Regression, Prediction and Shrinkage
Author(s): J. B. Copas
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 45, No. 3 (1983), pp. 311-354
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2345402
Accessed: 25-02-2019 01:36 UTC

J. R. Statist. Soc. B (1983),
45, *No.* 3, *pp.* 311–354

Regression, Prediction and Shrinkage

By J. B. COPAS

University of Birmingham, UK

[*Read before the* Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, January 12th, 1983*, Professor R. N. Curnow *in the Chair*]

SUMMARY

The fit of a regression predictor to new data is nearly always worse than its fit to the original data. Anticipating this shrinkage leads to Stein-type predictors which, under certain assumptions, give a uniformly lower prediction mean squared error than least squares. Shrinkage can be particularly marked when stepwise fitting is used: the shrinkage is then closer to that expected of the full regression rather than of the subset regression actually fitted. Preshrunk predictors for selected subsets are proposed and tested on a number of practical examples. Both multiple and binary (logistic) regression models are considered.

Keywords: PREDICTION; SHRINKAGE; MULTIPLE REGRESSION; BINARY REGRESSION; STEIN ESTIMATION; EMPIRICAL BAYES

1. INTRODUCTION

A major use of regression models is to predict. Thus, given data on a response variable y and associated predictor variables x_i, our aim is to find a function of the x_i's which is, in some sense, a good predicator of y. We will be assuming throughout that the x_i's at which future predictions are required are not specified in advance but will occur randomly over some population of values, and that the success of a predictor can be judged by its average performance over such a population. In the context of a given regression model there is of course a superficial similarity between this problem of finding a predictor and the more familiar problem of estimation, in the sense that a particular predicator implies a particular vector of regression coefficients and *vice versa*. However, the loss functions for the two problems are different, and so a method for achieving a good predictor may be quite inappropriate for other questions in regression analysis such as the interpretation of individual regression coefficients or testing hypotheses about them. For example, a good predictor may include variables which are "not significant", exclude others which are, and may involve coefficients which are systematically biased.

In discussing the fit of a proposed predictor we distinguish between *retrospective fit* (fit to the original data) and *prospective* or *validation fit* (fit to new data). Since any assessment of retrospective fit "uses the data twice", it is obvious that it gives too optimistic a picture of the validation fit likely to be obtained on new data. We use the term *shrinkage* to denote the amount by which validation fit falls short of retrospective fit (precise interpretations of these terms will appear later). It turns out that shrinkage is greatly affected by empirical model selection such as stepwise regression or optimal subset methods. For if selection is achieved by maximizing some statistical measure of (retrospective) fit, then the parameter estimates in the resulting model will be biased precisely on that account. This bias leads to an increased tendency to overpredict and hence to increased shrinkage. The approach of this paper is to investigate the nature of shrinkage and to see how it can be anticipated in advance. Shrinkage is illustrated in Section 2 by way of examples, and an algebraic description for the standard multiple regression model is given in Section 3. This

Present address: Professor J. B. Copas, Dept of Statistics, The University, Box 363, Birmingham, B15 2TT, UK.

 0035–9246/83/45311 $2.00

motivates "preshrunk" predictors in Section 4 which, under certain assumptions, give a uniformly lower prediction mean squared error (PMSE) than least squares (LS). These predictors are closely related to those in Stein (1960) and Stone (1974); see also the review by Draper and van Nostrand (1979).

A corresponding empirical Bayes (EB) approach is given in Section 5. Sections 6 and 7 discuss the shrinkage of subset regressions, where it is seen that PMSE is often *not* improved by using empirical subset selection, although, as the number of variables increases, so does the need for preshrinking. The analogous situation for binary regression (when y is a dichotomy) is sketched briefly in Section 8, similar extensions to the wider class of generalized linear models (Nelder and Wedderburn, 1972) being also possible but not pursued here. Finally, Section 9 revisits the examples of Section 2 and gives a brief account of two further case studies.

2. SHRINKAGE BY EXAMPLE

Two examples are given, one of multiple regression, the other of binary (logistic) regression.

Example 1. *Parametric cost model.* Here the problem is to predict in advance the cost of an industrial project on the basis of data from similar projects undertaken in the past. Noah *et al.*, (1973) reported data on 31 aeroplanes and proposed a cost model using multiple regression. In this context, 31 is a very large sample size and, for obvious reasons, one is usually faced with the problem of fitting a cost model to a much smaller sample size. To illustrate, 8 aeroplanes have been chosen at random, the remaining 23 cases being used for validating the model. Here, y is log (cost per unit weight) and the x_i's are characteristics such as speed, wing area, etc., again measured on logarithmic scales. The data are published in full in the reference cited.

Using stepwise regression, two x_i's were chosen (weight and speed), one being significant at the 5 per cent level, the other almost so. Fig. 1 plots observed y against $\hat{y}$, the predicted value. The fit of the 8 selected cases to the line $y = \hat{y}$ is reasonable, as expected. Also as expected, the scatter of the 23 new cases is larger. But there is clear evidence of a lower slope for the new cases; the leftmost 6 points are above $y = \hat{y}$, the rightmost 3 points are all below. Predictions tend to be too extreme, the plot suggesting that a predictor of the form $\bar{y} + K(\hat{y} - \bar{y})$, with $K < 1$, would give a smaller sum of squared errors. It is worth noting that the shape of the plot remained rather similar when the number of x_i's in the regression was increased to 3 and 4, and that the same conclusions were evident when the exercise was repeated for other random selections of 8 cases.

Example 2. *Psychopath prediction.* A follow-up study of psychopaths discharged from a psychiatric hospital (Copas and Whiteley, 1976) defined a binary response, y, taking the values 1 (at least one reconviction or readmission within 3 years) and 0 (otherwise). Using logistic regression fitted by maximum likelihood (ML) to 91 observations, $P(y = 1)$ was estimated as a function of a number of predictive factors, x_i, available at the time of admission. Using stepwise fitting, 6 factors were selected for the logistic predictor, including two interaction terms which were found to be important. After the model was fitted, a further 2 years' experience yielded a second sample of observations of comparable size to the first.

The performance of a predictor in the binary case cannot be displayed as a scatter diagram such as Fig. 1, and so Copas and Whiteley (1976) reported a simple analysis in which the patients were divided into groups according to their values of the predicted probability. The predictions fitted very well retrospectively, but tended to be too extreme in the validation sample. If $\hat{z}$ denotes the predicted logit of $P(y = 1)$, then too many patients with large $\hat{z}$ had $y = 0$ and too many patients with small $\hat{z}$ had $y = 1$. This is illustrated in Fig. 2 which shows estimates of the actual value of $z = \text{logit}\{P(y = 1)\}$ as a function of $\hat{z}$ using the non-parametric binary regression method of Copas (1982). This method is based on estimating $P(y = 1)$ by a ratio of density estimates

$$\frac{\Sigma\, y_i\, \phi(h^{-1}(\hat{z} - \hat{z}_i))}{\Sigma\, \phi(h^{-1}(\hat{z} - \hat{z}_i))},$$

where y_i and $\hat{z}_i$ are the observed values of y and $\hat{z}$ respectively for the ith case, ϕ is the density of

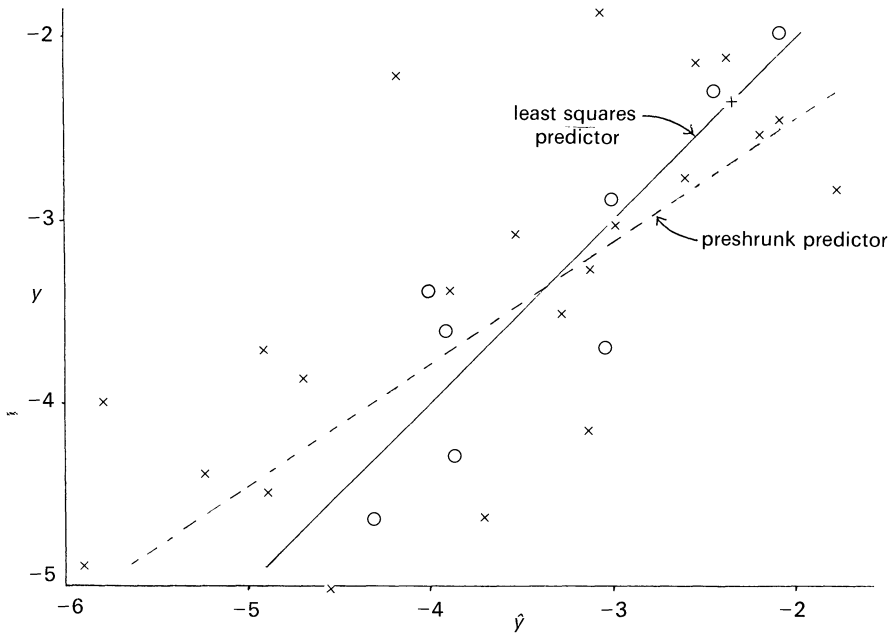

Fig. 1. Observed y against predicted $\hat{y}$ in construction sample (○) and in validation sample (×).

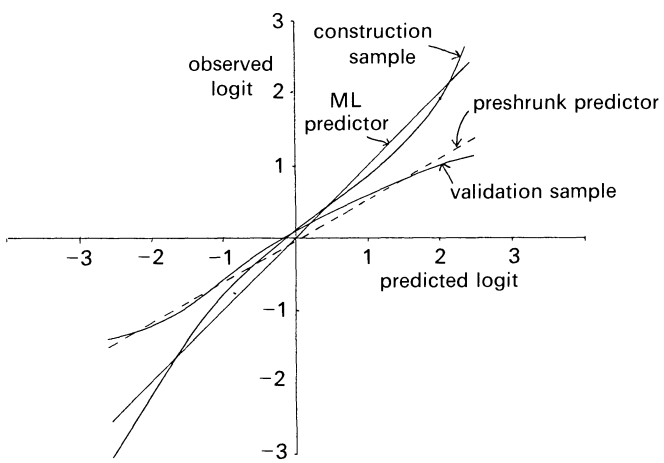

Fig. 2. Observed logit against predicted logit: non-parametric binary regressions.

$N(0, 1)$ and h is a suitably chosen smoothing constant. This was calculated separately for the two samples. The curve for the first sample is reasonably close to the line $z = \hat{z}$, indicating a good retrospective fit. The prospective curve is also reasonably straight, but is much flatter than the 45° line, indicating substantial shrinkage. Evidently, a predicted logit of the form $\bar{z} + K(\hat{z} - \bar{z})$ with $K < 1$ would give a better validation fit.

The dotted lines in Figs 1 and 2 will be described later in Section 9.

3. MULTIPLE REGRESSION AND SHRINKAGE

Let $\mathbf{x}$ be a vector of p predictive factors (or independent variables) and y a response (or dependent variable) given by the usual multiple regression model

$$y \mid \mathbf{x} \sim N(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}, \sigma^2).$$

For a sample of size n, called the *construction sample* (CS), we have the vector $\mathbf{y}$ of y's and the usual matrix $\mathbf{X}$ formed from the $\mathbf{x}$'s. For simplicity we will assume that the x_i's have been centred around their sample means so that the LS estimates are $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ and $\hat{\alpha} = \bar{y}$, the average of the y's. Let $\mathbf{V} = n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$, which is assumed to be of full rank. Throughout we will be conditioning on the observed $\mathbf{x}$'s in CS and so $\mathbf{V}$ can be taken as fixed.

We now suppose that, subsequent to CS, future values of $\mathbf{x}$ arise at random according to a multivariate probability distribution with mean $\mathbf{0}$ and variance–covariance matrix (var. matrix) $\mathbf{V}$, and that for each $\mathbf{x}$, y is generated from the same regression model as before. A sample of such future cases will be called a *validation sample* (VS). Note that multivariate normality of $\mathbf{x}$ will not usually be needed. This assumption that the *population* mean and var. matrix of future $\mathbf{x}$'s are the same as the *sample* mean and var. matrix of the $\mathbf{x}$'s in CS is a convenient idealization; an alternative model in which the $\mathbf{x}$'s in both CS and VS are generated randomly from the same (but unknown) distribution is mentioned in Section 4. The differences between the two approaches are unimportant if n is large.

Let $\hat{y}$ be the LS predictor

$$\hat{y} = \hat{\alpha} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}.$$

Then for a typical case $(y, \mathbf{x})$ in VS, the conditional bivariate distribution of $(y, \hat{y})^{\mathrm{T}}$ given CS (i.e. given $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$) will have mean $(\alpha, \hat{\alpha})^{\mathrm{T}}$ and var. matrix

$$\begin{pmatrix} \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} + \sigma^2 & \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} & \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}} \end{pmatrix}, \tag{3.1}$$

whose expectation over CS is

$$\begin{pmatrix} \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} + \sigma^2 & \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} \\ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} & \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} + n^{-1}p\sigma^2 \end{pmatrix}. \tag{3.2}$$

From (3.1), the LS slope of y on $\hat{y}$ is

$$K = \frac{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta}}{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}}}. \tag{3.3}$$

Now, from (3.2), the denominator of K is on average larger than the numerator; the expectation of K is in fact strictly less than one. We take (3.3) to be an index of shrinkage, $K = 1$ denoting no shrinkage, small K denoting substantial shrinkage. For given p, the distribution of K (with respect to variation in $\hat{\boldsymbol{\beta}}$) depends on a quantity δ given by

$$\delta^2 = \frac{\sigma^2}{n\,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta}}. \tag{3.4}$$

The distribution of K becomes more concentrated about $K = 1$ as $\delta \to 0$. Conversely, K is likely to be small if p is not small relative to n and/or the signal/noise ratio measured by $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta}/\sigma^2$ is small. These latter circumstances are characteristic of much research in the medical and social sciences.

We now introduce the orthogonalizing transformation given by the $p \times p$ matrix $\mathbf{M}$ with

$$\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{V}, \quad \mathbf{M}\mathbf{V}^{-1}\mathbf{M}^{\mathrm{T}} = \mathbf{I}, \tag{3.5}$$

and define $\boldsymbol{\xi} = \mathbf{M}\boldsymbol{\beta}$. Then $\mathbf{M}\hat{\boldsymbol{\beta}}$ can be expressed as $\boldsymbol{\xi} + \sigma n^{-\frac{1}{2}}\mathbf{u}$, where $\mathbf{u}$ is a vector of p i.i.d. $N(0, 1)$ deviates. Writing (3.3) and (3.4) in terms of $\boldsymbol{\xi}$ and $\mathbf{u}$, and noting that $\mathbf{u}^{\mathrm{T}}\mathbf{u} - (\boldsymbol{\xi}^{\mathrm{T}}\mathbf{u})^2(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi})^{-1}$ and $\boldsymbol{\xi}^{\mathrm{T}}\mathbf{u}(\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi})^{-\frac{1}{2}}$ are independently distributed as χ^2_{p-1} and $N(0, 1)$ respectively, shows that K is distributed as

$$\frac{1 + u\delta}{(1 + u\delta)^2 + \chi^2_{p-1}\delta^2}, \tag{3.6}$$

where u is $N(0, 1)$ and is independent of χ^2_{p-1}, as χ^2 deviate on $p - 1$ d.f.

The preshrunk predictors of the next section require that K be estimated. Consider the family of estimates

$$\hat{K}(k) = \frac{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}} - n^{-1}k\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}}} = \frac{F - p^{-1}k}{F}, \tag{3.7}$$

where

$$F = n\,\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}}/p\hat{\sigma}^2 \tag{3.8}$$

is the usual F-ratio and $\hat{\sigma}^2$ is the residual mean square. Although (3.1) to (3.3) suggest that k should be fixed at p, we leave k as a free argument of $\hat{K}$ as different values for it will be suggested in this and later sections. The work leading to (3.6) shows that $\hat{K}(k)$ is distributed as

$$1 - \frac{k\nu^{-1}\chi^2_\nu\delta^2}{(1 + u\delta)^2 + \chi^2_{p-1}\delta^2}, \tag{3.9}$$

where ν is the residual d.f. given by $\nu = n - p - 1$ and χ^2_ν is a χ^2 deviate on ν d.f. which is independent of both u and χ^2_{p-1}.

The denominator of both (3.6) and (3.9) is proportional to a non-central χ^2 deviate on p. d.f. with non-centrality parameter δ^{-2}. The result in Johnson (1959) and Kerridge (1965) shows that this distribution can be represented as a central χ^2 on $p + 2g$ d.f., where g has a Poisson distribution with mean $\frac{1}{2}\delta^{-2}$. It follows that

$$E(\hat{K}(k)) = 1 - E\left(\frac{k}{p - 2 + 2g}\right). \tag{3.10}$$

An extension of this argument following James and Stein (1962), see also Lemma 1 of Baranchik (1973), shows that

$$E(K) = 1 - E\left(\frac{p - 2}{p - 2 + 2g}\right). \tag{3.11}$$

Similar expressions for the higher moments of both K and $\hat{K}(k)$ can also be obtained—these show that moments of each quantity exist only up to order $\frac{1}{2}(p - 1)$. For the expectations to exist we must therefore have $p \geqslant 3$ which will be assumed throughout.

Immediate consequences of (3.10) and (3.11) are that $E(K) < 1$ and that $k = p - 2$ gives an unbiased estimate in the sense that $E(\hat{K}(p - 2) - K) = 0$. A series expansion of (3.10) in powers of δ^2 leads to the simple approximation

$$E(K) \simeq \frac{1 - 2\delta^2}{1 + (p - 4)\,\delta^2}. \tag{3.12}$$

Some idea of the magnitude of expected shrinkage is given by the values of $E(K)$, or equivalently of $E\{(\hat{K}(p - 2)\}$, given in Table 1, these values being found by direct numerical summation of the relevant Poisson series. Now (3.4) and (3.8) suggest that values of δ^2 and p correspond to an F-ratio of $F^* = 1 + (p\delta^2)^{-1}$, and so values of F^* are also shown to aid interpretation of the table. As expected, the higher the F^* the less the shrinkage. The practical value of

a regression with an F-ratio as small as the lower entries in Table 1 must be open to doubt, and if such cases are discounted it can be seen from the last column of the table that the simple approximation (3.12) is reasonably accurate.

TABLE 1
Expected shrinkage

δ^2	p	F^*	$E(K)$	(3.12)
0.01	5	21	0.970	0.970
	10	11	0.925	0.925
	20	6	0.845	0.845
0.05	5	5	0.858	0.857
	10	3	0.698	0.692
	20	2	0.513	0.500
0.10	5	3	0.735	0.727
	10	2	0.526	0.500
	20	1.5	0.341	0.308
0.20	5	2	0.554	0.500
	10	1.5	0.348	0.273
	20	1.25	0.203	0.143

Although we are examining shrinkage in terms of quantities such as K and $\hat{K}$ it is worth noting that many other measures of shrinkage are possible. Section 7 below studies shrinkage in terms of PMSE and in terms of correlation coefficients. Gardner (1972) discusses a "ratio bias" which measures the proportional increase in residual sum of squares when an old predictor is applied to new data, and a similar quantity is also examined in Nicholson (1960).

4. PRESHRUNK PREDICTORS

Now (3.1) shows that the linear function of $\hat{y}$ which is closest to y in the LS sense is $\alpha + K\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$, suggesting that y should be predicted by

$$\tilde{y} = \hat{\alpha} + \hat{K}\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}. \tag{4.1}$$

This predictor, assuming a suitable constant k is chosen in the argument of $\hat{K}$, might be called *preshrunk* in the sense that the average value of y for any given $\tilde{y}$ will be approximately equal to $\tilde{y}$. This contrasts with the LS predictor $\hat{y}$ which tends to overestimate large values of y and underestimate small values of y. Note that the fact that we are averaging over the future $\mathbf{x}$ is crucial to the argument; if we condition on $\mathbf{x}$ instead of on $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$ then $\hat{y}$ is the minimum variance unbiased predictor by the usual properties of least squares.

The overall PMSE of (4.1) is

$$E(y-\tilde{y})^2 = \sigma^2(1+n^{-1}) + E\{(\hat{K}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}(\hat{K}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}. \tag{4.2}$$

Using (3.7) and the fact that $\nu\hat{\sigma}^2/\sigma^2$ is distributed as χ^2 on ν d.f. and is independent of $\hat{\boldsymbol{\beta}}$, (4.2) becomes

$$\sigma^2\left(1+\frac{p+1}{n}\right) - \frac{2k\sigma^2}{n}E\{L(k)\}, \tag{4.3}$$

where

$$L(k) = 1 - (\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}})^{-1}\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}} + \frac{k\sigma^2}{2n}\left(1+\frac{2}{\nu}\right)\right).$$

Noting the similarity between $L(k)$ and K and $\hat{K}(k)$ in the last section, (4.3) can be written as

$$\sigma^2\left(1+\frac{p+1}{n}\right) - \frac{2k\sigma^2}{n}(p-2-\tfrac{1}{2}k(1+2\nu^{-1}))E\left(\frac{1}{p-2+2g}\right). \tag{4.4}$$

Now if $k = 0, \tilde{y} \equiv \hat{y}$ and (4.4) reduces to

$$\sigma^2 \left(1 + \frac{p+1}{n} \right), \tag{4.5}$$

which is the usual formula for the PMSE of LS (Seber, 1977, p. 369). Hence, comparing (4.4) with (4.5), the PMSE of $\tilde{y}$ is less than that of $\hat{y}$ provided

$$0 < k < \frac{2(p-2)}{1 + 2\nu^{-1}}, \tag{4.6}$$

and (4.4) is least when k is at the mid-point of this range, namely

$$\frac{p-2}{1 + 2\nu^{-1}}. \tag{4.7}$$

The difference between this and the value $k = p - 2$ suggested in the last section is small if ν is large, and the difference between the resulting PMSEs is even smaller owing to the quadratic dependence on k in (4.4). The value $k = p - 2$ always belongs to the improvement region (4.6) provided $\nu > 2$.

