STATISTICS IN MEDICINE, VOL. 15, 1987–1997 (1996)

EXPLAINED VARIATION FOR LOGISTIC REGRESSION

MARTINA MITTLBÖCK AND MICHAEL SCHEMPER

Section of Clinical Biometrics, Department of Medical Computer Sciences, Vienna University, Spitalgasse 23, A-1090 Vienna, Austria

SUMMARY

Different measures of the proportion of variation in a dependent variable explained by covariates are reported by different standard programs for logistic regression. We review twelve measures that have been suggested or might be useful to measure explained variation in logistic regression models. The definitions and properties of these measures are discussed and their performance is compared in an empirical study. Two of the measures (squared Pearson correlation between the binary outcome and the predictor, and the proportional reduction of squared Pearson residuals by the use of covariates) give almost identical results, agree very well with the multiple R^2 of the general linear model, have an intuitively clear interpretation and perform satisfactorily in our study. For all measures the explained variation for the given sample and also the one expected in future samples can be obtained easily. For small samples an adjustment analogous to R^2_{adj} in the general linear model is suggested. We discuss some aspects of application and recommend the routine use of a suitable measure of explained variation for logistic models.

1. INTRODUCTION

The squared multiple correlation coefficient R^2, also called the coefficient of determination, is well established in classical regression analysis. The main reason for the popularity of R^2 is its interpretability as the proportion of variation of the dependent variable which can be explained by a given regression model and predictor variables. It quantifies predictability and thus knowledge and gives the strength of a regression relationship.

In this paper we are not interested in the possible use of R^2 as a measure of goodness-of-fit for there are better tools and measures to check that (see Hosmer and Lemeshow[1]). However, in recommending R^2 as a measure of explained variation we briefly address two major criticisms which have been made against it (for example by Cox and Wermuth[2]). The first is that R^2 depends on the range and distribution of explanatory factors and the second is that with binary responses R^2 tends to be low even for an underlying perfect regression relationship.

As Korn and Simon[3] have pointed out, both criticisms apply only to R^2 as a measure of goodness-of-fit, not of explained variation. In the first case an explanatory factor with a 'large' estimated regression coefficient is of little prognostic value, if the factor has almost no variability in the population from which the sample was drawn. Therefore differences between values of R^2 among populations with the same assumed regression relationship but different spread of explanatory factors have a natural interpretation. In response to the second criticism, assume dose-response data are fitted perfectly by logistic[1,4,5] regression; knowledge of the particular dose given to an experimental unit reduces little of the uncertainty of an individual result, which is also affected by other unknown factors. Only if there exists a critical dose below/above which no/all experimental units respond should a measure of explained variation reach a value of one.

CCC 0277–6715/96/191987–11

Received October 1994
Revised November 1995

For the general linear model there exists only one suitable measure, multiple R^2, to quantify explained variation; there are several equivalent definitions[6] of this measure.

For logistic regression, today perhaps the most frequently used regression model after the general linear model, many different proposals have been made to measure explained variation. Major software packages include one or other of such measures and therefore they get cited in medical papers where logistic regression has been used for statistical analysis. We suggest that a more rational and unified choice from among the available candidate measures is desirable and have therefore conducted a comprehensive investigation. In the following section twelve measures of explained variation in logistic regression are reviewed and the properties which follow directly from the definitions are discussed.

2. SYSTEMATIC PRESENTATION OF MEASURES

We classify the measures to be discussed into three groups which will be dealt with separately in the sequence: squared correlation between observed and predicted outcomes; proportional reduction in dispersion of outcome, and likelihood-based measures.

We have a sample of observations (y_i, x_i), $i = 1, \ldots, n$, where $y_i = 0$ or 1, is the dependent variable and x_i is a corresponding covariate vector. We denote the estimates from a logistic regression by $\widehat{\text{Prob}}(y_i = 1 | x_i) = \hat{p}_i = \exp(\hat{\beta} x_i)/(1 + \exp(\hat{\beta} x_i))$ with $\hat{\beta}$ denoting the estimated parameter vector. Furthermore $\widehat{\text{Prob}}\,(y_i = 1) = \bar{p} = \sum_i y_i/n$.

