RM-75-63

BAYESIAN REGRESSION AND CREDIBILITY THEORY

William S. Jewell

November 1975

Abstract

The development of a Bayesian theory of regression requires special distributional assumptions and rather complicated calculations. In this paper, general formulae for predicting the mean values of the regression coefficients and the mean outcomes of future experiments are developed using the methods of credibility theory, a linearized Bayesian analysis originally used in actuarial problems. No special distributional assumptions on prior or error distributions are needed, and heteroscedastic errors in both the dependent and independent variables are permitted. The first group of formulae hold for arbitrary design matrices and dimensionality of input, since, as common in Bayesian methods, there are none of the usual problems of identifiability. However, in the event that the design matrix has full rank, the credibility results are equivalent to a linear mixture of the prior mean prediction and the classical (generalized) least-squares regression predictor; thus, the credibility result provides a bridge between full Bayesian methods and classical estimators. One can also find easily the preposterior covariance matrix for the credibility estimators, and it is shown that prior information and the results from prior experiments can be cascaded in a particularly intuitive manner. Many special applications of the credibility formulae are possible because of the generality of the assumptions.

Bayesian Regression and Credibility Theory

William S. Jewell

Introduction

Regression theory plays a fundamental role in statistical model-building, parameter estimation, and forecasting. In recent years, the need to incorporate prior information into these models has stimulated the development of Bayesian methods of regression analysis, particularly in the field of econometrics [8,20,21,22,24,32]. However, the resulting formulae are usually complex, and require quite stringent assumptions on the error likelihoods and on the prior distributions of parameters.

Credibility theory, which was developed for a variety of simple predictive problems in insurance [4,5,12,13,14,15,17], is a linearized Bayesian method for forecasting mean values which circumvents many of the difficulties of a full Bayesian analysis; furthermore, in many cases of practical interest, the simplified formulae are also exact. In this paper, which was stimulated by the initial work of Hachemeister and Taylor [10,25], we apply credibility ideas to the full range of Bayesian regression models.

1. Classical Multiple Regression

In the classical model of linear normal multiple regression [8,23], we assume that an n×1 random vector of observable *output variables*, $\tilde{y}$, satisfies the linear model

$$\tilde{y} = X\beta + \tilde{u} \tag{1.1}$$

where X is a known n×k matrix of observations on k independent variables, called the *data* or *design matrix*, β is a k×1 vector of unknown *regression coefficients*, and $\tilde{u}$ is an n×1 random vector of unobservable *error variables*. If we assume that $\tilde{u}$ is multinormally distributed, with zero mean and known covariance matrix C,

$$\mathscr{C}\{\tilde{u};\tilde{u}\} = \mathscr{C}\{\tilde{y};\tilde{y}\} = \mathscr{V}\{\tilde{y}\} = C \quad ,^{*} \tag{1.2}$$

then it is well known that the ordinary least-squares estimator of β from the n observations $\tilde{y} = y$, with design matrix X and covariance matrix C, is given by

$$\hat{\hat{\beta}}(y) = \left(X'C^{-1}X\right)^{-1} X'C^{-1}y \quad . \tag{1.3}$$

In particular, if one makes the assumption that C is diagonal, with common terms, then (1.3) has the simpler form $\hat{\hat{\beta}} = (X'X)^{-1}X'y$, and the common error variance need not be known. Many other classical results are available based upon the normality assumption (see, e.g., [8,22,23]).

---

*We define the (possibly non-square and unsymmetric) covariance matrix,

$$\mathscr{C}\{\tilde{w};\tilde{y}\} = \mathscr{E}\{\tilde{w}\tilde{y}'\} - \mathscr{E}\{\tilde{w}\}\,\mathscr{E}\{\tilde{y}'\} \quad ,$$

for any two conformable random vectors or scalars $\tilde{w}$ and $\tilde{y}$, and write $\mathscr{C}\{\tilde{y};\tilde{y}\} = \mathscr{V}\{\tilde{y}\}$, which is usually called the covariance matrix.

2. Bayesian Multiple Regression

For a full Bayesian analysis, it is convenient to replace (1.1) by an equivalent model in which the expected values of the outputs are linear functions of the known inputs, viz.

$$\mathscr{E}\{\tilde{y}|\theta\} = X\beta(\theta) \quad . \tag{2.1}$$

Here θ denotes an unknown parameter which controls all the parameters of the conditional density, or *likelihood*, of $\tilde{y}$, given θ, denoted by $p(y|\theta)$. The *conditional covariance* of y, given θ, will be taken as an arbitrary symmetric n×n matrix

$$\mathscr{V}\{\tilde{y}|\theta\} = \sum(\theta) \quad . \tag{2.2}$$

Given the fixed, but unknown, parameters $[\beta(\theta),\sum(\theta),\ldots]$, we assume in Bayesian analysis that a *prior density*, $p(\theta)$, or what is the same thing, a joint prior density, $p(\beta,\sum,\ldots)$, is available. Then, *a priori* (i.e. prior to data), we define the first two moments of the vector of regression coefficients as

$$\mathscr{E}\beta(\tilde{\theta}) = b \quad ; \quad \mathscr{V}\beta(\tilde{\theta}) = \Delta \quad , \tag{2.3}$$

and the prior expected value of the covariance matrix as

$$\mathscr{E}\sum(\tilde{\theta}) = \mathscr{E}\mathscr{V}\{\tilde{y}|\tilde{\theta}\} = E \quad .^{*} \tag{2.4}$$

From these definitions, we can also obtain the prior first two moments of the output variables, given X. From (2.2), the mean and covariance of the conditional mean output are

$$\mathscr{E}\{\tilde{y}\} = \mathscr{E}\mathscr{E}\{\tilde{y}|\tilde{\theta}\} = Xb \quad , \tag{2.5}$$

and

---

*We use the convention that a multiple conditional expectation

$$\mathscr{E}\mathscr{E}\mathscr{E}\{f(\tilde{a},\tilde{b},\tilde{c})|\tilde{b}|\tilde{c}\}$$

means the expectation of f first with respect to $p(a|b,c)$, followed by expectation with respect to $p(b|c)$, then using $p(c)$. Arguments may be multiple, and other operators, such as $\mathscr{V}$ and $\mathscr{C}$, may be used. If the order is unimportant, and only $\mathscr{E}$ operators are used, the above is, of course, $\mathscr{E}\{f(\tilde{a},\tilde{b},\tilde{c})\}$.

$$\mathcal{V}\mathcal{E}\{\tilde{y}|\tilde{\theta}\} = D = X\Delta X' \quad . \tag{2.6}$$

From the covariance of the mean and the mean covariance, we obtain the total covariance (1.2) of the output variables prior to data as

$$\mathcal{V}\{\tilde{y}\} = C = E + D = E + X\Delta X' \quad . \tag{2.7}$$

If multinormal and related densities are used for $p(y|\theta)$ and $p(\theta)$, these are the only moments of interest.

Now, suppose an n_1-dimensional experiment is run with design matrix X_1, resulting in a vector of outputs, $\tilde{y} = y_1$; we denote this by (n_1, X_1, y_1). Using the likelihood $p(y_1|\theta) = p(y_1|\theta, X_1)$, and the prior on the parameters, $p(\theta)$, we obtain the *posterior* (to the data) *density* $p(\theta|y_1) = p(\theta|y_1, X_1)$ in the usual way:

$$p(\theta|y_1) = \frac{p(y_1|\theta)p(\theta)}{\int p(y_1|\Phi)p(\Phi)d\Phi} \quad , \tag{2.8}$$

where, for convenience, we suppress the known design matrix, X_1.

From (2.8), the updated estimates of the parameters $\beta(\tilde{\theta})$, $\sum(\tilde{\theta}), \ldots$, are, in principle, available. For example, the expected value of the vector of regression coefficients posterior to the data is

$$\mathcal{E}\{\beta(\tilde{\theta})|y_1\} = \int \beta(\theta)p(\theta|y_1)d\theta \quad , \tag{2.9}$$

and the *predictive density* for a future experiment (n_2, X_2, y_2), with the same parameters, but independent outputs, is

$$p(y_2|y_1) = p(y_2|y_1, X_1, X_2) = \int p(y_2|\theta, X_2)p(\theta|y_1)d\theta \quad . \tag{2.10}$$

Because of the difficulty of carrying out (2.8)-(2.10) for arbitrary priors and likelihoods, most of the Bayesian regression literature makes the following additional assumptions:

(1) The likelihood, $p(y|\theta) = p(y|\theta,X)$, is multinormal for any experiment (n,X,y)--thus only the parameters $\tilde{\beta} = \beta(\tilde{\theta})$ and $\tilde{\Sigma} = \Sigma(\tilde{\theta})$ are involved, and (2.8) can be restated in terms of $p(\beta,\Sigma)$;

(2) Either the Ando-Kaufmann [1] Normal-Wishart natural-conjugate prior $p(\beta,\Sigma)$ is used to simplify the updating in (2.8);

(3) Or, $\tilde{\beta}$ and $\tilde{\Sigma}$ are assumed independent, $p(\beta,\Sigma) = p(\beta)p(\Sigma)$, and simple marginal densities are chosen, typically multinormal or non-informative (diffuse) for $\tilde{\beta}$, and inverse Wishart or non-informative for $\tilde{\Sigma}$.

