*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.

© 1983 Royal Statistical Society 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 $i$th case, $\phi$ is the density of

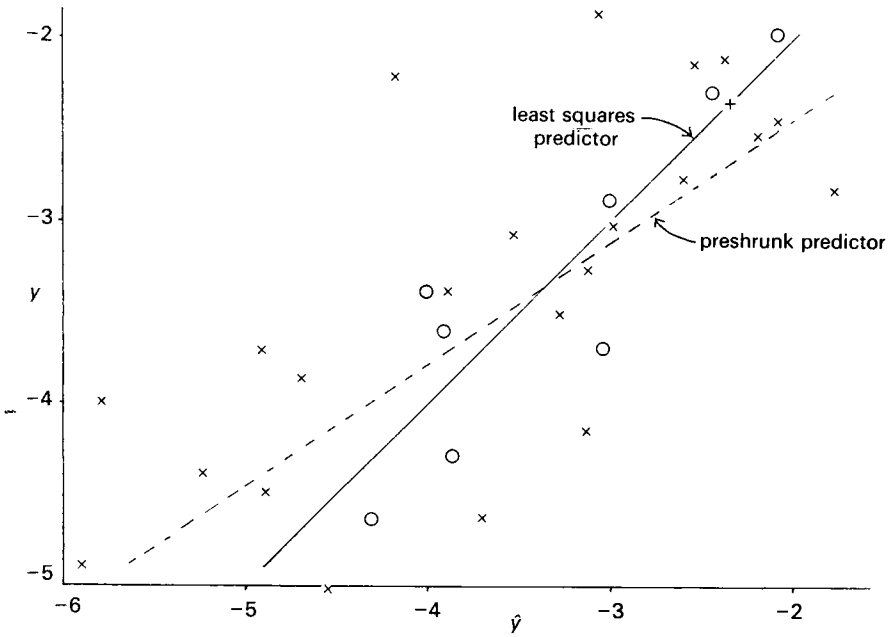


Fig. 1. Observed $y$ against predicted $\hat{y}$ in construction sample (○) and in validation sample (×).

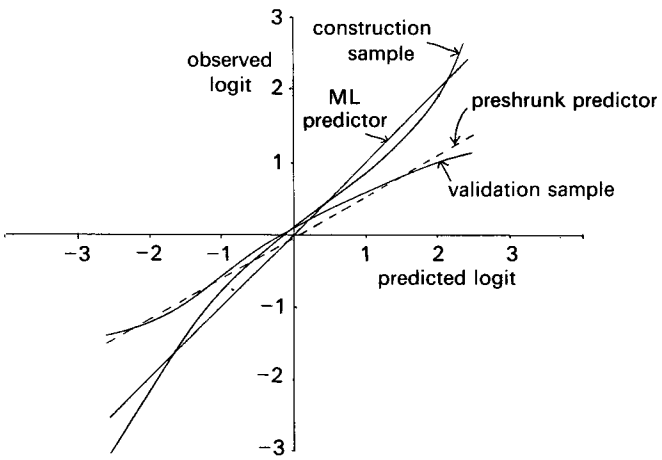


Fig. 2. Observed logit against predicted logit: non-parametric binary regressions.

$N(0, 1)$ 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 $\delta$ 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 $\chi^2_{p-1}$ 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 $\chi^2_{p-1}$, as $\chi^2$ 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 $\nu$ is the residual d.f. given by $\nu = n - p - 1$ and $\chi^2_\nu$ is a $\chi^2$ deviate on $\nu$ d.f. which is independent of both $u$ and $\chi^2_{p-1}$.

The denominator of both (3.6) and (3.9) is proportional to a non-central $\chi^2$ deviate on $p$. d.f. with non-centrality parameter $\delta^{-2}$. The result in Johnson (1959) and Kerridge (1965) shows that this distribution can be represented as a central $\chi^2$ 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 $\delta^2$ 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 $\delta^2$ 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*

| $\delta^2$ | $p$ | $F^*$ | $E(K)$ | (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 $\chi^2$ on $\nu$ 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 $\nu$ 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 $\sigma^2$ is unknown but is estimated by an independent mean square $\hat{\sigma}^2$ based on $\nu$ 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 $\sigma^2$. 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

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 $\alpha$ 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 $\nu$ 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) $\sigma^2$ 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 $\alpha$ 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 $\kappa$ is taken by the statistic $\hat{K}$ in (3.7).

