Statistical Methods in Medical Research 1997; **6**: 167–183

Using regression models for prediction: shrinkage and regression to the mean

JB Copas Department of Statistics, University of Warwick, Coventry, UK

The use of a fitted regression model in predicting future cases, either as a diagnostic tool or as an instrument for risk assessment is discussed. The regression to the mean effect implies that the future values of the response variable tend to be closer to the overall mean than might be expected from the predicted values. The extent of this *shrinkage* is studied for multiple and logistic regression models, and is found to be related to simple goodness-of-fit statistics of the original regression. Shrinkage is a particularly serious problem if the sample size is small and/or the number of covariates is large. Shrinkage of predictors is illustrated by two examples. A more general formulation is suggested.

1 Regression to the mean

One of the classic examples of *regression to the mean* is the inheritance of height. Tall fathers tend to have shorter sons, and short fathers tend to have taller sons. Here Y_1 (father's height) and Y_2 (son's adult height) are two random variables with (almost exactly) the same distribution, but the conditional expectation $E(Y_2|Y_1)$ is not Y_1 but a value closer to the overall mean. This is nicely illustrated by the famous data on heights discussed by Fisher.[1]

Regression to the mean takes a very simple form when, as well as having the same distribution, Y_1 and Y_2 are also jointly normal, for then

$$E(Y_2|Y_1) = \rho Y_1 + (1-\rho)\mu \tag{1.1}$$

where ρ is the correlation between Y_1 and Y_2 and μ is their common mean. This shows that the regression to the mean effect is more marked the lower is the correlation ρ. If Y_1 and Y_2 are not jointly normal, then the right-hand side of (1.1) is interpreted as the best (in the least squares sense) linear approximation to $E(Y_2|Y_1)$.

Regression to the mean is a direct consequence of statistical variability in Y_1 and Y_2. For suppose that $E(Y_2|Y_1) = Y_1$ for all values of Y_1. Then

$$\begin{aligned}\mathrm{Var}(Y_2) &= \mathrm{Var}(E(Y_2|Y_1)) + E(\mathrm{Var}(Y_2|Y_1))\\ &\geq \mathrm{Var}(Y_1)\end{aligned}$$

This contradicts the assumed equality of the distributions of Y_1 and Y_2, the only exception being when $Y_1 = Y_2$ with no statistical variation. In practice this will never happen.

Typically, Y_1 and Y_2 will be measurements of similar quantities, perhaps at different points of time or before and after some intervention. Then, arguably, Y_1 and Y_2 will have the same distribution under the null hypothesis that the intervenion has no effect. This example shows the importance of recognizing the regression to the

Address for correspondence: JB Copas, Department of Statistics, University of Warwick, Coventry CV4 7AL, UK. E-mail: jbc@stats.warwick.ac.uk

0962-2802(97)SM131RA

mean effect when analysing experimental data. For a plot of Y_1 against Y_2 will, from (1.1), give a linear regression with slope less than 1, which can all too easily be mistaken for an interaction between the treatment effect and the baseline measurement Y_1.

Another example, which we go on to study in this paper, arises in the use of regression. It is a more subtle consequence of the regression to the mean effect, but we hope to demonstrate that it is nonetheless of considerable practical importance. Take the standard multiple regression model with response variable Y distributed as

$$Y \sim N(\alpha + \beta^T X, \sigma^2)$$

where X is a vector of covariates. We suppose that over the population of interest vector X is distributed according to a multivariate normal distribution – without any loss of generality we assume the covariates are centred so that $E(X) = 0$, and let $\mathrm{Var}(X) = V$, say. We assume throughout that X is a vector of m components.

Now suppose that this multiple regression model is fitted to data consisting of n observations (y_i, x_i), $i = 1, 2, ..., n$. We assume that the x_is are covariate vectors chosen to be representative of the population in the sense that they have the same means, variances and covariances, so that $\sum x_i = 0$ and $\sum (x_i x_i^T)/n = V$. Either the x_is are chosen by some suitable design which ensures that these equations hold, or else the x_is are sampled randomly from the population of subjects in which case these moment equations will be approximately true if n is large enough. All regression calculations will be made conditional on the actual x_is in the data, so in either case we treat the x_is as fixed.

From standard theory of multiple regression the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are

$$\hat{\alpha} = \bar{y} \sim N\left(\alpha, \frac{\sigma^2}{n}\right)$$

$$\hat{\beta} = (nV)^{-1} \sum y_i x_i \sim N\left(\beta, \frac{\sigma^2}{n} V^{-1}\right)$$

with $\hat{\alpha}$ and $\hat{\beta}$ independent.

What we want to do is to predict the response at a new covariate vector X, assumed to be randomly chosen from the population. The least squares prediction is

$$Y_1 = \hat{\alpha} + \hat{\beta}^T X$$

We can write

$$Y_1 = \alpha + \beta^T X + \epsilon_1$$

where it is easy to show that

$$\mathrm{Cov}(\epsilon_1, \beta^T X) = E(\hat{\beta} - \beta)^T V \beta = 0$$

and so ϵ_1 and $\beta^T X$ are independent. Now let ϵ_2 be a completely independent random variable with the same distribution as ϵ_1, and put

$$Y_2 = \alpha + \beta^T X + \epsilon_2$$

Then Y_1 and Y_2 have the same distribution, are jointly normal, and so satisfy the regression to the mean formula (1.1). Also

$$\begin{aligned} \mathrm{Cov}(Y_1, Y_2) &= \beta^T V \beta \\ \mathrm{Var}(Y_1) &= E(\hat{\beta}^T V \hat{\beta}) + \mathrm{Var}(\hat{\alpha}) \\ &= \beta^T V \beta + \frac{m+1}{n}\sigma^2 \end{aligned}$$

Hence from (1.1)

$$E(Y_2|Y_1) = \rho Y_1 + (1-\rho)\alpha \tag{1.2}$$

where

$$\rho = \frac{\beta^T V \beta}{\beta^T V \beta + \left(\frac{m+1}{n}\right)\sigma^2} \tag{1.3}$$

Now let Y be the *actual* value of the response variable corresponding to this randomly chosen covariate vector X. Then

$$E(Y|Y_1) = E(Y_2|Y_1) = \rho Y_1 + (1-\rho)\alpha \tag{1.4}$$

with ρ as in (1.3). Thus, when predicting a new case, the distribution of the actual Y is not centred on the least squares predictor Y_1, but regresses towards the mean by an amount which depends on the value of ρ in (1.3). For a large sample size ($n \to \infty$) and/or a well fitting model ($\sigma \to 0$), ρ is close to 1 so the regression to the mean effect can be ignored. But if n is small and/or $||\beta||$ is small and/or m is large and/or σ is large, ρ can be substantially less than 1.

