H.G.B. Alexander Research Foundation
Graduate School of Business
University of Chicago

Bayesian Shrinkage Estimates and Forecasts of Individual and Total or Aggregate Outcomes

by

Arnold Zellner

University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637

Paper 0204 **August 2002**

Bayesian Shrinkage Estimates and Forecasts of Individual and Total or Aggregate Outcomes[1]

Arnold Zellner
University of Chicago

Abstract. Bayesian shrinkage à la Stein and others can improve estimation of individual parameters and forecasts of individual future outcomes. In this paper the issue of the impact of shrinkage on the estimation of sums or totals of individual parameters and of individual outcomes is analyzed. Quadratic and "balanced" loss functions will be employed. The latter are linear combination of "goodness of fit" and "precision of estimation" loss functions. Several examples will be analyzed in detail to illustrate general principles.

I. Introduction

In the paper, Zellner and Chen (2001), various shrinkage techniques were employed to forecast the annual rates of growth of real sales of eleven sectors of the U.S. economy year by year, 1970-1997 and utilized to derive annual forecasts of rates of growth of total U.S. real GDP, 1970-1997. On presenting this work at a meeting in Calcutta, J.K. Ghosh (2001) raised the question about how use of shrinkage techniques affects the properties of forecasts of totals, a problem that I thought had been analyzed in the literature since estimation and prediction of totals are "standard" problems. To my surprise, I found limited work on this problem in the literature; see, e.g. Cohen (1968), Brown (1990). Thus I decided to provide a brief review of some old and new approaches to shrinkage in estimating a vector of parameters and the effects of shrinkage on the estimation of the sum of the parameters within the framework originally employed by Stein (1956,1960,1962) and used by others, e.g., Lindley (1962), Lindley and Smith (1972), Sclove (1971, 1978), Zellner and Vandaele (1974), Berger (1985), Leonard and

---

[1] Research financed in part by the National Science Foundation, the CDC Investment Management Corp. and the H.G.B. Alexander Endowment Fund, Graduate School of Business, University of Chicago. E-mail:arnold.zellner@gsb.uchicago.edu http://gsbwww.uchicago.edu/fac/arnold.zellner/more
Thanks to the Editor, Associate Editor and referees for many helpful comments on the original draft of this paper.

Hsu (2001), Judge, et al. (1985), and Mittelhammer, et al. (2000), and then to go on to consider shrinkage within a forecasting context. That is, effects of the use of shrinkage forecasting techniques on forecasts of totals of individual outcomes, e.g., the sum of forecasts of sector annual real sales, as in Zellner and Chen (2001) and Zellner (2001), will be analyzed. It will be seen that when the additional prior information of the kind employed by Stein and others is introduced, shrinkage techniques yield improved estimates and predictions of not only individual outcomes but also of total outcomes. The result that shrinkage can affect the precision of forecasts of totals has implications for properties of "consensus" forecasts of totals derived from surveys of individuals' anticipated values, perhaps formed in accord with Muth's (1961) rational expectations hypothesis. And indeed in economic theory, it appears that effects of shrinkage in the formation of individual rational expectations on rational expectations of totals has not been analyzed.

In Section II, after stating the problem and indicating general solutions, a review of Stein's and others' approaches to Stein's original problem of estimating n means will be provided along with associated, optimal estimates of the sum of the n mean parameters. These problems will be analyzed using several different Bayesian and non-Bayesian approaches, including least squares, maximum likelihood, traditional Bayes, Bayesian method of moments, etc. Also, analysis of the problem of forecasting a new vector of observations and the sum of the elements of the new vector of observations will be presented. In Section III, the analyses are extended to apply to a set of regression or time series forecasting equations. Here shrinkage forecasts are presented for individual future outcomes and the sum of the future outcomes. In Section IV, some sampling properties of various shrinkage estimates and point predictions for totals will be presented. As will be seen, for the loss functions employed, there is evidence that IT PAYS TO SHRINK in estimating totals in many circumstances, but not all.

II. Stein's Means Problem

Stein analyzed the following n means estimation problem:

$$y_i = \theta_i + u_i \qquad i = 1, 2, ..., n \tag{1a}$$

or in vector notation,

$$y = \theta + u \tag{1b}$$

where $y' = \left(y_1, y_2, ..., y_n\right)$ is a vector of observations, $\theta' = \left(\theta_1, \theta_2, ..., \theta_n\right)$ is a vector of means and $u' = (u_1, u_2, ..., u_n)$ is a vector of error terms. Stein concentrated his attention on estimating θ using a quadratic loss function, namely, $(\hat{\theta} - \theta)'(\hat{\theta} - \theta),$ where $\hat{\theta}$ is an estimate of θ. Herein, we shall also use a balanced loss function, see e.g. Zellner (1994) and Dey, Ghosh and Strawderman (1998), that reflects both precision of estimation and goodness of fit, namely $L\left(\hat{\theta}, \theta\right) = w\left(y - \hat{\theta}\right)'\left(y - \hat{\theta}\right) + (1 - w)\left(\hat{\theta} - \theta\right)'\left(\hat{\theta} - \theta\right),$ where w is a given weight, $0 < w < 1,$ and again $\hat{\theta}$ is some estimate. While interest often centers on the estimation of the individual θ_i 's, many times there is also great interest in estimating their total, denoted by $T = \sum_{i-1}^{n} \theta_i,$ say relative to squared error loss, $L\left(\hat{T}, T\right) = \left(\hat{T} - T\right)^2$ or balanced loss, $L_b = w\left(\iota' y - \hat{T}\right)^2 + (1 - w)\left(\hat{T} - T\right)^2,$ where $\iota' = (1,1,...,1).$

