is as expected, given that $\widetilde{\beta} \to b_R$ as $n \to \infty$. Similarly, the following result is not surprising:

Theorem 8 Over-fitting the model reduces the asymptotic efficiency of the natural-conjugate Bayes estimator of β_2 unless $\Sigma_{12} = 0$; and under-fitting the model increases the asymptotic precision of this estimator unless $\Sigma_{12} = 0$.

Proof

From the Proof of Theorem 2, with $r = 0$ and $R = (I, \quad 0)$,

$$a.c.m.[\sqrt{n}(\widetilde{\beta}_2 - \beta_2)] = \sigma^2\Sigma_{22}^{-1}$$

$$a.c.m.[\sqrt{n}(\breve{\beta}_2 - \beta_2)] = \sigma^2[\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}')^{-1}\Sigma_{12}\Sigma_{22}^{-1}] .$$

Clearly, these two a.c.m.'s are equal iff $\Sigma_{12} = 0$. Otherwise, the matrix difference, $\{a.c.m.[\sqrt{n}(\breve{\beta}_2 - \beta_2)] - a.c.m.[\sqrt{n}(\widetilde{\beta}_2 - \beta_2)]\}$ is positive-definite.

■

So, the same condition of a zero asymptotic covariance between the two sub-sets of regressors applies here as in Theorem 7, and as for the corresponding results based on a diffuse prior p.d.f.[6].

4.3 *Finite-sample properties*

First, consider the over-fitted model. That is the restrictions $R\beta = r$ are valid but are not imposed. In this case, from Theorem 3, $Bias(\widetilde{\beta}) = JBias(\breve{\beta})$ and $V(\widetilde{\beta}) = JV(\breve{\beta})J'$. Taking into the particular from of *R* and *r* here, we have:

Theorem 9 When $\beta_1 = 0$, $Bias(\widetilde{\beta}_2) = Bias(\breve{\beta}_2) + W_{22}^{-1}W_{12}'Bias(\breve{\beta}_1)$.

(The proof follows immediately from Theorem 3 on making the appropriate substitutions.)

Accordingly, over-fitting the model *does* affect the bias of the natural-conjugate Bayes estimator of β_2, in general. This is in contrast to the situation under a diffuse prior p.d.f. – then, over-fitting the model leaves the (zero) bias of the estimator of β_2 unchanged. Of course, if $Bias(\breve{\beta}_1) = 0$, then $Bias(\widetilde{\beta}_2) = Bias(\breve{\beta}_2)$, but this is very unlikely to arise as it requires that $\overline{\beta}_1 = \beta_1$, which would be a remarkably fortuitous assignment of prior information. Alternatively, $Bias(\widetilde{\beta}_2) = Bias(\breve{\beta}_2)$ if $W_{12} = 0$, but from the proof of Theorem 6 we know that in this extreme case $\widetilde{\beta}_2 = \breve{\beta}_2$, anyway.

From Theorem 3 and Corollary 2, we know that over-fitting the model reduces the precision of the natural-conjugate Bayes estimator, in general, and from Theorem 4 we know that over-fitting the model adversely affects the matrix MSE of this estimator, regardless of the sample values, the choice of prior parameters, of the values of the model's parameters.

Now consider the under-fitted model.

Theorem 10 When $\beta_1 \neq 0$, $Bias(\widetilde{\beta}_2) = Bias(\breve{\beta}_2) + W_{22}^{-1}W_{12}'[Bias(\breve{\beta}_1) + \beta_1]$.

(The proof follows immediately from Theorem 3.)

So, under-fitting the model affects the bias of the natural-conjugate Bayes estimator of β_2 and it affects it to a degree that is different from that in the case when the model is over-fitted. Again, from Theorem 3 and Corollary 2, we see that under-fitting the model generally improves the precision of the natural-conjugate Bayes estimator of β_2, and we also have the following result:

Theorem 11 Under-fitting the model leads to an "improvement" in matrix MSE iff

$$\lambda = \beta_1'[MMSE(\breve{\beta}_1)]^{-1}\beta_1 \leq 1 .$$

Proof

From Theorem 5, the required condition is

$$\lambda = (r - R\beta)'[R\,MMSE(\breve{\beta})R']^{-1}(r - R\beta) \leq 1 .$$

Substituting with $r = 0$ and $R = (I, 0)$, and letting

$$MMSE(\breve{\beta}) \equiv \breve{M} = \begin{pmatrix} \breve{M}_{11} & \breve{M}_{12} \\ \breve{M}_{12}' & \breve{M}_{22} \end{pmatrix},$$

we obtain the desired result immediately.

