

Explained Variation for Logistic Regression – Small Sample Adjustments, Confidence Intervals and Predictive Precision

M. MITTLBÖCK and M. SCHEMPER

Department of Medical Computer Sciences
Vienna University, Vienna
Austria

Summary

The proportion of explained variation in logistic regression can be expressed by the multiple R^2 originally developed for the general linear model (cf. MITTLBÖCK and SCHEMPER (1996)). In this paper we present a detailed investigation of this measure in small samples and/or with many covariates and propose either of two adjustments, one being a direct analogue of R^2_{adj} of the general linear model, and the other being based on shrinkage. Furthermore, we explore the use of bootstrap confidence intervals and give a table of the expected variability of estimates of explained variation for samples of varying sizes. We recommend to quantify gains of predictive precision due to prognostic factors by both relative and absolute measures. For binary outcomes the components of the relative measure, R^2 , are suitable absolute measures of predictive precision. They are interpretable as average absolute residuals conditional on using prognostic factors and without such information. We motivate application of the presented measures by the statistical analysis of a study of physical characteristics of urine possibly related to the presence of calcium oxalate crystals.

Key words: Prediction error; Predictive accuracy; Proportion of explained variation, R^2 measures; Shrinkage

1. Introduction

Analyses of prognostic factors in studies of dichotomous outcomes are most often based on the logistic regression model. The results derived from such analyses comprise point estimates and confidence intervals of the relative risk associated with prognostic factors, and, corresponding p -values. By restricting analysis of a data set to these measures, a medical investigator may miss important additional information on the extent to which prognostic factors actually determine the outcome for a patient, or similarly, how much is known about the etiology of diseases or of their further development. Such questions can be suitably addressed by measures of the proportion of variation of a dependent variable explained by prognostic factors and by measures of predictive precision with and without prognostic factors.

MITTLBÖCK and SCHEMPER (1996) and MENARD (2000) reviewed several candidates for a suitable measure of explained variation for logistic regression which have been proposed or included in major software packages. MITTLBÖCK and SCHEMPER (1996) found that two of these – squared Pearson correlation between the binary outcome and the predictor, and, the proportional reduction of squared Pearson-type residuals by the use of covariates, R_{SS}^2 – give almost identical results, give approximately the same values as the multiple R^2 of the general linear model (GLM), when both models are suitable, and have an intuitively clear interpretation. Though either of these two measures can be recommended for practical use, for simplicity the discussion will now be limited to R_{SS}^2 .

KVALSETH (1985) proposed eight criteria for a “good” R^2 statistic for the linear model and concluded that R_{SS}^2 should be preferred over other R^2 definitions, also for non-linear regression models. MENARD (2000) applied Kvalseth’s criteria to logistic regression and, in particular, found that R_{SS}^2 violates criterion one: ‘ R^2 must possess utility as a measure of goodness of fit and have an intuitively reasonable interpretation’. In this context MENARD (2000) criticises that R_{SS}^2 depends on the base rate \bar{p} (1) and that it is not optimised by the fitting process (2). Furthermore, in an investigation of R^2 for binary responses, COX and WERMUTH (1992) criticise that R^2 tends to be low even for an underlying perfect regression relationship (3).

If R_{SS}^2 is viewed as a measure of goodness of fit, in the sense of systematic departures of observed and expected proportions, all three properties *are* disturbing. However, viewing R^2 as a measure of *explained variation* (KORN and SIMON, 1991), they are quite natural. *Explained variation* aims at quantifying how much the prediction error is reduced when using covariates compared to when not using them.

In response to Menard’s first criticism, the sensitivity of R_{SS}^2 to the base rate or prevalence of an outcome level also is seen as a strength because the practical importance of covariates, their “real-world value” (HILDEN, 1991; ASH and SHWARTZ, 1999), does depend on the prevalence of the problem in the target population. To be more specific, if \bar{p} is around 0.5 total variability is high and covariates may explain more of the uncertainty than if the outcome is already pretty much determined by \bar{p} , i.e., if \bar{p} is close to zero or one. This issue will be resumed in Section 4.