This contrasts with a *retrospective* comparison of actual and predicted values of the response variable. For the empirical covariance between Y and Y_1 based on the values observed in the data set is

$$\frac{1}{n}\sum y_i \hat{\beta}^T x_i = \hat{\beta}^T V \hat{\beta}$$

which is exactly the same as the retrospective estimate of the variance of predicted values

$$\frac{1}{n}\sum (\hat{\beta}^T x_i)^2 = \hat{\beta}^T V \hat{\beta}$$

Hence the best linear approximation of $E(Y|Y_1)$ based on the sample is

$$E_{\text{sample}}(Y|Y_1) \simeq Y_1 \tag{1.5}$$

as of course must be the case because of the usual unbiasedness property of least squares.

Comparing (1.4) and (1.5) we see that the least squares predictor appears to be correctly calibrated when judged retrospectively on the same set of data used in fitting the model, but is not well calibrated when tested on new data. This deterioration of fit from old data to new data is sometimes referred to as *shrinkage*. There is a sense in which the estimated model fits the data too well – shrinkage occurs because any unusual random features of the original data will be reflected in the predictions but not be replicated in a set of independent observations. The term *overfitting* is used to describe this feature of model fitting, and is particularly noticeable when the model is poorly fitting or inaccurately estimated, either through the residual variance being large, or the sample size small, or the number of regressor variables being too large relative to the sample size.

Another interpretation of equation (1.4) is to note that $E(Y|X)$, the true value of Y, can be 'estimated' by two 'estimates', α and Y_1, both being unbiased estimates in the sense that

$$E(\alpha - E(Y|X)) = E(Y_1 - E(Y|X)) = 0$$

Given two unbiased estimates of the same quantity, the standard procedure is to combine them together into a single estimate, weighting in inverse proportion to their variances. Here

$$\text{Var}(\alpha - E(Y|X)) = \text{Var}(\beta^T X) = \beta^T V \beta$$

and

$$\text{Var}(Y_1 - E(Y|X)) = \text{Var}(\hat{\alpha} - \alpha + (\hat{\beta} - \beta)^T X) = \frac{m+1}{n}\sigma^2$$

The corresponding weighted average estimate of the true value of Y is therefore just the shrinkage formula (1.4). This interpretation emphasizes that randomness in X is crucial to the argument – the above variances only arise because X is being averaged over a population. Thus we are predicting 'typical' cases over a population rather than predicting Y at any single fixed value of X.

Figure 1 shows a practical example of the above argument. The data consist of a population sample of 1500 individuals, the Scottish wing of the World Health Organization's *MONICA* project into the incidence of heart disease and related health indicators.[2] One variable of interest is body mass index (BMI), which after a logarithmic transformation we take to be Y, and we study the extent to which this is predicted by a range of covariates covering lifestyle, behavioural and psychometric measures (the vector X). In an exploratory analysis of these data, a multiple regression of log(BMI) on $m = 16$ covariates is fitted to a random subset of just 100 of the cases. The circles in Figure 1 are the 100 actual values of log(BMI) plotted against their least squares predictions. The least squares line fitted to these 100 points gives the solid line

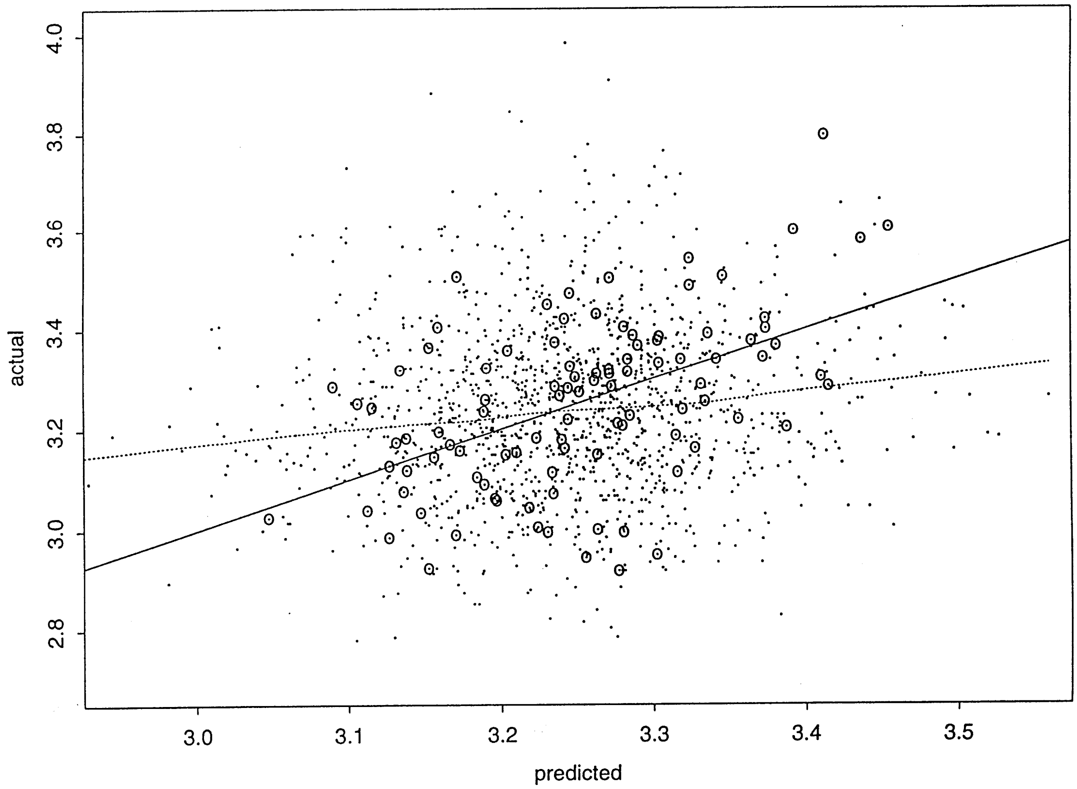

Figure 1 Calibration of predictions of log(BMI)