■

Note that $MMSE(\breve{\beta}_1) = \breve{M}_{11}$ depends on σ^2. The sample data, and *all* of the elements of $\overline{\beta}$, A and β. In particular, the value of λ *does* depend on the various factors relating to the wrongly omitted part of the model.

5. Relationship to "ridge" regression

It is well known that the "ridge" regression estimator (Hoerl and Kennard, 1970a, 1970b) can also be interpreted as a natural-conjugate Bayes estimator. For example, see Smith and Campbell (1980, p.78) and Loesgen (1990). Some related ideas are discussed by Swindel (1976), Plisken (1987) and Trenkler (1988).

The ridge estimator of β is:

$$\hat{\beta} = [C + X'X]^{-1}X'y , \tag{17}$$

where C is a known $(k \times k)$ positive-definite matrix of rank k. The "ordinary" ridge estimator arises when $C = cI$, for some chosen $c > 0$. Of course, this estimator is defined even if $rank(X) < k$. We see immediately that (17) is simply the Bayes estimator in (7), with $A = C$ and $\overline{\beta} = 0$. The ridge estimator of β "shrinks" the least squares estimator, b, towards the origin, while the Bayes estimator shrinks b towards the prior mean, $\overline{\beta}$.

There is a vast literature on ridge regression and its extensions which will not concern us here. An early survey of this literature is provided by Vinod (1978), and an excellent critique is given by Smith and Campbell (1980) and in the associated discussion. Interestingly, this literature makes very little mention of imposing explicit restrictions on the parameters when applying the ridge estimator. This possibility is hinted at by Gunst (1980), and Lee (1979), Ohtani (1985) and Uemukai (2010) consider some of the effects of under-specifying or over-specifying the regressor matrix when applying certain variants of ridge regression. Obviously, if we assign $A = C$ and $\overline{\beta} = 0$ in our Bayesian prior, then all of the results established in the present paper apply to ridge regression when valid or invalid exact linear restrictions of the form $R\beta = r$ are imposed when implementing this estimator of β. This makes all of our results directly applicable to the very large literature on ridge regression.

As an example, consider the condition $W_{12} = 0$, in sub-section 4.1, under which the restricted and unrestricted Bayes estimators, $\tilde{\beta}$ and $\breve{\beta}$, coincide. In the case of the "ordinary" ridge estimator for which $C = cI$, the condition $A_{12} = 0$ is always satisfied and so the equivalence of the restricted ridge estimator and $\hat{\beta}$ is assured simply if $X_1'X_2 = 0$, just as in the least squares case.

6. Conclusions

By working within quite a general framework for mixing exact restrictions with a more flexible prior p.d.f., we have shown how the sampling properties of a particular Bayesian regression estimator are affected when the model is mis-specified in terms of the regressor matrix. We have examined how the bias and matrix mean squared error of the estimator are affected by either over-specifying or under-fitting the model. A simple condition has been found under which under-fitting will improve the matrix mean squared error of the Bayesian estimator based on a natural-conjugate prior for the parameters of the model. Conversely, we have shown that over-fitting the model will never improve the matrix mean squared error of this estimator.

The approach that we have suggested for mixing exact and uncertain prior information about a regression model's parameters has a range of other applications beyond the analysis of under-specified or over-specified regressions. In addition, our results have implications beyond the natural-conjugate Bayes estimator. One example of this that we have provided is the case of the ridge regression estimator. Others include the "mixed" regression estimator of Theil and Goldberger (1961), for which our framework can be used to generalize the "weak" mean squared error results of Kadiyala (1986) to the matrix mean squared error case; the "prior integrated mixed" regression estimator proposed by Mittelhammer and Conway (1988); the "weighted mixed" estimator of Schaffrin and Toutenburg (1990), Toutenburg *et al.* (1998) and Heumann and Shalabh (2008); and variants of the Liu estimator (Liu, 1993; Hu *et al*., 2009; Zuo, 2009).

References

Berk, R. H. (1970). Consistency a posteriori. *Annals of Mathematical Statistics*, 41, 894-906.

Bester, A. and C. Hansen (2005). Bias reduction for Bayesian and frequentist estimators. Working paper, Graduate School of Business, University of Chicago.

Bolstad, W. M. (2004). *Introduction to Bayesian Statistics*. Hokoben, NJ: Wiley

Davies, R. B. (1980). The distribution of a linear combination of chi-squared random variables. *Applied Statistics*, 29, 323-333.

Giles, D. E. A. and A. C. Rayner (1979). The mean squared errors of the maximum likelihood and natural-conjugate Bayes regression estimators. *Journal of Econometrics*, 11, 319-334.