In response to Menard’s second criticism: in the fitting process of logistic regression a sum of weighted squared differences of observed and predicted outcome values (‘raw residuals’) is minimised (MCCULLAGH and NELDER, 1989; p. 6). While this approach is optimal with respect to efficiency of parameter estimation, the unequal weighting of raw residuals and thus of individuals harms the intuitive appeal of related measures of explained variation (WILLET and SINGER, 1988). We think that fitting a statistical model and judging its explanatory capacity for real life are distinct tasks. For the latter purpose agreement of fitting criteria and of measures of explained variation is not required. ZHENG and AGRETI (2000) point

out that R^2 -type measures for logistic regression based on the likelihood function are optimised by the fitting process but for the price of losing a natural scale of interpretation.

In response to the concern by COX and WERMUTH (1992), assume binomial data 40/100, 50/100 and 60/100 corresponding to values 0, 1 and 2, respectively, of a covariate. They are fitted perfectly by logistic regression, Hosmer and Lemeshow's goodness of fit test being totally insignificant. However, viewed on an individual-by-individual basis the observations are all zeros and ones while the predictions are either 0.4, 0.5 or 0.6. Thus knowledge of the particular covariate value reduces little of the uncertainty of an individual result. While the goodness of fit is perfect for this model, the explained variation is low. Only if there exists a critical value of a covariate below/above which no/all experimental units respond, should a measure of explained variation reach a value of one.

The purpose of this contribution is to present adaptations of estimates of explained variation by R^2_{SS} for use in small samples (Section 2). All unadjusted R^2 -type measures are biased and their bias increases with an increasing number of covariates and with decreasing sample size. In Section 3 confidence intervals are presented. Section 4 deals with absolute measures of predictive precision, which are components of the relative R^2 -type measures and improve their interpretation. In Section 5 the performance of the methods is investigated by means of a Monte Carlo study and Section 6 motivates application of the presented methods by means of an example.

2. Estimating R^2_{SS} in Small Samples with Many Covariates

We observe a sample of n observations $(y_i, \mathbf{x}_i), i = 1, \dots, n$, where $y_i \in \{0, 1\}$ denotes the dependent variable and \mathbf{x}_i a corresponding covariate vector. The estimates from a logistic regression are given by $\text{Pr ob}(y_i = 1 | \mathbf{x}_i) = \hat{p}_i = \exp(\hat{\beta}\mathbf{x}_i) / (1 + \exp(\hat{\beta}\mathbf{x}_i))$ with $\hat{\beta}$ denoting the estimated parameter vector. Furthermore, $\text{Pr ob}(y_i = 1) = \bar{p} = \sum y_i/n$.

Explained variation in logistic regression can be defined by the proportional reduction in dispersion of the dependent variable: $R^2_{SS} = 1 - \text{SSE}/\text{SST}$ where $\text{SST} = \sum (y_i - \bar{p})^2$ and $\text{SSE} = \sum (y_i - \hat{p}_i)^2, 1 \leq i \leq n$.

In the GLM the unsuitable property of inflation of R^2 -measures if the number of covariates k is large relative to a given sample size n can be avoided by using R^2 -adjusted, R^2_{adj} . For logistic regression we consider two different 'adjusted measures', which are motivated by the rationale underlying the R^2_{adj} of GLM:

One is $R^2_{SS,adj} = 1 - [\text{SSE}/(n - k - 1)]/[\text{SST}/(n - 1)]$ which is formally identical with the common definition of R^2_{adj} . The other formulation derives from $R^2_{adj} = R^2\hat{\gamma}$ of the GLM, where $\hat{\gamma} = (F - 1)/F$ is termed the shrinkage factor and F is the usual F -ratio statistic for testing whether any covariates are associated with the dependent variable y (cf. COPAS (1997), HARRELL et al. (1996), and VAN

HOUWELINGEN and LE CESSIE (1990)). Because shrinkage in logistic regression is usually estimated by $\hat{\gamma} = (\text{model } \chi^2 - k) / \text{model } \chi^2$, we suggest an R^2 -adjusted, better termed ' R^2 -shrunk', $R_{SS,shr}^2 = \hat{\gamma} R_{SS}^2$ in analogy to the definition of R_{adj}^2 given previously for the GLM. The model χ^2 is the total likelihood ratio χ^2 statistic for testing whether any of k covariates are associated with dependent variable y (cf. HARRELL et al. (1996)).