As is well known, in general relative to squared error loss, the point estimate that minimizes the posterior mean of the loss function is the posterior mean of the parameter being estimated. Thus, in estimating θ vis à vis squared error loss, the optimal estimate is $E\theta \big| D,$ the posterior mean, where D denotes given data and prior information. In estimating the total or sum of the means, T relative to quadratic loss, the optimal estimate is $ET \big| D = \iota' E\theta \big| D$, the sum of the posterior means of the elements of θ. The issue is how this last estimate depends, if at all, on shrinkage assumptions. Similarly, with respect to a balanced loss function, the optimal estimate of the total is $\hat{T} = w\iota' y + (1 - w) ET \big| D$ and the issue is how this estimate is affected by various shrinkage assumptions.

Similarly in terms of a future vector of observations, $y_f = \theta + u_f$, where u_f is a vector of future errors, given a quadratic loss function $L = \left(\hat{y}_f - y_f\right)'\left(\hat{y}_f - y_f\right),$ the optimal point prediction is the predictive mean $Ey_f \big| D = E\theta \big| D$, given that $Eu_f \big| D = 0.$

Similarly, for prediction of a future total, $\iota' y_f$ relative to squared error loss, the optimal point prediction is the predictive mean, $\iota' E y_f \left| D. \right.$ Again, the issue is whether or not this optimal predictive mean reflects shrinkage assumptions. By analyzing particular cases, this issue can be addressed.

2.1 Estimation of n Means and Their Total

Stein (1956, 1962) assumed that the u_i 's in (2.1) have been independently drawn from a normal distribution with mean = 0 and variance = 1. Note that variance is set equal to 1 since, without further information, it is not possible to estimate n means and a variance, n + 1 parameters with just n observations. With this assumption about the error terms, the implied density for the n y's is $f\left(y|\theta\right) \propto \exp\{-\left(y-\theta\right)'\left(y-\theta\right)/2\}$ and the maximum likelihood and least squares estimate of θ is just y, the vector of observations. With an improper uniform prior on the elements of θ, the posterior mean and modal value for θ is just y. Also, with these estimation approaches, the estimate of the total $T = \iota'\theta = \sum_{i=1}^{n} \theta_i$ is $\hat{T} = \iota' y = \sum_{i=1}^{n} y_i$.

What has not been noted about the above estimate of θ, namely y, is that it leads to the following estimate of the realized error vector, $\hat{u} = y - \hat{\theta} = y - y = 0,$ which is quite strange until one recognizes that the perfect fit results from "overfitting", namely using n parameters with just n observations. Is it possible to get an estimate of T without this implication? The answer is yes. For example, if we employ the Bayesian method of moments approach, see, e.g., Zellner (1997a, b) and Zellner and Tobias (2001), we sum both sides of (1a) to obtain, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \theta_i + n\sum_{i=1}^{n} u_i / n,$ where the y_i's are given observations and the u_i 's are given, realized error terms. On taking the subjective expectation of both sides and using the assumption $E\sum u_i / n = 0,$ we have $ET = E\sum_{i=1}^{n} \theta_i = \sum_{i=1}^{n} y_i.$ Note further that for a single given observation, $y_i = E\theta_i + Eu_i$ and unless there is a good reason for assuming that the subjective expectation of the i'th realized error term is equal to zero, we get the reasonable result that $E\theta_i$ is not necessarily equal to y_i.

2.2 Traditional Bayesian Approaches

As explained in Stein (1962), Zellner and Vandaele (1974) and Leonard and Hsu (2001), the following prior density for the θ_i's was employed by Stein to derive a shrinkage estimator: the θ_i's are assumed to be $iid\ N(0,\tau^2)$ that is $f(\theta) \propto \tau^{-n} \exp\{-\theta'\theta / 2\tau^2\}$. Then, using this prior and the above normal likelihood function, the posterior density for θ is:

$$p(\theta|D) \propto \text{ prior } X \text{ likelihood} \tag{2}$$

$$\propto \exp\{-\left[\theta'\theta/\tau^2 + (y-\theta)'(y-\theta)\right]/2\}$$

with mean,

$$E\theta|D = [1-1/(1+\tau^2)]y \tag{3}$$

and expected total

$$ET|D = E\iota'\theta|D = [1-1/(1+\tau^2)]\iota' y \tag{4}$$

Thus using Stein's normal prior for the means, we get the shrinkage estimates in (3) and (4) for the vector of means and the total of the means, respectively that are well known to be optimal relative to a quadratic loss function. Stein obtained an estimate of τ^2 from $\sigma_y{}^2 = \tau^2 + 1$, namely, $\hat{\tau}^2 = \hat{\sigma}_y{}^2 - 1 = y'y/n - 1$ that he inserted in (3) to obtain an approximate posterior mean, $\hat{\theta} = (1 - n/y'y)\,y$. Of course, it is possible to use a prior for

τ and to compute the exact posterior mean of θ. Further, it should be noted that an estimate in the form of (3) can be obtained in a "minimum mean square" approach by considering the class of estimators, $\tilde{\theta} = qy$ relative to quadratic loss. As Zellner and Vandaele show, the optimal value of q, namely $q^* = 1 - n/(\theta'\theta + n)$ can be approximated by $1 - n/y'y$ to yield precisely Stein's approximate estimate given above. Further, the associated optimal estimate for the total is $\hat{T} = (1 - n/y'y)\iota'y$ that is different from $\iota'y = \sum_i^n y_i$.