shown, which is just the diagonal line $y = \hat{y}$ as expected from (1.5). The prediction formula $\hat{\alpha} + \hat{\beta}^T X$ is then calculated for all 1500 values of X, and compared with the corresponding values of log(BMI) observed in the data. These are the dots shown in Figure 1. The dotted line on the graph is a scatter plot smoother (using the LOWESS method of Cleveland[3]) applied to the 1500 dots, suggesting a straight line with a *much lower slope* than the solid line. The actual values of BMI appear not to be clustered around the predicted values as one might expect, but regress towards the overall mean of the data.

Shrinkage is very substantial here because there is a large scatter in the data (σ large), and m is not particularly small relative to n.

2 Shrinkage to the mean in multiple regression

The assumption of multivariate normality of the covariates in the multiple regression in Section 1 is not necessary for the linear shrinkage formula (1.4), provided that, as mentioned above, (1.4) is interpreted as the best linear approximation to $E(Y|Y_1)$. An alternative argument for shrinkage is based on an analysis of the normal equations for least squares, which we now take up in this section. This approach also suggests how the shrinkage correlation can be estimated from the data, leading to improved prediction of future cases.

Defining the fitted values $\hat{y}_i = \hat{\alpha} + \hat{\beta}^T x_i$, the normal equations leading to the least squares estimates of α and β are

$$\sum (y_i - \hat{y}_i) = 0$$
$$\sum (y_i - \hat{y}_i) x_i = 0$$

Hence

$$\sum ((y_i - \hat{\alpha}) - (\hat{y}_i - \hat{\alpha}))(\hat{y}_i - \hat{\alpha}) = 0 \qquad (2.1)$$

The dotted line in our example in Figure 1 suggests that future values should not be predicted by $\hat{y}_i$ but by a linear function of the form

$$\hat{\alpha} + (1 - \delta)(\hat{y}_i - \hat{\alpha}) \qquad (2.2)$$

where the shrinkage multiplier $1 - \delta$ is some number between 0 and 1. Equation (2.2) leaves the overall mean $\hat{\alpha}$ the same, but shrinks all the predictions towards that overall mean. Suppose that y_i^* is an *independent* replication of the response variable y_i at the covariate vector x_i, and imagine regressing y_i^* on $\hat{y}_i$ in order to estimate the required value of δ. The normal equation for this simple regression is

$$\sum (y_i^* - \hat{\alpha} - (1 - \delta)(\hat{y}_i - \hat{\alpha}))(\hat{y}_i - \hat{\alpha}) = 0 \qquad (2.3)$$

Subtracting (2.1) from (2.3) gives

$$\sum (y_i^* - y_i + \delta(\hat{y}_i - \hat{\alpha}))(\hat{y}_i - \hat{\alpha}) = 0 \qquad (2.4)$$

As we only have available the single set of data (y_i, x_i), the y_i^*s are not observed, but we do know that $E(y_i^*) = \alpha + \beta^T x_i$. Hence, averaging over the y_i^*s in (2.4) gives

$$\sum (y_i - \alpha - \beta^T x_i)(\hat{y}_i - \hat{\alpha}) = \delta \sum (\hat{y}_i - \hat{\alpha})^2 \qquad (2.5)$$

The average value of the left-hand side of (2.5) is

$$\begin{aligned}\sum_i \mathrm{Cov}(y_i, \hat{\beta}^T x_i) &= \sum_i \mathrm{Cov}(y_i, n^{-1} \sum_j y_j x_i^T V^{-1} x_j) \\ &= n^{-1} \sigma^2 \sum_i x_i^T V^{-1} x_i \\ &= m\sigma^2\end{aligned}$$

Replacing the left-hand side of (2.5) by this expected value and solving for δ gives

$$\delta = \frac{m\sigma^2}{\sum (\hat{y}_i - \hat{\alpha})^2} \qquad (2.6)$$

Now $\sum(\hat{y}_i - \hat{\alpha})^2$ is simply the regression sum of squares in the original multiple regression, and σ^2 will in practice be taken as the residual mean square. Hence

$$\delta = \frac{1}{F}$$

where F is the F-ratio for the regression ANOVA. Therefore the predictor of y for a future case with covariate vector x is, from (2.2)

$$\hat{\alpha} + \frac{F-1}{F}\hat{\beta}^T x$$

Thus, as expected, if the regression is 'highly significant', F will be large and hence $(F-1)/F$ will be close to 1, meaning that the shrinkage predictor will be close to the elementary predictor $\hat{y}$. But if the regression is weak, with only a small multiple correlation between y and x, then F will likely be close to 1, and so $(F-1)/F$ will be small. Then the predictor collapses down to the overall mean, confirming that no useful prediction is possible.

For the data used in Figure 1, the F-ratio for the regression to the selected subset of 100 values of log(BMI) is 1.76, giving the estimate of $(1-\delta)$ as 0.433, confirming the very substantial regression to the overall mean already evident from the graph.