To motivate this particular estimate of $\kappa$ 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 $\sigma^2$ is replaced by $\hat{\sigma}^2$, the corresponding estimate of $\kappa$ is just $\hat{K}(p - 2)$ in (3.7). When $\sigma^2$ is unknown but

is given the usual vague prior distribution (the prior for $\boldsymbol{\beta}$ in (5.2) is then conditional on $\sigma^2$) the same estimate of $\kappa$ 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 $\kappa$ 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 $\kappa$ 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 $\beta_i$ 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 $\beta_i$'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 $\sigma$), with $\hat{\tau}_i \sim N(\tau_i, 1)$. Suppose that the component corresponding to $\tau_i$ 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 $\Sigma'$ denotes summation over $i \in J(\mathbf{y})$ and $\Sigma''$ 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

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

$\phi$ and $\Phi$ 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.

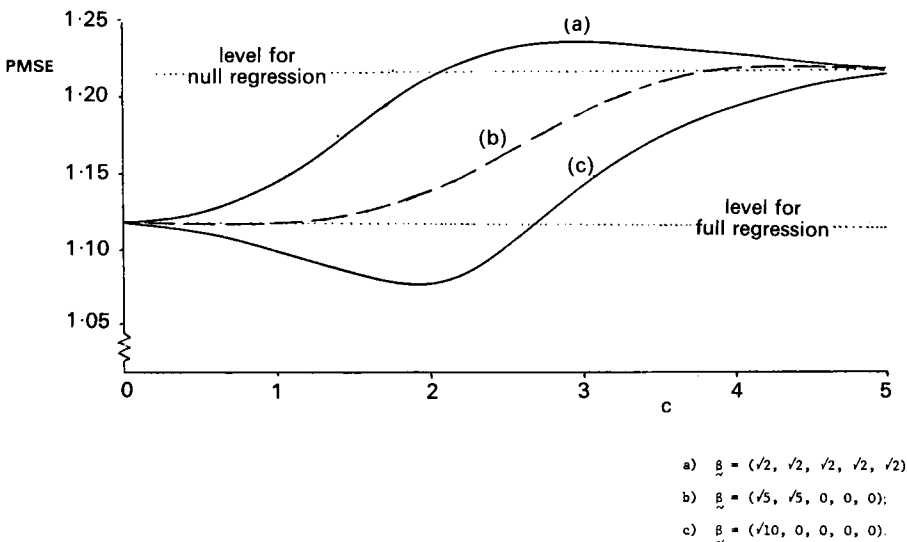


Fig. 3. Prediction mean squared error for selecting components with $|\hat{\tau}| > c$.
(a) $\boldsymbol{\beta} = (\sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{2})$; (b) $\boldsymbol{\beta} = (\sqrt{5}, \sqrt{5}, 0, 0, 0)$; (c) $\boldsymbol{\beta} = (\sqrt{10}, 0, 0, 0, 0)$.

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 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 $\tau_i$'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 $\tau$ 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, $c = 2$*

| $\tau$ | $C_2$ (*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

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

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 $\kappa$ 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 $\sigma^2$ 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 $\kappa$ 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 $s < p$, $\kappa$ 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 $\kappa$, and so the argument for estimation (7.4) is different. In fact (7.4) with $s = p$ and $\kappa$ 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 $\beta_i$'s, and also the $\hat{\beta}_i$'s, are assumed to be random samples from univariate normal distributions, implying that any configuration of $\beta_i$'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 $\beta_i$'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 $\kappa R_s$, 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 $\boldsymbol{\beta}_{(1)}^{\mathrm{T}} \mathbf{V}_{11} \hat{\boldsymbol{\beta}}_{(1)}$, 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 $\alpha$ is vague so that, *a posteriori*, $\alpha$ 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 $\sigma^2$ replaced by $e^2$. In particular, the graphical test following (5.8) is still available as a predictive check. The arguments for estimating $\kappa$ 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 $(1-\kappa)^{-1}\chi_p^2$ and so an unbiased estimate of $\kappa$ is