There are difficulties with all of these assumptions. For example, the Ando-Kaufmann prior is well known to be "thin"; that is, not all possible hyperparameters in $p(\beta,\Sigma)$ can be specified independently. And analysts are divided over the use of non-informative priors, although in some cases they follow from invariance or limiting arguments ([32], p. 226).

Also, computations made under these assumptions are distinctly untidy, involving much completion of the square, matrix manipulation, and multidimensional integration, particularly if the full posterior parameter density, $p(\beta,\Sigma|y_1)$, and its marginals are desired, or if the predictive density (2.10) is sought [21,30,32]. The only non-trivial relaxations of the normality assumption of which we are aware are the numerical trials of Box and Tiao ([3], Chapter 3) with the exponential power distribution.

In the sequel, we propose to follow a more modest course, by concentrating on (2.9) and the related problem of predicting the mean outcome of a future experiment, by using the linearized ideas of credibility theory. This almost distribution-free approach will greatly simplify the resulting formulae, and will provide an intuitively appealing bridge between classical and Bayesian regression techniques. And we shall see that in many cases of practical interest, the linearized credibility formulae are also exact Bayesian.

First we review the basic concepts of credibility theory.

3. Credibility Theory

Credibility theory is essentially linear least-squares applied to conditional distributions. Suppose that a p-dimensional random vector, $\tilde{w}$, is to be forecast from a single sample of an r-dimensional random vector, $\tilde{y} = y$, in the sense of finding a p-dimensional vector forecast function, $f(y)$, which minimizes the sum of the expected squared errors for each component

$$H = \int\int \sum_{i=1}^{p} [w_i - f_i(y)]^2 \, dP(w,y) = tr\mathscr{E}\{[\tilde{w} - f(\tilde{y})][\tilde{w} - f(\tilde{y})]'\} \quad . \tag{3.1}$$

It is known that the integrable functions f_i^0 which minimize (3.1) at value H^0 form the conditional mean vector,

$$f^0(y) = \mathscr{E}\{\tilde{w}|y\} \quad . \tag{3.2}$$

In many cases the exact conditional mean is difficult to calculate, and an approximate forecast vector, f, is acceptable. By completing the square, we find

$$\begin{aligned} H &= H^0 + \int \sum_{i=1}^{p} [f_i^0(y) - f_i(y)]^2 \, dP(y) \quad , \\ H^0 &= tr\,\mathscr{E}\mathscr{V}\{\tilde{w}|\tilde{y}\} \quad , \end{aligned} \tag{3.3}$$

so that any f can also be evaluated in terms of its fit to the conditional mean $f^0(y)$.

A convenient choice of an approximate forecast vector is a linear function of the observables,

$$f_i(y) = z_{i0} + \sum_{j=1}^{r} z_{ij} y_j \quad , \qquad (i = 1, \ldots p) \quad , \tag{3.4}$$

where the $p(r+1)$ coefficients $\{z_{ij}\}$, henceforth called *credibility coefficients*, are adjusted so as to minimize (3.1) or (3.3). It is well known that the optimal values of these coefficients are then given by rp *normal equations* of the form

$$\sum_{j=1}^{r} z_{ij} \mathscr{C}\{\tilde{y}_j; \tilde{y}_k\} = \mathscr{C}\{\tilde{w}_i; \tilde{y}_k\} \quad , \qquad \begin{matrix} (i = 1, \ldots p) \\ (k = 1, \ldots r) \end{matrix} \tag{3.5}$$

with the $\{z_{iO}\}$ determined so as to make the forecast (3.4) unbiased:

$$z_{iO} = \mathcal{E}\{\tilde{w}_i\} - \sum_{j=1}^{r} z_{ij}\,\mathcal{E}\{\tilde{y}_j\};\quad \mathcal{E}\{f_i(\tilde{y})\} = \mathcal{E}\{\tilde{w}_i\}\quad , \qquad (i = 1, \ldots p)\quad . \qquad (3.6)$$

Let z_O be the p-vector $[z_{iO}]'$, and Z the p×r matrix $[z_{ij}\,|\,j\neq O]$; then the optimal conditions (3.5)(3.6) can be written as

$$Z\mathcal{V}\{\tilde{y}\} = \mathcal{C}\{\tilde{w};\tilde{y}\}\quad , \qquad (3.7)$$

and

$$z_O = \mathcal{E}\{\tilde{w}\} - Z\mathcal{E}\{\tilde{y}\}\quad , \qquad (3.8)$$

so that the optimal linear forecast (3.4) is

$$f(y) = \mathcal{E}\{\tilde{w}\} + Z[y - \mathcal{E}\{\tilde{y}\}]\quad , \qquad (3.9)$$

and all attention can be focussed on finding the credibility matrix, Z, from (3.7). The minimal value of H is then easily shown to be

$$H = \mathrm{tr}[\mathcal{V}\{\tilde{w}\} - Z\mathcal{C}\{\tilde{y};\tilde{w}\}] \geq H^O\quad . \qquad (3.10)$$

Notice that each component in (3.1) is, in fact, minimized independently; we use matrix notation only for convenience.

In Bayesian problems, the joint distribution of $\tilde{w}$ and y is parametrized by a parameter θ which is not known. Therefore the optimal Z must be determined *a priori*, using measure $P(w,y) = \mathcal{E}P(w,y\,|\,\tilde{\theta})$. Thus, the covariances in (3.7) will, in general, consist of two terms similar to (2.7). One also looks for special forms of $\mathcal{V}\{\tilde{y}\}$ which will simplify the computation of Z in (3.7) [16].

In the insurance models which gave rise to credibility theory, there is an underlying sequence of p-dimensional random vectors $\{\tilde{x}_1, \tilde{x}_2, \ldots \tilde{x}_t, \tilde{x}_{t+1}, \ldots\}$, which are independent and identically distributed, given a fixed, but unknown,

"risk parameter," θ. The problem is to predict $\mathscr{E}\{\tilde{x}_{t+1}|x_1,x_2,\ldots x_t\}$, called the "experience-rated fair premium". Using the above analysis, it is easy to show that the optimal linearized approximation to the conditional mean is

$$\mathscr{E}\{\tilde{x}_{t+1}|x_1,x_2,\ldots x_t\} \approx f(x_1,x_2,\ldots x_t) = (I_p - Z_x)\mathscr{E}\{\tilde{x}\} + Z_x\left[\frac{1}{t}\sum_{u=1}^{t} x_u\right], \tag{3.11}$$

where I_p is the p×p unit matrix, and Z_x is the p×p optimal credibility matrix, given by

$$Z_x(E_x + tD_x) = tD_x \quad , \tag{3.12}$$

where E_x and D_x are the p×p matrix components of the covariance of a typical $\tilde{x}$, defined in a manner similar to (2.4) and (2.6) [13].

The original credibility formula was developed heuristically by American actuaries in the '20s for a one-dimensional version of (3.11), in which Z_x gives the weight, or "credibility," to be attached to the "experience" sample mean, $(\Sigma x_u/t)$, as opposed to the "manual fair premium" $\mathscr{E}\{\tilde{x}\}$. In the one-dimensional case, $0 \leq Z_x \leq 1$, and approaches unity as the "weight of evidence", t, becomes large. In the general (but nondegenerate) model, Z_x consists of p^2 rational functions of t, not restricted to [0,1]; however, $Z_x \to I_p$ as $t \to \infty$, showing that ultimately the sample mean of the ith component is "fully credible" for predicting the ith component of the next observation.

Although credibility theory was originally developed as an approximation theory for mean forecasts, it can also be used as an approximation theory for higher moments, or even for distributions [4,5,11].

Moreover, and perhaps more importantly, it also turns out to be an exact theory for forecasting the mean, when the likelihood is a member of the exponential family in which the sample mean is a sufficient statistic, and when a natural conjugate prior is chosen. For further details, see [12,13,14].