The term *shrinkage* was first used in connection with Stein's estimate of the multivariate normal mean. In his famous paper, Stein[4] considered the deceptively simple problem of estimating a set of means $\mu_1, \mu_2, \dots, \mu_m$ from independent observations $z_i \sim N(\mu_i, 1)$. Stein argued that in terms of the overall mean squared error risk $E\sum(\hat{\mu}_i - \mu_i)^2$, μ_i should be estimated not by z_i but by the shrinkage estimate

$$z_i\left(1 - \frac{m-2}{\sum z_i^2}\right) \tag{2.7}$$

Stein proved that this led to a uniform improvement in risk when compared with z_i (provided $m > 2$).

Mathematically, Stein's problem of estimating the multivariate normal mean is very similar to the problem of prediction in multiple regression. To see the equivalence, we need to choose a matrix B such that $B^T B = V$, and define the vectors

$$\mu = \frac{\sqrt{n}}{\sigma}B\beta, \qquad z = \frac{\sqrt{n}}{\sigma}B\hat{\beta}$$

Then $E(z) = \mu$ and $\mathrm{Var}(z) = BV^{-1}B^T = I$, the $m \times m$ identity matrix. Hence z and μ follow the simple Stein model. The Stein shrinkage factor in (2.7) is

$$1 - \frac{m-2}{z^T z} = 1 - \frac{(m-2)\sigma^2}{n\hat{\beta}^T V\hat{\beta}} = 1 - \frac{(m-2)\sigma^2}{\sum(\hat{y}_i - \hat{\alpha})^2} \tag{2.8}$$

This is exactly the same as the shrinkage factor suggested in (2.6), except for an adjustment of 2 to m, the number of covariates.

Further, turning to Stein's risk function, we have for a general estimate $\tilde{\mu}$ corresponding to general estimate $\tilde{\beta}$

$$
\begin{aligned}
E(\tilde{\mu}-\mu)^T(\tilde{\mu}-\mu) &= \frac{n}{\sigma^2}E(\tilde{\beta}-\beta)^T V(\tilde{\beta}-\beta) \\
&= \frac{n}{\sigma^2}E\sum(y_i^* - \hat{\alpha} - \tilde{\beta}^T x_i)^2 - n - 1
\end{aligned}
$$

Up to a linear transformation, this is just the expected sum of squares of errors when predicting the new observations y_i^*. Thus the uniform improvement in risk shown by Stein implies that the shrinkage regression predictor using the shrinkage factor in (2.8) has a lower prospective prediction mean squared error than the least squares predictor.

A further discussion of the relation between Stein estimation and shrinkage in regression, and a more careful approach to the somewhat heuristic arguments we have used here, can be found in Copas.[5] There the factors m in (2.6) and $m-2$ in the corresponding formula (2.8) are generalized to an adjustable positive constant k. Following Stein's development, it is shown that the prediction mean squared error is better than that of least squares for all positive values of k up to $2(m-2)$, with the choice $k = m - 2$ being best. Clearly this requires that $m > 2$. It also follows that the choice $k = m$ in (2.6) is still better than least squares provided $m > 4$.

The observation by Lindley[6] that the Stein estimate of the multivariate normal mean could be given a rather natural Bayesian interpretation, also implies that shrinkage prediction in regression can be thought of as a Bayes procedure – this too is discussed by Copas.[5] For an informal indication of the argument, suppose the regression is weak, with a relatively small value of the variance ratio F. In this case there must be serious doubt as to whether x plays *any* useful role in prediction, and the null regression with predictor $\hat{y} = \hat{\alpha}$ may be a plausible explanation of the data. On the other hand if nothing is assumed a priori about β, then the least squares predictor would be sensible. The shrinkage predictor is just the weighted average of these two predictors, with weights given by the degree of credence one could attach to the null value $\beta = 0$ in the light of the data. This is just a special case of 'Bayesian averaging' as a way of allowing for model uncertainty, as discussed by Draper.[7]

Although the arguments for shrinkage imply that we should expect shrinkage multipliers to lie between 0 and 1, it is possible that (2.8) may be negative. Sclove[8] shows that the 'positive part' version of the Stein estimate

$$
z_i \max\left(0, 1 - \frac{m-2}{z^T z}\right)
$$

gives a lower mean squared error risk than the basic form (2.7), and this argument extends to the regression problem. Thus if the estimated shrinkage factor is negative, it is best to predict at the overall mean $\hat{y} = \hat{\alpha}$.

3 Shrinkage to the mean in logistic regression

Regression to the mean, and hence shrinkage in regression, is a rather general phenomenon, and not just confined to examples with normally distributed data. One of the most widely used non-normal regression models in medical applications is logistic regression, for relating a binary response y to a vector of covariates x, which we now go on to consider in this section. We find that the amount of shrinkage in predicting new cases is very similar to the multiple regression case of Section 2.

Suppose that y_i, the binary response for the ith patient, is coded as 1 for *success* and 0 for *failure*, and let p_i be the probability that y_i is 1. These probabilities depend on covariates x_i, modelled by the usual logistic regression formula

$$\text{logit}(p_i) = \alpha + \beta^T x_i$$

where the logit function is

$$\text{logit}(p) = \log \frac{p}{1-p}$$

The maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ satisfy the nonlinear equations

$$\begin{aligned} \sum (y_i - \hat{p}_i) &= 0 \\ \sum (y_i - \hat{p}_i)\hat{\beta}^T x_i &= 0 \end{aligned} \tag{3.1}$$

where

$$\text{logit}(\hat{p}_i) = \hat{\alpha} + \hat{\beta}^T x_i$$

Following the argument in Section 2, suppose that y_i^* are independent replications at the same covariate vectors x_i, and suppose that these are predicted not by $\hat{p}_i$ but by $\tilde{p}_i$ given by the shrinkage predictor

$$\text{logit}(\tilde{p}_i) = \hat{\alpha} + (1-\delta)\hat{\beta}^T x_i$$

Then, if we knew the y_i^*s we could estimate δ by a logistic regression of y_i^* on $\hat{\beta}^T x_i$, for which the estimating equation is

$$\sum (y_i^* - \tilde{p}_i)\hat{\beta}^T x_i = 0$$

Of course the y_i^*s are not actually observed, but we do know that $P(y_i^* = 1) = p_i$. Thus taking expectations over the new data gives

$$\sum (p_i - \tilde{p}_i)\hat{\beta}^T x_i = 0 \tag{3.2}$$

In large samples we can assume that δ is small, in which case we have the first-order approximation

$$\tilde{p}_i \simeq \hat{p}_i - \hat{p}_i(1-\hat{p}_i)\delta\hat{\beta}^T x_i$$

Substituting into (3.2) leads to

$$\sum (p_i - \hat{p}_i + \delta\hat{p}_i(1-\hat{p}_i)\hat{\beta}^T x_i)\hat{\beta}^T x_i = 0 \tag{3.3}$$

