Appl. Statist. (1992)
41, *No.* 1, *pp.* 191–201

Ridge Estimators in Logistic Regression

By S. LE CESSIE† and J. C. VAN HOUWELINGEN

University of Leiden, The Netherlands

[Received January 1990. Revised November 1990]

SUMMARY

In this paper it is shown how ridge estimators can be used in logistic regression to improve the parameter estimates and to diminish the error made by further predictions. Different ways to choose the unknown ridge parameter are discussed. The main attention focuses on ridge parameters obtained by cross-validation. Three different ways to define the prediction error are considered: classification error, squared error and minus log-likelihood. The use of ridge regression is illustrated by developing a prognostic index for the two-year survival probability of patients with ovarian cancer as a function of their deoxyribonucleic acid (DNA) histogram. In this example, the number of covariates is large compared with the number of observations and modelling without restrictions on the parameters leads to overfitting. Defining a restriction on the parameters, such that neighbouring intervals in the DNA histogram differ only slightly in their influence on the survival, yields ridge-type parameter estimates with reasonable values which can be clinically interpreted. Furthermore the model can predict new observations more accurately.

Keywords: Cross-validation; Deoxyribonucleic acid histogram; Logistic regression; Predictive value; Ridge regression

1. Introduction

In biostatistics, logistic regression is a popular method to model binary data. However, unstable parameter estimates occur when the number of covariates is relatively large or when the covariates are highly correlated. In this paper it is shown how ridge estimators can be combined with logistic regression to improve the model in such situations.

As an example we consider the following clinical problem. For 81 patients with ovarian cancer the deoxyribonucleic acid (DNA) content of about 300 cancer cells was determined by DNA image cytometry. For each patient a histogram of the distribution of the DNA content of the cancer cells was made. The question was how the relation between survival and DNA content of cancer cells could be modelled by using the information on the whole DNA histogram.

The DNA value expresses the amount of DNA in a cell, where $1C$ corresponds to the amount of DNA in a haploid cell, a cell with one set of 23 chromosomes. To construct a DNA histogram, the range of DNA values is split into 37 classes, with class interval $0.2C$, except that the first class contains the fraction of cells with DNA values less than $0.9C$ and the last class the fraction of cells with DNA values greater than $7.9C$. An example of a DNA histogram is given in Fig. 1.

†*Address for correspondence*: Department of Medical Statistics, University of Leiden, PO Box 9512, 2300 RA Leiden, The Netherlands.

 0035–9254/92/41191 $2.00

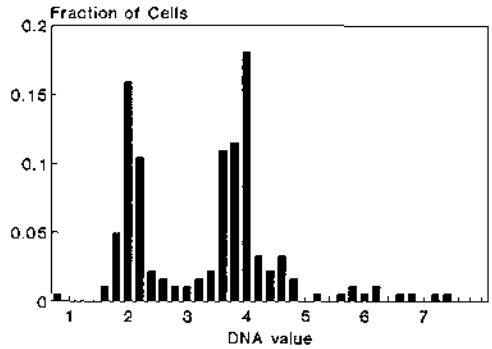

Fig. 1. Example of a DNA histogram for a patient with advanced ovarian cancer, based on about 300 cancer cells: the fractions of cells in the various categories of the histogram are used as covariates in the logistic regression model

In all histograms the DNA value $2C$ corresponds to the amount of DNA of a cell with two pairs of 23 chromosomes. If a person is healthy, their DNA histogram would have a large peak at $2C$ and a small peak at $4C$. Since empirical fractions are used, the histograms are mutually comparable. We shall denote the fraction of cells in class j for patient i as X_{ij}, where $\Sigma_j X_{ij} = 1$.

For 70 of the 81 patients, it was known whether they died within 2 years of diagnosis (28) or survived longer than 2 years (42) and in this paper we shall restrict our analysis to these 70 patients. The dichotomous variable Y is defined as $Y_i = 1$ if person i has died within 24 months and $Y_i = 0$ otherwise.