4. Credibility Applied To Regression

We now apply the above theory to three related Bayesian estimation problems, assuming that data from an (n_1, X_1, y_1) experiment is available:

(1) the estimation of the mean regression parameters posterior to the data;

(2) the prediction of the mean response in a future experiment (n_2, X_2, y_2);

(3) the estimation of the mean error variables in (1.1).

We shall show, with minor exceptions, that the three credibility estimates are equivalent, and related to the classical estimator (1.3).

4.1 Estimation of Regression Parameters

Suppose we wish to estimate $\mathscr{E}\{\beta(\tilde{\theta}) \mid y_1\}$ with credibility theory (X_1 is still fixed and known). Then in Section 3 we take $\tilde{w} = \beta(\tilde{\theta})$, $k = r$, and $\tilde{y} = \tilde{y}_1$, giving $\mathscr{E}\{\tilde{w}\} = b$, $\mathscr{E}\{\tilde{y}\} = X_1 b$,

$$\mathscr{C}\{\tilde{w};\tilde{y}\} = \mathscr{C}\{\beta(\tilde{\theta});\mathscr{E}\{\tilde{y}_1 \mid \tilde{\theta}\}\} = \Delta X_1' \quad ,$$

and, from (2.7),

$$\mathscr{V}\{\tilde{y}\} = C_{11} = E_{11} + X_1 \Delta X_1' \quad ,$$

where $E_{11} = \mathscr{E}\Sigma_{11}(\tilde{\theta})$ is the $n_1 \times n_1$ matrix of expected covariances of y_1 during the experiment.

From (3.7), the $k \times n_1$ credibility matrix

$$Z_\beta = \Delta X_1' C_{11}^{-1} = \Delta X_1' (E_{11} + X_1 \Delta X_1')^{-1} \qquad (4.1)$$

gives a linear, unbiased estimate of the posterior parameter vector

$$\mathscr{E}\{\beta(\tilde{\theta}) \mid y_1, X_1\} \approx f_\beta(y_1, X_1) = (I_k - Z_\beta X_1) b + Z_\beta y_1 \quad . \qquad (4.2)$$

Notice that no assumptions have been made about the distributions $p(y \mid \theta)$ and $p(\theta)$ (except for the existence of the

indicated moments), nor about the independence of the components of $\tilde{y}_1$, given θ. However, E_{11}^{-1} must exist for the inverse in (4.1) to be well defined, if no special assumptions are made about X_1 (see Section 4.3).

4.2 Prediction of Mean Response in Future Experiments

Now suppose we have in mind a well-defined future experiment (n_2, X_2, y_2), and the problem is to estimate $\mathscr{E}\{\tilde{y}_2|y_1\} = \mathscr{E}\{\tilde{y}_2|y_1, X_1, X_2\}$ by credibility theory. There are two possible cases, depending on whether

$$\Sigma_{21}(\tilde{\theta}) = \mathscr{C}\{\tilde{y}_2; \tilde{y}_1|\theta\} \quad ; \quad E_{21} = \mathscr{E}\{\Sigma_{21}(\tilde{\theta})\} \quad ;$$

are zero or not, i.e., whether knowledge of the parameter decouples the results of past and future experiments or not.

4.2.1 No Covariance Between Experiments

In most classical regression models, there is no covariance between past and future observations, given θ, either by assumption, or because there is a sufficient interval between the two experiments, even if, say, the error process has serial correlation.

For an exact Bayesian analysis, we have from (2.1) and (2.9):

$$\mathscr{E}\{\tilde{y}_2|y_1, X_1, X_2\} = X_2\mathscr{E}\{\beta(\tilde{\theta})|y_1, X_1\} \quad , \tag{4.3}$$

which shows the close relation between the two problems.

Similarly, because of the linearity of a credibility forecast, it follows that

$$\mathscr{E}\{\tilde{y}_2|y_1, X_1, X_2\} \approx f_{y_2}(y_1, X_1, X_2) = X_2 f_\beta(y_1, X_1)$$

$$= (X_2 - Z_{y_2}X_1)b + Z_{y_2}y_1 \quad , \tag{4.4}$$

where Z_{y_2} is the $n_2 \times n_1$ credibility matrix

$$Z_{y_2} = X_2\Delta X_1'(E_{11} + X_1\Delta X_1')^{-1} = X_2 Z_\beta \quad . \tag{4.5}$$

In other words, when there is no covariance between experiments, estimation of the regression coefficients by credibility is equivalent to estimation of future response.

4.2.2 Covariance Between Experiments

In the general case in which $\Sigma_{21}(\theta) \neq 0$, infrequently considered in the literature, the complete Bayesian analysis is more complicated, and one needs to replace the assumption $\mathscr{E}\{\tilde{y}_2|X_2;\theta\} = X_2\beta(\theta)$ by an equivalent assumption about $\mathscr{E}\{\tilde{y}_2|y_1,X_1,X_2,\theta\}$. This could be of arbitrary form, but if it is to be in agreement with the classical multinormal results, then we must choose the usual *regression of* y_2 *on* y_1 (see, e.g. [23]):

$$\mathscr{E}\{\tilde{y}_2|y_1,X_1,X_2,\theta\} = X_2\beta(\theta) + \Sigma_{21}(\theta)\,\Sigma_{11}^{-1}(\theta)[y_1 - X_1\beta(\theta)] \quad . \quad (4.6)$$

In an exact updating through (2.8), difficulty would arise from the possible covariance of the terms $\Sigma_{21}(\theta)$ and $\Sigma_{11}^{-1}(\theta)$ with each other, and with $\beta(\theta)$. However, if these terms have small covariances compared with those of $\beta(\theta)$, then one could with small error replace these terms by their expected values, and use the approximation

$$\mathscr{E}\{\tilde{y}_2|y_1,X_1,X_2,\beta(\theta)\} \approx X_2\beta(\theta) + E_{21}E_{11}^{-1}[y_1 - X_1\beta(\theta)] \quad (4.7)$$

to give an exact Bayesian updating:

$$\mathscr{E}\{\tilde{y}_2|y_1,X_1,X_2\} \approx X_2\mathscr{E}\{\beta(\tilde{\theta})|y_1\} + E_{21}E_{11}^{-1}[y_1 - X_1\mathscr{E}\{\beta(\tilde{\theta})|y_1\}] \quad . \quad (4.8)$$

In the credibility approximation, the formula in Section 4.2.1 is replaced by

$$\mathscr{C}\{\tilde{w};\tilde{y}\} = X_2\Delta X_1' + E_{21} \quad , \quad (4.9)$$

so that the new credibility matrix is

$$Z_{y_2} = (X_2\Delta X_1' + E_{21})(E_{11} + X_1\Delta X_1')^{-1} \quad , \quad (4.10)$$

and, after some algebra, we find

$$\mathcal{E}\{\tilde{y}_2|y_1,X_1,X_2\} \approx f_{y_2}(y_1,X_1,X_2) \equiv X_2 f_\beta(y_1,X_1) + E_{21}E_{11}^{-1}[y_1 - X_1 f_\beta(y_1,X_1)], \tag{4.11}$$

which is of the same form as (4.8). So, to the degree to which (4.7) may replace (4.6), we again have a simple relation between credibility estimates for the parameters and forecasts for future observations.

4.3 Relationship to Classical Regression Estimation

In classical regression, emphasis is placed upon having sufficient observations to fully identify all of the regression parameters, i.e., $n_1 \geq k$, and X_1 has full rank k; the necessity for this can be seen from the classical estimator (1.3).

On the other hand, in the Bayesian credibility model, it can be seen from (4.1)-(4.2) that the finiteness of b, E_{11}^{-1}, and Δ is sufficient to guarantee the existence of an estimator for $\tilde{\beta}$; one sample will revise the prior estimate of b, even if X_1 does not have full rank! In fact, if n_1 is small, the calculation of $(E_{11} + X_1\Delta X_1')^{-1}$ is particularly simple.

However, to relate our results to classical theory, we shall henceforth assume that $n_1 \geq k$, and $\text{rank}(X_1) = k$, and use the following result which Bodewig ([2] pp. 39, 218) attributes to H. Hemes, and which is also given by Tocher [29] (see also Lindley and Smith [19], pp. 6 and 34 for two later attributes).

Theorem. If α and β are $n \times k$ matrices, then

$$(I_n + \alpha\beta')^{-1} = I_n - \alpha(I_k + \beta'\alpha)^{-1}\beta' \quad , \tag{4.12}$$

whenever either of the indicated inverses exists.

The fact that the determinants of the two terms in parenthesis are identical shows that the existence of one inverse implies the existence of the other.