As, Good and Wallace have pointed out, if the problem of n means is considered in polar coordinates and one has a uniform prior on the radius vector and the n-1 angles, the implied prior for the means in rectangular coordinates, that is for the n means, is close to that which Stein indicates produces his shrinkage estimate. For details, see Stein (1962, p. 201) and Zellner and Vandaele (1974, p. 631).

Lindley (1962) noted in discussion of Stein's paper that Stein had assigned a prior mean of zero for the θ_i's in his normal prior density given above and suggested broadening the model to $y = \theta + u$, as above, and

$$\theta_i = \mu + e_i \qquad i = 1, 2, .., n \tag{5a}$$

Here Lindley has removed the zero mean prior assumption by introducing the mean parameter μ for which he assumes a uniform prior, and showed that the posterior mean is approximately:

$$E\theta|D = \bar{y}\iota + [1 - (n-3)/(y - \bar{y}\iota)'(y - \bar{y}\iota)](y - \bar{y}\iota) \tag{5b}$$

Note that with this estimate, the associated estimate of $T = \iota'\theta$ is $\hat{T} = \iota'y = \sum_i^n y_i$. Thus in this case, the total is estimated without shrinkage. Similarly, in the mean square error approach, mentioned above, if a broader class of estimators, namely, $\hat{\theta} = q_0\iota + q_0 y$ is considered, the approximately optimal estimator is

$\hat{\theta} = \bar{y}\iota + \left[1 - a / (y - \bar{y}\iota)'(y - \bar{y}\iota)\right](y - \bar{y}\iota)$ and the associated optimal estimator of the total is $\hat{T} = \iota' y$ for any value of the parameter a. If a = n – 3, the minimal MSE estimate is exactly equal to Lindley's estimate. See Zellner and Vandaele (1974, p. 634) for details. Thus here, as with the introduction of the extra parameter μ and its uniform prior, the estimate of the total does not involve shrinkage.

2.3 The Bayesian Method of Moments (BMOM) Approach

In the Bayesian method of moments approach, see, e.g., Mittelhammer, et al (2000), Zellner (1997a, b) and Zellner and Tobias (2001) no sampling assumptions are made about the given observation vector y or the error vector u. The parameter vector θ and the realized error vector u are treated as unknown parameters. Further, extra information regarding the parameters is introduced via a conceptual realized sample, e.g., $z = \theta + v,$ with z an nx1 vector of conceptual observations and v an nx1 vector of realized errors. On combining the observed and conceptual samples, we have: $y + z = 2\theta + u + v.$ Then, with $\iota' = (1,1,...,1), \iota' y + \iota' z = 2\iota'\theta + n(\bar{u} + \bar{v}),$ where $\bar{u} = \iota' u / n$ and $\bar{v} = \iota' v / n$ both assumed to have subjective means = 0, we have $E\iota'\theta \middle| D = (1/2)(\iota' y + \iota' z) = \left[1 - (1/2)(1 - \bar{z} / \bar{y})\right]\iota' y,$ where $\bar{z} = \iota' z / n$ and $\bar{y} = \iota' y / n$, is a shrinkage estimate of the total, $T = \iota'\theta$ when $\bar{z}$ is assigned a value to reflect our initial information, just as Stein assigned a value to the prior mean of the θ_i's. Thus, the BMOM approach has provided a shrinkage estimate of the total that is optimal relative to quadratic loss without making normal sampling assumptions and without a prior density for the parameters. With additional assumptions, higher order moments can be derived as in usual BMOM analyses and maxent post data densities for the parameters can be derived.

2.4 Model Uncertainty and Averaging Estimates

Above, with respect to (5a), Stein assumed $H_{o:}\ \mu = 0$ while Lindley assumed that μ is uniformly distributed. Given that posterior probabilities associated with these hypotheses are available, say p and $1-p$, the posterior odds $K = p/(1-p)$ and using a quadratic loss function, $L = (\hat{\theta}-\theta)'(\hat{\theta}-\theta)$, posterior expected loss

= $EL = pEL|H_o + (1-p)EL|H_a$ and the optimal point estimate is, with $\hat{\theta}_o = E(\theta|D,H_o)$ and $\hat{\theta}_a = E(\theta|D,H_a)$, the posterior means under H_o and H_a, respectively:

$$\hat{\theta} = p\hat{\theta}_o + (1-p)\hat{\theta}_a \tag{6}$$

or

$$\hat{\theta} = \hat{\theta}_o + [1-K/(1+K)](\hat{\theta}_a - \hat{\theta}_o)$$

as given in Zellner and Vandaele (1974). Clearly, if $K=0$, $\hat{\theta} = \hat{\theta}_a$, while if $K \to \infty$, $\hat{\theta} \to \hat{\theta}_o$. Further, the associated optimal estimate of the total relative to quadratic loss is $ET|D = E\iota'\theta|D = \iota'\hat{\theta}_o + [1-K/(1+K)](\iota'\hat{\theta}_a - \iota'\hat{\theta}_o)$ which in general is not a simple sum of the observations.