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

Note that $q^2$ is the asymptotic $\chi^2$ 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 $p(F-1) = 100.8$, $\delta^2$ 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 $\beta_i$'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

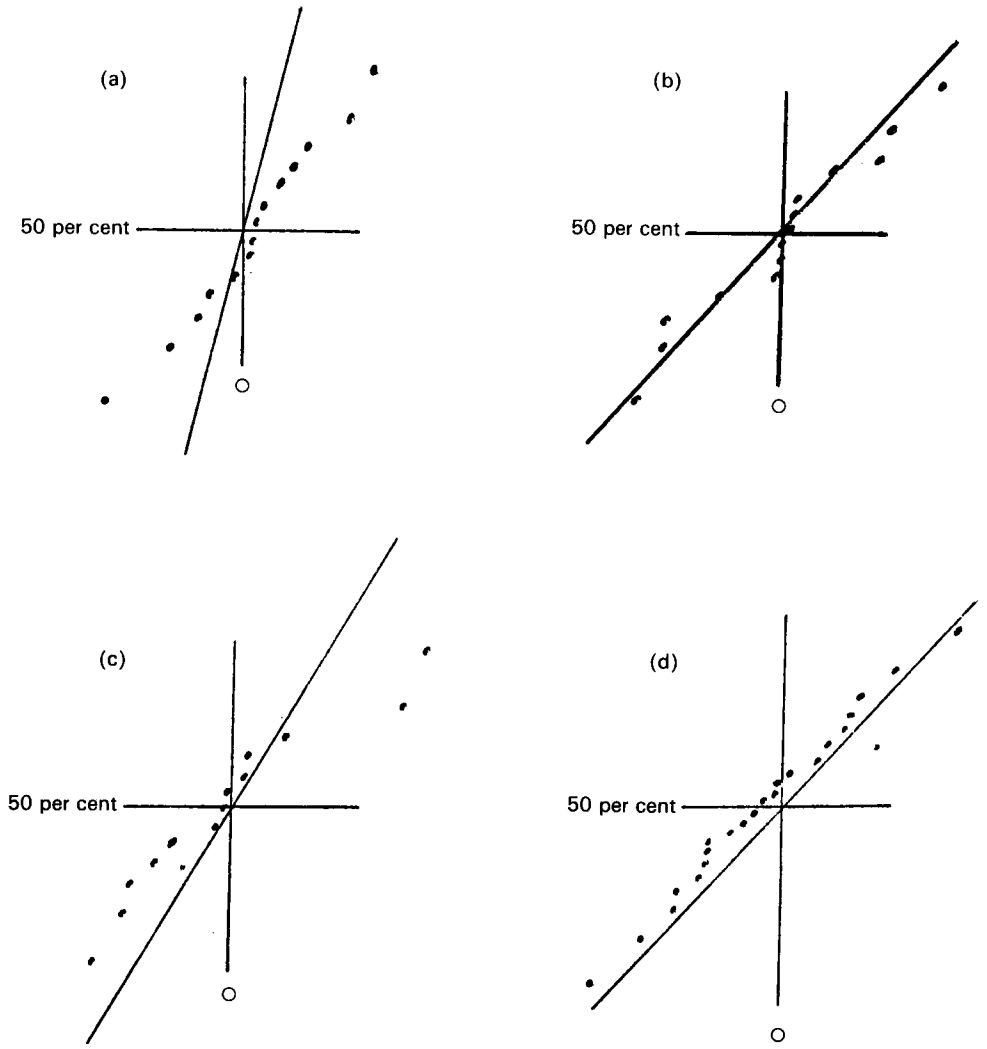


Fig. 4. Normal plots for fit of empirical Bayes model: lines give distribution expected when $\boldsymbol{\beta} = \mathbf{0}$.