Subtracting (3.3) from (3.1) gives

$$\sum (y_i - p_i)\hat{\beta}^T x_i = \delta \sum \hat{p}_i(1-\hat{p}_i)(\hat{\beta}^T x_i)^2$$
$$= \delta\hat{\beta}^T\Lambda\hat{\beta} \qquad (3.4)$$

where

$$\Lambda = \sum \hat{p}_i(1-\hat{p}_i)x_i x_i^T$$

Again in large samples we have the first-order linear approximation

$$p_i \simeq \hat{p}_i - \hat{p}_i(1-\hat{p}_i)(\hat{\alpha} - \alpha + (\hat{\beta}-\beta)^T x_i)$$

Thus using this, and also the original estimating equation (3.1), we can approximate the left-hand side of (3.4) as

$$\sum (y_i - \hat{p}_i + \hat{p}_i(1-\hat{p}_i)(\hat{\alpha}-\alpha+(\hat{\beta}-\beta)^T x_i))\hat{\beta}^T x_i$$
$$= \sum \hat{p}_i(1-\hat{p}_i)(\hat{\alpha}-\alpha)\hat{\beta}^T x_i + \hat{\beta}^T\Lambda(\hat{\beta}-\beta) \qquad (3.5)$$

As before we can assume that the covariates are appropriately centred, but in the context of logistic regression this means that the *weighted* sum of the x_is is zero, the weights being $\hat{p}_i(1-\hat{p}_i)$. Thus the first term in (3.5) is zero. With x_is thus centred, the asymptotic variance-covariance matrix of $\hat{\beta}$ is, from standard theory of logistic regression

$$\mathrm{Var}(\hat{\beta}) \simeq \Lambda^{-1}$$

Hence the expectation of the second term in (3.5) is

$$E\hat{\beta}^T\Lambda(\hat{\beta}-\beta) \simeq \mathrm{trace}(\Lambda\Lambda^{-1}) = m$$

Thus if the (unobserved) left-hand side of (3.4) is replaced by its expectation, the solution to equation (3.4) is simply

$$\delta = \frac{m}{\hat{\beta}^T\Lambda\hat{\beta}} \qquad (3.6)$$

Note that the denominator of δ in (3.6) is the standardized distance between the estimated $\hat{\beta}$ and zero, asymptotically equal to the *deviance* of generalized linear modelling.[9] The shrinkage predictor $\tilde{p}$ for the probability that $y = 1$ at a future covariate vector x is therefore

$$\mathrm{logit}(\tilde{p}) = \hat{\alpha} + \left(1 - \frac{m}{\mathrm{deviance}}\right)\hat{\beta}^T x$$

Thus, just as in the multiple regression case of Section 2, a weak model (low deviance) leads to substantial shrinkage. A strongly fitting model (high deviance), on the other hand, leads to little shrinkage, the probabilities for new observations then being just the usual fitted probabilities $\hat{p}$. Again we set $\tilde{p} = \hat{p}$ if the deviance is less than m.

A practical example demonstrating shrinkage in logistic regression is shown in Figure 2. Here we use the same data as in the example of Figure 1, but now we are predicting the probability that the subject will suffer ischaemic heart disease, but using the same covariates as before. Firstly, a logistic regression is fitted to the same randomly chosen subsample of 100 cases, which gives 100 fitted probabilities for these cases. These probabilities are divided into 10 groups from the lowest to the highest, and for each group we calculate the average fitted probability and the proportion of subjects who actually did report ischaemic heart disease. These are the circles shown in Figure 2. A logistic regression of the actual ys on the predicted logits for these 100 cases gives the solid line, which of course is just the identity $p = \hat{p}$. Secondly, the same prediction equation is applied to the *whole* sample and the procedure of dividing into 10 class intervals repeated. This gives the dots plotted on the graph, summarized by the logistic model given by the dotted line. The lower slope of the dotted line compared with the solid line is very noticeable, and it is clear that if the subset fit were to be used directly to predict the whole set of cases the calibration of the predictions would be very poor.

For the logistic regression on the subset of 100 cases, the deviance is 26.42 on 16 degrees of freedom (just significant at the 5% level). The value of (3.6) is then 0.61, suggesting 0.39 as the estimate of $(1 - \delta)$. The dashed line in Figure 2 shows the corresponding shrinkage predictor $\tilde{p}_i$. The proximity of the dashed line to the dots on