2.5 Balanced Loss Functions

Above, following Stein, we have employed a quadratic, precision of estimation loss function. As pointed out above, such a loss function does not reflect goodness of fit. For example, a quadratic "goodness of fit" loss function that is widely used in least squares approaches, etc. is $L_g = (\hat{\theta}-y)'(\hat{\theta}-y)$ for which the optimal estimate is $\hat{\theta}^* = y$. A balanced loss function, see e.g. Zellner (1994), Tsurumi (1991), and Dey, Ghosh and Strawderman (1998) is defined as follows for the present problem.

$$L(\hat{\theta},\theta) = w(y-\hat{\theta})'(y-\hat{\theta}) + (1-w)(\hat{\theta}-\theta)'(\hat{\theta}-\theta) \tag{7}$$

with $0 < w < 1$, where $\hat{\theta}$ is some estimate of θ. The value of $\hat{\theta}$ that minimizes posterior expected loss is given by:

$$\hat{\theta}^* = wy + (1-w)E\theta|D \tag{8}$$

where $E\theta|D$ is the posterior mean of θ For the total, $T = \iota'\theta$, a balanced loss function is: $L = w(\iota'y - \hat{T}) + (1-w)(\hat{T} - T)^2$ and the optimal point estimate of T is:

$$\hat{T}^* = w\iota'y + (1-w)ET|D = w\iota'y + (1-w)\iota'E\theta|D \tag{9}$$

Thus with the use of a balanced loss function, the optimal estimate of the total, T, will reflect the shrinkage estimation of the individual means in cases in which $E(\theta|D) \neq y.$

In summary, there are a number of circumstances in which shrinkage estimation for the individual means leads to a shrinkage estimate for the total of the means, e.g. Stein's assumption that $\mu = o$, use of an informative prior for μ, averaging over hypotheses, use of balanced loss functions, and also use of asymmetric loss functions, see e.g. Varian (1975), Berry et al. (1996) and Zellner (1986). Similarly, in the minimal mean square error approach, considering a class of estimators that is given by $\hat{\theta} = qy$ and evaluating the optimal value of q results in a shrinkage estimate of θ and of the total, $T = \iota'\theta.$ And since the optimal point prediction of the total $T = \iota'y_f$ relative to a quadratic predictive loss function is just $ET|D = \iota'Ey_f|D = E\iota'\theta|D,$ when there is shrinkage in estimating the sum of the means, there will be shrinkage in predicting the sum of the future values.

III. Shrinkage Estimation and Prediction for Multivariate Regression and Time Series Models

Here we assume that the observation vectors, $y_1, y_2, ..., y_m$ have been generated by the seemingly unrelated regression (SUR) model:

$$y_\alpha = X_\alpha \beta_\alpha + u_\alpha \qquad \alpha = 1,2,..,m \tag{10a}$$

$$nx1 \quad nxk \quad kx1 \quad nx1$$

or

$$y = Z\beta + u \tag{10b}$$

where $y' = \left(y_1', y_2', ..., y_m'\right), \beta' = \left(\beta_1', \beta_2', ..., \beta_m'\right), u' = (u_1', u_2', ...u_m')$ and Z is a block diagonal matrix with the X 's on the diagonal.

How shrinkage estimation of the β 's affects prediction of the future observation vector $w_f = Z_f \beta + u_f$ and the total of the elements of w_f, namely $T_f = \iota' w_f$ using shrinkage and non-shrinkage techniques will be considered below. Also, in some cases, it is of interest to consider estimation of the total of the elements of β, i.e., $J'\beta$. where $J' = (I, I, ..., I)$. Further, there is the problem of estimating the sum, $\sum_i X_i \beta_i$ and how shrinkage estimation of β affects the estimation of this sum.

An interesting special case of (10) is that of "identical units," say firms assumed to utilize the same production function or consumers assumed to have the same demand functions, namely, $\beta_1 = \beta_2 = \beta_m = \beta$, a case analyzed in de Alba and Zellner (1991) with respect to aggregation and disaggregation issues. In this case, (10) can be written as:

$$y = W\beta + u \tag{11}$$

where $y' = \left(y_1', y_2', ..., y_m'\right), W' = \left(X_1', X_2', ..., X_m'\right)$ and

$u' = \left(u_1', u_2', ..., u_m'\right)$.

A $1xm$ vector of outcomes, $y_f' = \left(y_{if}, y_{2f}, ..., y_{mf}\right)$ is given by

$$y_f = W_f \beta = u_f \tag{12}$$

with W_f given. Here the issue is whether shrinkage estimation of β will result in improved point predictions of y_f and of the total of $T_f = \iota' y_f$ relative to, say, diffuse prior or least squares or general least squares analyses of (11) and (12).

As regards sampling assumptions for the error terms in the above relations, it will be assumed that $Eu = 0$ and $Euu' = I$ or $Euu' = \sum,$ a positive definite symmetric matix. In the latter case, the systems in (10) and (11) can be transformed in the usual way, $Hy = HZ\beta + Hu,$ where H is such that $EHuu'H = H'\sum H = I.$ A similar transformation can be applied to (11), namely, $Hy = HWb + Hu.$ See Percy (1992, 1996) and Zellner and Vandaele (1974) for Bayesian analysis of the SUR model using such transformations.

Further note $Z'y = Z'Z\beta + Z'u$ and if $Z'Z = I,$ then the regression problem reduces to a "means problem," as pointed out by Sclove (1968). Also, he points out that the regression problem can be reduced to a means problem by reparameterization, namely, $y = ZPP'\beta + u$ and $P'Z'y = P'Z'ZP\theta + P'Z'u = \theta + e$ where $\theta = P'\beta$ and P is an orthogonal matrix such that $P'Z'ZP = I.$ While these transformations of the original problem are very interesting, below only the original parameterizations will be considered.