2.1. R^2-measures using the squared correlation of y and $\hat{p}$

There are six of these:

(i) Squared Pearson correlation (r^2)

$$r = \sum_i (y_i - \bar{p})(\hat{p}_i - \bar{p}) \Big/ \sqrt{\left\{\sum_i (y_i - \bar{p})^2 \sum_i (\hat{p}_i - \bar{p})^2\right\}}$$

$$= \left(\sum_i y_i \hat{p}_i - n\bar{p}^2\right) \Big/ \sqrt{\left\{n\bar{p}(1 - \bar{p}) \sum_i (\hat{p}_i - \bar{p})^2\right\}}.$$

(ii) Squared Spearman correlation (r_s^2)

$$r_s = \sum_i (R(y_i) - \bar{R})(R(\hat{p}_i) - \bar{R}) \Big/ \sqrt{\left\{\sum_i (R(y_i) - \bar{R})^2 \sum_i (R(\hat{p}_i) - \bar{R})^2\right\}}$$

where $R(z)$ denotes the rank of z and $\bar{R} = (n + 1)/2$.

(iii) Squared Kendall's $\tau_a(\tau_a^2)$

$$\tau_a = \sum_{i<j} \text{sign}(y_j - y_i)\,\text{sign}(\hat{p}_j - \hat{p}_i)/[n(n - 1)/2] \text{ where } \text{sign}(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0. \end{cases}$$

(iv) Squared Kendall's $\tau_b(\tau_b^2)$

$$\tau_b = \sum_{i<j} \text{sign}(y_j - y_i)\,\text{sign}(\hat{p}_j - \hat{p}_i) \Big/ \sqrt{\left\{\sum_{i<j} \text{sign}^2(y_j - y_i) \sum_{i<j} \text{sign}^2(\hat{p}_j - \hat{p}_i)\right\}}.$$

(v) Squared Somers' $D_{\hat{p}y}(D^2_{\hat{p}y})$

$$D_{\hat{p}y} = \sum_{i<j} \operatorname{sign}(y_j - y_i)\operatorname{sign}(\hat{p}_j - \hat{p}_i) \Big/ \sum_{i<j} \operatorname{sign}^2(y_j - y_i)$$

(vi) Squared Goodman and Kruskal's $\gamma(\gamma^2)$

$$\gamma = \sum_{i<j} \operatorname{sign}(y_j - y_i)\operatorname{sign}(\hat{p}_j - \hat{p}_i) \Big/ \left[\sum_{i<j} \operatorname{sign}^2(y_j - y_i)\operatorname{sign}^2(\hat{p}_j - \hat{p}_i) \right].$$

While r_s is just r calculated from a separate ranking for y and $\hat{p}$, measures (iii)–(vi) are related by having the same numerator and by being identical for completely untied samples, but using different standardizations, with τ_a penalizing for ties in either y or $\hat{p}$, $D_{\hat{p}y}$ only for ties in $\hat{p}$, and γ and τ_b in neither y nor $\hat{p}$. The measure τ_b is the closest analogue to r and r_s in the case of ties but gives generally lower values than either r or r_s (see Kendall and Gibbons,[7] Chapter 9, and Stuart[8]), the last two being very close in normally distributed samples. As γ does not account for ties that occur only in y or only in $\hat{p}$ it could happen that in a 2 × 2 table only one of the four cell frequencies is zero and γ is 1 or -1 although the prediction is not perfect. With dichotomous y the measure τ_a can never reach a value of 1, even if prediction is perfect, which makes it an inappropriate measure of explained variation.