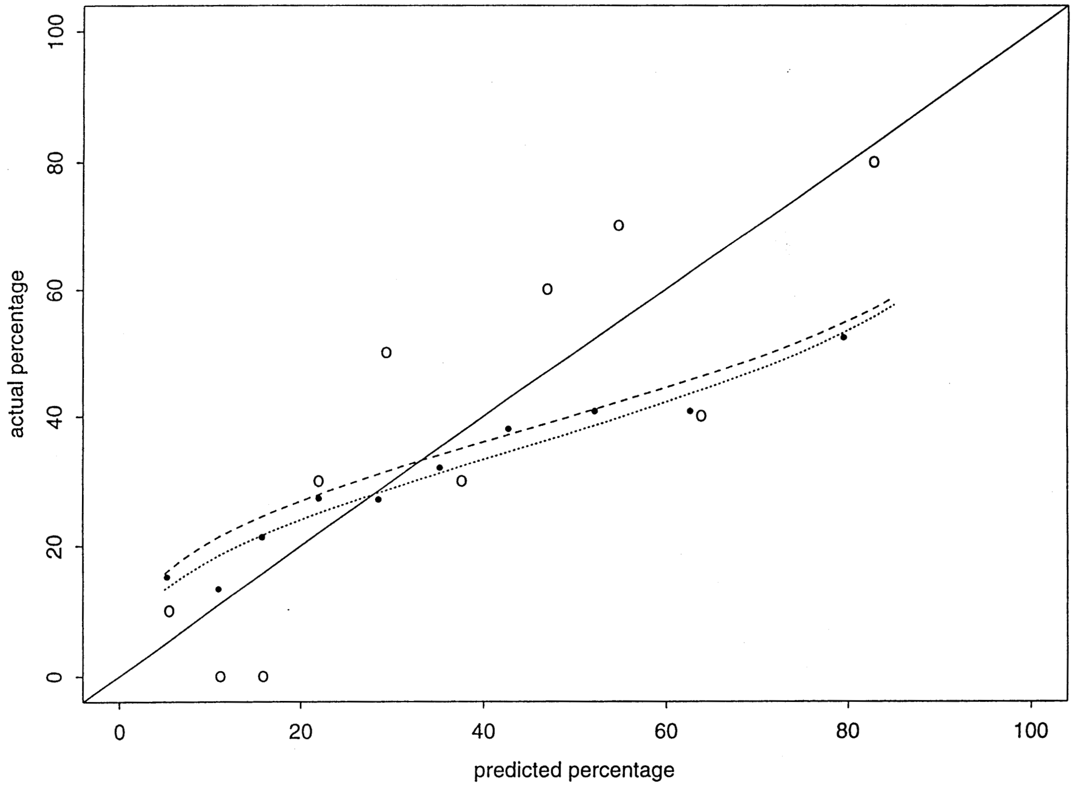

Figure 2 Calibration of predictions of the risk of heart disease

the graph confirms that the shrinkage predictor is reasonably well calibrated on the whole data set.

4 A more general setting

Figures 1 and 2 show that if we fit a regression (or logistic regression) of *new* values of y on their predictions based on *old* data, then we will obtain a regression slope which is typically less than 1, and substantially so if the fit of the original regression is poor. This slope can be interpreted as a measure of quality or usefulness of the predictions, in the sense that the lower is this slope the less the degree of discrimination achieved across the population of future patients, and hence the less useful the predictions will be as a diagnostic tool. It is helpful to separate out two distinct stages in this argument, firstly the model and its fitting procedure (e.g. logistic regression), and secondly the measure used to assess quality of future predictions (e.g. the slope just defined).

This can be cast in a more general setting as follows. Suppose a vector of n observations y has probability density function $f(y, \theta)$, where θ is a vector of unknown parameters. Usually there will be a vector of covariates for each observation, but we are assuming throughout that these are fixed and so can be subsumed within the notation for the distribution f. Let $E(y) = \mu$ and $\mathrm{Var}(y) = \Omega$, both being functions of θ as well as of any covariates. For example, we may have f defined as

$$\log f = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2}(y - X_M\theta)^T(y - X_M\theta) \tag{4.1}$$

This would be the usual model for multiple regression of y on the columns of matrix X_M with regression coefficients θ. In this case $\mu = X_M\theta$ and $\Omega = \sigma^2 I$.

We now assume that the quality of the fit of the model is judged by another function $H(y, \theta)$ – this is our assessment of the fit of a general data vector y to the model given by parameter value θ. For example, in the multiple regression case, H could be the residual sum of squares between y and μ – from (4.1) this would be equivalent to taking $H = \log f$. Alternatively, as suggested by Section 2 above, H could be defined to be the least squares slope of the regression of y on the predicted values from the model.

Once f and H are specified, we can use H to measure the fit of the *original* data vector y to the fitted values given by $\hat{\theta}$, giving $H(y, \hat{\theta})$, and then compare this with the expected value of $H(y^*, \hat{\theta})$, where y^* is an *independent replication* of y. The difference can be interpreted as a measure of *overfitting*. By suitable choices of f and H we can study different statistical models and different aspects of overfitting.

For ease of notation in what follows, we let $g(y, \theta) = \log f(y, \theta)$, and use suffices $\theta, y, \theta y$, ... to denote partial derivatives. When the arguments of g, H and their derivatives are not shown explicitly in the notation, these functions are assumed to be evaluated at $y = \mu$ and at the true value of θ.

Let $\hat{\theta}(y)$ be the maximum likelihood estimate of θ based on data y, given by

$$g_\theta(y, \hat{\theta}) = 0$$

Differentiating this gives

$$\hat{\theta}_y(y) = -g_{\theta\theta}^{-1} g_{\theta y}$$

and standard likelihood theory gives

$$\mathrm{Var}(\hat{\theta}(y)) \simeq -g_{\theta\theta}^{-1}$$

Now for values of y close to their expected value μ, and parameter vectors θ' close to the true value θ, we have

$$\begin{aligned} H(y, \theta') &\simeq H + (\theta' - \theta)^T H_\theta + \frac{1}{2}(\theta' - \theta)^T H_{\theta\theta}(\theta' - \theta) \\ &\quad + (y - \mu)^T H_y + \frac{1}{2}(y - \mu)^T H_{yy}(y - \mu) \\ &\quad + (y - \mu)^T H_{y\theta}(\theta' - \theta) \end{aligned} \tag{4.2}$$

Let $I(\theta, \theta')$ be the expected value of $H(y, \theta')$ when y is sampled from $f(y, \theta)$. Then

$$\begin{aligned} I(\theta, \theta') &\simeq H + (\theta' - \theta)^T H_\theta + \frac{1}{2}(\theta' - \theta)^T H_{\theta\theta}(\theta' - \theta) \\ &\quad + \frac{1}{2}\mathrm{trace}(\Omega H_{yy}) \end{aligned}$$

Thus

$$EI(\theta, \hat{\theta}(y)) \simeq H + \frac{1}{2}\mathrm{trace}(-g_{\theta\theta}^{-1} H_{\theta\theta} + \Omega H_{yy}) \tag{4.3}$$

This is the *prospective* expectation of H, i.e. the assessment of the average quality of the fitted model as judged by applying it to *independent* data.