Note that under the null hypothesis of no covariate effects $E(\text{model } \chi^2) = k$. If $\text{model } \chi^2 < k$ then $\hat{\gamma} < 0$ and therefore $R_{SS,shr}^2$ can become negative as well as $R_{SS,adj}^2$ and the R_{adj}^2 of the GLM. However, quantifying the explained variation of a totally insignificant model is of little interest. Therefore the possibility that $R_{SS,shr}^2$ and $R_{SS,adj}^2$ could become negative is not considered a disadvantage for application. As the conditions under which R_{adj}^2 was derived (normality and homoscedasticity of residuals, unweighted least-squares estimates of parameters) do not hold for logistic regression, the analogous adjustments used with $R_{SS,shr}^2$ and $R_{SS,adj}^2$ are only considered as approximate. Therefore, it was important to subject $R_{SS,shr}^2$ and $R_{SS,adj}^2$ to an extensive empirical study, which is presented in Section 5.

3. Confidence Limits for Measures of Explained Variation in Logistic Regression

While in the GLM analytic formulae are known for confidence intervals of R^2 (see HELLAND (1987)) no analogous solution seems achievable for logistic regression. We therefore considered the bootstrap (cf. SHAO and TU (1995) and DAVISON and HINKLEY (1997)) based on resampling of observation vectors, the 'paired bootstrap', and construction of confidence intervals by the percentile, the BC and the BCa methods. For an intended coverage of $(1 - \alpha)$, the simple percentile interval is obtained by determining the $\alpha/2$ -th and the $(1 - \alpha/2)$ -th percentile of N (say 1000) bootstrap replicates of $R_{SS,shr}^2$ or $R_{SS,adj}^2$. The empirical performance of the bootstrap procedure is investigated in Section 5.

4. Additional Information from the Components of R_{SS}^2

Usually, in applied statistics, interpretation of a relative measure requires knowledge of its absolute components. For example, when citing a relative risk of 2, say, this may refer to an increase of event rates from 10% to 20% but also applies to an increase from 0.1% to 0.2%. Often only the former situation will be considered of practical relevance. As the relative effect measure is the same in both very different situations, it is important to also consider corresponding absolute risks when reporting relative risk.

A related problem is the interpretation of explained variation by the relative measure R_{SS}^2 . If, for a certain study population, unconditional prediction is already

precise, i.e., if the denominator of R_{SS}^2 is small because of \bar{p} being close to 0 or 1 (homogeneous population), we cannot improve much by using further covariates. If \bar{p} is about 0.5 (heterogeneous population) and covariate effects are strong then the R_{SS}^2 will be high compared to the R_{SS}^2 for a homogeneous stratum within the cited population. For such a stratum no strong covariates may be available, but unconditional prediction can be reasonably precise. Thus, with increasing knowledge and increasing tendency to study more homogeneous subpopulations, a relative measure of explained variation, such as R_{SS}^2 , will tend to decrease. However, with increasing understanding of diseases, one should expect explained variation to increase. The obvious contradiction can be resolved by citing values of an absolute measure of prediction error for the null model without covariates and for the full model.

The absolute components of R_{SS}^2 , MSE (= SSE/n) and MST (= SST/n), are average squared residuals or distances on the probability scale and thus have no intuitive interpretation. Average absolute residuals, conditional on the use of covariates, $D_x = n^{-1} \sum |y_i - \hat{p}_i|$, or without such information, $D = n^{-1} \sum |y_i - \bar{p}|$, would be preferable.

Fortunately, for binary outcomes y it can be shown that

$$D = n^{-1} \sum |y_i - \bar{p}| = 2n^{-1} \sum (y_i - \bar{p})^2 = 2MST$$

and

$$D_x = n^{-1} \sum |y_i - \hat{p}_i| = 2n^{-1} \sum (y_i - \hat{p}_i)^2 = 2MSE.$$

This remarkable relationship of absolute and squared error with binary data can be verified by setting $y_i = 1$ and $y_i = 0$, $n\bar{p}$ and $n(1 - \bar{p})$ times, respectively.

Therefore, alternatively, $R_{SS}^2 = (D - D_x)/D$. Now the components of R_{SS}^2 measuring the prediction errors D and D_x are directly interpretable on the scale of proportions.

In Section 2 we have presented shrunk and adjusted versions of R_{SS}^2 . In small samples the MSE component of R_{SS}^2 is biased towards zero. One could either use an adjusted version of MSE ($MSE_{adj} = SSE/(n - k - 1)$) or obtain a properly inflated version of MSE, termed MSE_{shr} , $MSE_{shr} = MST(1 - \hat{\gamma}) - MSE \hat{\gamma}$ from $R_{SS,shr}^2 = \hat{\gamma} (MST - MSE)/MST$ assuming MST not being affected by shrinkage. Thus, for no shrinkage occurring ($\hat{\gamma} = 1$), $MSE_{shr} = MSE$ while from $\hat{\gamma} = 0$ $MSE_{shr} = MST$ follows. Therefore, a better estimate of the average conditional prediction error in small samples is $D_{x,adj} = 2MSE_{adj}$ or $D_{x,shr} = 2MSE_{shr}$.