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

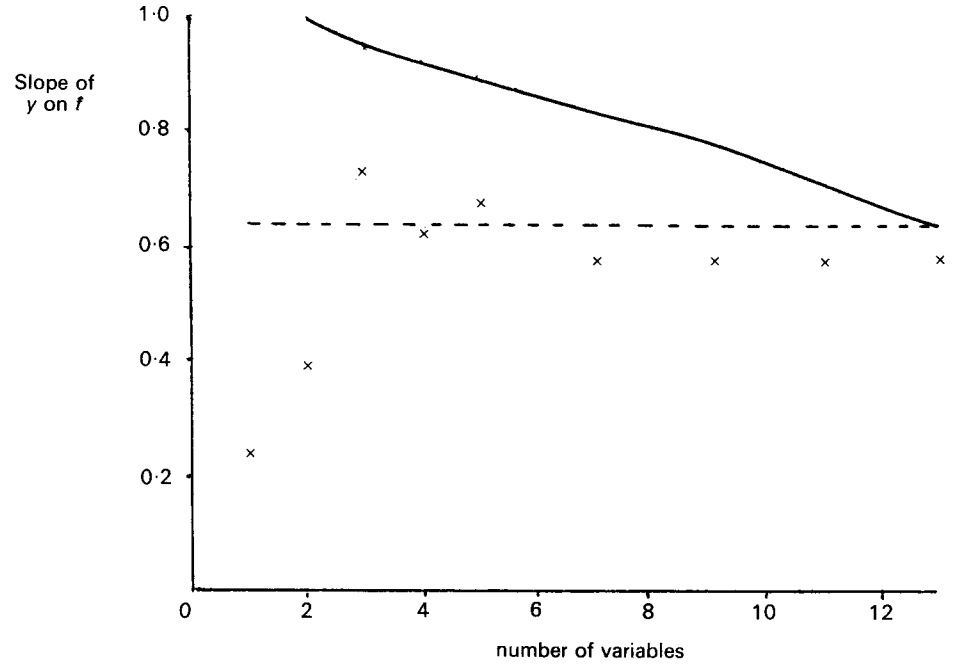


Fig. 5. Shrinkage in stepwise regression. × = observed; —— = predicted (selection ignored); – – – – = predicted (EB model).

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

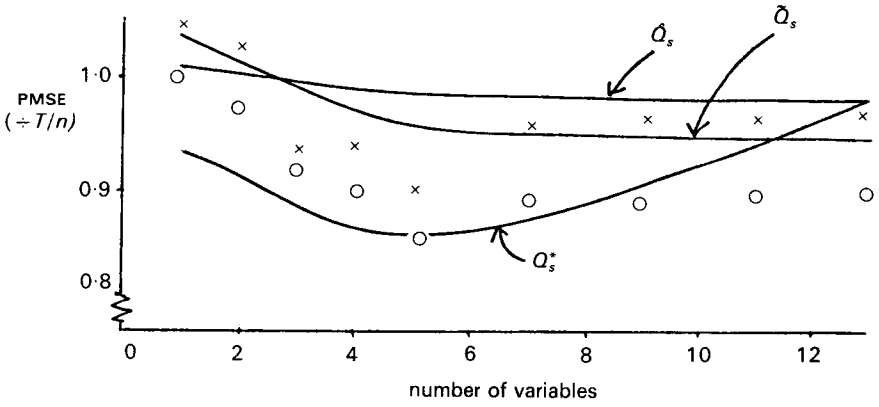


Fig. 6. Prediction mean squared error in stepwise regression (as proportion of total mean square). X = observed, least squares; O = observed, preshrunk.

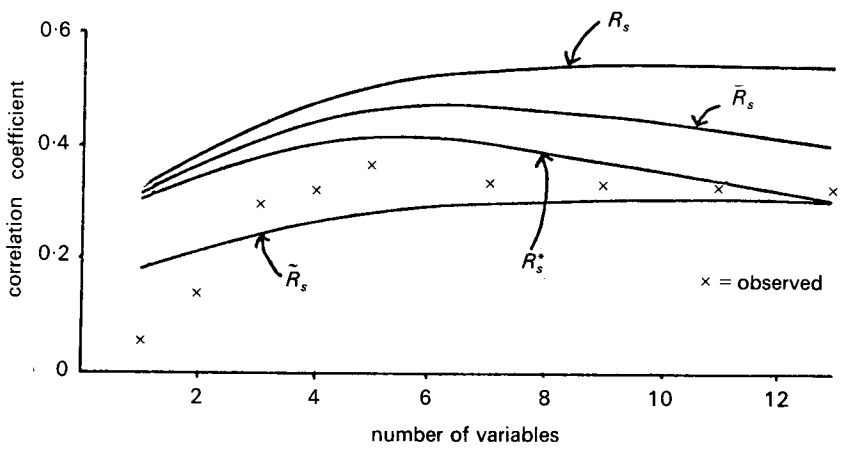


Fig. 7. Correlation coefficients in stepwise regression.