As τ_b is the geometric average of both asymmetric Somers' D's ($\tau_b = \sqrt{\{D_{y\hat{p}}D_{\hat{p}y}\}}$), D is an analogue of a regression rather than a correlation coefficient.

Logistic regression is a parametric tool and therefore all measures of ordinal association do not really correspond to it. Resulting inconsistent behaviour can be demonstrated by adding a weak continuous covariate to a model with a strong binary covariate. While the variation explained by this model will be only slightly increased, for example as measured by r^2, the explained variation by a measure of rank correlation can strikingly decrease. This is a consequence of forcing different ranks to slightly different values of the predictor, while respective ranks were tied in the single covariate model.

Furthermore, as pointed out by a referee, such measures cannot distinguish among models which, depending on the choice of different monotonically related link functions, have poorer or better fit. Also, for a model with a single covariate, non-linear monotonic transformations of this covariate will affect goodness-of-fit and parameter estimates and all measures of explained variation except those using rank information only.

Further discussion of the cited coefficients is given in Kendall and Gibbons,[7] Somers[9] and Goodman and Kruskal.[10]

Though squared tau-type measures (iii)–(vi) have been used occasionally (see for example Korn and Simon[11]) it is unclear in what sense they can be interpreted as a proportional reduction in variation.

Thus we already recognize qualitative discrepancies between some of the measures but will further investigate and comment on their suitability in the context of an empirical study in Section 3.

The program PROC LOGISTIC of SAS[12] uses the measures $D_{\hat{p}y}$, γ and τ_a to describe the association of y and $\hat{p}$.

2.2. R^2-measures based on a proportional reduction in dispersion of y

The general form of these measures[13,14] of the proportion of explained variation (PEV) is

$$\mathrm{PEV} = \left[\sum_i D(y_i) - \sum_i D(y_i|x_i) \right] \Big/ \sum_i D(y_i)$$

where $D(y_i)$ and $D(y_i|x_i)$ denote a measure of the distance of y_i from an unconditional or conditional (on a model and covariate vector x_i) central location parameter. The four measures of this subsection differ in their specification of $D(y_i)$ and $D(y_i|x_i)$:

(vii) Sums-of-squares $R^2 (R^2_{\mathrm{SS}})$
In the general linear model, PEV with $D(y_i) = (y_i - \bar{y})^2$ and $D(y_i|x_i) = (y_i - \hat{y}_i)^2$, reduces to the standard multiple $R^2 = 1 - \mathrm{SSE/SST}$ with $\mathrm{SST} = \sum_i D(y_i)$ and $\mathrm{SSE} = \sum_i D(y_i|x_i)$. We can still use squared residuals $D(y_i) = (y_i - \bar{p})^2$ and $D(y_i|x_i) = (y_i - \hat{p}_i)^2$ for logistic regression and arrive at

$$R^2_{\mathrm{SS}} = 1 - \mathrm{SSE/SST} = \left[2 \sum_i y_i \hat{p}_i - \sum_i \hat{p}_i^2 - n\bar{p}^2 \right] \Big/ (n\bar{p}(1 - \bar{p}))$$

(see Margolin and Light[15]).

(viii) Gini's concentration measure (R^2_{G})
With logistic regression this measure uses $D(y_i) = \bar{p}(1 - \bar{p})$ and $D(y_i|x_i) = \hat{p}_i(1 - \hat{p}_i)$ (see Haberman[16]), which is the expected variance under the logistic model. However, the explicit use of the binomial variation in R^2_{G} assumes the model to be correct. The measure simplifies to

$$R^2_{\mathrm{G}} = \left(\sum_i \hat{p}_i^2 - n\bar{p}^2 \right) \Big/ [n\bar{p}(1 - \bar{p})].$$