5. Empirical Investigations

The purpose of this section is to describe the performance of the procedures introduced in previous sections and, based on these, to make suggestions on their use in practice.

The topics covered by our Monte Carlo study are:

1. Unbiasedness of small sample adjustments
2. Coverage by bootstrap confidence intervals
3. Variability of estimates of R_{SS}^2

We have included the third item because the expected variability of estimates of R_{SS}^2 depending on sample size (and other factors) may be of interest when the effort involved in obtaining confidence intervals is not considered worth while.

In our simulation project we have investigated logistic regression models with dichotomous $[0,1]$ and with uniformly $(0-1)$ distributed covariates \mathbf{x} . Values of the dependent variable y , 1 and 0, were generated with probabilities

$$P = \exp\left(b_0 + \sum_j b_j x_j\right) / \left[1 + \exp\left(b_0 + \sum_j b_j x_j\right)\right] \quad \text{where } 1 \leq j \leq k,$$

and $1 - P$, respectively, using the random number generator G05CAF of NAG (1998). Values of the parameters b_0 and b_1 were chosen in such a way that the underlying population R_{SS}^2 would be 0, 0.1, 0.2, 0.4 and 0.6 and underlying $\bar{p} = 0.5$ and 0.9. For experiments with $k > 1$ covariates the parameters b_j ($2 \leq j \leq k$) were taken to be zero in the underlying population. Results for each experimental condition are based on 1000 simulated samples and, for investigations of the coverage of confidence intervals on 1000 bootstrap replications each. Due to the magnitude of the Monte Carlo study and in order not to confuse the reader we shall present only the most typical results in tables but comment on how further experimental results agree with these.

From Table 1 we learn that the performance of $R_{SS,adj}^2$ and of $R_{SS,shr}^2$ is equally satisfactory. The unadjusted estimate of R_{SS}^2 is consistently inflated and therefore not recommended. As shown by Table 1 it reaches a mean value of 0.15 for a population value of 0.0 if $k = 15$ and $n = 100$. No different conclusions were drawn from experiments with $\bar{p} = 0.9$ or with other types of covariates.

Coverage probabilities from one-sided lower (extending to $-\infty$) and upper (extending to ∞) confidence intervals are based on the paired bootstrap of $R_{SS,shr}^2$ using 1000 replications for the percentile method. Table 2 indicates that the observed coverage probabilities are not always close to the nominal ones – in particular for the case with many covariates – but still such confidence intervals permit a judgement of the range of underlying R_{SS}^2 -values compatible with a sample. Results do not change substantially with different covariate distributions, an underlying $\bar{p} = 0.9$ or with bootstrapping $R_{SS,adj}^2$ instead of $R_{SS,shr}^2$.

The reader will recognize that we present results for $R_{SS}^2 = 0.1$ instead of $R_{SS}^2 = 0.0$. In the presence of R_{SS}^2 -values close to zero, which implies a completely insignificant regression model, confidence intervals become anti-conservative. However, there is little use of a model – and of a confidence interval for R_{SS}^2 – which has no predictive or explanatory capacity at all.

Results for the BC and BCa methods are not given but these methods almost always were outperformed by the simpler percentile method.

Table 1
Comparison of estimates by R_{SS}^2 , $R_{SS,adj}^2$ and $R_{SS,shr}^2$

k	n	Population $R_{SS}^2 (\times 100)$											
		0		20		40		60					
5	100	5	0	0	24	20	20	43	40	39	63	61	58
5	200	3	0	0	22	20	20	42	40	39	61	60	59
5	500	1	0	0	21	20	20	41	40	40	60	60	60
10	100	10	0	1	28	20	19	46	41	37	65	61	57
10	200	5	0	0	24	20	20	43	40	39	62	60	58
10	500	2	0	0	22	20	20	42	40	40	61	60	59
15	100	15	0	1	33	22	19	53	42	40	—	—	—
15	200	8	0	0	26	20	19	45	41	39	64	61	57
15	500	3	0	0	22	20	20	42	40	39	61	60	59

Note: All estimates ($\times 100$) are given as triplets ($R_{SS}^2, R_{SS,adj}^2, R_{SS,shr}^2$) and are averages from 1000 simulated trials with expected $\bar{p} = 0.5$ and k dichotomous covariates. Due to the frequent occurrence of separation for $k = 15, n = 100$ and underlying $R_{SS}^2 = 0.6$ we do not present these results.