The probability that $Y_i = 1$, given the value of $X_i = (X_{i1}, \ldots, X_{i37})$, is denoted $p(X_i)$ and is modelled with the standard logistic regression model

$$p(X_i) = \exp\left(\sum_{j=1}^{37} \beta_j X_{ij}\right) \Big/ \left\{1 + \exp\left(\sum_{j=1}^{37} \beta_j X_{ij}\right)\right\},$$

or equivalently

$$\text{logit}\{p(X_i)\} = \sum_{j=1}^{37} \beta_j X_{ij},$$

without a constant since $\Sigma_j X_{ij} = 1$.

One problem is that the number of covariates is large compared with the number of observations and that the covariates are highly correlated. Overfitting and collinearity yield very unstable estimates and in our example some of the maximum likelihood estimates (MLEs) are infinite. A procedure to obtain more stable estimates is to pool neighbouring categories. However, determining the number of groups and the way of grouping is often rather subjective. An alternative approach, which does not suffer from these drawbacks, is to specify a restriction on the parameters β_j. In this example the DNA content is measured on a continuous scale and it is reasonable to assume that neighbouring categories in the DNA histogram differ only slightly in their influence on the outcome. This can be achieved by requiring that the difference between two successive parameters is small, i.e. $\Sigma(\beta_{j+1} - \beta_j)^2$ is restricted.

In ordinary linear regression finding a least square estimate subject to spherical restrictions on the parameters leads to ridge regression. In Section 2 extensions to logistic regression are made. It is shown how ridge estimators are used in the logistic regression model to obtain more realistic estimates for the parameters and to improve the predictive value of the model. How much the βs are restricted depends on the choice of the unknown ridge parameter. Various methods to determine the ridge parameter are discussed in Section 3. In Section 4, ridge regression is applied to the ovarian cancer data, to model the two-year survival probability for the ovarian cancer patients.

2. Ridge Estimators in Logistic Regression

In this section the approach of Duffy and Santner (1989) is followed to extend ridge regression theory in standard linear regression to logistic regression. The ridge estimator is derived as a restricted maximum likelihood estimator. A slightly different approach to define a ridge-type estimator is given by Schaefer *et al.* (1984). We shall demonstrate that both approaches are asymptotically equivalent.

Suppose that we have n observations (X_i, Y_i), where the Y_i are mutually independent binary (1–0) response variables, with $p(X_i)$ the probability of $Y_i = 1$ and X_i are d-dimensional row vectors of covariates. The probability function p follows the logistic regression model,

$$p(X_i) = \exp(X_i\beta)/\{1+\exp(X_i\beta)\},$$

with β a d-dimensional parameter vector. For the moment we assume that there is no constant term involved in this regression problem. Often the constant plays the role of a base-line and is treated differently from the other parameters. The treatment of the constant will be discussed in Section 4, where the ridge regression method is applied to the ovarian cancer data set.

The log-likelihood l of the data (X, Y) under this model is

$$l(\beta) = \sum_i [Y_i \log p(X_i) + (1 - Y_i) \log\{1 - p(X_i)\}].$$

Maximization of $l(\beta)$ yields the ordinary MLE $\hat{\beta}$ for β.

Duffy and Santner (1989) consider the maximization of the log-likelihood function with a penalty on the norm of β:

$$l^{\lambda}(\beta) = l(\beta) - \lambda\|\beta\|^2, \tag{2.1}$$

where $l(\beta)$ is the unrestricted log-likelihood function and $\|\beta\| = (\Sigma\beta_j^2)^{1/2}$, the norm of the parameter vector β. The maximizer of equation (2.1) is denoted $\hat{\beta}^{\lambda}$. The ridge parameter λ controls the amount of shrinkage of the norm of β. When $\lambda = 0$ the solution will be the ordinary MLE, whereas if $\lambda \to \infty$ the β_j all tend to 0.

A large number of explanatory variables and/or much correlation between the various explanatory variables give rise to unstable parameter estimates. Shrinking the βs towards 0 and allowing a little bias will stabilize the system and provide estimates with smaller variance. Therefore, for a good choice of λ, the estimate $\hat{\beta}^{\lambda}$ is expected to be on average closer to the real value of β than the unrestricted MLE, i.e. $\mathrm{MSE}(\hat{\beta}^{\lambda}) < \mathrm{MSE}(\hat{\beta})$. For the standard linear regression model there always exists a ridge

parameter $\lambda > 0$ for which the estimates have smaller mean-squared error than the minimum least square estimate. See Hoerl and Kennard (1971) or Draper and Smith (1983) for more detail on ridge regression for the ordinary least square situation.