(ix) Classification error $R^2 (R^2_{\mathrm{CER}})$
This is discussed by van Houwelingen and le Cessie[17] and by Makuch *et al.*[18] and is equivalent to Goodman and Kruskal's[10] λ. A cutpoint (usually 0·5) is chosen for the predictions $\bar{p}$ and $\hat{p}_i$ above and below which a case is classified as 1 and 0, respectively. Therefore

$$D(y_i) = \begin{cases} 1 & \text{if } |y_i - \bar{p}| > 0{\cdot}5 \\ 0{\cdot}5 & \text{if } |y_i - \bar{p}| = 0{\cdot}5 \\ 0 & \text{if } |y_i - \bar{p}| < 0{\cdot}5 \end{cases} \quad \text{and} \quad D(y_i|x_i) = \begin{cases} 1 & \text{if } |y_i - \hat{p}| > 0{\cdot}5 \\ 0{\cdot}5 & \text{if } |y_i - \hat{p}| = 0{\cdot}5 \\ 0 & \text{if } |y_i - \hat{p}| < 0{\cdot}5. \end{cases}$$

(x) Entropy-based $R^2 (R^2_{\mathrm{E}})$
Using the entropy of the binomial distribution the distance measures become $D(y_i) = -(y_i \log \bar{p} + (1 - y_i) \log(1 - \bar{p}))$ and $D(y_i|x_i) = -(y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i))$, named 'deviance residuals', leading to $\sum_i D(y_i) = -\log L(0)$ and $\sum_i D(y_i|x_i) = -\log L(\hat{\beta})$, respectively. Here $L(\hat{\beta})$ and $L(0)$ denote the likelihoods of the fitted and of the null model without covariates. Thus R^2_{E} measures the reduction in maximized log-likelihood. The entropy measure has already been used by Theil[19] and is equivalent to McFadden's[20] pseudo $R^2 = 1 - \log L(\hat{\beta}) / \log L(0)$.

The measures R^2_{SS} and R^2_{G} are generalizations of Goodman and Kruskal's[10] τ. However, R^2_{SS} uses the observed distances while R^2_{G} uses corresponding distances expected under the model. This is easily demonstrated.

Generally, probability limits can be given if the parameters of the statistical model are known and the optimal predictor is used; they can be expressed by $R^2 = 1 - \overline{f(p)}/f(\bar{p})$ with $f(p)$ as follows:

$$R^2_{\mathrm{SS}}, R^2_{\mathrm{G}} \text{ and } r^2\text{: } f(p) = p(1 - p)$$

$$R^2_{\mathrm{E}}\text{: } f(p) = -p \log p - (1 - p) \log(1 - p)$$

$$R^2_{\mathrm{CER}}\text{: } f(p) = \min(p, 1 - p).$$

More discussion of the use of these expected distances is found in van Houwelingen and le Cessie.[17] Thus, R^2_{SS}, R^2_G and r^2 relate the variance of the prediction to the variance of the dependent variable y in logistic regression, relatively low variance in prediction resulting in little variation explained by a model.

It will also be demonstrated in Section 3 that values of R^2_G and the theoretically more robust R^2_{SS} and r^2 are generally very close.

The classification error R^2_{CER} depends crucially on the chosen cutpoint and on $\bar{p}$, therefore we do not consider it to be a useful measure of explained variation.

The measures R^2_G and R^2_E are reported in the LOGLINEAR routine of SPSS.[21]

2.3. R^2-measures based on model likelihoods

We consider two of these:

(xi) Likelihood-Ratio $R^2(R^2_{LR})$

The measure $R^2_{LR} = 1 - [L(0)/L(\hat{\beta})]^{2/n}$ when applied to the general linear model is identical to the standard multiple R^2 but can also be applied to generalized linear models, such as the logistic model.[22,23] The term $T = [L(\hat{\beta})/L(0)]^{2/n}$ is the geometric mean improvement per observation produced by fitting the more elaborate model and using $R^2_{LR} = 1 - T^{-1}$ (see Cox and Snell,[4] p. 208–209). With a logistic model, however, R^2_{LR} cannot attain a value of one even if the model predicts perfectly. With $p_1 = 0$ or $p_1 = 1$ the model fits perfectly and residuals become zero but $R^2_{LR} = 0{\cdot}75$.