Summarising, the results permit a cautious recommendation of the bootstrap percentile method. Currently, there appears to be no more accurate or computationally less involved alternative. Improvements could be an area of further research.

However, as the use of R^2 -measures is mainly descriptive, for many practical situations only a rough idea of the variability of such measures may be needed. Therefore we supply such information for $R_{SS,shr}^2$ in Table 3. The estimators $R_{SS,shr}^2$

Table 2
Coverage probability ($\times 100$) of one-sided lower/upper $(1 - \alpha)\%$ bootstrap confidence intervals for R_{SS}^2 using estimates by $R_{SS,shr}^2$

k	$(1 - \alpha)\%$	Population $R_{SS}^2 \times 100$			
		10	20	40	60
1	99.5	100/99	100/100	99/100	100/99
1	97.5	98/98	98/99	97/97	98/96
1	95	95/96	95/96	95/94	96/93
5	99.5	100/99	100/99	100/99	100/99
5	97.5	100/95	100/96	99/94	99/96
5	95	99/92	98/92	97/93	96/93
10	99.5	100/97	100/98	100/98	100/98
10	97.5	100/90	100/96	99/94	99/96
10	95	100/84	98/92	98/90	98/93

Note: All estimates ($\times 100$) are based on 1000 simulated trials of $n = 200$ with expected $\bar{p} = 0.5$ and dichotomous covariates.

Table 3
 Expected standard deviations of estimates using $R_{SS,shr}^2$

n	\bar{p}	$k =$	Population $R_{SS}^2 (\times 100)$								
			5		20		40		60		
			1	10	1	10	1	10	1	10	
50	0.5		7	11	12	13	14	13	14	14	11
100	0.5		5	6	8	8	10	10	10	10	10
200	0.5		3	4	6	6	7	7	7	7	7
200	0.9		5	7	10	10	11	12	12	11	
500	0.5		2	2	4	4	5	4	4	4	4
500	0.9		3	4	6	6	7	7	7	7	7
1000	0.5		1	1	3	3	3	3	3	3	3
1000	0.9		2	2	4	4	5	5	5	5	5

Note: All estimates of standard deviations ($\times 100$) are based on 1000 simulated trials.

and $R_{SS,adj}^2$ are almost identical in this respect. We recognize the obvious effect of n , some effect of \bar{p} , and almost no effect of k . The variability produced by underlying R_{SS}^2 is analogous in shape to the variability of proportions, i.e., it is highest around 0.5. The effect of covariate distributions was negligible. In Table 3 results for $\bar{p} = 0.9$ and $n \leq 100$ are omitted due to frequent non-convergence in the fitting process due to separation.

6. An Example and Concluding Remarks

We motivate the use of explained variation in logistic regression for prognostic factor studies by a clinical example. A study of $n = 77$ completely documented urine specimens (ANDREWS and HERZBERG (1985)) was analysed to determine if certain physical characteristics of the urine might be related to the presence of calcium oxalate crystals. The investigated physical characteristics were:

(1) *specific gravity*, the density of urine relative to water; (2) *pH*, the negative logarithm of the concentration of the hydrogen ion; (3) *osmolarity* is proportional to the concentration of molecules in solution; (4) *conductivity* is proportional to the concentration of charged ions in solution; (5) *urea concentration* in millimoles per litre; and (6) *calcium concentration* in millimoles per litre. Some of these characteristics are highly correlated. For computational reasons *specific gravity* was multiplied by 100 and all variables were standardised to a mean of zero. Table 4 gives standard results by logistic regression analysis.

These results do not tell anything about the degree to which prediction errors are reduced if physical characteristics are taken into account. Such information is

Table 4
Standard analysis results of urine data

Factors	Odds Ratio	(95% Confidence Limits)	P
Specific Gravity	1.43	(0.92–2.21)	0.11
PH	0.61	(0.20–1.86)	0.38
Osmolarity	1.02	(0.98–1.05)	0.35
Conductivity	0.65	(0.40–1.06)	0.08
Urea Concentration	0.97	(0.94–1.00)	0.05
Calcium Concentration	2.19	(1.36–3.52)	0.001

provided by measures of explained variation and predictive accuracy. In our example $R_{SS,adj}^2 = 0.48$ and $R_{SS,shr}^2 = 0.45$, the latter meaning that the regression model can explain 45% of the variation of the dependent variable, *presence of calcium oxalate crystals*. The unadjusted large sample estimate $R_{SS}^2 = 0.52$ is inappropriate here.