Analogous to the unrestricted MLEs, $\hat{\beta}^\lambda$ may be obtained by the Newton–Raphson maximization procedure. The first derivative of $l^\lambda(\beta)$ is

$$\begin{aligned} U^\lambda(\beta) &= \sum_i X_i' \{Y_i - p(X_i)\} - 2\lambda\beta \\ &= U(\beta) - 2\lambda\beta, \end{aligned} \tag{2.2}$$

with $U(\beta)$ the derivative of the unrestricted log-likelihood. The negative of the matrix of second derivatives is

$$\Omega^\lambda(\beta) = \Omega(\beta) + 2\lambda I, \tag{2.3}$$

where $\Omega = X' V(\beta) X$ is the negative of the matrix of second derivatives of the unrestricted likelihood and $V(\beta)$ is an $n \times n$ diagonal matrix with $v_{ii} = p(X_i)\{1 - p(X_i)\}$.

Large sample properties of the restricted MLEs can be obtained by carrying out a Taylor series expansion of the first derivative of the penalized likelihood about the real parameter value β_0. This yields

$$U^\lambda(\hat{\beta}^\lambda) = U^\lambda(\beta_0) - (\hat{\beta}^\lambda - \beta_0)' \, \Omega^\lambda(\beta_0) + o(\|\hat{\beta}^\lambda - \beta_0\|).$$

Using equations (2.2) and (2.3) and $U^\lambda(\hat{\beta}^\lambda) = 0$ yields as a first-order approximation for $\hat{\beta}^\lambda$

$$\begin{aligned} \hat{\beta}^\lambda &= \beta_0 + \{\Omega(\beta_0) + 2\lambda I\}^{-1}\{U(\beta_0) - 2\lambda\beta_0\} \\ &= \{\Omega(\beta_0) + 2\lambda I\}^{-1}\{U(\beta_0) + \Omega(\beta_0)\beta_0\}. \end{aligned}$$

Analogously it can be shown that a first-order estimate for the unrestricted MLE is $\hat{\beta} = \beta_0 + \Omega^{-1}(\beta_0)\, U(\beta_0)$. Hence a first-order estimate of $\hat{\beta}^\lambda$ is

$$\hat{\beta}^\lambda = \{\Omega(\beta_0) + 2\lambda I\}^{-1}\, \Omega(\beta_0)\, \hat{\beta}.$$

Here we see that $\hat{\beta}^\lambda$ shrinks towards 0 if the value of the ridge parameter increases. Replacing $\Omega(\beta_0)$ by its estimate $\Omega(\hat{\beta})$ yields exactly the ridge-type estimate defined by Schaefer *et al.* (1984). Note that the estimate of Schaefer *et al.* is not defined if some of the unrestricted MLEs are infinite, which is so in the ovarian cancer example.

Under certain regularity conditions (Cox and Hinkley, 1974) $\hat{\beta}$ is asymptotically unbiased with covariance matrix $\Omega(\beta_0)^{-1}$. Then the asymptotic bias of $\hat{\beta}^\lambda$ becomes

$$E(\hat{\beta}^\lambda - \beta_0) = -2\lambda\{\Omega(\beta_0) + 2\lambda I\}^{-1}\beta_0$$

and the asymptotic variance of $\hat{\beta}^\lambda$ is

$$\{\Omega(\beta_0) + 2\lambda I\}^{-1}\, \Omega(\beta_0)\{\Omega(\beta_0) + 2\lambda I\}^{-1}.$$

Unfortunately, this approximation to the variance of $\hat{\beta}^\lambda$ cannot be used directly to construct approximate confidence intervals around $\hat{\beta}^\lambda$, since we have to take into account the bias of the estimate. Jackknife and bootstrapping might be possible methods to obtain more insight into the variability of $\hat{\beta}^\lambda$.