The PMSE of the optimum predictor (with k equal to (4.7)) can be deduced immediately from (4.4) and the corresponding expected shrinkage $E(K)$ discussed earlier, since it turns out that

$$E\left\{ L\left(\frac{p-2}{1 + 2\nu^{-1}} \right) \right\} = \tfrac{1}{2}(1 - E(K)).$$

Preshrunk predictors are closely related to so-called Stein estimates. The simplest situation in Stein estimation is that of a vector $\mathbf{T} = (T_1, T_2, \ldots, T_p)^{\mathrm{T}}$ of independent random variables, where T_i is $N(\mu_i, \sigma^2)$. James and Stein (1961) showed that when $p \geqslant 3$, the estimate

$$\hat{\boldsymbol{\mu}} = \left(1 - \frac{(p-2)\sigma^2}{\mathbf{T}^{\mathrm{T}}\mathbf{T}} \right) \mathbf{T}$$

gives a uniformly lower expected value of the quadratic loss function

$$(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \tag{4.8}$$

than does $\hat{\boldsymbol{\mu}} = \mathbf{T}$. If σ^2 is unknown but is estimated by an independent mean square $\hat{\sigma}^2$ based on ν d.f., James and Stein proposed

$$\hat{\boldsymbol{\mu}} = \left(1 - \frac{(p-2)\hat{\sigma}^2 \nu}{(\nu+2)\,\mathbf{T}^{\mathrm{T}}\mathbf{T}} \right) \mathbf{T}, \tag{4.9}$$

which likewise dominates maximum likelihood (ML). A simple interpretation of these estimates is to note that $E(\mathbf{T}^{\mathrm{T}}\mathbf{T}) = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} + p\sigma^2$, and so, on average, the ML vector is further away from the origin than the true vector, and so should be scaled by some factor less than one.

Returning to the regression problem, we have already seen, following (3.5), that $\hat{\boldsymbol{\xi}} = \mathbf{M}\hat{\boldsymbol{\beta}}$ has the spherical multivariate normal distribution $N(\boldsymbol{\xi}, n^{-1}\sigma^2\mathbf{I})$, and so $\hat{\boldsymbol{\xi}}$ has the same distribution as $\mathbf{T}$ but with $\boldsymbol{\xi}$ replacing $\boldsymbol{\mu}$ and $n^{-1}\sigma^2$ replacing σ^2. Applying (4.9) and then transforming back leads to $\boldsymbol{\beta}$ being estimated by

$$\tilde{\boldsymbol{\beta}} = \left(1 - \frac{(p-2)\,\hat{\sigma}^2 \nu}{n(\nu+2)\,\hat{\boldsymbol{\xi}}^{\mathrm{T}}\hat{\boldsymbol{\xi}}} \right) \hat{\boldsymbol{\beta}}.$$

It is easy to see that the scaling factor in this is precisely equal to $\hat{K}$ in (3.7) with k chosen as (4.7). The corresponding estimate of $\hat{\alpha} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ is then just $\tilde{y}$ in (4.1). It is important to note that

the loss function also transforms to

$$(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})^{\mathrm{T}}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{4.10}$$

where $\hat{\boldsymbol{\beta}}$ here stands for any arbitrary estimate of $\boldsymbol{\beta}$. By comparing with (4.2), we see that minimizing (4.10) is equivalent to minimizing PMSE.

Whether the Stein estimate is a mere mathematical curiosity or whether it is useful in practice has always been controversial. It seems nonsensical that an estimate of one parameter should be influenced by the data in apparently unrelated problems. The loss function (4.8), however, does impose an artificial link between the components in the sense that some compensation between the different estimation errors is allowed. Stein estimation is only justified when such compensation of errors makes practical sense. Typically, the method does better than ML in the majority of components, but worse (and often much worse) in a minority of components. This has led to a number of papers (starting with Efron and Morris, 1971) proposing modified Stein estimates which limit the possibility of gross errors. However, in the prediction problem of this paper such arguments are irrelevant, since compensation of errors in regression coefficients arises naturally through the use of PMSE.

Another point of contention about the Stein estimate is the arbitrary choice of origin towards which the shrinkage is directed; this has led to suggestions that the shrinkage should be made towards the grand mean rather than towards zero (Lindley, 1962). But again such considerations are unnecessary in the prediction situation, as zero is a natural origin for a regression coefficient (in fact it constitutes the standard null hypothesis in regression analysis).

Thus the estimation of shrinkage, a phenomenon clearly observed in practice (e.g. Fig. 1), provides a simple motivation for Stein estimation, a motivation which seems lacking in much of the literature on this topic. A notable exception is the important paper by Stone (1974), who shows how shrinkage estimates and predictors are suggested by cross-validation. Stone's multiple regression predictor does not quite correspond to those developed here but is similar to $\tilde{y}$ with k given by (4.7) if n is large relative to p. Hjorth and Holmqvist (1981) give an interesting case study of cross-validation in time series prediction with particular reference to order selection, which is related to the topic of Sections 6 and 7 below. Another argument leading to Stein-type estimates in regression is given by Narula (1974).

A useful modification to the Stein estimate which has been widely discussed in the literature is the so-called "positive part" estimate. In our context, this amounts to replacing $\hat{K}$ by $\hat{K}_+ = \max(\hat{K}, 0)$. Sclove (1968) shows that, in general, the estimate $\hat{K}_+\hat{\boldsymbol{\beta}}$ gives a lower expected value of the loss (4.10) than does $\hat{K}\hat{\boldsymbol{\beta}}$. With k as in (4.7), $\hat{K}$ is negative when F is less than k/p, which is in turn less than one, an eventuality which could hardly lead to useful prediction. If this happens, $\hat{K}_+$ is zero and all cases are predicted by the overall mean, which accords with common sense.

It has already been mentioned that methods leading to good predictors may be quite unsuitable for the problem of estimating $\boldsymbol{\beta}$. This is one reason why there has been such conflicting results from the simulation studies which have attempted to compare Stein and ridge methods with LS (e.g. Dempster *et al.*, 1977, and others reviewed in Draper and Van Nostrand, 1979). Most authors have been concerned with comparing estimates using loss functions such as (4.8); had attention been confined to (4.10), or PMSE, such simulations could only confirm that $\tilde{y}$ is uniformly better than LS.

Finally in this section we comment on the assumption that the mean and var. matrix of future samples $\mathbf{x}$ exactly match the corresponding sample moments in CS. An alternative formulation is to suppose that the $\mathbf{x}$'s in both CS and VS arise at random according to some unknown mean $\boldsymbol{\eta}$ and var. matrix $\mathbf{V}_0$, with the true regression line being $y = \alpha + \boldsymbol{\beta}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\eta})$. To see how this works out for LS, we temporarily abandon the convention of subtracting the means from the x_i's so that $\hat{y} = \hat{\alpha} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}(\mathbf{x} - \bar{\mathbf{x}})$, $\bar{\mathbf{x}}$ being the sample mean vector in CS. Averaging over future $\mathbf{x}$ and also over the

$\mathbf{x}$'s in CS, the overall PMSE of $\hat{y}$ is

$$\sigma^2 + E\{\alpha - \hat{\alpha} + \hat{\boldsymbol{\beta}}^T(\bar{\mathbf{x}} - \boldsymbol{\eta})\}^2 + E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{V}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{4.11}$$

Using the fact that $\hat{\alpha}$ $(= \bar{y})$ has mean $\alpha + \boldsymbol{\beta}^T(\bar{\mathbf{x}} - \boldsymbol{\eta})$ and is independent of $\hat{\boldsymbol{\beta}}$, (4.11) simplifies to

$$\sigma^2(1 + n^{-1})\{1 + E(\text{trace}\,(\mathbf{V}^{-1}\mathbf{V}_0)\}.$$

If we now additionally assume that $\mathbf{x}$ is multivariate *normal*, results on the distribution of the inverse of a sums-of-squares-and-products matrix (Anderson, 1958, p. 85) show that the above expected trace is $p/(\nu - 1)$. Hence the PMSE is

$$\sigma^2\left(1 + \frac{n(p+1) - 2}{n(\nu - 1)}\right). \tag{4.12}$$

This is greater than the corresponding formula (4.5) obtained earlier since we are now allowing for the sampling variability in $\mathbf{V}$. Note that (4.12) can also be obtained using the analysis given in Gardner (1972), and in Narula (1974). Now (4.12) is less than

$$\sigma^2\left(1 + \frac{p+1}{\nu - 1}\right)$$

which is like (4.5) but with n replaced by $\nu - 1$. This suggests that overall PMSEs in this more general setting may be similar to those calculated in the simpler situation but with n reduced by $(p + 1)$. This is negligible if n is large relative to p but can be important for small data sets. For example, if $\nu = 1$ the overall PMSE of LS is infinite.

Stein (1960) also obtained an expression equivalent to (4.12), and then went on to consider the admissibility of LS. Under the artificial restriction that $\boldsymbol{\eta}$ and α are known, Stein showed that when $p \geqslant 3$, a predictor similar to (4.1) dominates LS in the sense of overall PMSE. His results imply that the optimum value of k in (3.7) is $(p - 2)/(1 + 3\nu^{-1})$ if $\boldsymbol{\beta}$ is zero but $(p - 2)/(1 + 2(\nu + 1)^{-1})$ if $\boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta}$ is large. The value (4.7) is in between these two, but all approximate $(p - 2)$ if ν is large. (See also Baranchik (1973) who extended Stein's earlier paper). Thus, although the simplifying assumptions about $\boldsymbol{\eta}$ and $\mathbf{V}_0$ made in this paper lead to under-estimates of the prediction errors if n is small, the arguments for preshrinking remain.

5. EMPIRICAL BAYES PREDICTORS

The fact that Stein estimates can be motivated by a Bayesian argument is well known (Stein, 1962; Lindley, 1962). In the multiple regression context, suppose that $(\alpha, \boldsymbol{\beta})$ has some prior distribution, but that (for the moment) σ^2 is known. Then to minimize PMSE, y should be predicted by the posterior mean of $\alpha + \boldsymbol{\beta}^T\mathbf{x}$. This takes the preshrunk form

$$\hat{\alpha} + \kappa\,\hat{\boldsymbol{\beta}}^T\mathbf{x} \tag{5.1}$$

when the prior on α is vague and that on $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim N\left(\mathbf{0}, \frac{\sigma^2\kappa}{n(1-\kappa)}\,\mathbf{V}^{-1}\right), \tag{5.2}$$

where $0 < \kappa < 1$. Thus $\tilde{y}$ in (4.1) can be thought of as a Bayes predictor in which the role of κ is taken by the statistic $\hat{K}$ in (3.7).

To motivate this particular estimate of κ note that the marginal distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \sim N\left(\mathbf{0}, \frac{\sigma^2}{n(1-\kappa)}\,\mathbf{V}^{-1}\right), \tag{5.3}$$

and so the quantity $(p - 2)\sigma^2/(n\,\hat{\boldsymbol{\beta}}^T\mathbf{V}\hat{\boldsymbol{\beta}})$ is a (marginally) unbiased estimate of $1 - \kappa$. If σ^2 is replaced by $\hat{\sigma}^2$, the corresponding estimate of κ is just $\hat{K}(p - 2)$ in (3.7). When σ^2 is unknown but

is given the usual vague prior distribution (the prior for $\boldsymbol{\beta}$ in (5.2) is then conditional on σ^2) the same estimate of κ is also (marginally) unbiased, although in this case the previous results suggest that the slightly larger estimate given by k in (4.7) is to be preferred.

Since $\tilde{y}$ involves replacing a parameter of a prior distribution by a sample estimate, it is natural to describe it as an *empirical Bayes* (EB) predictor. This usage of the term is rather less specialized than the usual one (e.g. Maritz, 1970) in which the prior distribution is estimated from past occurrences of the same situation. However, (5.2) implies that the components of $\mathbf{M}\boldsymbol{\beta}$ are independent and identically distributed and so the essential feature of having independent repetitions of the same decision problem is present even within the confines of the one set of data in CS.

If κ were known the PMSE of (5.1) would be

$$\sigma^2 \left(1 + \frac{1 + p\kappa}{n} \right). \tag{5.4}$$

This formula applies both in the sense of posterior expectation (given CS) and also in the sense of overall PMSE. But if the role of κ is taken by the sample estimate $\hat{K}(k)$ then the corresponding expression

$$\sigma^2 \left(1 + \frac{1 + p\hat{K}(k)}{n} \right) \tag{5.5}$$

tends to underestimate the true PMSE of $\tilde{y}$ since it ignores the variability in $\hat{K}$. For the expectation of (5.5) is, from (3.10),

$$\sigma^2 \left(1 + \frac{p+1}{n} \right) - \frac{\sigma^2 kp}{n} \, E\left(\frac{1}{p - 2 + 2g} \right), \tag{5.6}$$

which is less than the overall PMSE of $\tilde{y}$ given in (4.4). A simple device for overcoming this is to adjust the value of k to be used in (5.5). Equating (5.6) with (4.4) leads to

$$k = \frac{(p-2)^2}{p(1 + 2\nu^{-1})} \, . \tag{5.7}$$

This is less than (4.7) but greater than $(p-4)/(1+2\nu^{-1})$, which is (4.7) with p taken to be two less than the actual value. Thus, whilst k in (4.7) should be used for the construction of the predictor, the Bayes formula (5.5) gives a better approximation to the overall PMSE if k is taken to be the somewhat smaller value of (5.7). The difference between these two values becomes less important the larger is p.

Since $\tilde{y}$ is guaranteed to give a lower PMSE than LS it can be argued that the question of whether one "believes" in the family of prior distributions in (5.2) is irrelevant. So far this Section has merely pointed out that *if* we start with (5.2) *then* there exists a reasonable EB argument which leads to $\tilde{y}$. However, the fact that a Bayesian argument involves averaging the risk function over the assumed prior distribution suggests that the improvement in $\tilde{y}$ over $\hat{y}$ is likely to be greatest when $\boldsymbol{\beta}$ is "typical" of (5.2). For instance, the zero mean of (5.2) indicates that one should expect about as many positive regression coefficients as negative ones, and that the sign of each β_i is uncertain prior to the data. In scientific experiments this would be absurd, but in many practical applications of regression analysis the nature of the contribution of each x_i may be a matter for speculation, and even when the sign of the marginal contribution is obvious from the context of the problem the sign of the partial coefficient may not be (e.g. aircraft cost is obviously positively correlated with wing area, but what is the partial correlation after correcting for weight?). The form of the var. matrix in (5.2) indicates that for any vector of constants, $\mathbf{d}$, the likely departure of $\mathbf{d}^{\mathrm{T}}\boldsymbol{\beta}$ from zero is proportional to the standard error with which it could be estimated from a set of data similar to CS.

Some test of the agreement between CS and (5.2) is given by transformation (3.5), which, together with (5.3), gives

$$\mathbf{M}\hat{\boldsymbol{\beta}} \sim N\left(\mathbf{0}, \frac{\sigma^2}{n(1-\kappa)}\mathbf{I}\right). \tag{5.8}$$

Thus the elements of $\mathbf{M}\hat{\boldsymbol{\beta}}$ should be a random sample of size p from a normal distribution with mean zero, which can be examined directly by a normal plot. This informal graphical test is somewhat akin to the "predictive checks" discussed in Box (1980). It is useful to compare the plot from (5.8) with the line corresponding to $N(0, n^{-1}\hat{\sigma}^2)$ representing the "noise level": the ratio of the variance (given by the slope of the plot) to $n^{-1}\hat{\sigma}^2$ is just F in (3.8). A marked departure of the plot from linearity can indicate the need for a transformation; for instance, in the aircraft example cost is dominated by an overall size effect but the plot is reasonable if y is scaled to be cost per unit weight (see Section 9).

It should be emphasized that this graphical test suffers from three disadvantages. Firstly, $\mathbf{M}$ is not unique for given $\mathbf{V}$, as it can be multiplied by an arbitrary orthogonal matrix. The shape of the plot is not invariant under such a multiplication, although the test is valid provided $\mathbf{M}$ is chosen on the basis of the x_i's and not the y's in CS. Secondly, if F is not much greater than one, the shape of the plot will be influenced more by the normality of the sampling distribution than by the configuration of the β_i's. Thirdly, if p is small, a normal plot is very insensitive as a test of goodness of fit. (For this reason, a half normal plot might be better in such cases.)

6. LEAST SQUARES AND EMPIRICAL SELECTION

The standard formula (4.5) for the PMSE of LS shows that improved prediction might be possible by reducing the dimension of the x_i's. Mallows' C_p (Gorman and Toman, 1966) in fact operates by estimating a simple transformation of (4.5) for various subsets, from which an optimum selection can be made. Many other methods of subset selection have also been proposed, some of which are asymptotically equivalent to C_p (Shibata, 1981), and nearly all statistical packages include at least one version of stepwise variable selection. But the usual properties of LS are invalid when a subset is selected on the basis of the data. In the simplest methods, for instance, x_i is selected if $|\hat{\beta}_i|$ exceeds some value (e.g. significant at some nominal level), and so is more likely to be selected if $|\hat{\beta}_i|$ overestimates $|\beta_i|$ than if $|\hat{\beta}_i|$ underestimates $|\beta_i|$. Thus the coefficients for a selected subset will be biased, as a result of which the usual measures of fit will be too optimistic, sometimes markedly so. Unfortunately the sampling properties of such methods are very complicated, although some tentative results for one aspect of the problem are given in Rencher and Pun (1980), who also reference other related work. In this Section we simplify the analysis by orthogonalizing the problem so that selection is made from amongst the principal components of the x_i's, this being equivalent to selection on the x_i's themselves only if $\mathbf{V}$ is diagonal.

Using the transformation (3.5), let $\boldsymbol{\tau} = n^{\frac{1}{2}}\sigma^{-1}\mathbf{M}\boldsymbol{\beta}$ and $\hat{\boldsymbol{\tau}} = n^{\frac{1}{2}}\sigma^{-1}\mathbf{M}\hat{\boldsymbol{\beta}}$. These are the vectors of standardized orthogonal regression coefficients (for given σ), with $\hat{\tau}_i \sim N(\tau_i, 1)$. Suppose that the component corresponding to τ_i is selected if $i \in J(\mathbf{y})$ and omitted otherwise. Then the PMSE of the resulting LS predictor is

$$\frac{\sigma^2}{n}\{n + 1 + E(\Sigma'(\hat{\tau}_i - \tau_i)^2 + \Sigma''\tau_i^2)\}, \tag{6.1}$$

where Σ' denotes summation over $i \in J(\mathbf{y})$ and Σ'' denotes summation over $i \notin J(\mathbf{y})$. If selection were fixed in advance, i.e. if $J(\mathbf{y})$ were independent of $\mathbf{y}$, then (6.1) would be least when J includes precisely those i with $\tau_i^2 > 1$. This suggests that $J(\mathbf{y})$ should be chosen to include i with $\hat{\tau}_i^2 > 1$ or, as $E(\hat{\tau}_i^2) = \tau_i^2 + 1$, with $\hat{\tau}_i^2 > 2$. More generally we set

$$J(\mathbf{y}) = \{i: |\hat{\tau}_i| > c\} \tag{6.2}$$

for some constant c, in which case (6.1) can be shown to equal

$$\frac{\sigma^2}{n}\ (n + 1 + \Sigma\ G_c(\tau_i)), \tag{6.3}$$

where

$$G_c(\tau) = 1 + (\tau^2 - 1)\,(\Phi(c + \tau) - \Phi(\tau - c)) + (c + \tau)\,\phi\,(c + \tau) - (\tau - c)\,\phi\,(\tau - c),$$

ϕ and Φ denoting the density and distribution functions of $N(0, 1)$ respectively. If $c \to 0$, when all variables are included, (6.3) tends to (4.5) as expected. If $c \to \infty$, when all variables are omitted, (6.3) tends to the total mean square $\sigma^2 n^{-1}(n + 1 + \boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\tau})$.

In order to compare (6.3) for different $\boldsymbol{\beta}$'s, it is sensible to keep these two limiting values fixed. For example, let $n = 50$, $p = 5$, $\sigma^2 = 1$ and $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} = 0.20$, these implying that $\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\tau} = 10$ and that the multiple correlation coefficient is about $\frac{1}{2}$. Within these constraints, (6.3) is minimized when $\boldsymbol{\tau}$ has one element $\sqrt{10}$ and the others zero (when a single component is most likely to be selected), and is maximized when $\boldsymbol{\tau}$ has all elements equal to $\sqrt{2}$ (when each component is equally likely to be selected). The corresponding values of (6.3) are shown as the solid lines in Fig. 3. Curves for other $\boldsymbol{\beta}$'s are somewhere in between, e.g. the dashed line is for $\boldsymbol{\tau}$ consisting of two elements equal to $\sqrt{5}$ with the others zero. It is clear from Fig. 3 that in many, perhaps most, situations LS with empirical selection gives a worse PMSE than fitting the whole regression, and can even be worse than omitting the regressors altogether. To be better than the full regression, we evidently need a $\boldsymbol{\tau}$ with very disparate elements, and a value of c which is not too large—in this case c should be no greater than about 2.7.

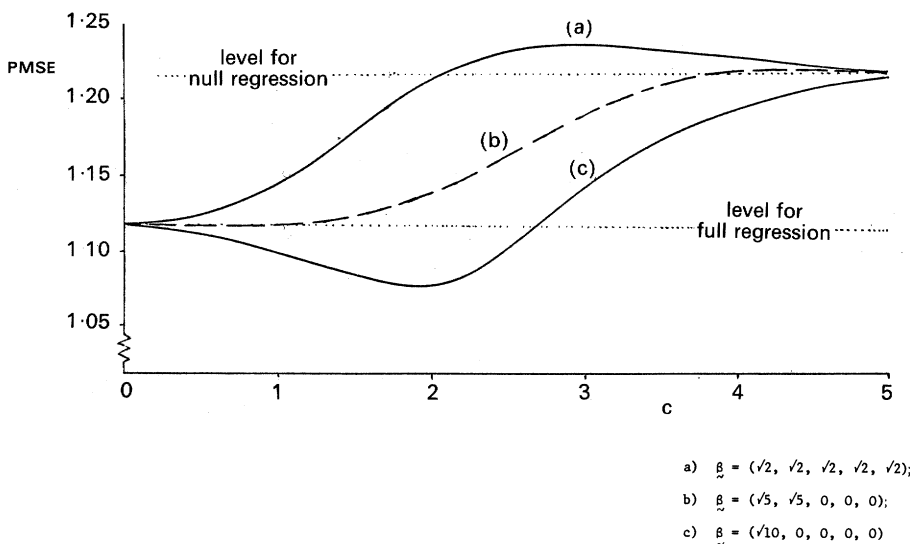

Fig. 3. Prediction mean squared error for selecting components with $|\hat{\tau}| > c$.
(a) $\boldsymbol{\beta} = (\sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{2})$; (b) $\boldsymbol{\beta} = (\sqrt{5}, \sqrt{5}, 0, 0, 0)$; (c) $\boldsymbol{\beta} = (\sqrt{10}, 0, 0, 0, 0)$.