If we apply this to C_{11}^{-1}, with $\alpha = X_1$ and $\beta' = \Delta X_1' E_{11}^{-1}$, we get

$$\begin{aligned} C_{11}^{-1} &= (E_{11} + X_1\Delta X_1')^{-1} \\ &= E_{11}^{-1}[I_n - X_1(I_k + \Delta X_1' E_{11}^{-1} X_1)^{-1}\Delta X_1' E_{11}^{-1}] \quad . \end{aligned} \tag{4.13}$$

Defining the two $k \times k$ matrices

$$\varepsilon_1^{-1} = X_1' E_{11}^{-1} X_1 \quad ; \tag{4.14}$$

$$z_1 = (I_k + \varepsilon_1 \Delta^{-1})^{-1} = \Delta(\Delta + \varepsilon_1)^{-1} = (\varepsilon_1^{-1} + \Delta^{-1})^{-1} \varepsilon_1^{-1} ; \tag{4.15}$$

we obtain finally

$$Z_\beta = z_1 \varepsilon_1 X_1' E_{11}^{-1} \quad ; \quad Z_\beta X_1 = z_1 \quad ; \tag{4.16}$$

and (4.2) and (4.5) become

$$f_\beta(y_1, X_1) = (I_k - z_1) b + z_1 \hat{\beta}_1(y_1) \quad ; \tag{4.17}$$

$$f_{y_2}(y_1, X_1, X_2) = X_2 [(I_k - z_1) b + z_1 \hat{\beta}_1(y_1)] \quad ; \tag{4.18}$$

with a k-dimensional vector estimator for $\hat{\beta}$ of

$$\hat{\beta}_1(y_1) = (X_1' E_{11}^{-1} X_1)^{-1} X_1' E_{11}^{-1} y_1 \quad . \tag{4.19}$$

This rearrangement requires $\text{rank}(\varepsilon_1^{-1}) = k$.

(4.17) is, from an aesthetic viewpoint, extremely satisfying, for it shows the familiar credibility mixing between the prior mean parameter vector, b, and a sample statistic, $\hat{\beta}(y_1)$, in a manner similar to the multidimensional credibility formula (3.11), and extensions of it to other sample statistics [12][13]. Only a small credibility matrix, z_1, need be calculated from (4.15), and its size depends only on the number of parameters to be estimated, not the number of data points. Of course, one must calculate E_{11}^{-1}, but this is needed in any regression problem, and is often assumed to be of diagonal form. There is an obvious parallel between (4.15) and (3.12).

There remains to explain the relation between the estimator $\hat{\beta}_1(y_1)$ in (4.19), and the classical estimator $\hat{\hat{\beta}}_1(y_1)$ in (1.3), for, as we know, the latter should be used with the total covariance $C_{11} = E_{11} + X_1 \Delta X_1'$. However, a simple calculation will show that the second term is annihilated in the

least-squares form, so that

$$\hat{\hat{\beta}}_1(y_1) \equiv \hat{\beta}_1(y_1) \quad , \tag{4.20}$$

and it is a matter of indifference how the estimator is calculated.

4.4 Estimation of Error Variables

After a regression model has been calibrated, it is often useful to verify the assumptions of the model by examining the residual vector, $y_1 - X_1 f_\beta(y_1, X_1)$.

One can also think of estimating the true value of the error variables, u_1, in (1.1) by using Bayesian analysis [33]. Using the credibility approach, we first find $\mathscr{E}\{\tilde{u}_1\} = 0$, $\mathscr{V}\{\tilde{u}_1\} = \mathscr{C}\{\tilde{u}_1; \tilde{y}_1\} = E_{11}$, and then find the mean estimate,

$$\begin{aligned} \mathscr{E}\{\tilde{u}_1 | y_1, X_1\} \approx f_{u_1}(y_1, X_1) &= (I_{n_1} - X_1 Z_\beta)(y_1 - X_1 b) \\ &= y_1 - X_1 f_\beta(y_1, X_1) \quad , \end{aligned} \tag{4.21}$$

which is exactly the vector of residuals! This might have been expected from first principles.

Perhaps it is worth pointing out that [6, Appendix 3]

$$\mathscr{C}\{\tilde{u}_1; f_{u_1}(\tilde{y}_1, X_1)\} = 0 \quad . \tag{4.22}$$

5. Estimation Error Covariances--Limiting Cases

It is of interest to compute the improvement in estimation to be expected from the credibility formulae.

For the regression parameters, let the estimation error covariance matrix be

$$\Phi_\beta(X_1) = \mathscr{E}\{[\beta(\tilde{\theta}) - f_\beta(\tilde{y}_1, X_1)][\beta(\tilde{\theta}) - f_\beta(\tilde{y}_1, X_1)]'\}$$

$$= \mathscr{V}\{\beta(\tilde{\theta}) - f_\beta(\tilde{y}_1, X_1)\} \quad , \tag{5.1}$$

because the estimator is unbiased, a priori.

By elementary calculations based on Sections 3.1 and 4, we find that the minimal "preposterior" value is the analog of the term in square brackets in (3.10):

$$\Phi_\beta(X_1) = (I_k - z_1)\Delta = z_1\epsilon_1 \quad . \tag{5.2}$$

Remember that only the diagonal terms of Φ are (independently) minimized in using (3.1), $H = tr\Phi$.

For the prediction of mean future response, we find in the no-covariance case of Section 4.2.1:

$$\Phi_{y_2}(X_1, X_2) = \mathscr{V}\{\tilde{y}_2 - f_{y_2}(\tilde{y}_1, X_1, X_2)\}$$

$$= E_{22} + X_2(I_k - z_1)\Delta X_2' = E_{22} + X_2 z_1 \epsilon_1 X_2' \quad . \tag{5.3}$$

The result with covariance between experiments is similar, with additional terms involving E_{21}.

The preposterior estimate of the covariance matrix of the residual vector (4.21) is

$$\Phi_{u_1}(X_1) = X_1 Z_\beta E_{11} = X_1 \Phi_\beta(X_1) X_1' \quad . \tag{5.4}$$

Without an initial experiment, the value of z_1 would be zero, and from (4.17)(4.18)(4.21) we would have to use the means, b, $X_2 b$ and y_1, as predictors, and (5.2)(5.3)(5.4) would be equal to the appropriate total prior covariance matrices,

Δ, $E_{22} + X_2 \Delta X_2'$, and 0, respectively.

Similarly, if the first experiment is performed under poor observational conditions, then the diagonal elements of E_{11} will be much larger than those of $X_1 \Delta X_1'$. We see directly that z_1 would be zero, and there would be a vote of "no confidence" in the estimator $\hat{\beta}_1(y_1)$, and b, $X_2 b$, and y_1 would again be the minimum-variance predictors for $\beta(\theta)$, y_2, and u_1, respectively.

However, conversely, if the diagonal elements of Δ are very large compared to those of ε_1, this means that our prior knowledge is very imprecise compared to the error conditions of the experiment; $\Delta^{-1} \to 0$ is the credibility equivalent of the "diffuse prior" assumptions often made in Bayesian analysis. In this case, we see that $z_1 \to 1$; "full credibility" is attached to the classical estimator $\hat{\beta}_1(y_1)$, and the prior mean, b, is given zero weight. There remain only the irreducible error covariances ε_1 in estimating $\beta(\theta)$, $E_{22} + X_2 \varepsilon_1 X_2'$ in predicting y_2, and $X_1 \varepsilon_1 X_1'$ in estimating u_1.

Also, if we consider experiments with increasing n_1, then, under certain natural conditions, such as:

(1) The elements of E_{11} are bounded, for all n_1;

(2) The design matrix, X_1, "fills out" a finite range of the x-axis in a stable manner, as n_1 increases;

it is easy to show that the elements of ε_1 in (4.14) are bounded by a function which diminishes as n_1^{-1}, that is, z_1 approaches I_k as n_1 increases (see, e.g., [18]). In practical terms, this means that an increasing number of initial sample points can reduce the preposterior covariance in estimating the regression parameter (5.2) as close to zero as desired; however, there will always be an irreducible covariance E_{22} in making forecasts (5.3). The covariance matrix $\Phi_u(X_1)$ in (5.4) continues to grow in dimension, and depends in a complicated manner upon the actual structure of X_1.

6. Random Design Matrices

In many applications, X_1 and/or X_2 must be considered as random, either as a result of an uncontrollable input, because the effective input cannot be precisely observed, or because of deliberate randomization. There are many special cases in the literature, (see, e.g.,[7,32]); we shall derive general credibility results, and indicate only a few of the possible specializations. Special attention must be paid to whether X_1, X_2, or both are random variables, so throughout this section we shall indicate the status of all inputs and outputs explicitly. We start with two simpler cases.