By similar reasoning, an asymptotic expression for the mean-squared error of $\hat{\beta}^\lambda$ can be obtained and for this expression it can be proved, analogous to the ordinary

least square situation (Hoerl and Kennard, 1971), that it attains its minimum at a value $\lambda^* > 0$.

3. Choice of Ridge Parameter

The ridge parameter can be chosen with or without (empirical) Bayesian arguments. We prefer to use a purely frequentist argument, based on minimizing an estimate of the prediction error of the model. Suppose that we have parameter estimates based on the data (X_i, Y_i), $i = 1 \ldots n$, and an estimate $\hat{p}(x)$ of the real probability function, based on the parameter estimates. We predict for a new observation, with covariate vector X_{new}, the probability that $Y_{\text{new}} = 1$ by $\hat{p} = \hat{p}(X_{\text{new}})$ and denote the real probability that $Y_{\text{new}} = 1$ by p. Various ways to define the error made by this prediction are discussed by Efron (1986) and by van Houwelingen and le Cessie (1990). We concentrate on three different measures to quantify the error of the prediction:

(a) classification or counting error

$$\begin{aligned} \text{CE} = 1 \quad & \text{if } Y_{\text{new}} = 1 \text{ and } \hat{p} < \tfrac{1}{2} \\ & \text{or } Y_{\text{new}} = 0 \text{ and } \hat{p} > \tfrac{1}{2}, \\ = \tfrac{1}{2} \quad & \text{if } \hat{p} = \tfrac{1}{2}, \\ = 0 \quad & \text{otherwise;} \end{aligned}$$

(b) squared error

$$\text{SE} = (Y_{\text{new}} - \hat{p})^2;$$

(c) minus log-likelihood error

$$\text{ML} = -\{Y_{\text{new}} \log \hat{p} + (1 - Y_{\text{new}}) \log(1 - \hat{p})\}.$$

The mean of all three measures is maximal if p is around $\frac{1}{2}$ and tends to 0 if p tends to 1 or to 0. This corresponds to the intuitive feeling that Y is more difficult to predict if the probability that $Y = 1$ is around $\frac{1}{2}$. The choice of the error measure depends mainly on the way that the model is used to predict future observations. The classification error corresponds to the prediction rule $\hat{Y}_{\text{new}} = 1$ if $\hat{p} > \frac{1}{2}$, $\hat{Y}_{\text{new}} = 0$ if $\hat{p} < \frac{1}{2}$ and a random assignment of $\hat{Y}_{\text{new}} = 1$ and $\hat{Y}_{\text{new}} = 0$ if $\hat{p} = \frac{1}{2}$. It indicates how well the model discriminates and is sensitive to the model predictions in the neighbourhood of $p = \frac{1}{2}$.

The other two measures consider the model predictions in the whole range of p-values. The squared error is intuitively appealing. It measures the Euclidean distance between Y_{new} and $\hat{p}$ and is a direct analogy of the squared error in the ordinary linear regression model.

The third measure ML, equal to $-\log \hat{p}$ if $Y_{\text{new}} = 1$ and equal to $-\log(1 - \hat{p})$ if $Y_{\text{new}} = 0$, is commonly used as an error measure for binary data. Summing the ML over all observations yields minus the log-likelihood of the data (Y, X), given the parameter vector β. Therefore in this paper this error measure is called the minus log-likelihood error. Names for twice the ML are the deviance and the entropy (Efron, 1978, 1986). Advantages of the ML are its relation to the log-likelihood function and the fact that it is not restricted to binary regression but can also be used in a more general setting. In van Houwelingen and le Cessie (1990), more properties of the ML are discussed and extensions are made to survival analysis.