In principle, curves such as those in Fig. 3 could be constructed for the more realistic procedure of selection based on the $\hat{\beta}_i$'s rather than on the $\hat{\tau}_i$'s. Simulations of situations with various $\mathbf{V}$'s and $\boldsymbol{\beta}$'s consistent with the above constraints have been attempted, and whilst the resulting values of PMSE do not always lie between the two solid lines in Fig. 3, the qualitative conclusion remains the same, namely that empirical selection can be better than the full regression but is often worse, sometimes considerably so.

We turn now to the shrinkage of predictors based on empirical selection. If selection is made on the basis of (6.2), the analogue of K in (3.3) is

$$K_1 = \frac{\Sigma' \hat{\tau}_i \tau_i}{\Sigma' \tau_i^2} \quad . \tag{6.4}$$

Had the dependence of $J(\mathbf{y})$ on $\mathbf{y}$ been ignored, this would have been estimated by the analogue of (3.7) for the subset regression actually fitted, namely $\hat{K}(k)$ with k taken as $s-2$, $s(=s(\mathbf{y}))$ being the number of i's in $J(\mathbf{y})$ and with the role of $\hat{\sigma}^2$ taken by the residual mean square for the reduced regression. This gives the estimate $\hat{K}_2$ where

$$\hat{K}_2 \Sigma' \hat{\tau}_i^2 = \Sigma' \hat{\tau}_i^2 - (s-2)\left(1 + \frac{\Sigma'' \hat{\tau}_i^2 - (p-2)}{n-s-1}\right). \tag{6.5}$$

From (6.2), $\Sigma'' \hat{\tau}_i^2 < c^2(p-s)$, and so

$$(\hat{K}_2 - K_1)\Sigma' \hat{\tau}_i^2 > \Sigma' (\hat{\tau}_i^2 - \hat{\tau}_i \tau_i - 1) + 2 - (n-s-1)^{-1}(s-2)(p-s)(c^2-1). \tag{6.6}$$

The expectation of the summation immediately after the inequality in (6.6) is equal to

$$c\,\Sigma\{\phi(c+\tau_i) + \phi(c-\tau_i)\},$$

which is positive. Hence, if $0 < c < 1$, the expectation of the right-hand side of (6.6) exceeds 2, and hence is certainly positive. If $c > 1$, this expectation is positive if n is sufficiently large relative to p. In either case the suggestion is that $\hat{K}_2$ overestimates K_1.

A very rough idea of the relative magnitudes of K_1 and $\hat{K}_2$ is given by replacing the random quantities in (6.4) and (6.5) by their expectations. We also assume that n is large so that the last term in (6.5) can be ignored. Now it can be shown that

$$E(s) = \Sigma A_c(\tau_i), \quad E(\Sigma' \hat{\tau}_i \tau_i) = \Sigma B_c(\tau_i), \quad E(\Sigma' \hat{\tau}_i^2) = \Sigma C_c(\tau_i),$$

where

$$A_c(\tau) = \Phi(-c-\tau) + \Phi(-c+\tau),$$
$$B_c(\tau) = \tau\{\tau A_c(\tau) - \phi(c+\tau) + \phi(c-\tau)\}$$

and

$$C_c(\tau) = (\tau^2+1)A_c(\tau) + (c-\tau)\phi(c+\tau) + (c+\tau)\phi(c-\tau).$$

Put

$$D_c(\tau) = B_c(\tau)/C_c(\tau), \quad H_c(\tau, p) = (C_c(\tau) - A_c(\tau) + 2p^{-1})/C_c(\tau).$$

Then making the appropriate replacements in (6.4) and (6.5) gives K_1 as a weighted average of $D_c(\tau)$ and $\hat{K}_2$ as a weighted average of $H_c(\tau, p)$ where both the averages are over $\tau = \tau_1, \tau_2, \ldots, \tau_p$ with weights $C_c(\tau)$. As a comparison the situation of the full regression ($c = 0$) gives K in (3.3) as the weighted average of $D_0(\tau) = \tau^2/(\tau^2+1)$ with weights $C_0(\tau) = \tau^2 + 1$.

Table 2 shows some values of these quantities for $c = 2$ (i.e. selecting only those coefficients which are significant at about the 5 per cent level). Since the same weights (second column) are used for both K_1 and $\hat{K}_2$, the third and fourth columns can be compared directly suggesting that shrinkage is greater than the nominal estimate based on the reduced regression, and substantially so if the τ_i's are small. The fourth column takes $p = \infty$, but as $H_2(\tau, p) > H_2(\tau, \infty)$ the difference will in fact be even greater for finite p. On the other hand, the third and fifth columns are much more nearly comparable, with $D_0(\tau)$ being the greater. However, the weights for K_1 and K are not equal; relatively more weight will be given to the larger values of τ for $D_2(\tau)$ than for $D_0(\tau)$, indicating that the difference between K_1 and K will be even less than the third and fifth columns suggest.

TABLE 2
Terms for weighted average approximations of shrinkage, $c = 2$

τ	C_2 (*weight for* D_2, H_2)	D_2 (*with selection*)	H_2 (*ignoring selection*)	D_0 (*without selection*)	C_0 (*weight for* D_0)
0	0.26	0	0.83	0	1.00
0.5	0.44	0.17	0.83	0.20	1.25
1.0	1.05	0.38	0.85	0.50	2.00
1.5	2.24	0.55	0.86	0.69	3.25
2.0	4.10	0.68	0.88	0.80	5.00
2.5	6.60	0.79	0.90	0.86	7.25
3.0	9.62	0.86	0.91	0.90	10.00
3.5	13.08	0.91	0.93	0.93	13.25
4.0	16.94	0.94	0.94	0.94	17.00

Obviously these calculations can only be regarded as crude approximations, although the bias inherent in replacing the expectation of a ratio by the ratio of the expectations is always in the same direction. In the case of K, for instance, the weighted average of $D_0(\tau)$ is just equal to $(1+p\delta^2)^{-1}$. to be compared with the approximation to $E(K)$ given by (3.12). For the situation of Fig. 3, these are 0.667 and 0.727 respectively, to be compared with the true value of 0.735 (Table 1). For larger values of p the approximation would be better. Doubtless, better approximations for K_1 and $\hat{K}_2$ could be developed, but are not pursued in this paper.

7. EMPIRICAL BAYES AND SUBSET SELECTION

When the model of Section 5 is applicable, the study of subset selection is greatly simplified since a Bayesian method conditions on the observed data and hence also conditions on the subset actually selected (compare the situation in sequential analysis where the likelihood function does not depend on the stopping rule). Suppose that $\mathbf{x}^T = (\mathbf{x}_{(1)}^T : \mathbf{x}_{(2)}^T)$, where $\mathbf{x}_{(1)}$ consists of s selected regressors and $\mathbf{x}_{(2)}$ consists of the $(p-s)$ omitted regressors. Thus, although this partition may have depended on $\mathbf{y}$ in CS, we can treat it as if it were fixed in advance. Let the corresponding partitions of $\boldsymbol{\beta}^T$ be $(\boldsymbol{\beta}_1^T : \boldsymbol{\beta}_2^T)$, of $\hat{\boldsymbol{\beta}}^T$ be $(\hat{\boldsymbol{\beta}}_1^T : \hat{\boldsymbol{\beta}}_2^T)$, and of $\mathbf{V}$ be the submatrices $\mathbf{V}_{ij}$ $(i, j = 1, 2)$. Then the true regression of y on $\mathbf{x}_{(1)}$ is

$$y = \alpha + \boldsymbol{\beta}_{(1)}^T \mathbf{x}_{(1)},$$

where $\boldsymbol{\beta}_{(1)}$ and its LS estimate $\hat{\boldsymbol{\beta}}_{(1)}$ are given by

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_1 + \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\beta}_2, \quad \hat{\boldsymbol{\beta}}_{(1)} = \hat{\boldsymbol{\beta}}_1 + \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\hat{\boldsymbol{\beta}}_2.$$

Our aim is to predict y for a future case using only the values of x_i in $\mathbf{x}_{(1)}$. (Note that this contrasts with Lindley (1968) who considers the different problem of deciding which components of $\mathbf{x}$ should have been measured in CS.) With the prior distribution (5.2), the posterior distribution of $\boldsymbol{\beta}$ is

$$N\left(\kappa\hat{\boldsymbol{\beta}}, \frac{\sigma^2\kappa}{n}\mathbf{V}^{-1}\right), \tag{7.1}$$

and so the posterior expectation of $\boldsymbol{\beta}_{(1)}$ is just $\kappa\hat{\boldsymbol{\beta}}_{(1)}$. Hence the Bayes subset predictor of y is

$$E(y \mid \mathbf{x}_{(1)}, \mathrm{CS}) = \hat{\alpha} + \kappa\hat{\boldsymbol{\beta}}_{(1)}^T\mathbf{x}_{(1)}. \tag{7.2}$$

In the last section it was suggested that the shrinkage of a subset regression is similar to that of the full regression. According to (7.2), the shrinkage represented by κ is the same for *all* subsets, no matter how they are selected.

Now

$$\text{Var}\,(y \mid \mathbf{x}_{(1)}, \alpha, \boldsymbol{\beta}) = \sigma^2 + \boldsymbol{\beta}_2^{\mathrm{T}}(\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12})\boldsymbol{\beta}_2$$

and so, using (7.1),

$$n\,\text{Var}\,(y \mid \mathbf{x}_{(1)}, \text{CS}) = \sigma^2(n+1) + n\kappa^2\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}(\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12})\hat{\boldsymbol{\beta}}_2 + \kappa\sigma^2(p - s + \mathbf{x}_{(1)}^{\mathrm{T}}\mathbf{V}_{11}^{-1}\mathbf{x}_{(1)}).$$

As before, we are concerned with the performance of a predictor for general use on future (as yet unspecified) values of $\mathbf{x}_{(1)}$, and so the relevant posterior PMSE of (7.2) is the average of this conditional variance over $\mathbf{x}_{(1)}$. Estimating σ^2 by $\hat{\sigma}^2$ (from the full regression in CS), and relating the quadratic form in $\hat{\boldsymbol{\beta}}_2$ to the sum of squares for the omitted variables, gives this average to be $(T/n)\widetilde{Q}_s$, where

$$\widetilde{Q}_s = \nu^{-1}(1 - R_p^2)(n + 1 + \kappa p) + \kappa^2(R_p^2 - R_s^2), \tag{7.3}$$

T is the total sum of squares, R_s is the multiple correlation of y on $\mathbf{x}_{(1)}$ and R_p is the multiple correlation of y on $\mathbf{x}$. By a similar calculation, the posterior PMSE of the LS predictor based on $\mathbf{x}_{(1)}$ is $(T/n)\hat{Q}_s$, where

$$\hat{Q}_s = \widetilde{Q}_s + (\kappa - 1)^2 R_s^2 . \tag{7.4}$$

Section 5 considered the case $s = p$, when it was proposed that κ be estimated by (3.7) or

$$\hat{K}(k) = 1 - \frac{k(1 - R_p^2)}{\nu R_p^2} . \tag{7.5}$$

It was suggested that k be taken as (4.7) for the purpose of constructing the preshrunk predictor, but as (5.7) for the purposes of estimating its PMSE. When $s < p$, κ in (7.2) should still be estimated from the full regression using (7.5) and (4.7), and it seems reasonable that the smaller value (5.7) should continue to be used for formula (7.3). The LS predictor, however, does not involve the estimation of κ, and so the argument for estimation (7.4) is different. In fact (7.4) with $s = p$ and κ taken as (7.5) equals

$$\nu^{-1}(1 - R_p^2)(n + p + 1) + (1 - \kappa)(k - p).$$

This should, on average, give (4.5), the PMSE of LS in the full regression. Thus we should take $k = p$, and it is reasonable that this value is also appropriate for calculating (7.4) when $s < p$. The resulting estimate of $\hat{Q}_s$ is just

$$1 - R_s^2 + 2\nu^{-1}(1 - R_p^2)(1 + pR_s^2/R_p^2). \tag{7.6}$$

Now if the subset $\mathbf{x}_{(1)}$ were fixed in advance, the PMSE of LS on $\mathbf{x}_{(1)}$ would be, in the usual sampling theory sense,

$$\sigma^2\left(1 + \frac{s+1}{n}\right) + \boldsymbol{\beta}_2^{\mathrm{T}}(\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12})\boldsymbol{\beta}_2 .$$

It is easy to show that an unbiased estimate of this is $(T/n)Q_s^*$, where

$$Q_s^* = 1 - R_s^2 + 2\nu^{-1}(s + 1)(1 - R_p^2). \tag{7.7}$$

Judging subsets by Mallows' C_p is equivalent to judging them by (7.7). However, (7.7) is for a *fixed* subset, whereas (7.6) derives from an entirely different argument which allows for the empirical solution of that subset. In fact (7.7) is less than (7.6) exactly when R_s^2/s exceeds R_p^2/p, or when the mean square for the additional variables in $\mathbf{x}_{(2)}$ is less than the mean square for $\mathbf{x}_{(1)}$ alone. This will always be the case with the usual subset selection procedures which include the "most significant" variables in $\mathbf{x}_{(1)}$. If, perversely, the "least significant" variables are selected, (7.7) will exceed (7.6).

Suppose that we have searched over all subset sizes $s = 1, 2, \ldots, p$ and that for each s some particular subset has been chosen (e.g. by forward selection). Usually, R_s^2 is a concave function of s, and so, as s increases, Q_s^* decreases and then increases, suggesting some intermediate subset size as optimum. However, this cannot occur in the EB formulation of subset selection. In particular $\tilde{Q}_s$ always decreases with s and so is smallest at $s = p$, which is only to be expected as $\kappa \hat{\boldsymbol{\beta}}$ is the true Bayes estimate of $\boldsymbol{\beta}$. Evidently, the benefits of finding a small subset are more than offset by the fact that that subset has to be empirically selected, and the inflation of variance caused by adding x_i's of low predictive value is catered for by preshrinking rather than by discarding variables. The more "noisy" is the regression, the greater should be this allowance for shrinkage. If, on the other hand, LS is to be used, the estimated PMSE in (7.6) either increases or decreases with s, according as to whether $F < 2$ or $F > 2$, where F is the variance ratio for the full regression. Thus, if $F < 2$, the best LS predictor is to ignore the x_i's altogether and predict all cases by $\bar{y}$, although even in this situation the preshrunk predictor using all x_i's will do better. Of course, these remarks do not reflect any *practical* advantages there might be in reducing subset size. Typically, R_s^2 will change relatively little as s approaches p, and so both (7.3) and (7.4) will be little affected by discarding some of the later variables (unlike (7.7) which would show a more marked change).

It must be stressed that the EB analysis rests on the family of prior distributions in (5.2), and that the discussion of subset selection depends more critically on the assumed normality than does the simpler situation of Section 5. For example, suppose that $\mathbf{V} = \mathbf{I}$. Then the β_i's, and also the $\hat{\beta}_i$'s, are assumed to be random samples from univariate normal distributions, implying that any configuration of β_i's with a heavy left or right tail has small probability both before and after observing the data. But Fig. 3 shows that it is precisely when there is a long-tailed empirical distribution of the β_i's that screening variables can improve PMSE—for example, the lowest curve in this figure corresponds to $\boldsymbol{\beta}$ with all elements zero except for one outlier. Vectors $\boldsymbol{\beta}$ which are more "plausible" according to (5.2) will come nearer the upper curve in Fig. 3, for which the full regression is optimum. Note that the parameter values used for Fig. 3 predict F to be about 3 and so the EB analysis shows that $s = p$ is optimum for both LS and $\tilde{y}$. If a heavy tailed prior distribution were to be assumed, however, $\boldsymbol{\beta}$'s near the lowest curve in Fig. 3 might be given sufficient weight to make selection worthwhile: this could happen, for example, if (perversely) a number of spurious regressors were introduced. Further research on robustness to normality is obviously needed, although it is perhaps better to make at least some attempt to quantify the effects of empirical selection which can be applied easily to standard regression analysis rather than to ignore it altogether.

Finally in this Section we note that the above analysis can also be expressed in terms of the correlation coefficient between predicted and observed values of y. Retrospectively, this correlation is just R_s. Prospectively, we need the correlation between y and $\hat{y}_{(1)} = \hat{\alpha} + \hat{\boldsymbol{\beta}}_{(1)}^{\mathrm{T}} \mathbf{x}_{(1)}$ for a future observation $(y, \mathbf{x}_{(1)})$, this correlation being the same whether or not the predictor is preshrunk. The posterior covariance between y and $\hat{y}_{(1)}$ is $\kappa \hat{\boldsymbol{\beta}}_{(1)}^{\mathrm{T}} \mathbf{V}_{11} \hat{\boldsymbol{\beta}}_{(1)}$, the posterior variance of $\hat{y}_{(1)}$ is $\hat{\boldsymbol{\beta}}_{(1)}^{\mathrm{T}} \mathbf{V}_{11} \hat{\boldsymbol{\beta}}_{(1)}$, and the variance of y is estimated by T/n. Hence the validation correlation is just κR_s, which is estimated by (taking $k = p$ as in estimating the PMSE of LS)

$$\tilde{R}_s = \frac{(n-1)R_p^2 - p}{\nu R_p^2} R_s. \tag{7.8}$$

This implies that the correlation shrinks by the same amount for all subsets. If, however, the effect of selection is ignored so that $\mathbf{x}_{(1)}$ is taken as fixed, we look to the sampling covariance of y and $\hat{y}_{(1)}$, which is $\boldsymbol{\beta}_{(1)}^{\mathrm{T}} \mathbf{V}_{11} \boldsymbol{\beta}_{(1)}$, whose (sampling) expectation is equal to that of $\hat{\boldsymbol{\beta}}_{(1)}^{\mathrm{T}} \mathbf{V}_{11} \hat{\boldsymbol{\beta}}_{(1)} - s\sigma^2/n$. With the same variances as above, the validation correlation is then estimated as

$$R_s^* = \frac{\nu R_s^2 - s(1 - R_p^2)}{\nu R_s}. \tag{7.9}$$

It is easy to see that the comparison of $\widetilde{R}_s$ with R_s^* is directly analogous to the comparison of $\widetilde{Q}_s$ with Q_s^* discussed above. In particular, R_s^* (ignoring selection) overestimates $\widetilde{R}_s$ (allowing for selection) when the "most significant" variables are retained but underestimates it if they are omitted. Both correlations are less than R_s.

A simple correction to the multiple correlation coefficient which is widely used in econometrics is $\bar{R}_s$ (Goldberger, 1964, p. 217) given by

$$\bar{R}_s^2 = \frac{(n-1)R_s^2 - s}{n-s-1}.$$

This is similar (but not identical) to the sum of terms up to order $O(n^{-1})$ in a series expansion of the minimum variance unbiased estimate of the multiple correlation for multivariate normal populations, as derived by Olkin and Pratt (1958). When $s = p$, $\bar{R}_p$ is the geometric mean of R_p and $\widetilde{R}_p = R_p^*$, and so is less than the former but not as small as the latter. For $s < p$, $\bar{R}_s$ exceeds the geometric mean of R_s and R_s^* whenever the mean square for the omitted variables is less than the residual mean square for the full set. Thus in realistic cases both R_s and $\bar{R}_s$ overestimate the correlation which is likely to be observed on validation, often substantially so. It is worth noting that both R_s and $\widetilde{R}_s$ increase as s increases, whereas R_s^* and $\bar{R}_s$ usually rise to a maximum and then decrease. In fact, Haitovsky (1969) points out that $\bar{R}_s$ takes its maximum for the largest subset in which all the "t-statistics" of the regression coefficients exceed unity.

8. SHRINKAGE AND BINARY REGRESSION

The above ideas are not confined to multiple regression but can be extended to the much wider class of "generalized linear models" (Nelder and Wedderburn, 1972). Brief consideration of just one case will be given here, that of binary regression.

Suppose that y is a binary response taking values 0 or 1 with

$$P(y = 1 \mid \mathbf{x}) = f(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \tag{8.1}$$

where f is a given function taking values in (0, 1) (e.g. logit or probit). As before, there are n cases in CS, with values of $\mathbf{x}$ equal to $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Then the information matrix for $(\alpha, \boldsymbol{\beta})$ is, in partitioned form,

$$\begin{pmatrix} \Sigma w_i & \Sigma w_i \mathbf{x}_i \\ \Sigma w_i \mathbf{x}_i & \Sigma w_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \end{pmatrix}, \tag{8.2}$$

where $w_i = f_i'^2/\{f_i(1-f_i)\}$, $f_i = f(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ and $f_i' = f'(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$. Let

$$e^2 = (f'(\alpha))^{-2} f(\alpha)(1 - f(\alpha)).$$

Then it follows that, when $\boldsymbol{\beta} = \mathbf{0}$, the asymptotic distribution of the ML estimates $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is multivariate normal with $\hat{\alpha}$ uncorrelated with the components of $\hat{\boldsymbol{\beta}}$, with the variance of $\hat{\alpha}$ equal to $n^{-1}e^2$, and with the var. matrix of $\hat{\boldsymbol{\beta}}$ equal to $n^{-1}e^2\mathbf{V}^{-1}$. If $\boldsymbol{\beta} \neq \mathbf{0}$, the relevant weighted sums in (8.2) are needed, but we will suppose that the degree of discrimination given by the data is sufficiently modest for the variation in the weights w_i to be ignored; i.e. we *assume* that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, n^{-1}e^2\mathbf{V}^{-1}) \tag{8.3}$$

and

$$\hat{\alpha} \sim N(\alpha, n^{-1}e^2),$$

with $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ independent. Essentially, this amounts to assuming that not too many of the probabilities f_i are close to 0 or 1.