6.1 X_2 Random and Independent of Fixed Initial Experiment

If the future design matrix X_2 is random, but independent of the fixed initial experiment (n_1, X_1, y_1), then the problem of estimating the regression parameters is unchanged from Section 4.1.

However, to predict the mean response of the second experiment, we must now calculate a credibility approximation to $\mathscr{E}\{\tilde{y}_2|y_1,X_1\} = \mathscr{E}\mathscr{E}\{\tilde{y}_2|y_1,X_1,\tilde{X}_2\}$. Assuming, for simplicity, unobservationally unrelated experiments, $\Sigma_{21}(\theta) = 0$, we have from (2.1) and Section 4.2.1.,

$$\mathscr{E}\{\tilde{y}_2\} = \mathscr{E}[\mathscr{E}\{\tilde{X}_2|\tilde{\theta}\}\cdot\beta(\tilde{\theta})] \quad , \tag{6.1}$$

and

$$\mathscr{C}\{\tilde{y}_2;\tilde{y}_1|X_1\} = \mathscr{C}\{\mathscr{E}\{\tilde{X}_2|\tilde{\theta}\}\cdot\beta(\tilde{\theta});X_1\beta(\tilde{\theta})\} \quad . \tag{6.2}$$

Since $\mathscr{V}\{\tilde{y}_1|X_1\} = E_{11} + X_1\Delta X_1'$ and $\mathscr{E}\{\tilde{y}_1|X_1\} = X_1 b$ still, the only effect in this case has been to modify the first term, $X_2\Delta X_1'$, in the definition of Z_{y_2} in (4.5) to the form in (6.2) and to change the z_0 term in (4.4).

An important special case is:

<u>Assumption I</u>. <u>Any random X is statistically independent of θ</u>. (6.3)

In this case, we see directly that $\mathscr{E}\{\tilde{y}_2\} = \mathscr{E}\{\tilde{x}_2\}b$ and $\mathscr{C}\{\tilde{y}_2;\tilde{y}_1|X_1\} = \mathscr{E}\{\tilde{X}_2\}\Delta X_1'$, that is, all the results of Section

4.2.1 apply with X_2 replaced by its expected value!

6.2 Estimation of Regression Parameters when X_1 is Random

If X_1 is random, then to estimate $\beta(\theta)$ we must use the joint density $p(y_1, X_1 | \theta)$ and generalize (4.2). For the mean outcome of the initial experiment,

$$\mathscr{E}\{\tilde{y}_1\} = \mathscr{E}\{\mathscr{E}\{\tilde{X}_1 | \tilde{\theta}\} \cdot \beta(\tilde{\theta})\} \quad , \tag{6.4}$$

but the covariance of y_1 now has three terms:

$$\mathscr{V}\{y_1\} = \mathscr{E}\{\Sigma_{11}(\tilde{X}_1, \tilde{\theta})\} + \mathscr{V}\{\mathscr{E}\{\tilde{X}_1 | \tilde{\theta}\} \cdot \beta(\tilde{\theta})\} + \mathscr{E}\mathscr{V}\{\tilde{X}_1 \beta(\tilde{\theta}) | \tilde{\theta}\} \quad , \tag{6.5}$$

where

$$\Sigma_{11}(X_1, \theta) = \mathscr{V}\{\tilde{y}_1 | X_1, \theta\} \tag{6.6}$$

shows explicitly the possible dependence of the conditional observational covariance both on the design X_1 and on θ. (For consistency, we shall assume in the next section that neither (6.4) nor (6.6) can, however, depend upon the future values (y_2, X_2).)

Since $\beta(\theta)$ is constant, given θ, there is still only one term in

$$\mathscr{C}\{\beta(\tilde{\theta}); \tilde{y}_1\} = \mathscr{C}\{\beta(\tilde{\theta}); \mathscr{E}\{\tilde{X}_1 | \tilde{\theta}\} \cdot \beta(\tilde{\theta})\} \quad . \tag{6.7}$$

This form and the first two terms in (6.5) are easily seen to be the generalizations of $\Delta X_1'$ and $E_{11} + X_1 \Delta X_1'$, respectively, as used in Section 4.1.

However, the last term in (6.5) is new, call it U. It has components

$$U_{tu} = \mathscr{E}\left\{ \sum_{i=1}^{k} \sum_{j=1}^{k} \beta_i(\tilde{\theta}) \beta_j(\tilde{\theta}) \mathscr{C}\{\tilde{x}_{ti}; \tilde{x}_{uj} | \tilde{\theta}\} \right\} \quad ,$$

$$(t,u = 1,2, \ldots k) \quad , \tag{6.8}$$

and thus contains information about the conditional covariances

between independent variables.

In many models, such as "errors-in-the-variables," or "target inputs" [7], successive inputs are independent, or have independent errors around fixed means, expressable as:

Assumption II. Rows of any random X are statistically independent. (6.9)

In this case, it follows that U is diagonal. Additionally, we point out that in many regression designs, the first column of X_1 is non-random (consisting entirely of 1's), so that the summations in (6.8) would begin with i = 2 and j = 2.

If Assumption I is taken also to apply to $\tilde{X}_1$,

$$\begin{aligned} &\mathscr{E}\{\tilde{y}_1\} = \mathscr{E}\{\tilde{X}_1\}b \quad ; \\ &\mathscr{V}\{\tilde{y}_1\} = \mathscr{E}\{\Sigma_{11}(\tilde{X}_1,\tilde{\theta})\} + \mathscr{E}\{\tilde{X}_1\}\Delta\mathscr{E}\{\tilde{X}_1'\} + U \quad ; \\ &\mathscr{C}\{\beta(\tilde{\theta});\tilde{y}_1\} = \Delta\mathscr{E}\{X_1'\} \quad ; \end{aligned} \tag{6.10}$$

and the main effect on the credibility estimate (4.1), apart from replacing X_1 by its mean value, and defining a more general average covariance E_{11}, is to add a diagonal matrix U to the covariance of $\tilde{y}_1$, with terms

$$U_{tt} = \sum_{i=1}^{k}\sum_{j=1}^{k}(\Delta_{ij} + b_i b_j)\mathscr{C}\{\tilde{x}_{ti};\tilde{x}_{tj}\} \quad ,$$

$$(t = 1,2, \ldots k) \quad . \tag{6.11}$$

This will change Z_β in an obvious manner, and we see that the estimator to be used in (4.17) becomes

$$\hat{\beta}(y_1) = \left[\mathscr{E}\{X_1'\}(E_{11} + U)^{-1}\mathscr{E}\{\tilde{X}_1\}\right]^{-1}\mathscr{E}\{\tilde{X}_1'\}(E_{11} + U)^{-1}y_1 \quad , \tag{6.12}$$

with the new interpretation of E_{11} from (6.10), and a new

$$\varepsilon_1^{-1} = \mathscr{E}\{X_1'\}(E_{11} + U)^{-1}\mathscr{E}\{\tilde{X}_1\} \tag{6.13}$$

used to define z_1 in (4.15).

6.3 General Case

In the general case when all inputs and outputs are random, we must work with the joint density $p(y_1,X_1,y_2,X_2|\theta)$, and be extremely careful about the assumptions of dependence and independence which are appropriate to the model under consideration. Different models may lead to different conditional decompositions of this joint density.

Usually the regression parameters are estimated after the initial experiment, so that the results of Section 6.2 apply. If both experiments are performed, then the total data may be pooled, and the same results apply with obvious modification (see Section 7).

Therefore the central problem of interest in credibility theory will be to predict $\mathscr{E}\{\tilde{y}_2|\tilde{y}_1\}$, for which we need: $\mathscr{E}\{\tilde{y}_1\}$, $\mathscr{E}\{\tilde{y}_2\}$, $\mathscr{V}\{\tilde{y}_1\}$ and $\mathscr{C}\{\tilde{y}_2;\tilde{y}_1\}$. (6.4) and (6.5) still apply because the data-gathering experiment is prior to the one for which the prediction is made. However, to compute $\mathscr{E}\{\tilde{y}_2\}$, we need an assumption such as (4.7) to specify a form for $\mathscr{E}\{\tilde{y}_2|y_1,X_1,X_2,\theta\}$. Given this, we then uncondition in any convenient way, say

$$\mathscr{E}\{\tilde{y}_2\} = \mathscr{E}\mathscr{E}\mathscr{E}\mathscr{E}\{\tilde{y}_2|\tilde{X}_2|\tilde{y}_1|\tilde{X}_1|\tilde{\theta}\} \quad , \tag{6.14}$$

using any other simplifications, such as Assumption I, which apply. Further reduction will need a careful analysis of the experimental conditions; for example

> Assumptions III(a) (b) or (c). The choice of the future design, $\tilde{X}_2$, given θ, depends only on (a) the past input, X_1; or (b) the past output, y_1; or (c) on both (X_1,y_1); (6.15)

III(a) might obtain if (X_1,X_2) were part of the same predetermined experimental design, or if errors in the independent variables were serially correlated; III(b) might be correct if the future input values depended upon the previous outputs, or perhaps on some estimator of $\beta(\theta)$, such as (4.2), as generalized in Section 6.2.