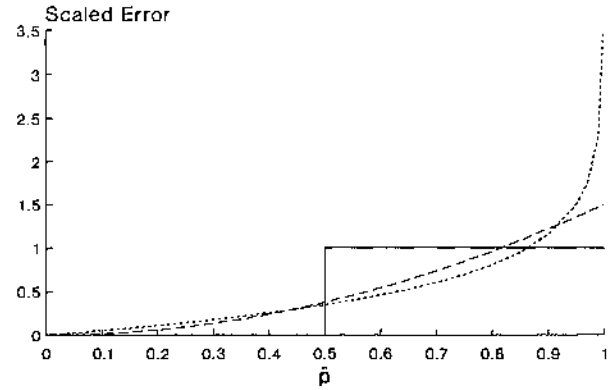

Fig. 2. Behaviour of the various prediction errors as functions of the predicted probability if $Y=0$: ——, CE; -------, ML; – – – –, SE

In Fig. 2 the various measures of the prediction error are compared. The error for one single new observation with $Y_{new}=0$ is given as a function of the predicted probability $\hat{p}$. The errors are scaled such that integrating the optimum error (the mean error when the real probability p is known) over p yields the same for all three errors. This means that the squared error is multiplied by $\frac{3}{2}$ and the minus log-likelihood is divided by 2. It is clearly seen that the SE and the ML differ little if $\hat{p}$ is not too far from 0 and that the classification error is a totally different criterion. In the neighbourhood of $\hat{p}=1$, SE tends to 1, while ML tends to infinity. The fact that impossible predictions, like $Y_{new}=0$ and $\hat{p}=1$, yield an infinitely large ML error is certainly another argument in favour of ML.

Ideally, we would have a validation set on which the predictive value of the logistic model could be compared for various values of λ and an optimal λ could be chosen such that the mean error rate is minimal. If a validation set is not available, a way to mimic the prediction is by cross-validation. In cross-validation the predictor for each observation is based on the other observations. Let $\hat{\beta}^{\lambda}_{(-i)}$ be the estimate based on all observations except (X_i, Y_i) and let $\hat{p}_{(-i)}(x)$ be the estimate of $p(x)$ based on $\hat{\beta}^{\lambda}_{(-i)}$. In this way it is possible to obtain an estimate for the mean prediction error, which can be minimized to obtain the ridge parameter. The cross-validated estimates of the mean of the three prediction errors defined above are

(a) the mean classification error

$$\mathrm{MCE}_{\mathrm{CV}} = n^{-1} \sum_i Y_i[\hat{p}_{(-i)}(X_i) < \tfrac{1}{2}] + (1 - Y_i)[\hat{p}_{(-i)}(X_i) > \tfrac{1}{2}] + \tfrac{1}{2}[\hat{p}_{(-i)}(X_i) = \tfrac{1}{2}],$$

where [ ] denotes the indicator function, i.e. [ ] = 1 if the proposition inside the brackets is true and [ ] = 0 if it is false,

(b) the mean-squared error

$$\mathrm{MSE}_{\mathrm{CV}} = n^{-1} \sum_i \{Y_i - \hat{p}_{(-i)}(X_i)\}^2$$

(c) and the mean minus log-likelihood

$$\mathrm{MML_{CV}} = -n^{-1} \sum_i [Y_i \log \hat{p}_{(-i)}(X_i) + (1 - Y_i) \log\{1 - \hat{p}_{(-i)}(X_i)\}].$$

Cross-validation is a time-consuming procedure. For each observation left out, the parameters are estimated again. Approximations for $\hat{\beta}^{\lambda}_{(-i)}$ can be made by following Cook and Weisberg (1982), who considered unrestricted MLEs. This yields as a one-step approximation for $\hat{\beta}^{\lambda}_{(-i)}$

$$\hat{\beta}^{\lambda}_{(-i)} = \hat{\beta}^{\lambda} - \frac{\{\Omega(\hat{\beta}^{\lambda}) + 2\lambda I\}^{-1} X_i' \{Y_i - \hat{p}(X_i)\}}{1 - h_{ii}}, \tag{3.1}$$

where $h_{ii} = v_i X_i \{\Omega(\hat{\beta}^{\lambda}) + 2\lambda I\}^{-1} X_i'$ and $v_i = \hat{p}(X_i)\{1 - \hat{p}(X_i)\}$. Hence we have an estimate for $\hat{\beta}^{\lambda}_{(-i)}$ without a need to re-estimate the parameters for each observation left out. Using arguments similar to those leading to equation (3.1) it can also be shown that $\mathrm{MSE_{CV}}$ can be approximated by

$$\mathrm{MSE_{ACV}} = n^{-1} \sum_i \frac{\{Y_i - \hat{p}(X_i)\}^2}{(1 - h_{ii})^2}. \tag{3.2}$$