The analogue of the prior distribution for $\boldsymbol{\beta}$ in (5.2) is

$$N\left(\mathbf{0}, \frac{e^2\kappa}{n(1-\kappa)}\mathbf{V}^{-1}\right),$$

from which the posterior distribution of $\boldsymbol{\beta}$ is, assuming (8.3),

$$N(\kappa\hat{\boldsymbol{\beta}}, \kappa n^{-1}e^2\mathbf{V}^{-1}).$$

The prior distribution for α is vague so that, *a posteriori*, α is independent of $\boldsymbol{\beta}$ with distribution $N(\hat{\alpha}, n^{-1}e^2)$. Note that these expressions are exactly the same as in the multiple linear regression case but with σ^2 replaced by e^2. In particular, the graphical test following (5.8) is still available as a predictive check. The arguments for estimating κ are also virtually the same as in Section 5. The quantity

$$q^2 = n e^{-2}\ \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}} \tag{8.4}$$

is, from (8.3), marginally distributed as $(1-\kappa)^{-1}\chi^2_p$ and so an unbiased estimate of κ is

$$1-\frac{p-2}{q^2}. \tag{8.5}$$

Note that q^2 is the asymptotic χ^2 statistic for testing significance of the regression, and can be deduced from the log-likelihood of the model or the "deviance" in Nelder and Wedderburn's terminology (Nelder and Wedderburn, 1972).

Specializing now to a probit model with $f = \Phi$, the predictive probability that $y = 1$ is the posterior expectation of $\Phi(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ which can be shown to be

$$\Phi\left(\frac{\hat{\alpha} + \kappa\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}}{(1+n^{-1}e^2(1+\kappa\mathbf{x}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{x}))^{\frac{1}{2}}}\right). \tag{8.6}$$

In contrast, the ML estimate of (8.1), or, in the terminology of Aitchison and Dunsmore (1975), the "estimative" probability that $y = 1$, is simply

$$\Phi(\hat{\alpha} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}). \tag{8.7}$$

Note that even when $\kappa = 1$ (vague prior), (8.6) is different from (8.7). In fact, if $\kappa = 1$, (8.6) always belongs to the finite interval $\Phi(\pm q)$ whereas (8.7) can give probabilities arbitrarily close to 0 or 1 for extreme values of $\mathbf{x}$. For example, if $p = 1$, q is just the "t-statistic" for the significance of the regression and the limits $\Phi(\pm q)$ correspond to the associated significance level—for instance $q = 2$ (5 per cent significance) gives the limits ($2\frac{1}{2}$ per cent, $97\frac{1}{2}$ per cent). This is closely related to the situation in discriminant analysis where one models the conditional distribution of $\mathbf{x}$ given y rather than the conditional distribution of y given $\mathbf{x}$ (or, in the terminology of Dawid, 1976, the "sampling paradigm" rather than the "predictive paradigm"). Aitchison and Dunsmore (1975) present a Bayesian analysis of the usual discriminant model and show that there is a similar effect of guarding against extreme predicted probabilities, even when the prior distribution is vague. This point is also emphasized in Aitchison *et al.* (1977).

As in multiple regression, preshrunk predictors can also be obtained by a sampling theory argument. Following Section 3, but now additionally assuming that $\mathbf{x}$ is multivariate *normal* $N(\mathbf{0}, \mathbf{V})$, the conditional distribution (over varying $\mathbf{x}$) of $\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ given a fixed value for $\hat{\alpha} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$ and given CS is

$$N(\alpha + K\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x},\ \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} - (\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\hat{\boldsymbol{\beta}})^2). \tag{8.8}$$

The resulting conditional probability that $y = 1$ is the expectation of (8.1) over (8.8). Arguments similar to those in Section 3 suggest that K be estimated by $\hat{K}$ in (8.5) and that the variance in (8.8) be estimated by $n^{-1}e^2(p-2)\hat{K}$. The conditional probability that $y = 1$ is therefore estimated

by

$$\Phi\left(\frac{\hat{\alpha} + \hat{K}\,\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}}{(1 + n^{-1}\, e^2 (p-2)\hat{K})^{\frac{1}{2}}} \right) . \tag{8.9}$$

If n is large, the contribution of the terms multiplying $n^{-1}e^2$ in the denominators in (8.6) and (8.9) are both negligible compared with the effect of the shrinkage introduced in the numerator, and so both probabilities are approximated by the simpler predictor

$$\Phi(\hat{\alpha} + \hat{K}\,\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}), \tag{8.10}$$

with $\hat{K}$ given by (8.4) and (8.5). Alternatively, (8.10) can be obtained directly from the EB argument by using a quadratic loss function on the probit scale rather than on the probability scale. Note that the linear predictor in (8.10) can be deducted from standard GLIM output (Baker and Nelder, 1978).

The probit model has been chosen because the expressions (8.6) and (8.9) can be evaluated explicitly. However, the assumption (8.3) can be made for any smooth response function f in (8.1), and the analogue of the simple predictor (8.10), namely

$$f(\hat{\alpha} + \hat{K}\,\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}), \tag{8.11}$$

still obtains with $\hat{K}$ given by (8.5). If f is the logistic function, (8.11) is of the preshrunk form suggested by the logistic regression example in Section 2.

It is worth noting that assumption (8.3) is related to the standard discriminant analysis model already mentioned. It is well known that for this model the logit of the group membership probability is a linear function of the Fisher discriminant, and that the discriminant function itself is equivalent to a multiple regression of y (given two arbitrarily coded values) on $\mathbf{x}$. Again, assume that there are not too many extreme predictions so that this regression is approximately homoscedastic. Then if the coefficients in the binary model are estimated by a discriminant analysis, they will behave in a similar way to ordinary LS estimates, (8.3) will be satisfied, and the shrinkage given in (8.5) will be exactly the same as (3.7) with $k = p - 2$. Further, the behaviour of subset regressions will follow the same pattern as in Sections 6 and 7. For binary regressions fitted using ML, however, regression coefficients for different subsets are not related by the usual formulae for LS, although it is reasonable to suggest that with a relatively low signal/noise ratio and with large n similar conclusions will apply.

9. EXAMPLES

Four illustrations are presented briefly. Questions of preliminary data analysis, choice of variables, etc. are not discussed; in each case it is simply assumed that a predictor is to be found and that the model being fitted is appropriate.

Example 1. *Parametric cost model* (continued from Section 2). The two x_i's defining the predictor in Fig. 1 were empirically selected from $p = 14$ possible explanatory variables. Obviously the full regression cannot be fitted to the 8 observations in CS, and so the method of Section 7 is not available. However, a rough idea of the likely shrinkage (for the purpose of checking the theory) can be obtained from the full regression on all 31 cases, for which $F = 8.20$. By noting that an unbiased estimate of $31\,\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta}/\sigma^2$ is $p(F-1) = 100.8$, δ^2 in (3.4) is estimated as 0.038 from which the approximate expected value of $\hat{K}$ in (3.12) is 0.67. The predictor $\tilde{y}$ in (4.1) with $\hat{K} = 0.67$ is shown as the dotted line in Fig. 1 and is quite close to the empirical regression line of y on $\hat{y}$ for the validation cases.

The normal plot for (5.8), again for the full regression on all 31 cases, is shown in Fig. 4a. The straight line gives the expected distribution for $\boldsymbol{\beta} = \mathbf{0}$. Evidently, the fit of the EB model is reasonable, and considerable variation in the β_i's above noise level is apparent.

Obviously the scatter of $(y, \hat{y})$ in Fig. 1 depends on the particular choice of the 8 cases that have been selected for CS, a dependence which can be studied by simulating over all $\binom{31}{8}$ possible

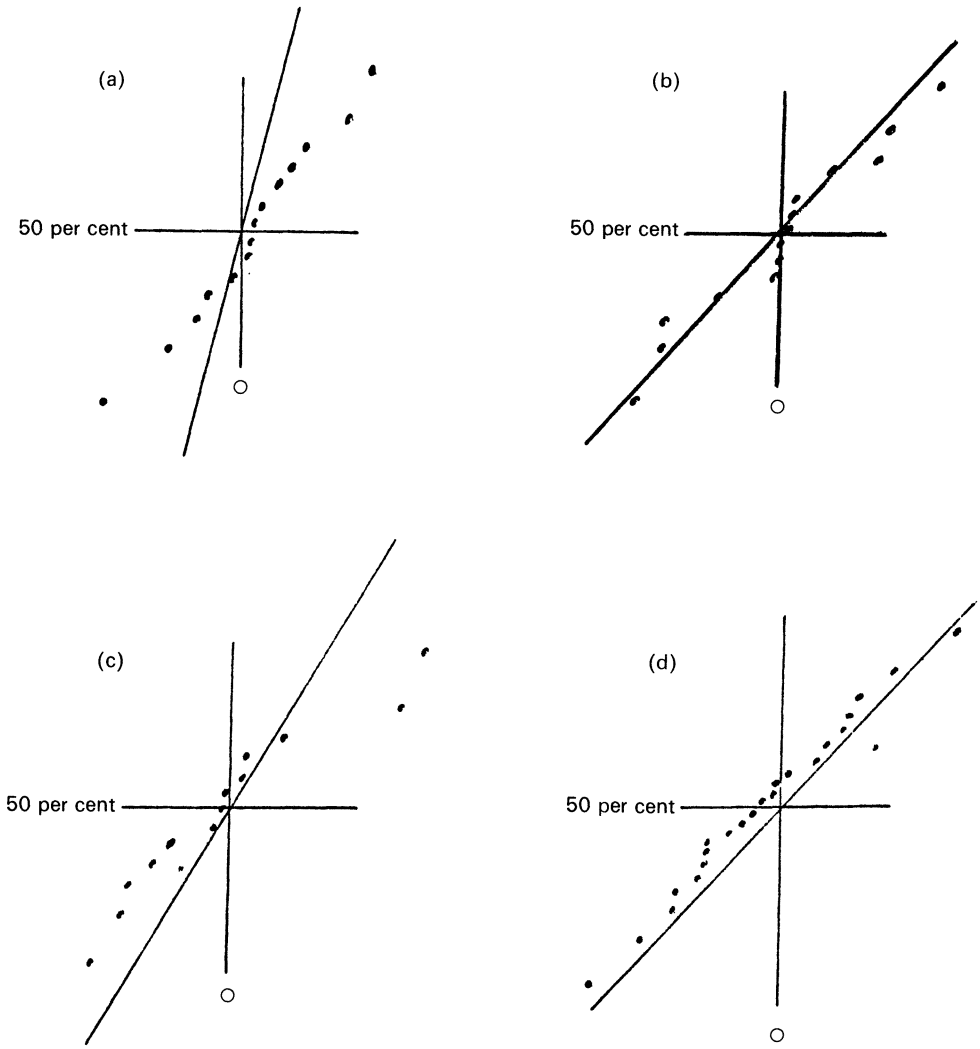

Fig. 4. Normal plots for fit of empirical Bayes model: lines give distribution expected when $\boldsymbol{\beta} = \mathbf{0}$.

sample splits. Suppose, for example, that $\hat{y}$ is the LS predictor calculated from CS which uses a fixed subset of $4x_i$'s (a subset chosen on the basis of the original CS). Let

$$K^* = \frac{\Sigma (y - \bar{y})(\hat{y} - \bar{y})}{\Sigma (\hat{y} - \bar{y})^2} \tag{9.1}$$

be the empirical slope of y on $\hat{y}$ for the 23 validation cases only. Then the median of K^* over the different sample splits is about 0.61, and $K^* < 1$ about 87 per cent of the time (and is negative about 2 per cent of the time).

Example 2. *Psychopath prediction* (continued from Section 2). The six x_i's used for the predictor in Fig. 2 were again empirically selected from $p = 14$ possible variables. Extreme predictions are rarely given in this example, and so shrinkage for the selected subset is assumed to follow from the full regression. The deviance of the full logistic regression gives $q^2 = 24.75$ on

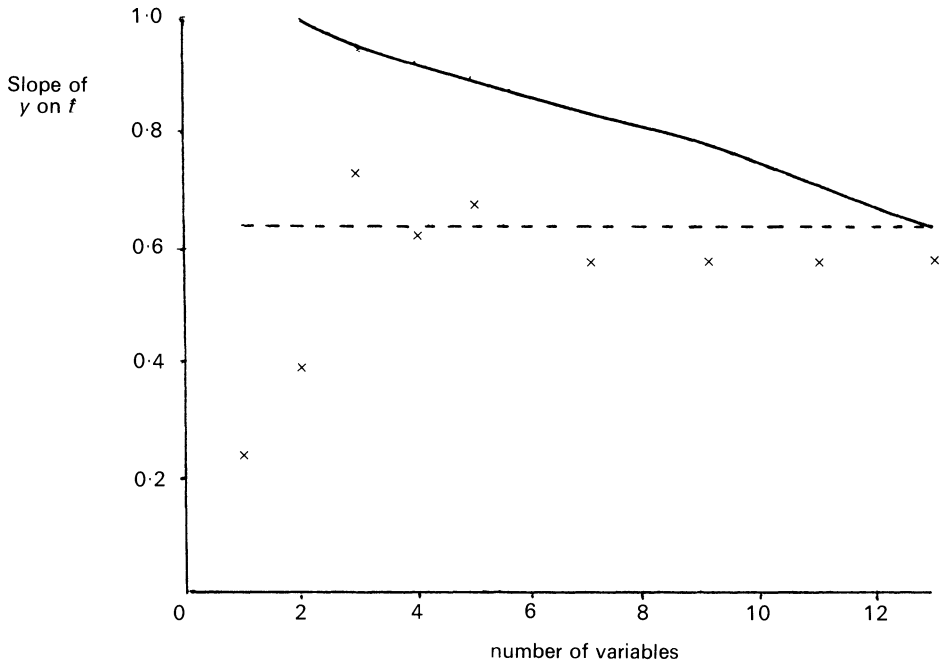

Fig. 5. Shrinkage in stepwise regression. × = observed; —— = predicted (selection ignored); – – – – = predicted (EB model).

14 d.f. from which $\hat{K}$ in (8.5) is 0.51. The logistic version of (8.11) gives the dotted line in Fig. 2, and corresponds closely to the actual validation results.

The components of $\mathbf{M}\hat{\boldsymbol{\beta}}$ give the normal plot in Fig. 4b, with the straight line corresponding to $\boldsymbol{\beta} = \mathbf{0}$ as before. The plot seems acceptable, although with some suggestion of too many very small orthogonal components of $\hat{\boldsymbol{\beta}}$. The closeness of the plot to the straight line confirms that only a very modest degree of prediction is possible.

Example 3. *Breast cancer prognosis.* Armitage *et al.*, (1969) reported a statistical study of prognosis in advanced breast cancer, including a multiple regression analysis of "mean clinical value" measured 3 months after treatment (y) on a number of prognostic variables known at the time of surgery (x_i's). Several different subset regressions using a stepwise method were discussed. To illustrate the results of the present paper, the data were divided into two halves, patients with even ages assigned to CS and patients with odd ages retained for validation. (The more natural procedure of dividing by order of entry into the study was not appropriate as there was evidence of instability in some of the x_i's over time.) This gave a CS with $n = 86$ and $p = 13$. Regressions were then fitted to CS by forward selection, and the LS predictor at each stage validated against the remaining cases.

At each of the subset sizes $s = 1$ (1) 5, 7 (2) 13, (9.1) was calculated for the validation data, the results being shown in Fig. 5. Also shown are the estimates $\hat{K}$ which would be appropriate if the selection effects were ignored, each being calculated from (3.7) using the appropriate subset value of F. Note that with this sample size it makes little difference whether k is chosen as $p - 2$ or (4.7). It is obvious from Fig. 5 that the selected regressions shrink much more than would be expected if the subsets were fixed in advance. Once s reaches 3 or 4, the shrinkage stabilizes to a value reasonably close to $\hat{K}$ for the full regression, as predicted in Sections 6 and 7.

The ratios of the average values of $(y - \hat{y})^2$ to T/n are shown as the crosses in Fig. 6, along with the various quantities defined in Section 7. Again, the behaviour of the predictors for small s is

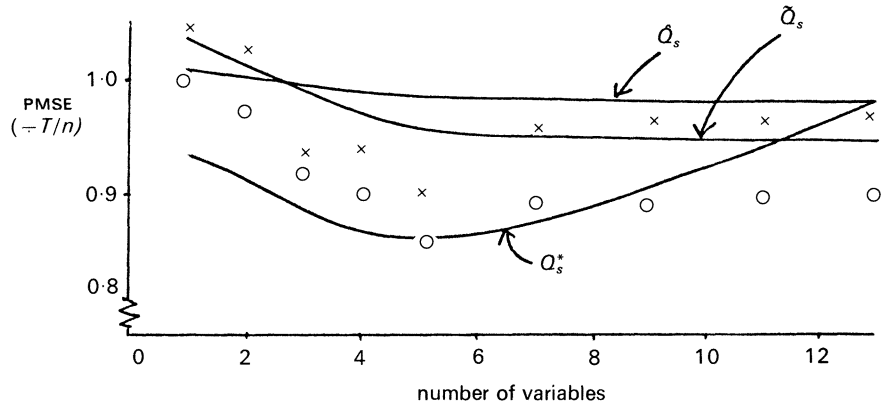

Fig. 6. Prediction mean squared error in stepwise regression (as proportion of total mean square). X = observed, least squares; O = observed, preshrunk.

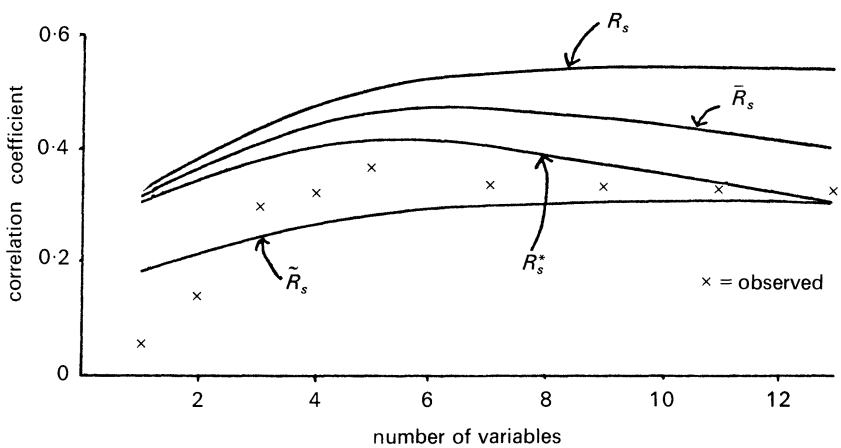

Fig. 7. Correlation coefficients in stepwise regression.

erratic, but agreement with $\hat{Q}_s$ in (7.6) is reasonable for the larger subset sizes. The quantities Q_s^* in (7.7), equivalent to Mallows' C_p, are substantial underestimates of the actual prediction errors. The preshrunk predictors (7.2) were also evaluated with $\hat{K}$ taken from the full regression using (3.7) and (4.7), and the average squared errors calculated as before – these are the circles in Fig. 6. In every case they are better than LS, but the agreement with $\tilde{Q}_s$ in (7.3) using k in (5.7) is not very good. Curiously, the circles agree much better with $\tilde{Q}_s$ using the usual value of k in (4.7) rather than (5.7), but this is presumably an artifact of these particular data. Needless to say, the absolute values of all these quantities are subject to substantial sampling variation, although comparisons between different values of s for the same data are perhaps more stable.

Fig. 7 shows the various correlation coefficients discussed in Section 7, along with the observed correlations between y and $\hat{y}$ in the validation sample. The fit to $\tilde{R}_s$ in (7.8) is reasonable for the larger values of s. The validation correlations for small subsets are clearly much worse than R_s^* in (7.9), the values predicted if selection is ignored.

The normal plot for assessing the fit of the EB model is shown in Fig. 4c, and seems reasonable.

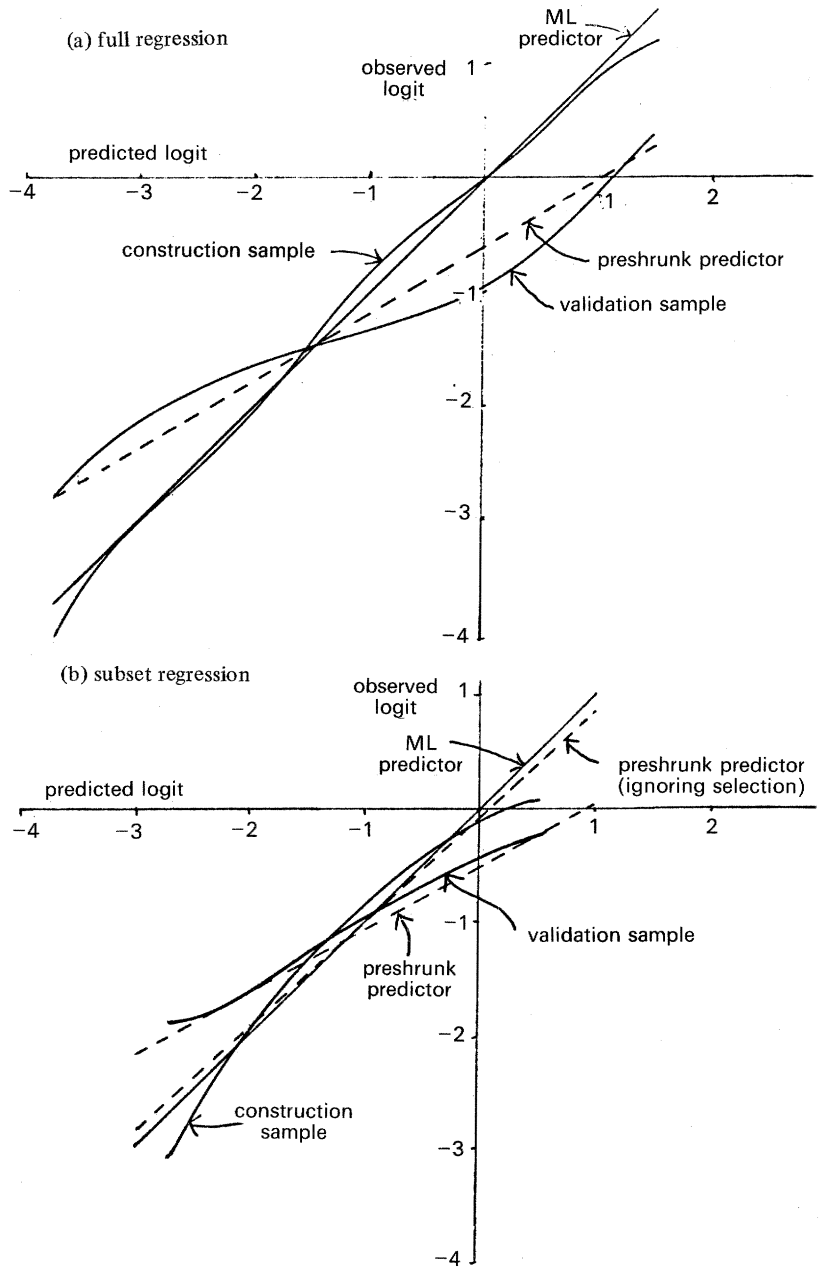

Fig. 8. Observed logit against predicted logit, full and subset regression.