We have applied the measures $R_{SS,adj}^2$ and $R_{SS,shr}^2$ to the analysis of several clinical studies of binary outcomes at the Vienna University Medical School and found that values of 0.4 and above already indicate “high” explained variation. Often, despite highly significant covariates and impressive estimates of their relative risk, explained variation will be substantially lower.

Two-sided 95% confidence limits for R_{SS}^2 obtained according to the bootstrap of Section 3 are 0.31 and 0.71. The width of this confidence interval is in agreement with the expected standard deviations of estimates from $R_{SS,shr}^2$ of Table 3.

We also cite the components of $R_{SS,shr}^2$ and thus quantify the estimated prediction error with the use of prognostic factors ($D_{x,shr} = 0.27$) and without ($D = 0.49$). The prediction error is reduced from 0.49 to 0.27, taking into account the small sample size.

We think that reporting both relative ($R_{SS,adj}^2$ or $R_{SS,shr}^2$) and absolute measures (D and D_x) of predictive accuracy is helpful in summarising the degree to which individual outcomes are determined by prognostic factors within a given model.

For routine evaluation of predictive accuracy and explained variation in logistic models a SAS macro is available at

‘<http://www.akh-wien.ac.at/imc/biometrie/evlogist.htm>’.

Acknowledgement

The authors wish to thank both referees for their helpful comments.

References

- ANDREWS, D. F. and HERZBERG, A. M., 1985: *Data. A Collection of Problems from Many Fields for the Student and Research Worker*. Springer Verlag, New York.
- ASH, A. and SHWARTZ, M., 1999: R^2 : A useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine* **18**, 375–384.
- COPAS, J. B., 1997: Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research* **6**, 167–183.
- COX, D. R. and WERMUTH, N., 1992: A Comment on the Coefficient of Determination for Binary Responses. *The American Statistician* **46**, 1–4.
- DAVISON, A. C. and HINKLEY, D. V., 1997: *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.
- HARRELL, F. E., LEE, K. L., and MARK, D. B., 1996: Tutorial in Biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.
- HELLAND, I. S., 1987: On the Interpretation and use of R^2 in regression analysis. *Biometrics* **43**, 61–69.
- HILDEN, J., 1991: The area under the ROC curve and its competitors. *Medical Decision Making* **11**, 95–101.
- KORN, E. L. and SIMON, R., 1991: Explained Residual Variation, Explained Risk, and Goodness of Fit. *The American Statistician* **45**, 201–206.
- McCULLAGH, P. and NELDER, J. A., 1989: *Generalized Linear Models*. Chapman & Hall, London.
- MENARD, S., 2000: Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician* **54**, 17–23.
- MITTLBÖCK, M. and SCHEMPER, M., 1996: Explained variation for logistic regression. *Statistics in Medicine* **15**, 1987–1997.
- NAG, 1998: *Nag Fortran Library Manual-Mark 18*. Numerical Algorithms Group Ltd., Oxford.
- SHAO, J. and TU, D., 1995: *The Jackknife and Bootstrap*. Springer, New York.
- VAN HOUWELINGEN, J. C. and LE CESSIE, S., 1990: Predictive value of statistical models. *Statistics in Medicine* **9**, 1303–1325.
- WILLET, J. B. and SINGER, J. D., 1988: Another cautionary note about R^2 : Its use in weighted least-squares regression analysis. *American Statistician* **42**, 236–238.
- ZHENG, B. and AGRESTI, A., 2000: Summarizing the predictive power of a generalized linear model. *Statistics in Medicine* **19**, 1771–1781.

MARTINA MITTLBÖCK and MICHAEL SCHEMPER
 Section of Clinical Biometrics
 Department of Medical Computer Sciences
 Vienna University
 Spitalgasse 23
 A-1090 Vienna
 Austria
 Email: Martina.Mittlboeck@akh-wien.ac.at

Received, January 2001
 Revised, June 2001
 Revised, December 2001
 Accepted December 2001