The *retrospective* expectation of the quality measure is $H(y, \hat{\theta}(y))$, when the same data y is used both to estimate the model and to assess its quality. From (4.2) this is approximately

$$H + E(y - \mu)^T H_{y\theta}(\hat{\theta}(y) - \theta) + \frac{1}{2}\mathrm{trace}(-g_{\theta\theta}^{-1} H_{\theta\theta} + \Omega H_{yy}) \tag{4.4}$$

Thus the difference between (4.3) and (4.4), a measure of the extent of overfitting, is approximately

$$\begin{aligned} E(y - \mu)^T H_{y\theta}(\hat{\theta}(y) - \theta) &\simeq -E(y - \mu)^T H_{y\theta} g_{\theta\theta}^{-1} g_{\theta y}(y - \mu) \\ &= -\mathrm{trace}(\Omega H_{y\theta} g_{\theta\theta}^{-1} g_{\theta y}) \end{aligned} \tag{4.5}$$

We illustrate the overfitting measure (4.5) by working out the details for the multiple regression model (4.1). Assume that the model has an intercept term, so the first column of X_M is 1, the vector of n ones, and assume that the other covariates are centred so that $1^T X_M = (n, 0, \ \dots \ 0)$. The corrected sum of products of y and $X_M\theta$ is $yJX_M\theta$, where

$$\mathcal{J} = I - n^{-1}11^T$$

and the corrected sum of squares of the elements of $X_M\theta$ is $\theta^T X_M^T \mathcal{J} X_M \theta$. Hence if we define $H(y, \theta)$ to be the least squares slope of the simple regression of y on the expected values $X_M\theta$, then

$$H(y, \theta) = \frac{y^T \mathcal{J} X_M \theta}{\theta^T X_M^T \mathcal{J} X_M \theta}$$

from which we obtain

$$H_{y\theta} = \frac{\mathcal{J} X_M}{\theta^T X_M^T \mathcal{J} X_M \theta} - 2\frac{\mathcal{J} X_M \theta\theta^T X_M^T \mathcal{J} X_M}{(\theta^T X_M^T \mathcal{J} X_M \theta)^2} \tag{4.6}$$

We now evaluate (4.5) in two parts corresponding to the two terms in (4.6). Firstly we find

$$\text{trace}(\mathcal{J} X_M (X_M^T X_M)^{-1} X_M^T) = \text{trace}((X_M^T X_M)^{-1} X_M^T \mathcal{J} X_M) = m$$

where, as before, m is the number of (non-intercept) covariates. This uses the fact that both the first row and first column of $X_M^T \mathcal{J} X_M$ are vectors of zeros, and that the rest of the elements of this matrix form the inverse of the corresponding submatrix of $(X_M^T X_M)^{-1}$. For the second term to be evaluated, we find

$$\begin{aligned}\text{trace}(\mathcal{J} X_M \theta\theta^T X_M^T \mathcal{J} X_M (X_M^T X_M)^{-1} X_M) &= \theta^T X_M^T \mathcal{J} X_M (X_M^T X_M)^{-1} X_M^T \mathcal{J} X_M \theta \\ &= \theta^T X_M^T \mathcal{J} X_M \theta\end{aligned}$$

Therefore from (4.6) we obtain

$$-\text{trace}(\Omega H_{y\theta} g_{\theta\theta}^{-1} g_{\theta y}) = \sigma^2 \frac{m-2}{\theta^T X_M^T \mathcal{J} X_M \theta}$$

The retrospective value $H(y, \hat{\theta}(y))$ is simply

$$\frac{y^T \mathcal{J} X_M \hat{\theta}}{\hat{\theta}^T X_M^T \mathcal{J} X_M \hat{\theta}} = \frac{y^T \mathcal{J} X_M (X_M^T X_M)^{-1} X_M^T y}{y^T X_M (X_M^T X_M)^{-1} X_M^T \mathcal{J} X_M (X_M^T X_M)^{-1} X_M^T y} = 1$$

using similar matrix manipulations to before. Of course this is just as expected, since retrospectively least squares gives the best calibration of y. Hence the expected prospective slope is

$$1 - \frac{(m-2)\sigma^2}{\theta^T X_M^T \mathcal{J} X_M \theta} \tag{4.7}$$

This is very similar to the shrinkage factor obtained in Section 2, with the numerator $(m - 2)$ as proposed by Stein. The only difference is that in (4.7) the denominator is the regression sum of squares calculated for the true regression, whereas in (2.8) the denominator is the ordinary empirical regression sum of squares. However, asymptotically, (2.8) and (4.7) will be the same up to the accuracy of the approximations being used.

5 Discussion

There are many uses of regression in medical and other applications, including data description (a concise description of how the values of y in the data are related to the values of x), explanation (seeing which covariates influence y) and prediction (for diagnosis or risk assessment). In this last case we are essentially making statements about new data (future patients) on the basis of old data (the sample on which the model is fitted). The regression to the mean effect implies, in the sense we have been discussing in this paper, that the values of y for the new patients will be closer to the overall mean than we would expect from an uncritical application of least squares or maximum likelihood.

For multiple regression we suggest calculating $100/F$ as a rough estimate of the shrinkage percentage – this follows from (2.6), with F equal to the usual F-ratio in the analysis of variance of the regression. If this percentage $100/F$ is at all large then least squares predictions are likely to be poorly calibrated, and this is a warning of the dangers of using a regression predictor which is based on a small sample size and/or a large number of covariates. The situation is much the same for logistic regression, where here the appropriated estimated percentage shrinkage is ($100m$/deviance) from (3.6).