The regression coefficient at $s = 1$ is highly significant ($t = 3.25$), and even at $s = 4$ all coefficients are significant at the 5 per cent level. Although it would appear that LS subset prediction should be useful, this is not so when the effects of selection are taken into account. In fact Figs 6 and 7 show that when $s = 1$ and 2 the average squared errors of $\hat{y}$ on validation are

worse than ignoring the prognostic variables altogether, and that the validation correlations are only about one third of the retrospective values. Of course, these remarks do not apply to the regressions actually fitted in the cited paper, since we have halved the sample size for the purpose of this example. The effects of shrinkage and selection for the full data would be much less severe, as can be seen by recalculating the relevant statistics for all 163 cases.

Note that such a drastic reduction in correlation is not unusual in applications, other examples being in Simon (1971) and Gardner (1972).

Example 4. *Absconding from Borstal training*. As an example with a much larger sample size, data were available on over 2000 borstal trainees admitted to open borstals in 1977–78. For each case, the data recorded whether the trainee had absconded during sentence ($y = 1$) or not ($y = 0$), together with the values of $p = 22$ predictor variables covering social and criminological factors. To illustrate the theory, logistic regressions were fitted using $n = 500$ randomly selected cases and then validated on the remainder of the data.

A non-parametric regression of observed logit on the ML predicted logit for the full model is shown in Fig. 8a, this graph being constructed in the same way as Fig. 2. The deviance of the fit gave $q^2 = 50.2$ on 22 d.f., leading to the value 0.602 for $\hat{K}$ in (8.5). As Fig. 8a shows, the associated preshrunk predictor gives a reasonable fit to the validation data. Note that a multiple regression of y on $\mathbf{x}$ gave $F = 2.25$, which results in almost exactly the same estimate of K.

A stepwise analysis of CS suggested that most information was contained in just four x_t's, and Fig. 8b shows the corresponding graph for a subset regression with $s = 4$. The lines for two preshrunk predictors are shown, the first with $\hat{K} = 0.931$ which is the value of (8.5) when calculated from the subset regression, the second with $\hat{K} = 0.602$ as in Fig. 8a. It is clear that the second fits well, but that the first is a gross underestimate of the observed shrinkage.

The normal plot for testing the EB model is given in Fig. 4d. The plot is acceptably straight, and the apparent location shift from zero is not significant ($t = 1.03$). The closeness of the plot to the null line shows that the predictive power of the x_t's is very modest, as is already evident from the statistics quoted. The analysis of the full data set would, of course, contain much more information.

ACKNOWLEDGEMENTS

I am greatly indebted to Mr David Dench for his help in the computational aspects of this paper, and to the Prison Department's Young Offender Psychology Unit for permission to use the data of Example 4. I am also grateful to referees for their helpful comments on an earlier version of this paper. Part of this work was supported by a research grant from the Social Science Research Council.

REFERENCES

Aitchison, J. and Dunsmore, I. R. (1975) *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.

Aitchison, J., Habbema, J. D. F. and Kay, J. W. (1977) A critical comparison of two methods of statistical discrimination. *Appl. Statist.*, **26**, 15–25.

Anderson, T. W. (1958) *Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Armitage, P., Copas, J. B. and McPherson, K. (1969) Statistical studies of prognosis in advanced breast cancer. *J. Chron. Dis.*, **22**, 343–360.

Baker, R. J. and Nelder, J. A. (1978) *The GLIM System, Release* 3. Oxford: Numerical Algorithms Group.

Baranchik, A. J. (1973) Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Ann. Statist.*, **1**, 312–321.

Box, G. E. P. (1980) Sampling and Bayes' inference in scientific modelling and robustness (with Discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Copas, J. B. (1982) Plotting p against x. *Appl. Statist.*, **32**, 25–31.

Copas, J. B. and Whiteley, J. S. (1976) Predicting success in the treatment of psychopaths. *Brit. J. Psychiat.*, **129**, 388–392.

Dawid, A. P. (1976) Properties of diagnostic data distributions. *Biometrics*, **32**, 647–658.

Dempster, A. P., Schatzoff, M. and Wermuth, N. (1977) A simulation study of alternatives to ordinary least squares. *J. Amer. Statist. Assoc.*, **72**, 77–106.

Draper, N. R. and Van Nostrand, R. C. (1979) Ridge regression and James–Stein estimation: review and comments. *Technometrics*, **21**, 451–466.
Efron, B. and Morris, C. (1971) Limiting the risk of Bayes and empirical Bayes estimators, part I: the Bayes case. *J. Amer. Statist. Assoc.*, **66**, 807–815.
Gardner, M. J. (1972) On using an estimated regression line in a second sample. *Biometrika*, **59**, 263–274.
Goldberger, A. S. (1964) *Econometric Theory*. New York: Wiley.
Gorman, J. W. and Toman, R. J. (1966) Selection of variables for fitting equations to data. *Technometrics*, 8, 28–51.
Haitovsky, Y. (1969) A note on the maximization of $\bar{R}^2$. *The Amer. Statistician*, **23**, 20–21.
Hjorth, U. and Holmqvist, L. (1981) On model selection based on validation with applications to pressure and temperature prognosis. *Appl. Statist.*, **30**, 264–274.
James, W. and Stein, C. (1961) Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium*, *Vol.* 1, pp. 361–379.
Johnson, N. L. (1959) On the extension of the connection between Poisson and χ^2 distributions. *Biometrika*, **46**, 352–363.
Kerridge, D. (1965) A probabilistic derivation of the non-central χ^2 distribution. *Aus. J. Statist.*, 7, 37–9 (corrig., 7, 114).
Lindley, D. V. (1962) Contribution to the discussion of Stein (1962).
Maritz, J. S. (1970) *Empirical Bayes Methods*. London: Methuen.
Narula, S. C. (1974) Predictive mean square error and stochastic regressor variables. *Appl. Statist.*, **23**, 11–18.
Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.
Nicholson, G. E. (1960) Prediction in future samples. In *Contributions to Probability and Statistics: Essays in Honour of Harold Hotelling* (I. Olkin, ed.), pp. 322–330. Stanford: Stanford University Press.
Noah, J. W., Daniels, J. M., Day, C. F. and Eskew, H. L. (1973) Estimating aircraft acquisition costs by parametric methods. United States Navy FR-103-USN.
Olkin, I. and Pratt, J. W. (1958) Unbiased estimation of certain correlation coefficients. *Ann. Math. Statist.*, **29**, 201–211.
Rencher, A. C. and Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53.
Sclove, S. L. (1968) Improved estimates for coefficients in linear regression. *J. Amer. Statist. Assoc.*, **63**, 596–606.
Seber, G. A. F. (1977) *Linear Regression Analysis*. New York: Wiley.
Shibata, R. (1981) An optimal selection of regression variables. *Biometrika*, **68**, 45–54.
Simon, F. H. (1971) *Prediction Methods in Criminology*, London: HMSO.
Stein, C. (1960) Multiple regression. In *Contributions to Probability and Statistics: Essays in Honour of Harold Hotelling* (I. Olkin, ed.), pp. 424–443. Stanford: Stanford University Press.
——— (1962) Confidence sets for the mean of a multivariate normal distribution (with Discussion). *J. R. Statist. Soc.* B, **24**, 265–296.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147.

DISCUSSION OF PROFESSOR COPAS'S PAPER

Dr I. R. Dunsmore (University of Sheffield): We have been privileged tonight to hear an impressive paper presented in the usual lucid manner with which we associate Professor Copas.

This paper, I believe, will be seen as a very important one in that it ties together very neatly many of the ideas which have been tossed around over the past few years of prediction and shrinkage in the regression model. A wide variety of models is incorporated and the theory is applied in four very practical examples. The intricate theory which surrounds the problem of shrinkage is expounded most clearly and there can be few quibbles from a theoretical point of view with much of the paper.

Like a Christmas stocking the paper is full to the brim with classical goodies. However, Christmas stockings have a habit of providing not only goodies but also some surprise packages and the occasional beautifully wrapped present which turns out to be a disappointment. Following the tradition of the Society that the proposer should proceed to find holes in the paper and emphasize them, I had to search hard to find any of these bogus offerings, but have with great difficulty found perhaps one or two.

The main point of concern lies in the basic assumptions made in the first paragraph, namely that "the x_i's at which future predictions are required are not specified in advance but will occur randomly over some population of values and that the success of a predictor can be judged by its average performance over such a population".

Suppose I go to the doctor with some complaint and ask him to predict the time y to remission. He will take some explanatory measurements $\mathbf{x}$ and provide some prediction for y. What I am interested in is a prediction for my $\mathbf{x}$, not for any other $\mathbf{x}$ that I might have had – but did not. Nor am I really interested in his necessarily using a predictor which is "best" over all possible $\mathbf{x}$'s. Perhaps rather selfishly, but I believe justifiably, I want the best prediction for my $\mathbf{x}$. Does it necessarily follow that the best predictor for my $\mathbf{x}$ should take the same form as for some other $\mathbf{x}$? Of course this can pose problems for the esteem of the doctor or his friendly statistician. Because we are concerned with actual observations the goodness or otherwise of the prediction will eventually become apparent. In this case the statistician will not be able to hide behind the screen provided by averaging over all possible future $\mathbf{x}$'s.

The central model of interest seems to be $p(y \mid \mathbf{x}, \alpha, \boldsymbol{\beta})$. Within the classical framework the simplistic $p(y \mid \mathbf{x}, \hat{\alpha}, \hat{\boldsymbol{\beta}})$ is as has been illustrated tonight, not without difficulties. Perhaps some tolerance region or relative likelihood or predictive likelihood approach could be developed to incorporate the information within CS. Within the Bayesian context the model converts naturally to the important predictive distribution given by

$$p(y \mid \mathbf{x}, \mathrm{CS}) = \int p(y \mid \mathbf{x}, \alpha, \boldsymbol{\beta})\, p(\alpha, \boldsymbol{\beta} \mid \mathrm{CS})\, d\alpha\, d\boldsymbol{\beta}\,.$$

This is the keystone for any predictions about y. Any optimality properties should be centred around this predictive distribution, perhaps through a formulation which explores the consequences $\mathrm{Var}(\hat{y}(\mathbf{x}), y)$ of the differences of the predictions $\hat{y}(\mathbf{x})$ from y.

In the normal multiple linear regression case here $p(y \mid \mathbf{x}, \mathrm{CS})$ could take a form of Student distribution. Various cases have been considered and general prior distributions can be accommodated in a straightforward manner; see, for example, Raiffa and Schlaifer (1961), Geisser (1965) and Zellner and Chetty (1965). In Section 5, under the heading Empirical Bayes, Professor Copas provides a rather apologetic Bayesian approach. To justify use of his previously derived preshrunk predictor $\hat{\alpha} + \kappa\, \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$ he assumes a very specific prior for $\boldsymbol{\beta}$ which not only depends on the data in CS but is also more "compact" than the marginal distribution on $\hat{\boldsymbol{\beta}}$. He then produces the throwaway line that whether one believes in his family of prior distributions or not is irrelevant. Similar problems occur in the binary regression model in Section 8. A range of prior distributions can be envisaged which will incorporate both the LS and the preshrunk predictors.

A possible criticism of such predictive approaches might be the strong model dependence; the robustness of the methods needs further investigation.

The second, but related, concern is with the whole concept of retrospective fit and shrinkage from a practical point of view. Retrospective fit is at the centre of attention in shrinkage problems purely by definition. Dare I suggest that retrospective fit is something of a red herring? Since we know that it is not reliable, why should we use it at all? In the empirical model selection problem is it sensible to achieve the reduction in dimension by maximizing some statistical measure of retrospective fit? Perhaps, like Father Christmas, the idea has been around for some time – and we all know what is supposed to happen if we do not believe in Father Christmas. Validation fit is the sensible measure on which to concentrate, and some form of predictive distribution seems the obvious candidate around which to define such a measure of fit and criteria of goodness. Perhaps the moral may be (with suitable amendment to a well-known Biblical quotation): He who sets his hands to prediction, and looks back too much at the CS, is not fit for the task.

It is perhaps something of an omission in the paper that no reference has been made to Geisser, who over the past 20 years has published many notable papers on the concepts of prediction (or predictivism, as he calls it); for example, Geisser (1975) discusses the problem of sample re-use or cross-validation, and in both of Geisser (1971, 1980) he discusses the general concepts and principles of prediction. It seems apt to close with a quote from his 1980 paper: "– decisions and utilities should be functions of observables and not parameters. In other words, the utility should be defined in terms of taking action a on the basis of potential observation y rather than on the assumption of the truth of parameter value θ".

Tonight we have seen a most important paper packed with detail and it indeed gives me great pleasure to propose the vote of thanks.

Professor M. Stone (University College London): This paper is properly describable as a *tour de force*: it ranges widely and hits its targets with judicious accuracy, exposing their underground

connections. It may also be regarded as a set of finely crafted exercises in the delicate art of bootstrapping, that is, *using the data to decide how to use the data.* (The "Efron (1979) bootstrap" is a fascinating, but special, example of this art.) The case for bootstrapping was dramatically demonstrated in the television reporting on the night of the 1964 UK General Election: the systematic drift in the forecast of the final outcome, as new voting results came in, must have been implicitly equivalent to a systematic bias in the ability of the model to predict the new results. A specialization of the bootstrap (bootlace?) could easily have been used to pull together the two sets of separated eyelets, that is, the graphs of new results and the associated predictions against time.

This unrealized application would have been an example of the bootstrap philosophy invoked to patch up the deficiencies of a model, by making another iteration on the input/output data of the statistical black box. Its application in the present paper is, however, to improve, rather than correct, a prediction procedure, which it does by shifting the balance between bias and variance away from that given by least squares prediction.

The term "shrinkage" used in association with this shift could be misleading, especially if we were to forget Professor Copas's clear statement of it as the phenomenon of poorer prospective than retrospective fit. Reference to K in $\tilde{y} = \bar{y} + K(\hat{y} - \bar{y})$ as "an index of shrinkage" and the later remark that zero is "a natural origin for a regression coefficient" induces one to think of "shrinkage" as necessarily referring to the reduction towards zero of the size of the regression coefficients used for prediction. However, zero plays no essential role in the bias/variance balance theory: we are dealing primarily with an exploitable statistical artefact rather than with any tendency of the real world to favour zero, or at least small, coefficients.

If, in Example 1, we were to take $\boldsymbol{\beta}_0$ as our origin for $\boldsymbol{\beta}$, the scatter plot of $y - \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$ against $\hat{y} - \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$ would be a shuffle of the points in Fig. 1 parallel to the line of unit slope. The shuffled scatter plot could be expected to show some measure of the shrinkage phenomenon noted for Fig. 1, but would then deliver as preshrunk predictor of y the expression $\tilde{y}_0 = \tilde{y} + (1 - K)\,\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$ in place of $\tilde{y}$. When $\tilde{y}_0$ is written in the form $K(\bar{y} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}) + (1 - K)(\bar{y} + \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$, we see clearly the trade-off between variance (in $\hat{\boldsymbol{\beta}}$) and bias (in $\boldsymbol{\beta}_0$), and note that the regression coefficients $K\hat{\boldsymbol{\beta}} + (1 - K)\boldsymbol{\beta}_0$ could be much *bigger* than $\boldsymbol{\beta}$. The analysis associated with the choice and estimation of K is unaffected as long as we replace $\boldsymbol{\beta}$ by $\boldsymbol{\beta} - \boldsymbol{\beta}_0$. It follows that, unless we invoke real world arguments in favour of $\boldsymbol{\beta}_0 = \mathbf{0}$, we cannot evade the question of choice of $\boldsymbol{\beta}_0$. Equation (4.4) shows that our maximum gain over LS prediction will be obtained when $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ is small: the ideal choices would be $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$ and $K = 0$, for then $\tilde{y} = \bar{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$!

It is tempting to consider whether the bootstrap philosophy can be used to guide the choice of $\boldsymbol{\beta}_0$, that is, to "estimate" it from the data. Without prior restrictions on $\boldsymbol{\beta}_0$, I do not think this is a feasible proposal, if we use either the model-based approach of this paper or the method of cross-validatory choice to which Professor Copas has generously referred. To try to start the cross-validatory machinery, Fig. 1 must be seen as an example of "23-out-at-a-time (once) cross-validatory assessment". Symmetrizing and simplifying this to the standard "1-out-at-a-time ($n = 31$ times)", let $\hat{y}_{\backslash i}$ denote the least squares prediction of y_i when the ith case is excluded. We have the well-known formula for the prediction error

$$y_i - \hat{y}_{\backslash i} = r_i/(1 - A_{ii}),$$

where $\mathbf{A} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ and r_i is the ith standard residual. (Note that the sample covariance of $\{(y_i - \hat{y}_{\backslash i}, \hat{y}_{\backslash i});\ i = 1, \ldots, n\}$ with weights $1 - A_{ii}$ is $-\Sigma A_{ii} r_i^2/(1 - A_{ii})$, which is the 1-out-at-a-time expression of the shrinkage feature of Fig. 1.) With $\bar{y}_{\backslash i}, \hat{\boldsymbol{\beta}}_{\backslash i}$ standing for LS estimates of α, $\boldsymbol{\beta}$ with $(y_i, \mathbf{x}_i)$ omitted, respectively, a cross-validatory assessment of $\tilde{y}_0$ is

$$\Sigma\,[y_i - \bar{y}_{\backslash i} - K\hat{\boldsymbol{\beta}}_{\backslash i}^{\mathrm{T}}\mathbf{x}_i - (1 - K)\,\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i]^2$$

which reduces to

$$\Sigma\,[(1 - K)(y_i - \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) + Kr_i/(1 - A_{ii})]^2. \qquad (*)$$

Minimizing (*) with respect to K for fixed $\boldsymbol{\beta}_0$ would give a predictor close to $\tilde{y}_0$. To extend cross-validatory choice to $\boldsymbol{\beta}_0$ as well would put it close to $\hat{y}$! This outcome should not surprise us, since the unrestricted "parameter" $\boldsymbol{\beta}_0$ is then acting as a surrogate for $\boldsymbol{\beta}$. Cross-validatory choice

works best when we have only a small number of last-ditch "parameters" that we cannot see how to fix.

The snapping of the bootstraps over the $\boldsymbol{\beta}_0$ question is, in fact, more reassuring than disappointing. For we know that bootstrapping must have limited effectiveness. (I would give a practical demonstration of the truth of the associated adage if I had not recalled Professor Anscombe's mortification when, in a public debate with Mr G. Spencer-Brown about the nature of randomization, he rolled a six-sided cylindrical die three times, only to get *three sixes*! If my luck ran out too, I would find myself levitating.)

With prior restrictions on $\boldsymbol{\beta}_0$, something useful might emerge but this is not the place to explore the idea. Otherwise, $\boldsymbol{\beta}_0$ has to be chosen independently of the data – beforehand, if you like – if we are to reap the benefit of preshrinking. This comes unattractively close to the once-propagated view that you must specify significance tests before looking at data, but it may be better viewed as a nudge in the Bayesian direction already explored by Professor Copas.

To finish, a comment and a query. The fact that the simplest way of getting the postulated distribution of $\mathbf{x}$ is to put probability $1/n$ on each $\mathbf{x}_i$ in the data set emphasizes that the paper provides a useful enhancement of the methodology associated with Mallows' C_p. The query is to ask for an extension to greater realism by allowing "errors in the x's", if that is possible.

If I have not commented on the other riches in this paper, especially the work on subset selection, it is only because of lack of time and digestive juices and the fact that the author has done such a good job on everything he has touched upon. I have great pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Dr T. Subba Rao (UMIST): Professor Copas must be congratulated for presenting an interesting paper. I considered the following problem in time series estimation, and the similarity between the results is interesting.