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.

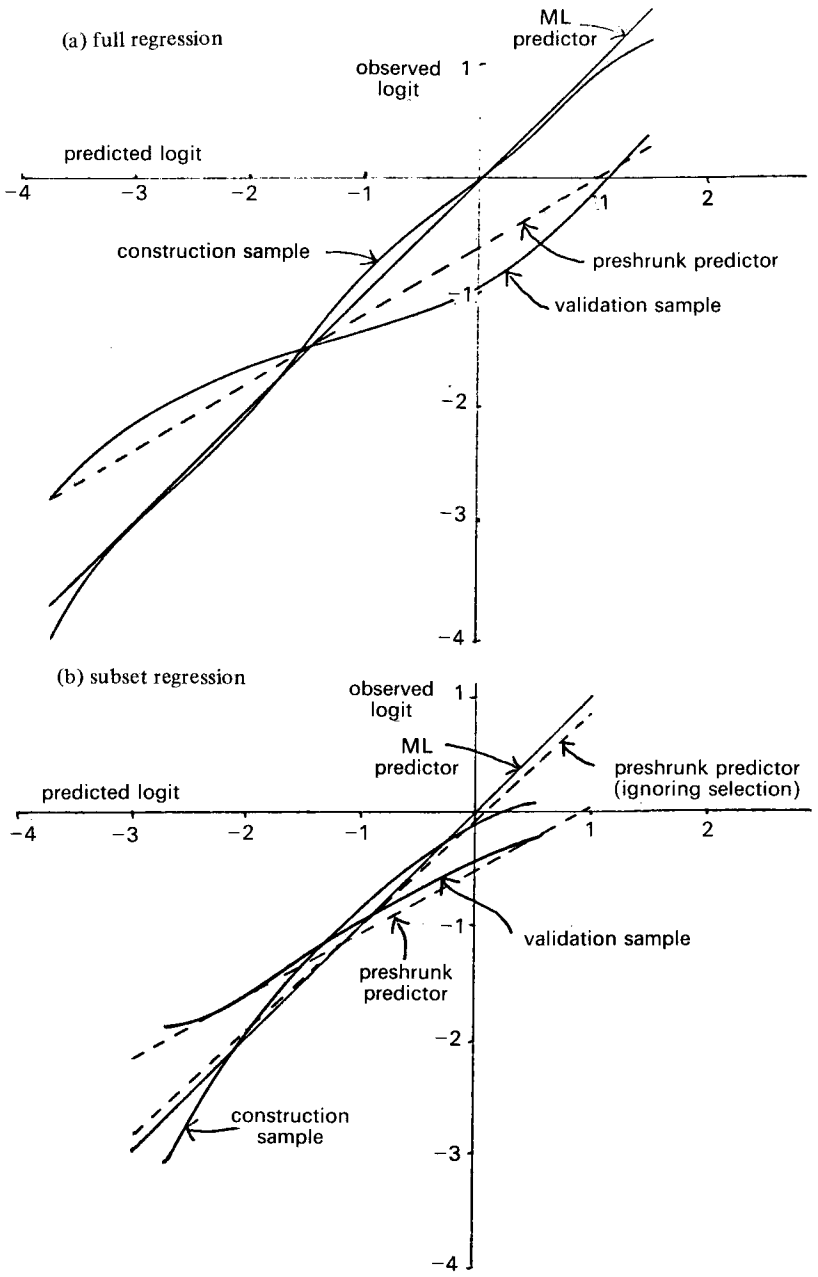


Fig. 8. Observed logit against predicted logit, full and subset regression.

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 $x_i$'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 $x_i$'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 $\chi^2$ distributions. *Biometrika*, **46**, 352–363.
Kerridge, D. (1965) A probabilistic derivation of the non-central $\chi^2$ 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".