Giles, J. A. and D. E. A. Giles (1993). Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys*, 7, 145-197.

Heumann, C. and Shalabh (2008). Weighted mixed regression estimation under biased stochastic restrictions. In Shabh and C. Heumann (eds.), *Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg*, 401-415.

Hoerl, A. E. and R. W. Kennard (1970a). Ridge regression: biased estimation of nonorthogonal problems. *Technometrics*, 12, 55-67.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: applications to non-orthogonal problems. *Technometrics*, 12, 69-82.

Howlader, H. A. and G. Weiss (1988). On the estimation of the Cauchy parameters. *Sankhyā*: *The Indian Journal of Statistics*, *B*, 50, 350-361.

Hu Y., X. Chang, and D. Liu (2009). Improvement of the Liu estimator in weighted mixed regression. *Communications in Statistics – Theory and Methods*, 38, 285-292.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal random variables. *Biometrika*, 48, 418-426.

Johnston, J. (1972). *Econometric Methods*. New York: McGraw-Hill.

Kadiyala, K. (1986). Mixed regression estimator under misspecification. *Economics Letters*, 21, 27-30.

Lee, B. M. S. (1979). On the use of 'improved' estimators in econometrics. Unpublished Ph.D. dissertation, Department of Statistics, Australian National University.

Liu, K. (1993) "A new class of biased estimate in linear regression,"

Communications in Statistics-Theory and Methods, 22, 393-

402.

Loesgen K-H. (1990). A generalization and Bayesian interpretation of ridge-type estimators with good prior means. *Statistical Papers*, 31, 147-154.

Mittelhammer, R. C. and R. K. Conway (1988). Applying mixed estimation in econometric research. *Journal of the American Journal of Agricultural Economics*, 70, 859-856.

Ohtani, K. (1985). Small sample properties of the generalized ridge regression predictor under specification error. *Economic Studies Quarterly*, 36, 53–60.

Pliskin, J. L. (1987). A ridge-type estimator and good prior means. *Communications in Statistics – Theory and Methods*, 16, 3429-3437.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, New York: Wiley.

Rao, C. R. and N. Shinozaki (1978). Precision of individual estimators in simultaneous estimation of parameters. *Biometrika*, 65, 23-30.

Rao, P. S. R. S. (2009). Performance of the empirical Bayes estimator for fixed parameters. Research report, Program in Statistics, University of Rochester.

Schaffrin, B, and H. Toutenberg (1990). Weighted mixed regression. *Zeitschrift für Angewandte Mathematik und Mechanik*, 70, 735-738.

Swindel, B. F. (1976). Good ridge estimators based on prior information. *Communications in Statistics – Theory and Methods*, 5, 1065-1075.

Theil, H. and A. S. Goldberger (1961). On pure and mixed statistical estimation in economics. *International Economic Review,* 2**,** 65–78.

Toro-Vizcarrondo, C. and T. D. Wallace (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63, 558-572.

Toutenburg, H., V. K. Srivastava and B. Schaffrin (1998). Efficiency properties of weighted mixed regression estimation. Sonderforschungsbereich 386, Paper 122, Institut für Statistik, Ludwig-Maximillians Universitāt, München.

Trenkler, G. (1988). Some remarks on a ridge-type estimator and good prior means. *Communications in Statistics – Theory and Methods*, 17, 4251-4256.

Uemukai, R. (2010). Small sample properties of a ridge regression estimator when there exist omitted variables. *Statistical Papers*, in press.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.

Zuo, W. (2009). A new stochastic restricted Liu estimator in weighted mixed regression. Second International Symposium on Computational Intelligence and Design, IEEE Computer Society.

Footnotes

* I am very grateful to the late Arnold Zellner for his on-going encouragement over a period of nearly forty years, and for his helpful comments on a much earlier version of this work.

1. The derivation of $\tilde{\beta}$ is directly analogous to that for the restricted MLE (restricted least squares estimator (*e.g.*, Johnston, 1972, pp. 157-158).

2. Note that $\text{p}\lim(J) = J^*$, as in Theorem 2.

3. We are assuming that the model is correctly specified in all other respects.

4. See Zellner (1971, pp. 386-388), and note that in earlier printings equation (B.48) the terms should be subtracted, not added.

5. We have also used the fact that $\nu_0, c_0 > 0$ for $p(\sigma)$ to be proper, and that $\nu_0 > 2$ is required for (16) to be defined.

6. For a more general discussion of the consequences of model mis-specification for the asymptotic properties of Bayes estimators, see Berk (1970).