Consider a stationary time series $\{X_t\}$ generated from the model $X_t = \Sigma_{j=1}^{p} a_j X_{t-j} + e_t$, where e_t is a sequence of i.i.d. random variables and each is $N(0, \sigma_e^2)$. Alternatively this model can be written in the form $\mathbf{x}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{e}_t$, $X_t = \mathbf{H}^{\mathrm{T}}\mathbf{x}_t$, where $\mathbf{x}_t$ is a vector of state variables and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & \ldots & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \ldots & \ldots \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix}, \quad \mathbf{H}^{\mathrm{T}} = (0, 0, \ldots, 1).$$

Then the optimal h-step ahead predictor is given by $\hat{X}_n(h) = \mathbf{H}^{\mathrm{T}}\mathbf{A}^h\mathbf{x}_n$, where $\hat{X}_n(h) = E(X_{n+h} \mid X_n, X_{n-1}, \ldots)$. When $h = 1$, $\hat{X}_n(1) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_n$, where $\mathbf{H}^{\mathrm{T}}\mathbf{A} = \boldsymbol{\beta}^{\mathrm{T}} = (a_p, a_{p-1}, \ldots, a_1)$. Now suppose we have a realization of size $n + M$, and we use the first n observations to fit the model and use the estimated model to predict the next M observations. Let $\hat{\boldsymbol{\beta}}$ be the least squares estimate of $\boldsymbol{\beta}$, and we know $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, $\mathrm{var}(\hat{\boldsymbol{\beta}}) = \frac{\sigma_e^2}{n}\boldsymbol{\Gamma}^{-1}$, where $\boldsymbol{\Gamma} = E(\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}})$. Let $\tilde{\boldsymbol{\beta}}$ be any other estimator of $\boldsymbol{\beta}$. Define the loss function

$$L(\tilde{\boldsymbol{\beta}}) = \frac{1}{M}\sum_{j=n}^{n+M-1}(X_{j+1} - \tilde{\tilde{X}}_j(1))^2,$$

where $\tilde{\tilde{X}}_j(1) = \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j$. Thus the risk is given by

$$\Phi(\tilde{\boldsymbol{\beta}}) = E\{L(\tilde{\boldsymbol{\beta}})\} = E\{(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{C}_M(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\} + \sigma_e^2,$$

where

$$\mathbf{C}_M = \frac{1}{M}\sum_{j=n+1}^{n+M}\mathbf{x}_j\mathbf{x}_j' \cdot \operatorname*{plim}_{M\to\infty}\mathbf{C}_M = \boldsymbol{\Gamma}.$$

If $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$, then $\Phi(\tilde{\boldsymbol{\beta}}) = \sigma_e^2\{1 + (p/n)\}$, a well-known result, which is used for order

determination. Is it possible to find a $\tilde{\boldsymbol{\beta}}$ such that $\Phi(\tilde{\boldsymbol{\beta}}) < \Phi(\hat{\boldsymbol{\beta}})$? This is where we can use the James–Stein result. Suppose we define the shrinkage estimator

$$\tilde{\boldsymbol{\beta}} = \left[1 - \frac{(p-2)\,S}{n(n-p+2)} \quad \frac{1}{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Gamma}\,\hat{\boldsymbol{\beta}}}\right]\hat{\boldsymbol{\beta}} \quad ,$$

where $S = \Sigma_{t=1}^{n}(X_t - \hat{a}_1 X_{t-1} \ldots - \hat{a}_p X_{t-p})^2$, $X_0 = \ldots = X_{-p} = 0$, then it can be shown (we are assuming $\boldsymbol{\Gamma}$ is known, but this can be relaxed).

$$E(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) = \frac{\sigma_e^2}{n}\left[p - \frac{2(n-p)(p-2)}{n-p+2}\,E\left(\frac{1}{p-2+2g}\right)\right].$$

Hence for $p > 2$, $\Phi(\tilde{\boldsymbol{\beta}}) < \Phi(\hat{\boldsymbol{\beta}})$. The time series analysis differs from classical regression when we wish to use this $\tilde{\boldsymbol{\beta}}$ to predict more than one step. Denote by $\tilde{\mathbf{A}}$ the matrix obtained from $\mathbf{A}$ by replacing a_p by $\tilde{\beta}_p$, a_{p-1} by $\tilde{\beta}_{p-1}$, etc. Let $\tilde{X}_n(h) = \mathbf{H}^{\mathrm{T}}\tilde{\mathbf{A}}^h\mathbf{x}_n$ be the h-step ahead predictor, and let

$$\Phi(\tilde{\tilde{\boldsymbol{\beta}}}, h) = E\,\Sigma\,(X_j - \tilde{\tilde{X}}_j(h))^2/(M+n-1)$$

$$= E\,[\mathbf{H}^{\mathrm{T}}(\tilde{\mathbf{A}}^h - \mathbf{A}^h)\,\boldsymbol{\Gamma}\,(\tilde{\mathbf{A}}^h - \mathbf{A}^h)\,\mathbf{H}] + \sigma_e^2 \sum_{j=0}^{h-1} \mathbf{H}^{\mathrm{T}}\mathbf{A}^j\mathbf{A}^{\mathrm{T}j}\mathbf{H}.$$

We can use the expression by Subba Rao (1980) or Yamamoto (1976) to obtain an expression for $\Phi(\tilde{\boldsymbol{\beta}}, h)$, provided an expression for $E(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\,\boldsymbol{\Gamma}\,(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'$ is known. I am not aware of an expression. The interesting thing is to find this variance-covariance matrix, and to show $\Phi(\tilde{\boldsymbol{\beta}}, h) < \Phi(\hat{\boldsymbol{\beta}}, h)$ for all $h > 1$ and $p > 2$. If this inequality is true, it can be used to improve the predictors.

D. P. J. Laycock (UMIST): The results in Egerton and Laycock (1982) can be used to give an exact expression—with $\frac{1}{2}p$ terms—for (3.11) as an alternative to Professor Copas's approximation (3.12). For p even, we find

$$E(K) = 1 - (\tfrac{1}{2}p - 1)\left\{\sum_{j=1}^{\frac{1}{2}p-1}\left((-1)^{j+1}\,(2\delta^2)^j(\tfrac{1}{2}p-2)!/(\tfrac{1}{2}p-j-1)!\right)\right.$$

$$\left. + (-1)^{\frac{1}{2}p-1}\,(2\delta^2)^{\frac{1}{2}p-1}e^{-\frac{1}{2}\delta^{-2}}(\tfrac{1}{2}p-2)!\right\}$$

with a similar expression for p odd. When for example $p = 10$, as used in Table 1, this becomes

$$E(K) = 1 - 8\delta^2 + 48\delta^4 - 192\delta^6 + 394\delta^8\,(1 - e^{-\frac{1}{2}\delta^{-2}}).$$

This expression has the particular advantage over (3.12) of being valid for large δ, and hence for high noise/signal ratios, which are "characteristic of much research in the medical and social sciences", as pointed out by Professor Copas.

Unpublished work in the Ph.D. thesis of Egerton (1979) examined the generalization of (4.2) to the case where the expectation E is over some arbitrary, but known, distribution F. This analysis suggested that an adaptation of Bhattacharya's (1966) *multiple* shrinkage estimator should be used whenever

$$\int \mathbf{x}\mathbf{x}^{\mathrm{T}}\,dF(\mathbf{x})$$

has full rank (greater than two). It should be possible to further modify this estimator, using the shrinkage factor in Alam (1973), so as to produce an *admissible* estimator for the loss function (4.2). The complexity of such a multiple shrinkage estimator would mitigate against its adoption in practice. It would be interesting to see numerical comparisons between it and Professor Copas's more easily applied methods.

A derivation of shrinkage estimators as empirical Bayes estimators is given by Oman (1981). His approach has the advantage of not using (5.2)—which equation seems to imply that one's prior beliefs about $\boldsymbol{\beta}$ depend on the design of the experiment constructed to elicit information about $\boldsymbol{\beta}$. He follows Chen (1979) in assuming a joint multivariate normal distribution for y, $\mathbf{x}$ and a Wishart prior for $\boldsymbol{\Sigma}^{-1}$. For pure shrinkage, and in the notation of this paper, he suggests

$$K(k)^{-1} = \left(1 + \frac{k}{n}\right)\left(1 + \frac{\alpha}{n}\,\hat{\boldsymbol{\beta}}^{\mathrm{T}} V\,\hat{\boldsymbol{\beta}}\right),$$

where $k \geqslant p + 1$ and $\alpha = -k\{1 + (k/n)\}^{-1}\,(\mathbf{y}^{\mathrm{T}}\mathbf{y} + k\sigma^2)^{-1}$.

Professor J. A. Anderson (University of Newcastle-upon-Tyne): Professor Copas is to be congratulated on an important and stimulating contribution to regression theory. His insights into the relationship between retrospective fit and prospective fit and subset selection are most illuminating.

As a potential user of the techniques introduced here, several questions and comments arise.

It is as well to recall that this paper is concerned only with problems where the construction and validation data sets are sampled in the same way. This specifically excludes the important case where the construction set has selected $\mathbf{x}$-values (e.g. in calibration) or partially selected $\mathbf{x}$-values (e.g. in a stratified medical study) while the validation sample has $\mathbf{x}$-values drawn randomly from some population. Secondly, attention is focussed on improved predictive point estimates. In practice, predictive intervals will often be required but there is no mention of these in the paper.

Several preshrunken estimators are introduced in the paper. A major advantage of these is that they give predictive estimators at little extra cost in terms of computing and time. However, their development necessarily depends on assumptions. In the linear regression case, normal, independent residuals are assumed. The practitioner will need some reassurance of robustness before he can be confident of using expressions like (4.1) with $\hat{K}$ given by, say (3.7) with $K = p - 2$.

The assumptions seem to be more stringent and less natural in the case of binary regression in Section 8. There the sampling theory justification for the preshrunk predictive estimators seems to need the $\mathbf{x}$-values to have a multivariate normal distribution. In many cases, this will be far from true. For example, in two-group discrimination the standard set of assumptions leads to $\mathbf{x}$ having a mixture of two multivariate normal distributions. In many other cases, some of the x_j will be categorical, again negating the multivariate normal assumption. Wisely, Professor Copas has concentrated on weak dependence in the binary regression case. Even moderately strong dependence often leads to complete separation of the sample points and infinite maximum likelihood estimates of $\boldsymbol{\beta}$. Even the methods of this paper are unlikely to have sufficient power to shrink these back to finiteness.

In some situations shrunken estimators are clearly desirable but the question remains of the acceptability of the preshrunken estimators advocated here. It may be that data-based methods for "estimating" K will usually be required.

Professor D. M. Titterington (University of Glasgow): Tonight's wide-ranging paper has neatly exemplified the links among several shrinkage methodologies. A parallel illustration can be drawn from the field of density estimation. Non-parametric estimates of a univariate density function $f(\cdot)$ are typically "derivatives" of a smoothed version of the empirical distribution function provided by a construction sample, $\mathrm{CS} = (x_1, \ldots, x_n)$. As in the present example there are two aspects to the smoothing: in what manner to smooth (e.g. smooth $\hat{\boldsymbol{\beta}}$ towards zero or, in the density estimation context, use a kernel-based method) and to what extent to smooth (e.g. choice of κ in (5.1) or choice of kernel smoothing parameter). The starting point is a predictive indicator of performance

$$d(\delta_y, \hat{f}), \tag{1}$$

in which δ_y is the Dirac δ on y, $\hat{f}$ is a density estimate and $d(\cdot, \cdot)$ is a "distance" measure.

One approach is to take $\hat{f}(\cdot) = \hat{f}_\kappa(\cdot \mid \mathrm{CS})$, some prescription, such as the kernel approach, with smoothing parameter κ, to envisage y as a validating sample of size 1 and to average (1)

over y and the CS. This often reduces, as far as choice of κ is concerned, to

$$E_{CS}\, d\{f(\cdot), \hat{f}_\kappa(\cdot \mid CS)\}. \tag{2}$$

The most familiar version of this gives mean integrated squared error. The minimizing $\hat{\kappa}$ is a function of the true $f(\cdot)$ and, in practice, some empirical-Bayes adaptation is required.

Alternatively, we might look at

$$n^{-1} \sum_{i=1}^{n} d(\delta_{x_i}, \hat{f}). \tag{3}$$

Minimization of (3) gives perfect retrospective fit with no smoothing. Smoothing is sought either by choosing a prescription $\hat{f}_\kappa(\cdot \mid CS)$ and selecting κ by cross-validation (Habbema *et al.*, 1974) or by adding a roughness penalty. The type of penalty function determines the manner of smoothing. Silverman (1982), for instance, smooths towards one of a class of maximum entropy densities. The degree of smoothing here is usually chosen by generalized cross-validation or by a goodness-of-fit constraint (Good and Gaskins, 1980); see Golub *et al.* (1979) for a generalized cross-validation attack on tonight's problem.

The version of all this for a finite sample space of s cells corresponds to smoothing relative frequencies, $\mathbf{r}$. Fienberg and Holland (1973) consider convex smoothing towards the uniform distribution, $s^{-1}\mathbf{1}$:

$$\hat{\mathbf{f}}(\kappa) = (1-\kappa)\,\mathbf{r} + \kappa s^{-1}\mathbf{1} \tag{4}$$

They follow the path to (2) and use empirical Bayes to choose κ. Stone (1974) and Titterington (1980) discuss cross-validatory choice and the latter gives a kernel formulation for (4). The estimator (4) can also be shown to be, approximately, a minimum penalized distance estimator. Suppose $\tilde{\mathbf{f}}(\kappa)$ solves

$$\min_{\mathbf{f}} \left\{ (\mathbf{r}-\mathbf{f})^{\mathrm{T}}\,(\mathbf{r}-\mathbf{f}) + \tfrac{1}{2}\,\kappa s^{-1} \sum_{i=1}^{s} \sum_{j=1}^{s} (f_i - f_j)^2 \right\},$$

subject to $\mathbf{f}$ being probabilities. Then $\tilde{\mathbf{f}}(\kappa) = \hat{\mathbf{f}}(\kappa) + o(\kappa)$. For this example, generalized cross-validatory choice of κ is also asymptotically optimal in consistently estimating the minimizer of $E_{CS}(\mathbf{f}-\hat{\mathbf{f}}(\kappa))^{\mathrm{T}}(\mathbf{f}-\hat{\mathbf{f}}(\kappa))$.

Professor R. L. Plackett (University of Newcastle-upon-Tyne): This is a very fine paper, and an interesting pointer to the direction of future developments in regression analysis for both continuous and discrete data. Among many good things, I particularly liked the illuminating comments on loss functions in Section 4.

The point is made in Section 1 that assessment of retrospective fit "uses the data twice". If I have counted correctly, the data are actually used three times, because they are first inspected with a view to deciding whether or not to transform any of the variables, and whether squares, products, or other functions should go into the full regression. However much objectivity is attempted, there is always a stage where subjective assessments have to be made. Thus Figs 4a and 4c are described as giving reasonable fits, whereas both seem to have more curvature than in Fig. 4d, say.

Standard methods for the analysis of residuals certainly use the data more than once, but some allowance is made for this and I wonder whether they are now being rejected in favour of cross-validatory methods. When the vectors $\mathbf{x}$ in both CS and VS arise at random, an assumption of multivariate normality, of the type made in Section 4, enables all the parameters to be estimated and predictions can be made on this basis. Presumably the predictions are much the same as from the two-stage validation process. Is there any loss of efficiency in checking that the model is correct?

I have a few detailed comments:

(i) Patnaik (1949) showed that the non-central χ^2 distribution is closely approximated by a

multiple of central χ^2 and this should lead to explicit formulas in place of (3.10), (3.11) and equations elsewhere.

(ii) The assumption is made just before (8.3) that the variation in the weights w_i can be ignored. If we turn to the first example of probit analysis given by Finney (1971), we see (p. 62) that the weights range from 16.8 to 29.2, or by a factor of 2. Without following this through Section 8, the approximation does seem rather rough.

(iii) The variation in $\mathbf{x}$ envisaged in Section 8 can, so to speak, be transferred to the other end: that is, y has a more variable distribution than specified here. Several models which achieve this objective have been proposed, and a comparison of predictions would be worthwhile.

Finally, here is a point of etymology.

The concept of "shrinkage" is already present in the term "regression", since Galton pointed out – in effect – that sons are shrunk in height compared with their fathers when tall.

Professor Murray Aitkin (University of Lancaster): The choice of a better predictor than the full least squares regression is an important problem, and this paper gives an interesting approach. Stepwise selection methods have intractable sampling properties, as Professor Copas notes. One way of avoiding this difficulty is to use the simultaneous inference approach which considers all possible subsets. Scheffé-type procedures can be fairly easily constructed; the theory was set out with examples in Aitkin (1974). That paper also considered mean square prediction error, and the selection of variable subsets which did not increase the MSPE. Of course in practice we cannot *know* that the MSPE has not been increased, but it is possible to *test* the hypothesis that it has not, using a simultaneous test valid for all subsets selected in any way. Aitkin considered the cases where each new $\mathbf{x}$ in the validation sample is fixed, or has a uniform distribution over the fixed $\mathbf{x}_i$ in the construction sample (which is equivalent to Professor Copas's model in Section 3 as Professor Stone noted), or has the same multivariate normal distribution as the $\mathbf{x}_i$ in the construction sample.

With sufficient real data, we can estimate the true MSPE from subset or other modified predictors. The first example considered by Professor Copas raises difficulties. Would we seriously attempt to predict Y from 14 predictors with a sample of 8? Rencher and Pun make clear the gross overestimation of R^2 which occurs with singular models. The simultaneous approach makes it clear that the construction sample must be larger than the number of predictors.

To assess the performance of the STP, I have reversed the sample-splitting used by Professor Copas: the last 23 observations will constitute the construction sample used to predict costs for the validation sample of the first 8 observations.

Table D1 summarizes a backward elimination from the full regression, with the construction sample $\hat{\sigma}^2$ and R^2, and the validation sample MSPE and cross-validation R^2, given for each step.

The shrinkage of R^2 on cross-validation is substantial, and the MSPE is four to ten times as large as the "unbiased" estimate of σ^2, for the early steps in backward elimination. At step 8, after the omission of variable 1, $R^2_{(CV)}$ increases and the MSPE decreases substantially, and at the next step $R^2_{(CV)}$ reaches its maximum in this sequence, and the MSPE its minimum, very close to the RMS estimate of σ^2. From here on until step 12, R^2 does *not* shrink on cross-validation, and $\hat{\sigma}^2$ and the MSPE are very similar.

Thus the five-variable model using variables 7, 10, 11, 13 and 14 might be considered for prediction. Further elimination of variables might also be considered, since the MSPE does not increase much for the two-variable model using 7 and 13.

The two-variable model using variables 1 and 2 considered by Professor Copas does very well on cross-validation, as he notes. In fact, it does much better on cross-validation than it does in the original sample, because I have reversed the two samples. For the construction sample, $R^2_{(C)} = 0.542$ and $\hat{\sigma}^2 = 0.468$ while on cross-validation, $R^2_{(CV)} = 0.780$ and MSPE = 0.279.

The simultaneous test makes clear just how parsimonious one can be in the prediction of y. Table D2 shows the lower limits of $R^2_{(CV)}$ for an MSPE-adequate (α) and an R^2-adequate (α) subset, as these were defined in Aitkin (1974).

The severe criterion of $\alpha = 0.05$ would allow *all* variables to be omitted without significantly increasing the MSPE. Indeed, Table D1 shows that the sample MSPE is smaller for the constant

TABLE D1

Backward elimination from the full regression for the second sample of 23, and cross-validation R^2 and MSPE for the first sample of 8

Step	*Variable omitted (t)*	R^2 *(C)*	$\hat{\sigma}^2$	R^2 *(CV)*	*MSPE*
0	–	0.913	0.223	0.646	0.895
1	4 (0.00)	0.913	0.198	0.646	0.896
2	9 (0.13)	0.913	0.179	0.650	0.826
3	8 (0.32)	0.912	0.164	0.683	0.809
4	12 (0.41)	0.910	0.153	0.684	0.965
5	6 (1.27)	0.898	0.160	0.662	1.240
6	5 (1.41)	0.883	0.131	0.581	2.153
7	2 (2.13)	0.845	0.212	0.623	2.095
8	1 (2.04)	0.802	0.253	0.781	0.707
9	3 (1.57)	0.772	0.275	0.818	0.288
10	10 (1.17)	0.753	0.281	0.696	0.387
11	11 (2.00)	0.698	0.325	0.715	0.397
12	14 (4.06)	0.437	0.576	0.578	0.382
13	13 (2.99)	0.185	0.794	0.383	0.539
14	7 (2.18)	–	0.930	–	0.733

TABLE D2

α	*MSPE-adequate* (α) *limit*	R^2*-adequate* (α) *limit*
0.50	0.593	0.754
0.25	0.422	0.666
0.10	0.172	0.536
0.05	0.00	0.419

model than for the full model. A less severe $\alpha = 0.25$ leaves the two-variable model using 7 and 13 as a minimal adequate subset. The weaker criterion of R^2-adequacy (0.25) would require the addition of variable 14 to this subset. Professor Copas's subset of variables 1 and 2 is another minimal MSPE-adequate (0.25) set of two variables.

Plots of observed against fitted y for the three-variable and five-variable subsets show little if any overprediction. Professor Copas's two-variable subset shows *under*prediction. Does this simply correspond to the overprediction in his Fig. 1 when the samples are reversed? If so, is overprediction a real effect?

Dr P. J. Brown (Imperial College, London): Using an ingeniously new and intuitively appealing notion of prediction shrinkage, Professor Copas has been able to give new life to the shrinkage estimator originally proposed by Charles Stein. I would like to examine this theoretical underpinning in the present context.

Let us assume, to avoid unnecessary clutter, that $\sigma^2 = 1$ and $\alpha = 0$. Suppose we have m future observations to predict, where

$$Y_j = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j + \epsilon_j, \quad j = 1, \ldots, m.$$

In the parametric cost model $m = 23$. Taking quadratic loss and averaging over ϵ, for any estimator $\tilde{Y}_j$ of Y_j

$$\begin{aligned} E\,\Sigma\,(\tilde{Y}_j - Y_j)^2/m &= 1 + \Sigma\,(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_j)^2/m \\ &= 1 + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \end{aligned}$$

where $\mathbf{W} = \Sigma\,\mathbf{x}_j\mathbf{x}_j^{\mathrm{T}}/m$. Now the least squares estimator $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution,

$N(\boldsymbol{\beta}, \mathbf{V}^{-1}/n)$, with $\mathbf{V} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ and $\hat{\boldsymbol{\mu}} = \mathbf{V}^{\frac{1}{2}}\hat{\boldsymbol{\beta}}$ is distributed $N(\boldsymbol{\mu}, \mathbf{I}/n)$ with $\boldsymbol{\mu} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\beta}$ and $\mathbf{V}^{\frac{1}{2}}$ a symmetric matrix square root. Thus the prediction loss is

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\,\mathbf{A}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \text{ with } \mathbf{A} = \mathbf{V}^{-\frac{1}{2}}\mathbf{W}\,\mathbf{V}^{-\frac{1}{2}}$$

$$= \Sigma\, L_i(\hat{\eta}_i - \eta_i)^2,$$

where $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} = \mathrm{diag}\,(L_1, \ldots, L_s)$, $\mathbf{Q}^{\mathrm{T}}\boldsymbol{\mu} = \boldsymbol{\eta}$, $\mathbf{Q}$ orthogonal, $0 < L_1 \leqslant \ldots \leqslant L_s$, with $s \leqslant \min(m, p)$. Typically even when $m > p$ the L_i may be quite dispersed, each tending to one only as m, n both tend to infinity, the case essentially considered by Professor Copas. This diversity of L_i in the prediction loss structure implies that in general Stein shrinkage will not dominate least squares. The PMSE of this paper is all too forgiving of inevitable random differences between $\mathbf{V}$ and $\mathbf{W}$, for even when the $\mathbf{X}_i$ come independently from the same normal $(\mathbf{0}, \boldsymbol{\Gamma})$ then $n\mathbf{V}$ and $m\mathbf{W}$ are Wishart on n and m degrees of freedom. Now, Stein shrinkage will dominate least squares if (Brown and Zidek, 1980,), and only if (Thisted and Morris, 1980), $s \geqslant 3$ and

$$L_s \leqslant [2s/(s+2)]\bar{L}.$$

If for example $s = 3$ then $L_s \leqslant (6/5)\bar{L}$ is required, so that $\mathbf{A}$ needs to be close to the identity matrix. The above two papers consider a large class of estimators, including adaptive ridge, and give conditions for their dominance of least squares. Berger (1980) emphasizes a range of other criteria for good estimation and for Professor Anderson's benefit obtains confidence regions.