Replacing h_{ii} by its average value $n^{-1} \Sigma h_{ii}$ yields a criterion, equivalent to the generalized cross-validation criterion of Golub *et al.* (1979) in the ordinary least square situation:

$$\mathrm{MSE_{GCV}} = n^{-1} \frac{\Sigma_i \{Y_i - \hat{p}(X_i)\}^2}{(1 - n^{-1} \Sigma h_{ii})^2}.$$

Golub *et al.* (1979) argue that this criterion is invariant under rotation in the model space. It is disputable whether rotation invariance is a necessary condition. $\mathrm{MSE_{GCV}}$ is less sensitive to the influential observations, the observations with a large value of h_{ii}. Both the $\mathrm{MSE_{ACV}}$ and the $\mathrm{MSE_{GCV}}$ are easy-to-compute estimates of the mean-squared error. In the application to the ovarian cancer data set, the two criteria differed only slightly.

4. Example

We return to the ovarian cancer data set. To write the restriction on the βs, $\Sigma(\beta_{j+1} - \beta_j)^2$, in the form of the restricted maximum likelihood problem we transform the covariates by defining

$$Z_{ij} = 1 - \sum_{k=1}^{j} X_{ik}, \qquad j = 0 \ldots 36.$$

Then the model becomes

$$\mathrm{logit}\{p(X_i)\} = \gamma_0 + \sum_{j=1}^{36} \gamma_j Z_{ij},$$

with $\gamma_0 = \beta_1$ and $\gamma_j = \beta_{j+1} - \beta_j$, $j = 1 \ldots 36$. Since the constant γ_0 plays the role of a base-line value, it is not allowed to shrink with the other parameters. Therefore we maximize the restricted likelihood with penalty $\Sigma_{j=1}^{36} \gamma_j^2$

$$l^{\lambda}(\gamma) = l(\gamma) - \lambda \sum_{j=1}^{36} \gamma_j^2. \tag{4.1}$$

It is straightforward to modify the formulae derived in Sections 2 and 3 for the alternative treatment of the constant and to obtain the estimates for γ by the Newton–Raphson maximization procedure.

The various estimates of the mean prediction error, discussed in Section 3, are optimized to obtain an estimate of the optimal ridge parameter. In Fig. 3 the cross-validated estimates of the various prediction errors are given as a function of the ridge parameter.

The cross-validated estimate of the classification error is small for the smallest values of λ. The classification error takes the same value for a whole range of λ and does not have one well-defined minimum. This is an argument against the use of this criterion in this situation. The question arises about whether there are enough observations to make MCE_{CV} a usable criterion. However, to compare this criterion with the other cross-validation criteria, we choose as optimal ridge parameter for this criterion, $\lambda = 0.005$.

Three different estimates for the mean-squared error were considered: the cross-validation criterion with parameters estimated again for each observation left out (MSE_{CV}), the approximation (3.2) of this criterion (MSE_{ACV}) and the generalized cross-validation criterion (MSE_{GCV}). The optimal choice of λ differed slightly, yielding as values for λ 0.065 (MSE_{ACV}), 0.05 (MSE_{CV}) and 0.035 (MSE_{GCV}). Finally, the MML_{CV} criterion yielded the largest value of the ridge parameter, 0.095.