For the RHS of (3.7), repeated application of the principle of conditional covariance leads to

$$\begin{aligned}\mathscr{C}\{\tilde{y}_2;\tilde{y}_1\} = {} & \mathscr{E}\mathscr{E}\mathscr{E}\{\Sigma_{21}(\tilde{X}_2;\tilde{X}_1;\tilde{\theta})\,|\,\tilde{X}_1\,|\,\tilde{\theta}\} \\ & + \mathscr{E}\mathscr{C}\{\mathscr{E}\mathscr{E}\{\tilde{y}_2\,|\,\tilde{X}_2\,|\,\tilde{X}_1,\tilde{\theta}\};\tilde{X}_1\beta(\tilde{\theta})\,|\,\tilde{\theta}\} \qquad (6.16) \\ & + \mathscr{C}\{\mathscr{E}\mathscr{E}\mathscr{E}\{\tilde{y}_2\,|\,\tilde{X}_2\,|\,\tilde{X}_1\,|\,\tilde{\theta}\};\mathscr{E}\{\tilde{X}_1\,|\,\tilde{\theta}\}\cdot\beta(\tilde{\theta})\}\ ,\end{aligned}$$

where the arguments of $\Sigma_{21}(X_2,X_1,\theta)$ show that the covariance of observational errors between $\tilde{y}_2$ and $\tilde{y}_1$ can now depend upon both inputs; one possible term in (6.16) is missing because we still assume $\mathscr{E}\{\tilde{y}_1\,|\,X_2,X_1,\theta\} = X_1\beta(\theta)$. Further simplification depends upon using forms such as (4.7), and clarifying the experimental relationships between $\tilde{\theta}$, $\tilde{X}_1$, and $\tilde{X}_2$.

7. Prior Information and Prior Experiments

The distinction between prior information, in the usual Bayesian sense, and the information obtained as the result of a prior experiment is not clear-cut. Suppose we have given prior information (b,Δ) about $\beta(\theta)$, and the matrix of observation error covariances E for any (n,X). A first experiment (n_1,X_1,y_1) then provides a further estimate of $\beta(\theta)$, which supplements our knowledge prior to the performance of a second experiment (n_2,X_2,y_2); thus, there is total prior information $(b,\Delta;E_{11};n_1,X_1,y_1)$ as input to the second stage. On the other hand, we know that the estimation of $\beta(\theta)$ after two experiments can be regarded as a combined single experiment, and it is interesting to examine further the relationship between these two viewpoints.

To estimate $\mathscr{E}\{\beta(\tilde{\theta})\,|\,y_1,X_1;y_2,X_2\}$, we form the enlarged versions of (2.1)(2.2):

$$\mathscr{E}\left\{\begin{bmatrix}\tilde{y}_1\\ \tilde{y}_2\end{bmatrix}\,\middle|\,\theta\right\} = \begin{bmatrix}X_1\\ X_2\end{bmatrix}\beta(\theta) \quad ; \tag{7.1}$$

$$\mathscr{V}\left\{\begin{bmatrix}\tilde{y}_1\\ \tilde{y}_2\end{bmatrix}\,\middle|\,\theta\right\} = \begin{bmatrix}\Sigma_{11}(\theta) & O\\ O & \Sigma_{22}(\theta)\end{bmatrix} \quad ; \tag{7.2}$$

where we have assumed the two experiments are observationally independent, and the design matrices are fixed. Then, following the analysis of Section 4.1, we find an enlarged Z_β-type $k \times (n_1 + n_2)$ credibility matrix, $Z_{1,2}$, for the combined experiment,

$$Z_{1,2} = \Delta\begin{bmatrix}X_1' & X_2'\end{bmatrix}\begin{bmatrix}E_{11} + X_1\Delta X_1' & X_1\Delta X_2'\\ X_2\Delta X_1' & E_{22} + X_2\Delta X_2'\end{bmatrix}^{-} , \tag{7.3}$$

which is then used in the estimate:

$$\mathscr{E}\{\beta(\tilde{\theta})\,|\,y_1,X_1;y_2,X_2\} \approx f_\beta(y_1,X_1;y_2,X_2) = \left(I_k - Z_{1,2}\begin{bmatrix}X_1\\ X_2\end{bmatrix}\right)b + Z_{1,2}\begin{bmatrix}y_1\\ y_2\end{bmatrix} . \tag{7.4}$$

If we define individual Z_β-type matrices for each of the experiments individually,

$$Z_i = \Delta X_i' (E_{11} + X_i \Delta X_i')^{-1} \quad , \qquad (i = 1,2) \, , \tag{7.5}$$

then the combined credibility matrix can be written in a simpler form:

$$Z_{1,2} = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} I_{n_1} & X_1 Z_2 \\ X_2 Z_1 & I_{n_2} \end{bmatrix}^{-1} \tag{7.6}$$

$$= \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} (I_{n_1} - X_1 Z_2 X_2 Z_1)^{-1} & -(I_{n_1} - X_1 Z_2 X_2 Z_1)^{-1} X_1 Z_2 \\ -(I_{n_2} - X_2 Z_1 X_1 Z_2)^{-1} X_2 X_1 & (I_{n_2} - X_2 Z_1 X_1 Z_2)^{-1} \end{bmatrix} .$$

Further simplification requires the assumption of full rank for X_1 and X_2, and the definitions (see (4.14) (4.15):

$$\varepsilon_i^{-1} = X_i' E_{ii}^{-1} X_i \quad ; \qquad Z_i = \Delta (\Delta + \varepsilon_i)^{-1} \quad ; \qquad (i = 1,2). \tag{7.7}$$

After repeated use of (4.12) and (4.16), the result finally simplifies to

$$Z_{1,2} = \left[(I_k - z_2)(I_k - z_1 z_2)^{-1} Z_1 \;\; (I_k - z_1)(I_k - z_2 z_1)^{-1} Z_2 \right]. \tag{7.8}$$

Defining the individual classical estimators for each experiment

$$\hat{\beta}_i(y_i) = \varepsilon_i X_i' E_{ii}^{-1} y_i \quad , \qquad (i = 1,2) \quad , \tag{7.9}$$

we obtain finally the combined-experiment estimate ,

$$f_\beta(y_1, X_1; y_2, X_2) = (I_k - z^{(1)} - z^{(2)}) b + z^{(1)} \hat{\beta}_1(y_1) + z^{(2)} \hat{\beta}_2(y_2) \quad , \tag{7.10}$$

where

$$z^{(1)} = (I_k - z_2)(I_k - z_1 z_2)^{-1} z_1; \quad z^{(2)} = (I_k - z_1)(I_k - z_2 z_1)^{-1} z_2. \tag{7.11}$$

This formula can then be rearranged so as to display a new prior mean, $b^{(2)}$, which is used as input to the second experiment, together with the credibility matrix $z^{(2)}$, in the "single-stage" formula

$$f_\beta(y_1, X_1; y_2, X_2) = (I_k - z^{(2)}) b^{(2)} + z^{(2)} \hat{\beta}_2(y_2) \quad . \tag{7.12}$$

Then, we find that

$$\begin{aligned} b^{(2)} &= b + (I_k - z^{(2)})^{-1} z^{(1)} [\hat{\beta}_1(y_1) - b] \\ &= (I_k - z_1) b + z_1 \hat{\beta}_1(y_1) = f_\beta(y_1, X_1) \quad , \end{aligned} \tag{7.13}$$

is just the usual first-stage credibility prediction (4.2) or (4.17), which becomes the mean input for the second experiment.

We may further clarify (7.12) by seeing what equivalent regression coefficient covariance, say $\Delta^{(2)}$, is used as input to the second experiment to find the credibility coefficient in the usual way as

$$z^{(2)} = \Delta^{(2)} (\Delta^{(2)} + \varepsilon_2)^{-1} \quad . \tag{7.14}$$

We find

$$\Delta^{(2)} = (\varepsilon_1^{-1} + \Delta^{-1})^{-1} = z_1 \varepsilon_1 = \Phi_\beta(X_1) \quad , \tag{7.15}$$

which is just the preposterior estimate of the error covariance (5.2) after the first experiment!