The problem with dominating estimators is that typically they do not do enough to shrink least squares, at least in *a priori* likely regions of the parameter space. As a general recipe I am sceptical about the sort of shrinkage in tonight's paper corresponding as it does to the prior distribution (5.2) with its data dependence. This is somewhat unnatural despite mitigating invariance arguments. If one must have a recipe then adaptive ridge certainly has a more natural prior. It shrinks the β_i differently, depending on their precision and influence. One cannot take refuge in the thought that the parameter space does not matter. With respect to the above prediction loss, Stein shrinkage, as implemented in tonight's paper, does not usually dominate least squares.

Dr A. J. Lawrance (Birmingham University): I have read this paper with much admiration; it addresses the problem of regression prediction under the assumption that future data will be "rather similar" to past data; this will often be necessary and sensible when the aim is prediction.

I claim no past or present research activity in this area, or familiarity with the literature, an admission of ignorance possibly, but historically no deterrent to the expression of interest at our meetings. Thus, I have some general comments and questions:

While accepting that a good retrospective fit is not required for good prediction, I would welcome a little more discussion as to why the least square predictions tend to be too extreme.

I am a little suspicious that the prediction problem is exacerbated for data with outlying points (non-robust data sets) and by the least squares method of estimation. This leads me to enquire whether anything is known about the prediction problem in relation to robust regression retrospective fits; if so, is it so acute?

Claims for shrinkage are investigated using PMSE; I wonder about higher order effects in connection with non-symmetry in the distributions of the prediction variables. Properties of the shrinkage estimators at extreme points in the predictor space would seem to be of interest; there is a paper by Matloff (1982) dealing with James–Stein regression estimation in prediction.

My last comment relates to the detection and characterization of data points of the construction sample which will be influential in prediction. Much has been developed on influence in relation to fit, but I have recently seen work by Johnson and Geisser (1982, 1983) concerned with prediction; this is in the Bayesian mode, but should have other parallels. The concern here is with predictive destributions and the way these change as data points are deleted. Cook-type statistics are shown to pertain to predictive means, and further statistics concerning predictive variance are derived.

I feel that the basic thoughts expressed in this paper will be of great value far outside the present contexts. I want to add my congratulations to Professor Copas.

Dr B. W. Silverman (University of Bath): May I add my congratulations to Professor Copas for a fascinating paper and an excellent presentation.

I should like to add to Professor Titterington's remarks about the connections between smoothing methods and the problems discussed in tonight's paper. The problem of estimating a curve or function f non-parametrically may be viewed as an inference problem in a high (or infinite) dimensional space. One way of seeing this is to expand the unknown curve $f(\cdot)$ as a Fourier series $\Sigma \beta_\nu \phi_\nu(\cdot)$; estimating f is then equivalent to the multivariate problem of estimating the unknown coefficients, or parameters, β_ν. In some circumstances a different series expansion of f is appropriate, but the general principle remains the same, whether f is a density function, a hazard rate, a regression curve, or some other curve of interest. Even methods of smoothing which do not explicitly involve series expansions can be regarded in this way. For example, it is very well known that convolving the data with a smoothing kernel corresponds precisely to a Fourier series smoothing method where the raw sample Fourier transform is multiplied by the Fourier transform of the kernel. For a discussion in the context of density estimation see, for example, Watson and Leadbetter (1963), who give an explicit form for the optimum function by which the raw sample Fourier transform should be multiplied. (This optimum unfortunately depends on the unknown density.) The penalized maximum likelihood approach referred to by Professor Titterington is also approximately of this form; see, for example, the later sections of Silverman (1982) where it is shown that it is most convenient to think in terms of a *generalized* Fourier series in this case.

The main difference between smoothing and Professor Copas's shrinkage is that, in smoothing, it is appropriate to shrink the sample (generalized) Fourier coefficients more and more the further one goes along the series, while Professor Copas's parameter estimates are all shrunk by the same factor K. In other words, in the smoothing case the high frequency "noise" components of the sample are damped out more strongly than are the low frequency "signal" components.

Focus attention on non-parametric regression, where the object is to fit a model of the form

$$y = f(u) + \epsilon,$$

to a set of data (u_i, y_i), $i = 1, \ldots, n$. Assume we are interested in values of u uniformly distributed on a fixed finite interval; then low prediction error corresponds precisely to good estimation of f in the mean integrated square error sense.

A natural approach is to define a set of functions ϕ_ν, $1 \leqslant \nu \leqslant n$, which are orthonormal with respect to the design points u_i; for example a set of orthogonal polynomials of increasing degree could be used. Express $f = \Sigma \beta_\nu \phi_\nu$. The raw unbiased estimates of the parameters β_ν are $\Sigma_i y_i \phi_\nu(u_i)$ and to obtain a good estimate of f one would shrink these estimates by a suitably chosen filter κ_ν; this would have κ_ν decreasing from one to a value near (or equal) zero as ν increased.

Why is it inappropriate to use a constant shrinkage factor K? One obvious explanation is the prior belief that f is a smooth function, in other words that the true coefficients β_ν are small for large ν. But how is the problem influenced by the fact that the set of possible $\mathbf{x}$ values at which future predictions $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ are of interest is of a very particular form? To predict y for a given value u, we require an estimate of $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\phi}(u)$ where $\boldsymbol{\phi}(u)$ is the vector of coefficients $\phi_\nu(u)$. Thus all $\mathbf{x}$ values of interest lie on a one-dimensional curve C, parametrized by u, in n-dimensional space. The high co-ordinates of a particle travelling on this curve oscillate rapidly and this strange form of the curve C no doubt affects the appropriate choice of shrinkage factors κ_ν. For instance, if the new value u at which a prediction is required is subject to some error, then the $\phi_\nu(u)$ for large ν will be subject to gross error, suggesting that a large degree of shrinkage is appropriate for the corresponding β_ν.

Are there other multiple regression problems, not immediately connected with smoothing, where different amounts of shrinkage should be applied to the various parameter estimates? Certainly this is a possibility that should be taken into account.

The following contributions were received in writing, after the meeting.

Dr C. Chatfield (Bath University): The author has made an impressive contribution to conventional regression methodology. However, I really cannot let him get away with Example 1. I realize that this example is primarily meant to illustrate the shrinkage phenomenon. Nevertheless, the innocent reader may go away with the impression that it is acceptable to fit a regression model to the construction sample, even though the number of observations (8) is

less than the number of variables (14) and even though one variable (x_{11}) is actually constant in the construction sample. Most statisticians would, I think, agree that this is not acceptable.

The author states that 31 is a "very large sample size" in the context of cost models. However, this sort of situation is one where I would be very dubious about using any regression model for prediction even with a much larger sample size. What is the population the model refers to? If we build a new aircraft to a new design of different material, perhaps in a different country, can we use the model to predict? I doubt it. Indeed I doubt if regression models should ever be used to predict in this sort of situation, though no doubt they are. Rather I would suggest that, if regression models are to be used at all, they should be used for exploratory purposes to indicate to the engineer which variables are likely to be important to look at when preparing cost-estimates. This exploratory use of regression is in my view the most important one for many situations. My doubts about conventional regression methodology arise because too much emphasis is placed on fitting the "best equation" to a single set of data, rather than trying to find a relationship which describes data collected under different conditions, by different people or whatever. I am also very dubious about some of the assumptions typically made in regression. For example, the author assumes (line 4) that the x_i's "occur randomly over some population of values''. This is arguably often untrue. I do not pretend to follow all of the mathematics in this paper, but it is worth remembering that its validity depends crucially on the assumptions on which it is based.

In his excellent presentation, the author reminded us that multiple regression is not only one of the most used statistical techniques but also one of the most abused. A major task for the statistical profession is to try to find ways of shrinking this abuse.

Dr A. J. Miller (CSIRO, Melbourne, Australia): At a rough guess, about 10^5 data sets per day are used as input to multiple regression packages around the world. Subset selection is probably used in most of these cases. The results in Section 6 of this paper will be a nasty shock to many of the users of these packages, though the results have broadly been known for many years.

Suppose that we believe it is reasonable to assume that the relationship between a variable $\mathbf{Y}$ and a set of k available predictors $\mathbf{X}^{\mathrm{T}} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k)$ is

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the unknown residual variation, ϵ, is randomly sampled from some distribution which is common to all samples and has finite variance. If we estimate the regression coefficients for a subset, X_p, of $p < k$ variables, chosen *a priori*, using least squares, then we are estimating a new set of regression coefficients, β_p, in the model

$$\mathbf{Y} = \mathbf{X}_p\,\boldsymbol{\beta}_p + \boldsymbol{\eta},$$

where the unknown residual variation, $\boldsymbol{\eta}$, is $\boldsymbol{\varepsilon}$ augmented by that part of the variation of $\mathbf{Y}$ in the $(k-p)$-dimensional subspace of $\mathbf{X}$ orthogonal to $\mathbf{X}_p$, and

$$\boldsymbol{\beta}_p = (\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1}\mathbf{X}_p\mathbf{X}\,\boldsymbol{\beta}.$$

If we denote the least-squares estimate of this vector by $\hat{\boldsymbol{\beta}}_p$, then $E(\hat{\boldsymbol{\beta}}_p) = \boldsymbol{\beta}_p$ only if the subset of variables is chosen *a priori* and, as John Copas points out in Section 6, the expected values of the sample regression coefficients are often much larger in absolute value than the corresponding elements in $\boldsymbol{\beta}_p$ if the same data are used for both subset selection and for estimation. In derivations of mean squared errors of prediction in subset selection, the difference between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_p$ (augmented with $k-p$ zeroes) is usually called "bias"; let us call it "omission" bias and use the term "selection" bias to indicate the difference between $\boldsymbol{\beta}_p$ and $E(\hat{\boldsymbol{\beta}}_p)$ when selection and estimation are from the same data. (N.B. the selection bias is then a function of the selection procedure and stopping rule used, and is liable to be larger when an exhaustive search has been carried out for the best-fitting subsets than when a simple forward selection has been used.) In practice when there is keen "competition" for selection between many subsets of variables which all give similar residual sums of squares, the selection bias is typically of the order of one to two "standard errors" in the regression coefficients where the "standard errors" are the usual ones obtained from the square roots of the diagonal elements of

$$\sigma_p^2(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1},$$

where σ_p^2 is the mean square of the true residuals, $\boldsymbol{\eta}$. This applies except for any "dominant"

variables, that is variables which are present in all subsets which fit reasonably well. As the usual estimate of σ_p^2 is also biased (too small) due to the overfitting which occurs in subset selection, the selection bias is often three or four times the sample estimate of the appropriate standard error.

Because of selection bias, it is not surprising that it is often better to use all variables (or sometimes even none) rather than a stepwise procedure for prediction. This result was already known to some through the work of Ken Berk (1978). Also the supposed minima of Mallows' C_p and of Akaike's Information Criterion are often closer to being local *maxima* of the mean squared error of prediction-versus-p curve when selection and estimation are from the same data; this is as suggested by Fig. 3 in the paper.

What can be done to improve predictions from subset selection procedures? One method is to use independent data sets for selection and estimation. This guarantees unbiased estimates (i.e. of β_p), but is inefficient. In any case, what should we do if there is some other subset, or many other subsets, which fit the estimation set of data much better than the chosen one?

A crude alternative method is that of using some kind of shrinkage such as that suggested in Section 6. In this case, a small amount of shrinkage reduces the bias in most of the regression coefficients. Shrinkage methods tend though to shrink all of the regression coefficients, we want a method which has little effect upon the regression coefficients of the dominant variables while substantially reducing the others. A further alternative is to estimate the selection bias and hence obtain almost unbiased estimates; shrinkage can then be applied to obtain a good estimator for prediction. The estimation of selection bias is always feasible using Monte Carlo methods, and is sometimes feasible using analytic methods, as John Copas has shown. Another method which the discussor is currently investigating is that of using maximum likelihood *conditional upon selection* to obtain estimates of the regression coefficients.

In view of the 10^5 or so multiple regressions per day, research into the properties of the estimates from subset selection procedures is long overdue.

Professor J V Zidek (University of British Columbia): The argument for shrinking regression coefficient estimators which derives from equations (3.2), (3.3) and (4.1) is novel and compelling. It, like that of Stein (1956), casts doubt on the relevance of the large sample theory of Fisher in problems where the number of parameters, p, is large. And while a large p alternative to Fisher's theory has not yet emerged there are a great many examples where superior alternatives to large-sample based procedures have been exhibited. Stein's method (Stein, 1973) has made the production of such alternatives quite routine (see Ghosh *et al.*, 1983 for a survey of recent results in discrete cases).

Although the PMSE analysis of Section 4 establishes that the preshrunken (PSP), like the LS predictor (LSP), is minimax its worth, if any, derives not from this weak qualification, but rather from that of the LSP. However, the status of the latter has declined in recent years. The data are now transformed, outliers and influential observations are eliminated and the LSP is robustified. One wonders if the arguments of Section 3 or 4 would support the preshrinking of these more realistic alternatives to the LSP.

It is well known that PSP's relative superiority to LSP is greatest when n is small and the signal-to-noise ratio is near 0. So the expansion given in (3.12) might better have been about the point $\delta^2 = \infty$. Its leading term would then be $E_{\mu} \| \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \|^2$ with $\boldsymbol{\mu} = \mathbf{0}$ when the problem is put into the canonical form in which $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I})$ is observable (to avoid unnecessary technical complication, $\sigma^2 = 1$ is assumed). And with respect to this leading term it is easily shown that among estimators of the form $\hat{K}\mathbf{X}$, $\hat{K} = (1 - k/\| \mathbf{X} \|^2)$, the best has $k = p - 2$. In fact this local superiority of the James–Stein estimator holds over the larger class where $\hat{K} = (1 - T(\| \mathbf{X} \|) \| \mathbf{X} \|^{-2})$ and T is a non-decreasing, non-negative function which is bounded above by $2(p - 2)$, i.e. over a broad class of minimax alternatives to LSP.

This local superiority is gained at the loss of significant improvement over the LSP outside a small neighbourhood of the origin. By relaxing this requirement more practical alternatives to LSP may be achieved.

The author's unrealistic assumption that $\mathbf{V} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ results in a great technical simplification of the problem. Baranchik in the work referred to by Professor Copas avoids it. However, his result requires a deep and ingenious argument. And its multivariate counterpart remains an unsolved problem. A result of Haff (1977) shows that replacing $\mathbf{V}$ by its seemingly equivalent

alternative $\mathbf{X}^T\mathbf{X}/n$ may have surprising consequences in a covariance estimation problem: the performance ranks of two estimators, as determined by their MSE's are reversed.

As Zidek (1978) and van der Merwe and Zidek (1980) argue, what Professor Copas calls preshrinking might more accurately be called filtering. This can be more clearly seen in the multivariate case where, in equation (4.1), $\hat{\mathbf{K}}$ is a matrix which operates on the right of $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}}\hat{\mathbf{K}} = (\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{Y}^*$ where $\mathbf{Y}^* = \mathbf{Y}\hat{\mathbf{A}}\mathbf{F}\hat{\mathbf{A}}^{-1}$, $\hat{\mathbf{A}}$ estimates the transformation which takes $\mathbf{Y}$ to its canonical form, $\mathbf{F} = \text{diag}\{F_1 \geqslant \ldots \geqslant F_n\}$, $F_i = [1 + \alpha - \alpha/\hat{\rho}_i^2]_+$ and ρ_i^2 is the ith canonical correlation coefficient. F filters out those canonical Y-variates, i, which have negligible canonical correlations. In the univariate problem treated here $\hat{\mathbf{A}} = \mathbf{I}$, and $F = [1 + \alpha - \alpha/\text{corr}^2(\mathbf{Y}, \hat{\boldsymbol{\beta}}^T\mathbf{X})]_+$.

The **Author** replied later, in writing, as follows.

It is a pleasure to thank the contributors to the discussion for their kind reception of the paper, and for their constructive and penetrating remarks. Their comments serve not only to expose the strengths and weaknesses of the paper but also to highlight possible further work.

Dr Dunsmore raises two general points which repay careful thought. Firstly, he questions the assumption made at the very start of the paper that predictions are to be judged in the context of a population of future $\mathbf{x}$'s and not just at some specific $\mathbf{x}$. To pursue the analogy of the doctor and the patient, all I can say is that the paper is written from the doctor's point of view, and not the patient's! No doubt the doctor will feel he is doing a better job if he cures 95 per cent of patients rather than only 90 per cent, even though a particular patient (Dr Dunsmore) might do better in the latter situation than in the former. As explained in the paper, preshrunk predictors do better than LS for most $\mathbf{x}$'s at the expense of doing worse at a minority of $\mathbf{x}$'s. Perhaps if we think our symptoms are unusual we should seek a consultant who is prepared to view our complaint as an individual research problem rather than rely on the blunt instrument of conventional wisdom. And the predictors discussed in the paper are indeed blunt instruments, since they depend on $\mathbf{x}$ only through the LS predictor $\hat{y}$. In considering the averaging over $\mathbf{x}$, note that there may be no clear distinction between what is a "dependent" variable and what is an "independent" variable, for example Dr Dunsmore's $\mathbf{x}$ might be the value of y at some previous occurrence of the complaint. Averaging over such observable quantities (considering data that might have been observed but were not) is of course at the heart of the classical statistical method.

The second point concerns the relationship between the "classical goodies" in the paper (to quote Dr Dunsmore's words) and the predictive approach, or "Bayesian goodie" as he would no doubt describe it. There is a sense in which all prediction problems have Bayesian overtones, since a predictor involves two layers of uncertainty, the variability of y given$(\alpha, \boldsymbol{\beta})$, and the inferential uncertainty in the value of $(\alpha, \boldsymbol{\beta})$. Preshrunk predictors accommodate the first in terms of ordinary sampling variability but allow for the second by shrinking (shrinkage arises only because of uncertainty in the parameter estimates). Some procedures, for example tolerance intervals, employ a somewhat uneasy juxtaposition of these two layers of uncertainty, and a complete synthesis is only obtained by assuming comparability of these types of uncertainty, which is in essence the Bayesian argument.

Dr Dunsmore asks if he may dare suggest that retrospective fit is a red herring. Certainly! The message of the paper is that methods of assessing model fit are tantamount to retrospective fit, but that if we are interested in prediction as our specific objective then we must do more—forecast the validation fit that would be obtained were we to obtain new data.

Professor Stone discusses the choice of origin towards which shrinkage is made. As he says, there seems to be no way of choosing $\boldsymbol{\beta}_0$ on the basis of bias and variance alone. However, the zero origin does seem to arise in a natural way from the approach in Section 3 of the paper. The graph of y against $\hat{y}$ (Fig. 1) does not depend on the particular parametrization of the model (for given y) although, as is pointed out, it is not invariant under the transformation from y to $y - \boldsymbol{\beta}_0^T\mathbf{x}$. It might be said, therefore, that zero is the "natural" origin for $\boldsymbol{\beta}$ only in as far as y is the "natural" definition of the dependent variable and the joint distribution of y and $\hat{y}$ is a "natural" way of assessing validation fit.

Professor Stone's comment that the maximum gain over LS will be obtained if $(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ is small suggests we look at $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\mathbf{V}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. This is least when $\boldsymbol{\beta}_0 = p^{-1}\mathbf{M}^{-1}\mathbf{1}\mathbf{1}^T\mathbf{M}\hat{\boldsymbol{\beta}}$. Since the resulting quadratic form is less than $\hat{\boldsymbol{\beta}}^T\mathbf{V}\hat{\boldsymbol{\beta}}$, the estimated

value of K will be less (i.e. *more* shrinkage) using this $\boldsymbol{\beta}_0$ as origin rather than zero as origin. This $\boldsymbol{\beta}_0$ corresponds to the average of the regression components in orthogonal space, and so the proposal is equivalent to Lindley's shrinkage towards the grand mean in the Stein problem (Lindley, 1962).

The query is raised about errors in $\mathbf{x}$. In a prediction problem which is concerned with the distribution of y *conditional* on the observed $\mathbf{x}$, an error structure in $\mathbf{x}$ should make no difference. Although the paper is conditioning on $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$ and not $\mathbf{x}$ itself, this conclusion still holds, at least in the following very special case. Suppose we have a structural relationship model in which each individual $\mathbf{x}$ coresponds to a true value $\mathbf{x}_T$ and an error var. matrix $c\mathbf{V}$ for some known scalar c. Assume that $\mathbf{x}_T$ has mean zero and var. matrix $\mathbf{V}$. Then the vector of covariances of y with $\mathbf{x}$ is $\mathbf{V}\boldsymbol{\beta}$ and the overall var. matrix of $\mathbf{x}$ is $(1+c)\mathbf{V}$. Hence an estimate of $\boldsymbol{\beta}$ is $(1+c)\hat{\boldsymbol{\beta}}$ which, as expected, has larger components than $\hat{\boldsymbol{\beta}}$. This might suggest the "structural" predictor $y^* = \bar{y} + (1+c)(\hat{y} - \bar{y})$. However the shrinkage is then more, because the LS slope of y on y^* is now $(1+c)^{-1}$ times the LS slope of y on $\hat{y}$. Thus if we apply this new shrinkage to the structural slope, we end up with the same estimate $\hat{\boldsymbol{\beta}}$ as before.

I am grateful to Dr Subba Rao for showing how shrinkage methods apply to time series. This is a particularly attractive application since the central assumption made in the paper, namely that future $\mathbf{x}$'s are random and follow the same distribution as the past values, arises naturally as part of the model itself, rather than having to be imposed from outside. The motivation for Stein predictors given in Section 3 of the paper also applies in the time series context. Using Dr Subba Rao's notation, the principle is to predict X_{n+h} by its conditional expectation given values of the series to date, and so summarizing the previous experience by the least squares predictor $\hat{X}_n(h)$ leads us to consider $E(X_{n+h} \mid \hat{X}_n(h), \hat{\boldsymbol{\beta}})$. With $h = 1$ this is just $K\hat{X}_n(1)$ where K is $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\beta} / \hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}$. This quantity K, which has expectation strictly less than one, is estimated by the shrinkage factor given in Dr Subba Rao's contribution. As he says, the details for $h = 1$ are very similar to those of the regression problem, but the case $h > 1$ is less clear and will certainly be a worthwhile area for research.