As above, we shall use quadratic and balanced loss functions, namely:

(13) $$L\left(\hat{\beta}, \beta\right) = \left(\hat{\beta} - \beta\right)' Q \left(\hat{\beta} - \beta\right) \qquad Q = \text{ given } pds \text{ matrix} \tag{13}$$

and

$$L\left(\hat{\beta}, \beta\right) = w\left(y - Z\hat{\beta}\right)'\left(y - Z\hat{\beta}\right) + (1-w)\left(\hat{\beta} - \beta\right)' Q\left(\hat{\beta} - \beta\right) \tag{14}$$

$0 < w < 1$

The quadratic predictive loss function that is employed is:

$$L\left(\hat{y}_f, y_f\right) = \left(\hat{y}_f - y_f\right)' Q_f \left(\hat{y}_f - y_f\right) \tag{15}$$

with Q_f a given pds matrix. The totals that shall be considered are given by $T = J'\beta,$, where $J' = (I, I, ..., I)$ and $\beta' = \left(\beta_1', \beta_2', ..., \beta_m'\right)$ and $T_f = \iota' y_f$, assuming that the corresponding elements of the β 's and of y_f are measured in identical units. The associated loss functions for T and T_f are:

$$L\left(\hat{T}, T\right) = \left(\hat{T} - T\right)' Q\left(\hat{T} - T\right) \tag{16}$$

and

$$L\left(\hat{T}_f, T_f\right) = (\hat{T}_f - T_f)^2 \tag{17}$$

As above, optimal point estimates and point predictions relative to quadratic loss functions are posterior means and means of predictive densities, that is $E\beta\,|D$ and $ET_f\,|D,$ where D denotes given data and prior information. With respect to the balanced loss function in (14), the optimal point estimate is

$$\beta = w\hat{\beta} + \left(1 - w\right) E\beta\,|D \tag{18}$$

where $\hat{\beta} = \left(Z'Z\right)^{-1} Z'y$ and $E\beta\,|D$ is the posterior mean of β given the model in (10), data and prior information. For the model in (11), the optimal estimate is as given in (18) with $\hat{\beta} = (W'W)^{-1} W'y$ and posterior mean $E\beta\,|D$ given the model in (11).

Given these general results, the results provided by alternative approaches to estimating the totals will be presented.

3.1 Least Squares and Maximum Likelihood Approaches

The least squares (LS) and maximum likelihood (ML) approaches yield the following estimates and predictions. For the SUR model in (10) and the restricted model in (11), using appropriate assumptions regarding the error term covariance matrix,

$$\hat{\beta} = (Z'Z)^{-1} Z'y \text{ and } \hat{T} = J'\hat{\beta} \text{ for model in (10),} \tag{19a}$$

with

$$\hat{\beta} = (W'W)^{-1} W'y \text{ and } \hat{T} = \iota'\hat{\beta} \text{ for model in (11)}$$

and

$$\hat{y}_f = Z_f \hat{\beta} \text{ and } \hat{T}_f = \iota' y_f = \iota' Z_f \hat{\beta} \quad \text{for model in (10),} \tag{19b}$$

with

$$y_f = W_f \hat{\beta} \text{ and } \hat{T}_f = \iota' \hat{y}_f = \iota' W_f \hat{\beta} \text{ for model in (11).}$$

These estimates and predictions are also posterior and predictive means produced by a "diffuse" prior Bayesian approach based on a normal likelihood function and by the BMOM approach. For example, assuming that the given data satisfy, $y = Z\beta + u$, where u is a vector of realized error terms, we have $\hat{\beta} = (Z'Z)^{-1} Z'y = \beta + (Z'Z)^{-1} Z'u$. Under the usual BMOM assumption, $Z'Eu|D = 0$, we have $E\beta|D = (Z'Z)^{-1} Z'y$ as the post data mean for β given the data without using a likelihood function, prior density and Bayes' theorem. For further BMOM results for multivariate regression and simultaneous equations estimation, see van der Merwe and Viljoen (1998) and Zellner (1998) that also includes a review of Monte Carlo experimental evidence regarding the sampling properties of BMOM and alternative estimates.

3.2 Estimation and Prediction with Added Prior Information

Just as Stein added an informative prior for his n-means, as explained above, it is possible to introduce extra information regarding the regression coefficients. See also, Leonard and Hsu (2001, p. 271) for a closely related example involving m regression relations. That is, a straightforward generalization of Stein's prior will be introduced for the parameter vector in the SUR model as in Zellner and Vandaele (1974, p. 643):

$$f(\beta) \sim N(b, \Omega^{-1}) \tag{20}$$

Using this prior and a normal likelihood function for the SUR model, the posterior mean is given by:

$$E\left(\beta \middle| D\right) = \left(Z'Z + \Omega\right)^{-1}\left(Z'Z\hat{\beta} + \Omega b\right) \quad (21)$$

$$= b + \left(Z'Z + \Omega\right)^{-1} Z'Z\left(\hat{\beta} - b\right)$$

where $\hat{\beta} = \left(Z'Z\right)^{-1} Z'y$ is the LS or GLS or ML estimate. Thus there is shrinkage toward the prior mean b. Further in estimation of the total, with $J' = \left[I, I...I\right]$ and $J'\beta = \beta_1 + \beta_2 + ... + \beta_m$, we have

$$J'E\beta \Big| D = J'b + J'\left(Z'Z + \Omega\right)^{-1} Z'Z\left(\hat{\beta} - b\right) \quad (22)$$

and there will in general be shrinkage toward $J'b$, the prior mean for the sum of the coefficient vectors.