To summarize, we can view the two experiments (n_1, X_1, y_1) (n_2, X_2, y_2):

(1) Either as a combined experiment in which the prior information b and Δ is used in (7.10) to form an estimate of $\beta(\theta)$;

(2) Or as a two-stage process in which b and Δ are used in the first experiment to form $f_\beta(y_1, X_1)$ and $\Phi_\beta(X_1)$, and these values are then used as the prior vector mean and matrix covariance of the regression coefficients for the independent second experiment, forming an estimate of $\beta(\theta)$ using (7.12)(7.14).

The extension to multiple cascaded experiments is obvious. Also, it follows that, prior to both experiments, our estimate of the final covariance matrix is

$$\Phi_\beta(X_1, X_2) = (\varepsilon_2^{-1} + \Phi_\beta^{-1}(X_1))^{-1} = (\Delta^{-1} + \varepsilon_1^{-1} + \varepsilon_2^{-1})^{-1} \quad .$$

In other words, the total final precision is estimated, prior to any experiment, to be the sum of the prior precision plus the observation precision of each experiment.

We now examine several special cases of interest.

7.1 Imprecise Experimental Results

If the first experiment is performed under poor observational conditions, we expect the diagonal elements of E_{11} to be large compared to those of $X_1 \Delta X_1'$. Under these conditions, $z_1 \to 0$, $z^{(2)} \to z_2$, and the results of the first experiment are ignored, with b and Δ used directly as inputs to the second stage. Similar remarks apply to imprecise results in the second experiment; and, of course, if both experiments have high observational variances, then the best forecast is just b.

7.2 Diffuse Prior Information

If, on the other hand, the prior variances of the regression coefficients are very large compared to the imputed covariances ε_1 and ε_2 due to observational error, then z_1 and z_2 approach unity, and we see from (7.14) (7.15), or by careful limits in (7.11), that $z^{(i)} \to (\varepsilon_1^{-1} + \varepsilon_2^{-1})^{-1}\varepsilon_i^{-1}$, (i = 1,2), and

$$f_\beta(y_1, X_1; y_2, X_2) = (\varepsilon_1^{-1} + \varepsilon_2^{-1})^{-1}\left[\varepsilon_1^{-1}\hat{\beta}_1(y_1) + \varepsilon_2^{-1}\hat{\beta}_2(y_2)\right] \quad . \quad (7.16)$$

In other words, the prior information is ignored as the diagonal elements of Δ become large (the prior becomes "diffuse"), and the resulting estimate weights the classical estimators from each experiment in the familiar proportional-to-precision manner. A formula similar to (7.16) is given by sampling theory arguments in the "mixed-estimation" method of Goldberger and Theil [8, Section 5-6][9][27][28].

Alternatively, we may regard this case as one in which a prior mean $\hat{\beta}_1(y_1)$ and a prior covariance ε_1 are used as input to the second experiment.

7.3 Direct Estimate of Regression Parameters

If the first experiment provides a direct measurement of the regression parameters, $\beta(\theta)$, then $n_1 = k$, $X_1 = I_k$, and for consistency, we could call $y_1 = b_1$ a new estimate of b, with covariance of observation errors, $\varepsilon_1 = \Delta_1$, say. Then, the credibility matrix in this special first experiment is $z_1 = \Delta(\Delta + \Delta_1)^{-1}$, the mean input (7.13) to the second experiment is

$$b^{(2)} = (\Delta^{-1} + \Delta_1^{-1})^{-1}\left[\Delta^{-1}b + \Delta_1^{-1}b_1\right] \quad , \tag{7.17}$$

and the covariance matrix input (7.15) is

$$\Delta^{(2)} = (\Delta^{-1} + \Delta_1^{-1})^{-1} \quad . \tag{7.18}$$

In other words, if there are two prior estimates of the regression parameters, then they should be combined in the usual proportional-to-precision manner, and then used as input.

7.4 Similar Experiments

If the design matrix, X, of the two experiments is the same, then the common $z = \Delta(\Delta + \varepsilon)^{-1}$, with $\varepsilon^{-1} = X'E^{-1}X$, and the forecast (7.10) can be written

$$f_\beta(y_1;y_2;X) = \varepsilon(2\Delta + \varepsilon)^{-1}b + 2\Delta(2\Delta + \varepsilon)^{-1}\left[\tfrac{1}{2}(\hat{\beta}(y_1) + \hat{\beta}(y_2))\right] \quad , \tag{7.19}$$

with an obvious definition of the common function $\hat{\beta}(y)$. In this form, the analogy with the many-sample credibility forecast (3.11)(3.12) is obvious, and the extension to t *similar*

experiments is immediate :

$$f_\beta(y_1;y_2;\ldots y_t;X) = [I_k - z(t)]b + z(t)\left[\frac{1}{t}\sum_{i=1}^{t}\hat{\beta}(y_i)\right] \quad , \qquad (7.20)$$

with a new credibility matrix

$$z(t) = t\Delta(t\Delta + \varepsilon)^{-1} \quad . \qquad (7.21)$$

7.5 Repeated Dissimilar Experiments

For completeness, we give the general formulae corresponding to (7.10)(7.11), when t *dissimilar* experiments $(n_1,X_1,y_1)(n_2,X_2,y_2) \ldots (n_t,X_t,y_t)$ are performed. In an obvious extension of notation ,

$$f_\beta((y_i,X_i);i=1,2,\ldots,t) = \left[I_k - \sum_{i=1}^{t} z^{(i)}\right]b + \sum_{i=1}^{t} z^{(i)}\hat{\beta}_i(y_i) \quad , \qquad (7.22)$$

where the $z^{(i)}$ are the solutions of

$$\left[z^{(1)}z^{(2)} \cdots z^{(t)}\right] = [1\ 1 \cdots 1]\begin{bmatrix} z_1^{-1} & I_k & \cdots & I_k \\ I_k & z_2^{-1} & \cdots & I_k \\ \vdots & \vdots & & \vdots \\ I_k & I_k & & z_t^{-1} \end{bmatrix}^{-1} \quad . \qquad (7.23)$$

The prior-to-experiments estimate of the final covariance of the estimator error is

$$\Phi_\beta(X_1,X_2,\ldots,X_t) = \left(\Delta^{-1} + \sum_{i=1}^{t}\varepsilon_i^{-1}\right)^{-1} \quad ; \qquad (7.24)$$

that is, the final precision is estimated to be the sum of the prior precision plus all of the observational precisions. Of course, as indicated earlier, it is probably easier to compute (7.22) in the recursive manner suggested earlier in this section.

8. Related Work

There are two papers which originated the application of credibility theory to regression problems. In a multidimensional model, with elaborate notation based on practical considerations, Hachemeister [10] has given prediction formulae equivalent to (4.18)(4.19); however, his derivation appears to require the assumption of heteroscedastic error terms, i.e.

$$\Sigma(\theta) = \sigma^2(\theta) I_n \quad , \tag{8.1}$$

or of the sample-mean generalization in which the ith diagonal term of $\Sigma(\theta)$ is $\sigma^2(\theta)/P_i$, where P_i is the "volume" of the ith sample.

He also gives a credibility result for a homogeneous estimator, i.e., with $z_{i0} = 0$ in (3.4), and the remaining credibility coefficients constrained to give an unbiased estimator. For models of this type, one usually has collateral data [17] from similar experiments performed on other risks, with independent values of θ.

Taylor's first paper [25] concentrates on the two-parameter, homogeneous estimator model, using essentially the same assumptions as Hachemeister, but with a simplified unbiasedness constraint. In a later paper [26], Taylor generalizes both the homogeneous and inhomogeneous versions of (4.18) to Hilbert spaces, and shows various special cases.

Turning to exact Bayesian regression results based upon multinormal likelihoods, Raiffa and Schlaiffer [22] give formulae equivalent to (4.17) for the cases in which (1) $\sigma^2(\theta) = \sigma^2$ is a known constant, and the prior on $\beta(\theta)$ is multinormal (b,Δ); (2) $(\sigma^2(\theta),\beta(\theta))$ are inverse-Gamma-multinormally distributed. Other models by Tiao, Zellner, and Chetty [29][30][32][34] concentrate on the use of a diffuse prior density, $p(\beta,\sigma^2) \propto \sigma^{-1}$, or its multidimensional equivalent [32, Chapter 8]; thus, after one experiment, $\hat{\beta}_1(y_1)$ is "fully credible," or after two experiments, results similar to (7.16) are obtained. Of course, since these are exact Bayesian results, the complete posterior distributions of the parameter are available--usually some variation of the multivariate-t density.