Both Dr Laycock and Professor Plackett show how (3.12) can be improved as an approximation to (3.10). The exact expression for the risk of the Stein estimate given in Egerton and Laycock's paper is not widely known but is a most useful advance. Happily the values of Dr Laycock's formula agree exactly with the entries in Table 1! Professor Plackett's idea is to replace the non-central χ^2 in (3.6) by a central χ^2 with the same mean and variance. A little algebra leads to the approximation

$$E(K) = \frac{1 + (p-2)\,\delta^2}{1 + 2(p-2)\,\delta^2 + p(p-2)\,\delta^4}.$$

This works remarkably well over all values of δ. If δ is small it agrees with (3.12) to order δ^2; if δ is large it works well because the relevant non-centrality parameter is then small. Dr Laycock mentions the advantage of knowing the risk at large values of δ. Hopefully, however, the value of δ will not be too large in any realistic application. For example, if the overall regression is significant at the 5 per cent level for the hypothesis that $\boldsymbol{\beta}$ is zero (surely a weak requirement if we are expecting useful predictors), then a simple approximation to the F-distribution at large ν shows that the corresponding estimate of δ^2 cannot be more than about $\{0.85 + 1.64\sqrt{(2p-1)}\}^{-1}$. For moderate values of p this is quite small (0.13 at $p = 10$).

Drs Laycock, Brown and Miller discuss other predictors that might be used instead of the Stein estimate, particularly when the distribution of future $\mathbf{x}$'s does not equal that of the past values. It will be interesting to follow up Dr Laycock's suggestion for an *admissible* version of the Stein predictor and to see what improvement in prediction mean squared error is possible (perhaps rather little?). Of course Dr Brown is right in saying that Stein does not dominate least squares for all possible distributions of future $\mathbf{x}$'s. Trivially, the mathematics is unchanged if the variances of the $\mathbf{x}$'s are scaled up or down by a constant provided the correlations remain the same, but if the mean of future $\mathbf{x}$'s changes dramatically then shrinkage might be quite the wrong thing to do. My contribution to the discussion of the paper which Dr Brown himself recently read to the Society (Brown, 1982) suggests how predictors might be updated to react to changes in the distribution of future values. Most alternatives to Stein (e.g. ridge) abandon the convention of uniform shrinkage and introduce different amounts of shrinkage in respect of different β_i's

(roughly, the principle is to shrink in inverse proportion to the accuracy with which each coefficient is determined). However there is no simple formulation in which such estimates dominate least squares or the Stein predictor. In fact, under the assumptions in the paper, ridge can be worse (and sometimes substantially worse) than Stein. For in the notation following (3.5), the (simple) ridge estimator of the ith orthogonal regression coefficient ξ_i is of the form $\lambda_i(\lambda_i+\lambda)^{-1}\hat{\xi}_i$, which has risk $\Sigma(\lambda_i+\lambda)^{-2}(\lambda_i\xi_i^2\lambda^2+\lambda_i^2)$, where λ_i is the corresponding eigenvalue of $\mathbf{V}$ and $\hat{\xi}_i$ is the least squares estimate of ξ_i. By contrast the Stein estimate of ξ_i is of the form $K\hat{\xi}_i$ which has risk $pK+(1-K)^2\Sigma\lambda_i\xi_i^2$. It may well happen (e.g. Jollife, 1982) that the regression coefficients with small eigenvalues make the *greatest* contribution to the regression, perhaps $\xi_i^2=\xi^2/\lambda_i$ for some ξ^2. Then the best value of K is $(1+\xi^2)^{-1}\xi^2$ for which the risk is $(1+\xi^2)^{-1}p\xi^2$. It is easy to show that there is no value of the ridge constant λ such that the risk of the ridge estimate is as small as this except when the λ_i's are all equal (in which case all the coefficients would be shrunk by the same amount in any case).

It is with great sadness at this point in writing my reply to the discussion that I note that Professor Anderson's comments will have been the last of his many contributions to the ordinary meetings of our Society. His remarks are typically penetrating and to the point. Firstly, he stresses the need for prediction intervals instead of mere point predictions. The mathematics needed for such an interval is somewhat akin to that in Section 8 of the paper. A sampling theory argument stemming from (8.7) suggests $\tilde{y}\pm u^*\hat{\sigma}(1+n^{-1}k\hat{K})^{\frac{1}{2}}$, u^* being the appropriate percentage point of $N(0,1)$, and k playing the same role as in (3.7). The empirical Bayes argument suggests $\tilde{y}\pm u^*\hat{\sigma}(1+n^{-1}(1+\hat{K}\mathbf{x}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{x}))^{\frac{1}{2}}$. These are in fact rather similar when it is noted that $E(\mathbf{x}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{x})=n^{-1}p$. These may be compared with the conventional tolerance interval $\hat{y}\pm u^*\hat{\sigma}(1+n^{-1}(1+\mathbf{x}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{x}))^{\frac{1}{2}}$. Thus the centre of the prediction interval is shrunk as for the point predictor, and some shrinkage in the width of the interval is also indicated. For practical purposes it is probably adequate to take the conventional tolerance interval and simply recentre it at the value of the shrunken predictor. The point about robustness of the estimate of K is well taken and certainly needs more research. Professor Anderson's other comments relate to the binary regression model of Section 8. Normality of $\mathbf{x}$ is needed to obtain the second term in the denominator in (8.8), but the contribution of this term is small and is omitted in the simpler predictor (8.9). Thus (8.9) assumes normality no more than does the linear regression predictor (4.1). The case of complete separation of the sample mentioned by Professor Anderson goes beyond the scope of Section 8, as the asymptotic approximations made in that section will break down. A re-working of the material to cover this case could in principle be done, but it would not lead to the simple scalar shrinkage discussed in the paper. An analogy is the problem of censored observations in life testing; if for example x is exponentially distributed with mean θ, but is observed to be censored at t, then the likelihood is $\exp(-t/\theta)$ which increases to a constant as $\theta\to\infty$. Similarly, with complete separation in binary regression the likelihood of $(\alpha,\boldsymbol{\beta})$ will increase to a finite constant as the parameter tends to infinity in some direction in the parameter space. Given a suitable family of prior distributions, therefore, the empirical Bayes predicted probability, which is just the posterior expectation of (8.1), will always be strictly bounded away from 0 and 1.

Professor Titterington and Dr Silverman discuss the link between the paper and density estimation. This is a striking analogy in that in both problems the need for shrinkage comes about only when fit to new data is considered. If there is no prior information about $f(\cdot)$ then, when judged by retrospective fit, there is no reason to move away from the delta functions ML estimate. Similarly for least squares in regression. But in both cases the estimates are demonstrably inadequate for validation fit. Unlike regression, however, there does not seem to be any natural way of estimating the desired degree of smoothing using the data alone, except by a contrived argument such as cross-validation or roughness penalties. Note the similarity between (4) of Professor Titterington's remarks and Lindley's modification of the Stein estimate for the multivariate normal mean—the uniform distribution towards which shrinkage is directed is just the average of the individual cell probabilities. Dr Silverman goes on to discuss the orthogonal series decomposition of the unknown function $f(\cdot)$. As he says, a uniform flattening over an arbitrarily large number of components cannot be sensible. However, the situation of Dr Silverman's $f=\Sigma\beta_\nu\phi_\nu$ is not really covered by the paper. Firstly, for the assumptions of the paper to hold we need p fixed with n fairly large relative to p, and so we cannot consider a given set of data

and let the number of components increase until a perfect fit is obtained. Secondly, there is a basic requirement in Stein shrinkage that we have variation in $\mathbf{x}$ over $p \geqslant 3$ dimensions; here $\mathbf{x}$, the vector of ϕ_v's, is characterized by a curve in one dimension only. For example the components of $\mathbf{x}$ may be uncorrelated but cannot be statistically independent or follow a multivariate normal distribution. Thirdly, (5.2) would not be a sensible prior distribution since (unlike ordinary multiple regression) the x_i's occur in a natural order and we would almost certainly expect low values for the high frequency components. In this respect there is a similarity with Dr Subba Rao's time series model in which observations further back in time become progressively less plausible as explanatory variables. For Dr Silverman's last point, see the contributions by Drs Laycock and Brown.

Professor Plackett gives us a timely reminder that however sophisticated a statistical method is, there always remains a subjective element necessary when that method is applied in practice. There is a paradox in the situation he mentions when all squares and products, etc. can be entertained as possible predictors; the "full regression" cannot be defined in the sense envisaged in the paper and one is back to Dr Silverman's example with an arbitrarily large number of high frequency components. Concerning Professor Plackett's second point, each of the examples uses a validation sample in order to show that the predictions of the characteristics of validation fit, made using the construction sample alone, do correspond to what would be observed if new data *were* available. If both samples were available at the time of the analysis then presumably a model (such as the one he mentions) would be fitted to all available data—still of course with a recalculated shrinkage factor! The suggestion that cross-validation leads to loss of efficiency is an interesting general point. Perhaps there is a kind of uncertainty principle here that if one has full efficiency one cannot check that the model is correct, and if one checks that the model is correct then one cannot have full efficiency. Professor Plackett's excellent suggestion about the Patnaik approximation has already been mentioned (and works very well). Regarding his second detailed comment, the fit of a binary regression seems remarkably robust to mis-specifications in the weights, and even with the example cited from Finney's book the recalculation using equal weights (inherent in (8.3)) does not give answers very different from the correctly weighted solution. In Finney's example, the correct standard deviations of $\hat{\alpha}$ and $\hat{\beta}$ are 0.0917 and 0.466 with correlation 0.068. Replacing the weights w_i by the constant average weight $\bar{w}$ gives these figures as 0.0914, 0.423 and zero respectively. (Of course the equal weights approximation in (8.3) is for the var. matrix of the estimates only—the estimates themselves are still calculated by maximum likelihood.) In any case the fit of Finney's data to the probit model goes far beyond the "weak dependence" assumed in Section 8 of the paper. The deviance is 95.05 on 1 d.f. and so even if shrinkage were appropriate (it is not, as $p = 1$) the value of K would be so close to unity it would make no essential difference. Curiously, there is a sense in which the equal weights approximation gives a closer agreement to asymptotic theory than does the ML solution, for the value of $\hat{\beta}^2/\mathrm{var}(\hat{\beta})$ is 80.3 for ML and 97.6 for equal weights, the latter being much closer to the deviance (95.1) than the former.

A number of interesting points are made by Professor Aitken. His simultaneous testing approach to subset selection is perhaps complementary to that of the paper. To ask whether a subset has a performance measure which is significantly better or worse than some other subset is of course invoking a different argument from the question of which subset seems "best" on the basis of all available evidence. However, as his calculations demonstrate, Professor Aitken's method shows the substantial influence of sampling error on the choice of subset. The calculation of the limits for adequate subsets should be mandatory for all those practitioners of stepwise regression analysis who become obsessive about *the* best subset. As Professor Aitken points out, his method, as well as that in the paper, requires $n > p$. Any theory based on the fit of the full regression breaks down if $n \leqslant p$, but there is no reason why sensible subset regressions should not also exist in this case. If a schoolboy in the physics laboratory measures the resistance of a piece of wire by taking a few readings of voltage (y) and current (x_1), is his experiment invalidated by the taking of a few other (almost) spurious readings such as time of day (x_2), temperature (x_3), humidity (x_4), etc? Any analysis of the enhanced data should recover the close correlation between y and x_1. The case $n < p$ is not an anomaly but a common occurrence in applied work (a large clinical study with multiple measurements and a longitudinal design may have hundreds of different variables, but, all too often, not hundreds of patients), and there is a clear need for a theory to cover this case also. Shrinkage will be more evident in such a theory; the larger is the

number of possible subsets the greater is the plausibility that the good fit of one particular subset may be a manifestation of chance rather than a real effect. Professor Aitken then raises the interesting question of reversing the roles of CS and VS. Let K^* in (9.1) be calculated for predicting from CS to VS, and then for predicting from VS to CS. If $p = 1$ these two values are reciprocal to each other and so an overprediction in one direction is exactly matched by an underprediction in the reverse direction. But this is not so for larger p. For $p \geqslant 3$, both values of K^* have expectation less than one, and preliminary calculations along the lines of Section 3 suggest that the expectation of the geometric mean of the two is also less than one. It must also be remembered that all these statistical measures of fit and shrinkage are themselves subject to large sampling errors when sample sizes are small (e.g. the cross-validation statistics in Professor Aitken's table are based on just eight observations). It turns out that the coefficient of variation of K in (3.3) and (3.6) is $\delta^2 + O(\delta^4)$, so for Example 1 (with δ^2 estimated as 0.038) the quantity K for the situation in Fig. 1 will have mean 0.67 with standard deviation 0.13.

Dr Lawrance asks for more discussion of why LS tends to overpredict. A possible further argument is as follows. Suppose a finite number of new values are predicted using $\hat{\boldsymbol{\beta}}$, where the $\mathbf{x}$'s are following some multivariate distribution, and consider the largest predicted value $\hat{y}_m$. If $\hat{\boldsymbol{\beta}} \equiv \boldsymbol{\beta}$, $\hat{y}_m$ is equally likely to be less than or greater than the associated true value y_m. What if $\mathrm{var}(\hat{\beta}_i) > 0$? Intuition suggests that $\hat{y}_m$ is more likely to overestimate y_m than underestimate y_m – if one accepts this then presumably one must also accept the need for shrinkage. As in all other aspects of the paper, the random nature of the future $\mathbf{x}$'s is crucial. To bring in Professor Plackett's point of etymology, this is a kind of regression effect (in the Galton sense) arising from the averaging over the set of $\mathbf{x}$'s consistent with the given $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$ (this being the set involved in the maximization in the above argument). I am afraid I have nothing to offer Dr Lawrance on his interesting question of higher order effects. With squared error it is the second-order moments of the $\mathbf{x}$'s that are important, but with other loss functions it might well be that skewness leads to some undesirable non-linear effects which could undermine the advantages of shrinking.

Dr Lawrance's other points, also raised by Professor Zidek, concern outlying observations and robust methods. If $\hat{\boldsymbol{\beta}}$ is replaced by some robust regression estimate $\hat{\boldsymbol{\beta}}^*$, the shrinkage slope K in (3.3) is still defined in the analogous way; if $\hat{\boldsymbol{\beta}}^*$ is unbiased with positive definite var. matrix then it would still seem that $E(K) < 1$, provided p is sufficiently large. If this can be estimated by an alternative version of (3.7) then a pre-shrink predictor is still obtained, but whether this is better than the predictor using $\hat{\boldsymbol{\beta}}^*$ directly is not clear. For reasonable robust methods the size of shrinkage will perhaps be broadly similar to that for least squares. The situation is different, however, in binary models. Preliminary work shows that robust methods tend to overpredict *more* than ML and so the shrinkage actually *increases*. This is illustrated by the problem of estimating a binomial probability p given s successes out of n trials. A robust estimate (the special case of a possible robust method for categorical data models) would maximize the modified log likelihood, $sf(\log p) + (n - s) f(\log (1 - p))$ where f is some concave influence function. It is easy to show that $|\hat{p} - \frac{1}{2}| > |s/n - \frac{1}{2}|$, i.e. the robust estimate will be closer to 0 or 1 than the ML estimate. In a saturated binary regression model, therefore, the slope coefficients will be numerically *larger*, and this presumably extends to general unsaturated models.

Dr Chatfield reprimands me for using Example 1; my only excuse is the one he mentions in the third sentence of his remarks. I fear that a misunderstanding of this example is not the only potential hazard to confront the "innocent reader" of my paper! Although I agree with everything he says about this particular example (dare I mention that I have seen parametric cost models being fitted to rather less than eight observation?), having $n < p$ with one x_i constant throughout does not necessarily make the data "unacceptable" for regression analysis. If in the example mentioned earlier it happened that temperature (x_3) stayed constant, would that invalidate the exercise? (Actually a constant temperature will give a *better* fit between y and x_1 as the usual conditions of Ohm's law will then be satisfied.)

I am grateful to Dr Miller for his clear discussion of subset selection, particularly his separation of "selection bias" and "omission bias", and the indications he gives of the sizes of these biases that arise in practice. Some remarks on differential shrinkage across regression coefficients are given above. Dr Miller's proposed maximum likelihood method (allowing for selection) looks promising and I look forward to reading of further developments.

Finally, Professor Zidek asks about the effect of preliminary transformations and the removal

of outliers (other points raised by Professor Zidek have already been mentioned). For any given model there is always a positive probability of the data exhibiting some unusual feature, and in practice unusual features are smoothed out by transformation or other device. Hence, averaging over a multitude of possible unusual features, retrospective fit will tend to look better than the degree of fit one should expect from the model. When the predictors come to be applied to new data (which will also exhibit unusual features in due proportion) the deterioration of fit (i.e. shrinkage) will therefore tend to be greater. The argument is the same as in subset selection; the more scope there is for producing a good retrospective fit, the greater will be the shrinkage.

REFERENCES IN THE DISCUSSION

Aitkin, M. A. (1974) Simultaneous inference and the choice of variable subsets in multipleregression. *Technometrics*, **16**, 221–227.

Alam, K. (1973) A family of admissible minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.*, **1**, 517–525.

Battacharya, P. K. (1966) Estimating the mean of a multivariate normal population with general quadratic loss function. *Ann. Math. Statist.*, **37**, 1819–1824.

Berger, J. (1980) A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.*, **8**, 716–761.

Berk, K. N. (1978) Sequential PRESS forward selection, and the full regression model. *Proc. Statist. Comput. Section, Amer. Statist. Ass.*, 309–313.

Brown, P. J. (1982) Multivariate calibration (with Discussion). *J. R. Statist. Soc.* B, **44**, 287–321.

Brown, P. J. and Zidek, J. V. (1980) Adaptive multivariate ridge regression. *Ann. Statist.*, **8**, 64–74.

Chen, C. (1979) Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *J. R. Statist. Soc.* B, **41**, 235–248.

Efron, B. (1979) Bootstrap methods:another look at the jack-knife. *Ann. Statist.*, **7**, 1–26.

Egerton, M. F. (1979) Estimation for non-linear functional relationships and general linear regression models. Ph.D. Thesis, University of Manchester.

Egerton, M. F. and Laycock, P. J. (1982) An explicit formula for the risk of James–Stein estimators. *Can. J. Statist.*, **10**, 199–205.

Fienberg, S. E. and Holland, P. W. (1973) Simultaneous estimation of multinomial probabilities. *J. Amer. Statist. Ass.*, **68**, 683–691.

Finney, D. J. (1971) *Probit Analysis*, 3rd ed. Cambridge: University Press.

Geisser, S. (1965) Bayesian estimation in multivariate analysis. *Ann. Math. Statist.*, **36**, 150–159.

——— (1971) The inferential use of predictive distributions. In *Foundations of Statistical Inference* (B. P. Godambe and D. S. Sprott, eds), pp. 456–469. Toronto: Holt, Rinehart and Winston.

——— (1975) The predictive sample re-use method with applications. *J. Amer. Statist. Ass.*, **70**, 320–328.

——— (1980) A predictivistic primer in Bayesian analysis in econometrics and statistics. In *Essays in Honor of Harold Jeffreys* (A. Zellner, ed.), pp. 363–381. Amsterdam: North-Holland.

Ghosh, M, Hwang, J. T. and Tsui, K-W. (1983) Construction of improved estimators in multiparameter estimation for discrete exponential families (with Discussion). *Ann. Statist.*, in press.

Golub, G. H., Heath, M. and Wahba, G. (1979) Cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

Good, I. J. and Gaskins, R. A. (1980) Density estimation and bump-hunting by the penalized maximum likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Ass.*, **75**, 42–55.

Habbema, J. D. F., Hermans, J. and Van der Broek, K. (1974) In *Compstat 1974* (G. Bruckmann, ed.), pp. 101-110. Vienna: Physica-Verlag.

Haff, L. R. (1977) Minimax estimators for a multinormal precision matrix. *J. Multiv. Anal.*, **7**, 374–385.

Johnson, W. and Geisser, S. (1982) Assessing the predictive influence of observations. In *Statistics and Probability: Essays in Honour of C. R. Rao* (G. Kallianpur *et al.*, eds), pp. 343–358.

——— (1983) A predictive view of the detection and characterization of influential observations in regression analysis. *J. Amer. Statist. Ass.*, in press.

Jolliffe, I. T. (1982) A note on the use of principal components in regression. *Appl. Statist.*, **31**, 300–303.

Matloff, N. S. (1982) James–Stein regression estimation in a prediction context. *Comm. in Statist. (Simul. and Comp.)*, **11**, 589–601.

Oman, S. D. (1981) A unified empirical Bayesian interpretation of ridge and Stein-like estimators in the context of multiple linear regression. Technical report, Department of Statistics, the Hebrew University of Jerusalem.

Patnaik, P. B. (1949) The non-central χ^2 – and F-distributions and their applications. *Biometrika*, **36**, 202–232.

Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory*. Harvard: University Press.

Silverman, B. W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, **10**, 795–810.

Stein, C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berk. Symp. Math. Statist. Prob.*, **1**, 197–206.

——— (1973) Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. Asymp. Statist.*, 345–381.

Stone, M. (1974) Cross-validation and multinomial prediction. *Biometrika*, **61**, 509–515.

Subba Rao, T. (1980) A simplified deviation of the asymptotic mean square error for an AR model with estimated coefficients. Technical Report No. 135, Mathematics Department, UMIST.

Thisted, R. A. and Morris, C.N. (1980) Theoretical results for adaptive ordinary ridge regression estimators. Technical Report No. 94, University of Chicago.

Titterington, D. M. (1980) A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22**, 259–268.

Van der Merwe, A. and Zidek, J. V. (1980) Multivariate regression analysis and canonical variates. *Can. J. Statist.*, **8**, 27–39.

Yamamoto, T. (1976) Asymptotic mean square prediction error for an AR model with estimated coefficients. *Appl. Statist.*, **25**, 123–127.

Zellner, A. and Chetty, V. K. (1965) Prediction and decision problems in regression models from the Bayesian point of view. *J. Amer. Statist. Ass.*, **60**, 608–616.

Zidek, J. (1978) Deriving unbiased risk estimators of multinormal mean regression coefficient estimators using zonal polynomials. *Ann. Statist.*, **6**, 769–782.