Similarly, with respect to prediction, relative to quadratic predictive loss, the optimal point prediction is given by:

$$Ey_f \Big| D = Z_f E\beta \Big| D \quad (23)$$

with $E\beta | D$ given in (21). Then the predictive mean of the total is given by:

$$ET_f \Big| D = E\iota' y_f \Big| D = \iota' Z_f E\beta \Big| D \quad (24)$$

Thus it is seen that in predicting the total in this case, there will be shrinkage toward the prior mean prediction of the total, $\iota' Z_f b$.

3.3 BMOM Analysis with a Conceptual Sample

In this section, the regression problems in Section 3.2 will be analyzed using the Bayesian method of moments (BMOM) approach. For earlier BMOM analyses of univariate and multivariate regression problems see Green and Strawderman (1996), van der Merwe et al. (1998, 2001), Zellner (1997a, b, 1998) and Zellner and Tobias (2001).

The given observed y and realized error vector u are assumed to satisfy the relation, $y = Z\beta + u$, as in (10c). Further, our prior information will be represented by a conceptual sample of data, y_c, assumed to satisfy, $y_c = Z_c\beta + v$, where Z_c is a given matrix, β is the regression coefficient vector and v is a vector of unobserved realized errors. Then we can write,

$$\begin{pmatrix} y \\ y_c \end{pmatrix} = \begin{pmatrix} Z \\ Z_c \end{pmatrix} \beta + \begin{pmatrix} u \\ v \end{pmatrix} \tag{25}$$

On multiplying both sides of (23) by $(Z' \vdots Z_c')$, , and taking post-data expectation of both sides, we have

$$Z'y + Z_c' y_c = \left(Z'Z + Z_c' Z_c \right) E\beta \big| D \tag{26}$$

$$+Z'Eu \Big| D + Z_c' Ev \big| D$$

where D denotes the given sample, conceptual data and prior information. The prior information involves the assumptions that the data satisfy (25), that there are no left out variables or outliers or errors in the variables. Under these assumptions, it is assumed that $Z'Eu \big| D = 0$ and $Z_c' Ev \big| D = 0,$ and with these assumptions, (26) yields:

$$E\beta \big| D = (Z'Z + Z_c'Z_c)^{-1}(Z'y + Z_c'y_c) \tag{27}$$

$$= \bar{\beta}_c + \left(Z'Z + Z_c'Z_c \right)^{-1} Z'Z \left(\hat{\beta} - \bar{\beta}_c \right)$$

where $\hat{\beta} = \left(Z'Z \right)^{-1} Z'y$ and $\bar{\beta}_c = \left(Z_c'Z_c \right)^{-1} Z_c' y_c$.

On assigning values for $\bar{\beta}_c$ and $Z_c'Z_c$ (25) is an operational, optimal estimate relative to quadratic loss that involves shrinkage of $\hat{\beta}_c$ toward $\bar{\beta}_c$, the prior mean vector. Further, pursuing a g-prior approach, see Zellner (1994), we can take $Z_c'Z_c = gZ'Z,$ with g>0. Under this specification, (27) becomes:

$$E\beta\big|D = \bar{\beta}_c + \left[1/(1+g)\right]\left(\hat{\beta} - \bar{\beta}_c\right) \tag{28}$$

a shrinkage estimate with again shrinkage toward the prior mean.

As regards prediction of the total, $T_f = \iota' y_f$, where $y_f = Z_f\beta + u_f$, for given Z_f and assuming that $Eu_f\,\big|D = 0$, we have

$$ET_f\,\big|D = \iota' Z_f E\beta\big|D \tag{29}$$

and when either (27) or (28) is substituted in (29), the optimal point predictions for T_f are obtained and in both cases incorporate shrinkage effects.

A similar BMOM analysis using a conceptual sample can be used in connection with the model in (11) and is left as an exercise for the reader. Also, above only first moment assumptions have been utilized. An expanded BMOM analysis permitting realized error terms to be correlated has been performed using a Gibbs Sampler approach to compute post data means for coefficients in the SUR model in van der Merwe and Viljoen (1998) that can be expanded to incorporate a conceptual sample, as above, to introduce shrinkage coefficient estimates and predictions.

Having shown that there are many situations in which shrinkage estimation of individual parameters and shrinkage prediction of individual future outcomes results in shrinkage estimates and prediction of totals, in the next section attention is directed at analyzing some risk properties of alternative estimators and predictors.

IV. Consideration of Risks of Alternative Estimators and Predictors

In terms of the n means problems, the following lemmas define sufficient conditions under which shrinkage estimators and predictors for totals uniformly dominate non-shrinkage estimators and predictors in terms of risk relative to quadratic loss functions.

Lemma 1: In estimation of a vector of means, θ, if a shrinkage estimator $\hat{\theta}_s$ uniformly dominates another estimator $\hat{\theta}$ in terms of risk relative to a quadratic loss function, and if $E\left(\hat{\theta}_s-\theta\right)\left(\hat{\theta}_s-\theta\right)'$ and $E\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)'$ are diagonal matrices, then $\hat{T}_s=\iota'\hat{\theta}_s$ will uniformly dominate $\hat{T}=\iota'\hat{\theta}$ relative to quadratic loss, $L\left(\tilde{T},T\right)=(\tilde{T}-T)^2$.