The minimum values of the various cross-validation criteria can be interpreted as follows. If the real probability p is known, $MSE = p(1-p)$, and $MML = -\{p \log p + (1-p) \log(1-p)\}$. Comparing this with the minimum values of the MML and MSE obtained by cross-validation, we see that they both correspond to p-values of about 0.35 and 0.65. This gives an indication of how well the groups with $Y=1$ and $Y=0$ are separated. Without any parameters the cross-validated MSE

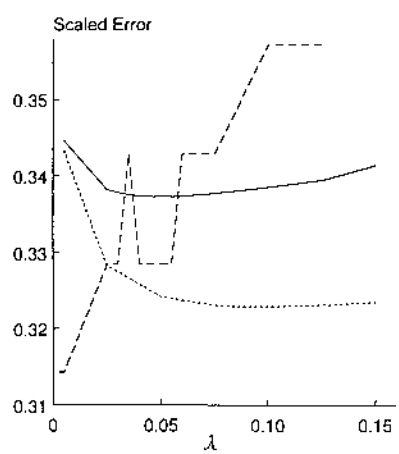

Fig. 3. Cross-validated estimates of the mean of the various error measures as functions of the ridge parameter λ: ———, MSE; – – – –, MCE; --------, MML

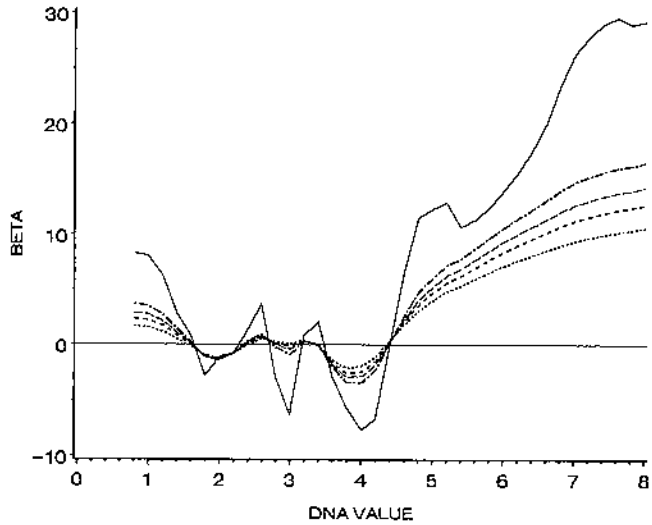

Fig. 4. Plot of the original parameters β (on the horizontal axis are the midpoints of the classes of the DNA histogram, corresponding to β_j): --------, $\mathrm{MML}_{\mathrm{CV}}$ ($\lambda = 0.095$); – – – –, $\mathrm{MSE}_{\mathrm{ACV}}$ ($\lambda = 0.065$); ——, $\mathrm{MSE}_{\mathrm{CV}}$ ($\lambda = 0.05$); – – – –, $\mathrm{MSE}_{\mathrm{GCV}}$ ($\lambda = 0.035$); ———, $\mathrm{MCE}_{\mathrm{CV}}$ ($\lambda = 0.005$)

would be 0.247 and MML would be 0.688, both corresponding to p-values of 0.45 and 0.55.

For the various ridge parameters, obtained by minimizing the various cross-validation criteria, the corresponding parameter estimates γ_j are calculated and transformed back to the original parameters β_j. In Fig. 4, β_j is plotted against the class midpoint of the DNA histogram, corresponding to category j. Clearly the curves become smoother if the ridge parameter increases, because the parameters $\gamma_j = \beta_{j+1} - \beta_j$ shrink towards 0.

The estimates seem reasonable from a practical point of view. $\Sigma_j \beta_j X_{ij}$ can be seen as a prognostic index for person i and higher values of it correspond to lower survival probabilities. We see that the βs are negative for DNA values around $2C$ and $4C$. Hence, patients with a relatively high number of cells with DNA contents of $2C$ or $4C$ have better survival probabilities. This corresponds to the fact that for a healthy person most of the cells have a DNA value of $2C$ or $4C$. If there are relatively more cells with an abnormal cell content, the risk of dying will increase.

To see the influence of the various estimates of the ridge parameter on the model predictions, the predicted probability of dying within 2 years is calculated for all patients by using the various estimated values of the ridge parameter. The mean absolute differences between the predicted probabilities are given in Table 1.

From the ridge parameters obtained by $\mathrm{MML}_{\mathrm{CV}}$ and the three mean-squared error estimates, we see that the mean predicted probabilities differ little. The predictions obtained by using the classification ridge parameter can differ considerably.