In [32, p. 240], Zellner takes an "informative" prior which is slightly more general than the usual natural-conjugate prior for the multinormal; his likelihood is multivariate, with homeoscedastic errors, which can be reinterpreted as

single-variate with arbitrary $\Sigma(\theta)$. By expanding the resulting posterior density for the regression parameters, he finds from the leading normal term a mean estimate which is "a 'matrix weighted average' of the prior mean...and the least-squares quantity $\hat{\beta}$ whose weights are the inverse of the prior covariance C and the sample covariance matrix." This is, of course, just our result (4.17)(4.18)(5.2), gotten as an approximation for arbitrary likelihood and prior densities.

We have also indicated that, using sampling theory arguments, Goldberger and Theil [8][9][27][28] have obtained formulae similar to (7.16), except that, since $\sigma_i^2(\theta)$ $(i = 1,2)$ in ε_1, ε_2 are unknown, they propose substituting various reasonable sample estimates.

9. Exact Results

It can be seen from the above that the credibility formulae presented here are exact when the likelihood is multinormal, and the prior is from a natural conjugate family. However, there are additional cases in which the credibility results are exact, based upon the Koopmans-Pitman-Darmois exponential-type families, and their (suitably enriched) natural conjugate priors. (See [12][13][14] for exact results for the model of (3.11).) These will be reported in a separate paper.

10. Extensions

Many of the topics which are considered as extensions in classical works on regression are already covered by our basic model, since no special assumption about the error covariance matrix $\Sigma(\theta)$ has been made; for example, error terms may be autocorrelated. Multivariate regression models are already "serially" included, and it remains only to translate them into the usual "parallel" notation. And, by following the discussion in Section 6, a variety of random input models may be elaborated; for example, successive inputs may follow a "random shocks" process [15].

There are many interesting regression models in which the design matrix is not of full rank. In these cases, (4.2) and (4.4) are still viable, even though the classical estimators do not exist. Or one may add additional constraints, based upon external considerations, until the problem is "identifiable," in the classical sense. The particular problem of estimating flows in a network will be the topic of a future report.

For a simple linear regression, one can also talk about problems of inverse regression; that is, given y, what was

the input x? These questions arise in various problems of measurement, and a detailed study of instrument calibration and measurement using credibility methods may be found in [18].

BIBLIOGRAPHY

[1] Ando, A. and Kaufman, G.M. "Bayesian Analysis of the Independent Multinormal Process--Neither Mean Nor Precision Known." J. Amer. Statist. Assoc., 60, pp. 347-358 (1965).

[2] Bodewig, E. Matrix Calculus (2nd Edition). North-Holland, Amsterdam (1959).

[3] Box, G.E.P. and Tiao, G.C. Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, Massachusetts (1973).

[4] Bühlmann, H. "Experience Rating and Credibility." ASTIN Bulletin, 4, Part 3, pp. 199-207 (July, 1967).

[5] _____ Mathematical Methods in Risk Theory. Springer-Verlag, New York (1970).

[6] Cox, D.R. and Hinkley, D.V. Theoretical Statistics. Chapman and Hall, London (1974).

[7] Florens, J.-P., Mouchart, M. and Richard, J.-F. "Bayesian Inference in Error-in-Variables Models." J. of Multivariate Analysis, 4, No. 4, pp. 419-452. (1974).

[8] Goldberger, A.S. Econometric Theory. J. Wiley & Sons, New York (1964).

[9] _____ "Efficient Estimation in Overidentified Models: An Interpretive Analysis." Chapter 7 in Structural Equation Models in the Social Sciences, A.S. Goldberger and O.D. Duncan (Eds.), Seminar Press, New York (1973).

[10] Hachemeister, C.A. "Credibility for Regression Models with Application to Trend." Proceedings of Actuarial Research Conference on Credibility Theory, Berkeley, California, September, 1974. Academic Press, New York (1975).

[11] Jewell, W.S. "The Credible Distribution." ORC 73-7, Operations Research Center, University of California, Berkeley (August, 1973). ASTIN Bulletin, 7, Part 3, pp. 237-269 (March, 1974).

[12] _____ "Credible Means are Exact Bayesian for Simple Exponential Families." ORC 73-21, Operations Research Center, University of California, Berkeley (October, 1973). ASTIN Bulletin, 8, Part 1, pp. 77-90 (September, 1974).

[13] _____ "Exact Multidimensional Credibility." ORC 74-14, Operations Research Center, University of California, Berkeley (May, 1974). Mitteilungen der Vereinigung Schweizerischer Versicherungs-mathematiker, 74, No. 2, pp. 193-214 (1974).

[14] _____ "Regularity Conditions for Exact Credibility." ORC 74-22, Operations Research Center, University of California, Berkeley (July, 1974). To appear in ASTIN Bulletin.

[15] _____ "Model Variations in Credibility Theory." ORC 74-25, Operations Research Center, University of California, Berkeley (August, 1974). Proceedings of Actuarial Research Conference on Credibility Theory, Berkeley, California, September, 1974. Academic Press, New York (1975).

[16] _____ "Two Classes of Covariance Matrices Giving Simple Linear Forecasts." RM-75-17, International Institute for Applied Systems Analysis, Laxenburg, Austria (May, 1975). To appear in Scandinavian Actuarial Journal.

[17] _____ "The Use of Collateral Data in Credibility Theory: A Hierarchical Model." RM-75-24, International Institute for Applied Systems Analysis, Laxenburg, Austria (June, 1975). To appear in Giornale dell' Istituto Italiano degli Attuari.

[18] Jewell, W.S. and Avenhaus, R. "Bayesian Inverse Regression and Discrimination: An Application of Credibility Theory." RM-75-27, International Institute for Applied Systems Analysis, Laxenburg, Austria (June, 1975).

[19] Lindley, D.V. and Smith, A.F.M. "Bayes Estimates for the Linear Model." J. Royal Statist. Soc., (B), 34, pp. 1-41 (1972).

[20] Malinvaud, E. Statistical Methods of Econometrics (2nd Revised Edition). North-Holland, Amsterdam (1970).

[21] Morales, J.A. Bayesian Full Information Structural Analysis. Springer-Verlag, Berlin (1971).

[22] Raiffa, H. and Schlaiffer, R. Applied Statistical Decision Theory. Harvard Business School, Boston (1961).

[23] Rao, C.R. Linear Statistical Inference and its Applications. J. Wiley & Sons, New York (1965).

[24] Rothenberg, T.J. Efficient Estimation with A Prior Information. Yale University Press, New Haven, Connecticut (1973).

[25] Taylor, G.C. "Credibility for Time-Heterogeneous Loss Ratios." Research Paper No. 55, MacQuarie University, Sydney, July, 1974. Proceedings of Actuarial Research Conference on Credibility Theory, Berkeley, California, September, 1974. Academic Press, New York (1975).

[26] _____ "Abstract Credibility." MacQuarie University, Sydney, and Herriot-Watt University, Edinburgh (February, 1975).

[27] Theil, H. "On the Use of Incomplete Prior Information in Regression Analysis." J. Amer. Statist. Assoc., 58, pp. 401-414 (1963).

[28] Theil, H. and Goldberger, A.S. "On Pure and Mixed Statistical Estimation in Economies." Intern. Econ. Rev., 2, pp. 65-78 (1961).

[29] Tiao, G.C. and Zellner, A. "Bayes Theorem and the Use of Prior Knowledge in Regression Analysis." Biometrika, 51, pp. 219-230 (1964).

[30] Tiao, G.C. and Zellner, A. "On the Bayesian Estimation of Multivariate Regression." J. Royal Statist. Soc., (B), 26, pp. 277-285 (1964).

[31] Tocher, K.D. "Discussion on Mr. Box and Dr. Wilson's Paper." J. Royal Statist. Soc., (B), 13, pp. 39-42 (1951).

[32] Zellner, A. An Introduction to Bayesian Inference in Econometrics. J. Wiley & Sons, New York (1971).

[33] _____ "Bayesian Analysis of Regression Error Terms." J. Amer. Statist. Assoc., 70, pp. 138-144 (1975).

[34] Zellner, A. and Chetty, V.K. "Prediction and Decision Problems in Regression Models from the Bayesian Point of View." J. Amer. Statist. Assoc., 60, pp. 608-616 (1965).