Proof: By assumption $E\left(\hat{\theta}_s-\theta\right)'\left(\hat{\theta}_s-\theta\right)<E\left(\hat{\theta}-\theta\right)'\left(\hat{\theta}-\theta\right)$ for all θ. Then

$E\left(\hat{T}_s-T\right)^2=E\iota'\left(\hat{\theta}_s-\theta\right)\left(\hat{\theta}_s-\theta\right)'\iota=E\left(\hat{\theta}_s-\theta\right)'\left(\hat{\theta}_s-\theta\right)$ and

$E\left(\hat{T}-T\right)^2=E\iota'\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)'\iota=E\left(\hat{\theta}-\theta\right)'\left(\hat{\theta}-\theta\right)$ and thus

$E\left(\hat{T}_s-T\right)^2<E\left(\hat{T}-T\right)^2$ given that $\hat{\theta}_s$ uniformly dominates $\hat{\theta}$ relative to quadratic loss, as assumed.

Lemma 2: In prediction of y_f, a future vector of observations if the shrinkage predictor $\tilde{y}_f$ uniformly dominates a non-shrinkage predictor $\hat{y}_f$ relative to quadratic loss, $L\left(y_f^*,y_f\right)=\left(y_f^*-y_f\right)'\left(y_f^*-y_f\right)$ and if $E\left(\tilde{y}_f-y_f\right)\left(\tilde{y}_f-y_f\right)'$ and $E\left(\hat{y}_f-y_f\right)\left(\hat{y}_f-y_f\right)'$ are diagonal matrices, then $\tilde{T}_f=\iota'\tilde{y}_f$ will uniformly dominate $\hat{T}_f=\iota'\hat{y}_f$ in terms of risk relative to quadratic loss.

Proof: By assumption $E\left(\tilde{y}_f-y_f\right)'\left(\tilde{y}_f-y_f\right)<E\left(\hat{y}_f-y_f\right)'\left(\hat{y}_f-y_f\right)$. Since

$E\left(\tilde{T}_f-T_f\right)^2=E\iota'\left(\tilde{y}_f-y_f\right)\left(\tilde{y}_f-y_f\right)'\iota=E\left(\tilde{y}_f-y_f\right)'\left(\tilde{y}_f-y_f\right)$ and

$E\left(\hat{T}_f-T_f\right)^2=E\iota'\left(\hat{y}_f-y_f\right)\left(\hat{y}_f-y_f\right)'\iota=E\left(\hat{y}_f-y_f\right)'\left(\hat{y}_f-y_f\right)$, given the diagonality assumptions, it is the case that $E\left(\tilde{T}_f-T_f\right)^2<E\left(\hat{T}_f-T_f\right)^2$.

Thus, there are many instances in which shrinkage estimators and predictors for totals will uniformly dominate non-shrinkage total estimators.

V. Summary and Conclusions

Herein, it has been demonstrated that shrinkage estimates for totals of individual parameters and for totals of future outcomes relative to quadratic loss are simple sums of individual shrinkage estimates and point predictions in cases in which extra prior information about the parameters is introduced. This was demonstrated in the case of the n-means problem and multivariate regression models using traditional Bayesian, and Bayesian method of moments, minimal mean square error, and model-averaging approaches. Further, similar results were obtained using balanced loss functions. Thus, use of informative shrinkage assumptions can lead to improved estimation and prediction of totals.

With respect to forecasting 11 economic sectors' outputs and the sum of the sectors' outputs, this was done using one equation for each sector's output. For example, there was an output equation for agriculture, mining, construction, etc. An interesting issue to be analyzed in future work is what happens when we disaggregate by region and by industry. In this case we may have several equations for each of our 11 industrial sectors and can consider using regional and national shrinkage assumptions. Much work remains to be done in analyzing these problems theoretically and in applying expanded shrinkage models in practice to evaluate their predictive performance.

References

Berger, J.O. (1985), Statistical Decision Theory and Bayesian Analysis (2nd ed.).New York: Springer-Verlag.

Berry, D. A., Chaloner, K. M. and Geweke, J. K. (eds.) (1996), Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner. New York: Wiley.

Brown, L.D. (1990), An Ancillarity Paradox Which Appears in Multiple Linear Regression, Annals of Statistics, 18, No. 2, 471-493.

Cohen, A. (1965), Estimates of Linear Combinations of the Parameters in the Mean Vector of a Multivariate Distribution , Annals of Mathematical Statistics, 36, No. 1, 78-87.

Dey, D.K, Ghosh, M. and Strawderman, W.E. (1999), On Estimation with Balanced Loss Functions. Statist. Probab. Lett., 45, 97-101.

De Alba, E. and Zellner, A. (1991), Aggregation, Disaggregation, Predictive Precision and Modeling. manuscript, H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago.

Ghosh, J. K. (2001), Discussion in Das, K. (2001), A Report on the Third Soumitra Kumar Chakravarti Memorial Lecture. Calcutta Statist. Assoc. Bull., 51, 1-10.

Green, E. and Strawderman, W.E. (1996), A Bayesian Growth and Yield Model for Slash Pine Plantations. J. Appl. Stat., 23, 285-299.