For a properly calibrated predictor, we suggest using the shrinkage predictor (2.2) with $(1-\delta)$ estimated by (2.8), rather than the least squares predictor. For logistic regression the shrinkage predictor corresponds to (3.6) – by analogy with multiple regression the numerator m here is probably better replaced by $(m-2)$, but this has not been proved.

As has been emphasized, the essence of the shrinkage argument as presented in this paper is that we are *predicting new random cases*. Our arguments are of no direct relevance to the first two objectives of regression analysis listed at the start of this section, data description and explanation. In particular, we are not suggesting that $(1-\delta)\hat{\beta}$ is a 'better' estimate of β than $\hat{\beta}$, or suggesting that the analysis of covariance is not a valid method of estimating a treatment effect in the presence of covariates.

We have assumed throughout that the set of covariates is fixed. In practice, particularly when analysing observational data where a large number of covariates have been measured, the covariates actually used in model fitting may be chosen on the basis of an exploratory analysis of the data. If the same set of data is used both to select the covariates and to fit the model, then the values of the coefficients will be biased and the shrinkage of the predictor will be even more marked. The reason for this is intuitively clear. If a regression coefficient is by chance overestimated (in absolute value) then it will be more likely to be selected than if it happened to be underestimated, hence the covariates which end up being selected for the model are likely to have coefficients which are too large. This means that the least squares predictor will give too wide a range of values, high values of $\hat{y}$ will be too high and low values of $\hat{y}$ will be too small.

Unfortunately there seems to be no satisfactory theory of the predictive properties of subset selection in regression. The problem is discussed in detail by Miller,[10] and it is clear from his monograph that although stepwise regression algorithms are available in most statistical packages and from a practical perspective they seem attractive to use, the question of whether selecting covariates prior to model fitting is sensible from a statistical point of view is far from clear. Copas[5] proposes an empirical Bayes approach which allows for covariate selection, suggesting that a subset predictor shrinks by an amount equal to the shrinkage of the *whole* regression. Even if F for the selected subset regression is large, the F ratio for the full regression, including the covariates which were candidates for selection but were not chosen, may be quite small. In particular if m/n is at all close to 1, prediction is likely to be useless even if the predictor is based on just one or two covariates which seem to have a large effect. Copas[5] illustrates this with a medical example.

An assumption we have made throughout, stated explicitly in Section 1 but only implicitly in Sections 2–4, is that the validation of a predictor is based on future observations made over the same population of covariates. This is therefore a minimalist approach to shrinkage, allowing only for the statistical effects of overfitting. If in practice a predictor is used in diagnosis or risk assessment for future cases, its calibration will also be affected by any changes in the population. This could arise from the fact that a model is fitted to a study sample which may have no pretence of being representative of the population of patients at large, or may reflect changes in morbidity or treatment over time. Copas and Jones[11] consider multiple regression when the variance–covariance matrix of the covariates differs beween the old and the new data. In terms of prediction mean squared error, the shrinkage predictor remains better than least squares provided the two matrices are not too different – see the cited paper for details.

There is a very large literature on nonstandard estimation in multiple regression, some of which covers shrinkage estimation related to the prediction methods discussed here. The usual emphasis is on estimation of the regression parameter β, especially with a relatively large number of covariates which may be highly intercorrelated, when the standard errors of components of $\hat{\beta}$ can be very large. Techniques such as ridge regression, principal components regression and latent root regression have been advocated as better than least squares, although none of these methods appears to have been widely used in practice. The recent paper by Breimen and Friedman[12] gives a review of some of these ideas and goes on to discuss how they can be extended to multivariate models using canonical variates. Much less attention appears to have been given to shrinkage prediction in logistic regression and other non-normal regression models.

Section 4 suggests a more general setting for the study of overfitting in statistical models, of which shrinkage of predictors is a special case. The formula obtained assumes that all the mathematical functions involved are smooth, admitting simple linear approximations. This needs to be generalized, for example if subset selection models are to be studied then the estimated regression vector would not be a continuous function of the observations.

Acknowledgements

I am grateful to Professor Hugh Tunstall-Pedoe for access to the data used in the examples in this paper. Research support from the ESRC (Programme in Risk and Human Behaviour, grant ref. L211252009) is also acknowledged.

References

1 Fisher RA. *Statistical methods for research workers*, 13th edn. London: Oliver & Boyd, 1958.

2 Pajak A, Kuulasmaa K. Geographical variation in the major risk factors of coronary heart disease in men and women aged 35–64 years. *World Health Statistics Quarterly* 1988; **41**: 115–40.

3 Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 1979; **74**: 829–36.

4 Stein C. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability* 1956; Vol. 1: 197–206.

5 Copas JB. Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society B*, 1983; **45**: 311–54.

6 Lindley DV. Contribution to the discussion of the paper by Stein (1962). [Stein, C. Confidence sets for the mean of a multivariate normal distribution (with discussion). *Journal of the Royal Statistical Society B,* 1962; **24**: 265–96.]

7 Draper D. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B*, 1995; **57**: 45–98.

8 Sclove SL. Improved estimation for coefficients in linear regression. *Journal of the American Statistical Association* 1968; **63**: 596–606.

9 McCullagh P, Nelder JA. *Generalized linear models*, 2nd edn. London: Chapman & Hall, 1989.

10 Miller AJ. *Subset selection in regression*. London: Chapman & Hall, 1990.

11 Copas JB, Jones MC. On the robustness of shrinkage predictors in regression to differences between past and future data. *Journal of the Royal Statistical Society B*, 1986; **48**: 223–37.

12 Breimen L, Friedman JH. Predicting multivariate responses in multiple linear regression (with discussion). *Journal of the Royal Statistical Society B* 1997; **59**: 3–54.