5. Discussion

As in ordinary linear regression, ridge regression is a good method for obtaining more stable parameter estimates for the logistic regression model. The type of

TABLE 1
Mean absolute differences between the predicted probabilities

	MSE_{GCV} ($\lambda = 0.035$)	MSE_{CV} ($\lambda = 0.05$)	MSE_{ACV} ($\lambda = 0.065$)	MML_{CV} ($\lambda = 0.095$)
MCE_{CV} ($\lambda = 0.005$)	0.056	0.066	0.073	0.083
MSE_{GCV} ($\lambda = 0.035$)		0.011	0.018	0.030
MSE_{CV} ($\lambda = 0.05$)			0.008	0.018
MSE_{ACV} ($\lambda = 0.065$)				0.012

restriction on the βs depends on the kind of covariates. In histogram data it is often reasonable to assume an underlying smooth structure and to require that differences between successive parameters be small. The same type of restriction can be used when covariates correspond to repeated measurements at consecutive time points.

Insight into the biological background of the histogram could be used to define relevant statistics directly from the DNA histogram. For example, in a previous analysis of these data by Rodenburg *et al.* (1987), two explanatory variables were derived from the DNA data: ploidy, a variable indicating whether a DNA histogram looks normal or abnormal, and the percentage of cells with a DNA value more than 5*C*. The variable ploidy is obtained by visual inspection, which can lead to observer bias. Furthermore in this approach the histogram of 37 classes is reduced to a histogram of two classes and one could wonder whether the data are not too reduced and whether the cut-off point at 5*C* is the optimal choice. The advantage of analysing DNA histograms by ridge regression is that the information from the entire histogram is used and that the only assumption made beforehand is the type of restriction on the parameters.

The choice of the ridge parameter depends on the cross-validation criterion which is used. In the ovarian cancer example we saw that the shrinkage of the parameters did not improve the classification error. There is no obvious relationship between the classification error and $MSE(\hat{\beta}^{\lambda})$ and it looks as though the fact that the mean-squared error of $\hat{\beta}^{\lambda}$ is smaller for a good choice of $\lambda > 0$ does not influence the behaviour of this criterion much. Apparently, a model with unrealistically large parameters can still be a good model for discriminating.

The use of one of the approximations instead of the exact cross-validated mean-squared error did not much change the outcomes in our small data set. These approximations become very useful with large data sets.

The minus log-likelihood and the squared error are reasonably equivalent. Although the squared error is intuitively more appealing, the minus log-likelihood has advantages, in that it is easy to extend to other situations and prevents impossible predictions, which makes it in our opinion the most suitable criterion.

Acknowledgement

We wish to thank Dr Diane Duffy for her helpful comments on an earlier draft of the paper.

References

Cook, R. D. and Weisberg, S. (1982) *Residuals and Influence in Regression*, sect. 5.4. London: Chapman and Hall.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Draper, N. R. and Smith, H. (1983) *Applied Regression Analysis*. New York: Wiley.

Duffy, D. E. and Santner, T. J. (1989) On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communs Statist. Theory Meth.*, **18**, 959–980.

Efron, B. (1978) Regression and ANOVA with zero–one data: measures of residual variation. *J. Am. Statist. Ass.*, **73**, 113–121.

——(1986) How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Ass.*, **81**, 461–470.

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

Hoerl, A. E. and Kennard, R. W. (1971) Ridge regression: biased estimates for nonorthogonal problems. *Technometrics*, **12**, 55–67.

van Houwelingen, J. C. and le Cessie, S. (1990) Predictive value of statistical models. *Statist. Med.*, **9**, 1303–1325.

Rodenburg, C. J., Ploem-Zaaijer, J. J., Cornelisse, C. J., Mesker, W. E., Hermans, J., Heintz, P. A. M., Ploem, J. S. and Fleuren, G. J. (1987) Use of DNA image cytometry in addition to flow cytometry for the study of patients with advanced ovarian cancer. *Cancer Res.*, **47**, 3938–3941.

Schaefer, R. L., Roi, L. D. and Wolfe, R. A. (1984) A ridge logistic estimate. *Communs Statist. Theory Meth.*, **13**, 99–113.