James, W. and Stein, C. (1961), Estimation with Quadratic Loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1 Berkeley, CA: University of California Press, 361-379.

Judge, G. G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. and Lee, T. (1985), The Theory and Practice of Econometrics (2nd ed.). New York: Wiley.

Leonard, T. and Hsu, J. S. J. (2001), Bayesian Methods. Cambridge: Cambridge University. Press.

Lütkepohl, H. (1986), Comparisons of Predictors for Temporally and Contemporaneously Aggregated Time Series. Internat. J. Forecasting. 2, 461-475.

Lindley, D.V. (1962), Discussion on Professor Stein’s Paper. J. R. Stat. Soc. Ser. B, 24, 285-287.

Lindley, D.V. and Smith, A.F.M. (1972), Bayes Estimates for the Linear Model. J. R. Stat. Soc. Ser. B , 34, 1-41.

Mittelhammer, R.C., Judge, G.G., and Miller, D.J. (2000), Econometric Foundations. Cambridge: Cambridge University Press.

Muth, J. (1961), Rational Expectations and the Theory of Price Movements. Econometrica, 29, 315-335.

Percy, D.F. (1992), Prediction for Seemingly Unrelated Regressions. J. R. Stat. Soc. Ser. B Stat., 54, 243-252.

Percy, D.F. (1996), Zellner’s Influence on Multivariate Linear Models, in Berry, D.A. et al. (1996), cit. supra, 203-213.

Sclove, S. L. (1968), Improved Estimators for Coefficients in Linear Regression. J. Amer. Statist. Assoc., 63, 596-606.

Sclove, S.L. (1971), Improved Estimation of Parameters in Multivariate Regression. Sankhya, Series A, 33, 61-66.

Stein, C. (1956), Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution" in Proceedings. of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1 Berkeley, CA: University of California Press, 197-206.

Stein, C. (1960), Multiple Regression, in I. Olkin (ed.), Contributions to Probability and Statistics in Honor of Harold Hotelling. Stanford: Stanford University Press.

Stein, C. (1962), Confidence Sets for the Mean of a Multivariate Normal Distribution. J. R. Stat. Soc. Ser B 24, 265-296.

Tsurumi, H. (1990), Comparing Bayesian and non-Bayesian Limited Information Estimators, in Geisser, S., Hodges, J.S., Press, S.J. and Zellner, A.,(eds.), Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard. Amsterdam: North-Holland, 179-207.

Van der Merwe, A. and Viljoen, C. (1998), Bayesian Analysis of the Seemingly Unrelated Regression Model. ms., Department of Mathematical Statistics, University ofthe Free State, Bloemfontein, S.A., presented to the annual meeting of the South African Statistical Association, November 1998.

Van der Merwe, A.J., Pretorius, A.L., Hugo, J. and Zellner, A. (2001), Traditional Bayes and the Bayesian Method of Moment Analysis for the Mixed Linear Model with an Application to Animal Breeding. South African Statist. J., 35, 19-68.

Varian, H. R. (1975), A Bayesian Approach to Real Estate Assessment, inFienberg, S.E. and Zellner, A., (eds.), Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage. Amsterdam: North-Holland, 195-208.

Zellner, A. (1986a), Bayesian Estimation and Prediction Using Asymmetric Loss Functions. J. Amer. Statist. Assoc., 81, 446-451.

Zellner, A. (1986b), On Assessing Prior Distributions and Bayesian Analysis with g-Prior Distributions, in Goel, P.K. and Zellner, A., (eds.), Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Amsterdam: North-Holland, 233-243.

Zellner, A. (1994), Bayesian and Non-Bayesian Estimation Using Balanced Loss Functions, in Gupta, S.S. and Berger, J.O., (eds.), Statistical Decision Theory and Related Topics. V New York: Springer-Verlag, Ch. 28, 377-390.

Zellner, A. (1997a), The Bayesian Method of Moments (BMOM): Theory and Applications, Fomby, T. and Hill, R., (eds.), Advances in Econometrics, Vol. 12. Greenwich, CT: Jai Press, 85-105.

Zellner, A. (1997b), Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers, invited contribution to Economists of the Twentieth Century Series, M. Perlman and M. Blaugh, (eds.), Cheltenham, UK: Edward Elgar Publ., Ltd.

Zellner, A. (1998), The Finite Sample Properties of Simultaneous Equations'Estimates and Estimators: Bayesian and non-Bayesian Approaches, in Klein, L.R., (ed.), Annals Issue of the J. Econometrics in Honor of Carl F. Christ, 83, 185-212

Zellner, A. (2001), The Marshallian Macroeconomic Model, invited paper presented at conference honoring Ryuzo Sato and published in Nagishi, T., Ramachandran, R.V. and Mino, K., (eds.), Economic Theory, Dynamics and Markets: Essays in Honor of Ryuzo Sato. Boston: Kluwer Academic Publishers, 19-29.

Zellner, A. and Vandaele, W.A. (1974), Bayes-Stein Estimators for k-means, Regression and Simultaneous Equation Models, in Fienberg, S.E. and Zellner, A., (eds.), Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage. Amsterdam: North-Holland, 627-653.

Zellner, A. and Tobias, J. (2001), Further Results on Bayesian Method of Moments Analysis of the Multiple Regression Model. Internat. Econom. Rev., 42, No.1, 121-140.

Zellner, A. and Chen, B. (2001), Bayesian Modeling of Economies and Data Requirements. Macroeconomic Dynamics, 5, 673-700.