(xii) Likelihood-Ratio R^2 modified (R^2_{CU})

Since R^2_{LR} cannot attain a value of one, it has been suggested by Cragg and Uhler[24] and later by Nagelkerke[25] to use $R^2_{CU} = R^2_{LR}/U$ where $U = 1 - [L(0)]^{2/n}$ is the maximum value that can be obtained by R^2_{LR}. Although the measure R^2_{CU} can reach a value of one the correction appears cosmetic as it can only force R^2_{CU} to 100 per cent for complete agreement and there is no indication why the scaling of the intermediate values of R^2_{CU} should be adequate.

R^2_{LR} and R^2_{CU} do not have good interpretability in terms of the p_i, in particular if compared with R^2_{SS}, R^2_G or r^2, and it appears that deeper understanding of more adequate likelihood based measures is still missing.

We have already mentioned McFadden's[20] pseudo R^2 which also belongs to this section and is identical to the entropy measure of (x).

The measures R^2_{LR}, R^2_{CU} and R^2_E are provided by the program GAUSS[26] and R^2_{LR}, R^2_{CU} are new features in SAS[12] 6·10.

3. EMPIRICAL COMPARISONS OF MEASURES

The performance of the measures described in the previous section was explored and compared under various conditions with a data set of size $n = 50{,}000$ generated according to a logistic model with a single continuous or dichotomous covariate; all measures were calculated for each data set. We did not investigate multiple covariates; we believe that the information from them carried by the predictor can be reasonably represented by the extremes of a single dichotomous or continuous covariate x.

With dichotomous x, values of 0 and 1 were perfectly balanced while with continuous x, a systematic sample from a uniform distribution (0, 1) was taken. In both cases values of the dependent variable y, 1 and 0, were generated with probabilities $P = \exp(b_0 + b_1x)/(1 + \exp(b_0 + b_1x))$ and $1 - P$, respectively, using the random number generator G05CAF of

Table I. Estimated percentage of explained variation for generated data sets according to various measures

R^2_{GLM}	r^2	R^2_{SS}	R^2_G	R^2_{CER}	R^2_E	r^2_s	τ^2_a	τ^2_b	$D^2_{\hat{p}y}$	γ^2	R^2_{LR}	R^2_{CU}
Continuous covariate												
0	0	0	0	0	0	0	0	0	0	0	0	0
12	12	12	12	30	9	12	4	8	16	16	12	16
—	25	25	25	25	21	24	7	16	36	36	23	32
—	50	50	50	51	46	44	12	29	70	70	43	61
—	75	75	75	75	72	57	16	38	92	92	58	83
—	100	100	100	100	100	77	25	54	100	100	75	100
Dichotomous covariate												
0	0	0	0	0	0	0	0	0	0	0	0	0
18	18	18	18	43	14	18	5	18	18	52	17	23
25	25	25	25	50	19	25	6	25	25	64	23	31
36	36	36	36	60	28	36	9	36	36	78	32	42
—	50	50	50	71	40	50	12	50	50	89	42	57
—	75	75	75	87	65	75	19	75	75	98	59	79
—	100	100	100	100	100	100	25	100	100	100	75	100
Continuous covariate, misspecified model												
—	24	24	25	26	20	25	8	16	35	35	24	32
—	55	55	55	60	49	51	16	34	72	72	47	65

R^2_{GLM} is only given for those experimental conditions for which analysis by a general linear model is appropriate. The symbols for the measures are those in Section 2.

NAG.[27] Values of the parameters b_1 and $b_0 = -b_1/2$ were chosen in such a way that the whole range of possible values for explained variation would be suitably covered and $\bar{p}$ in the underlying population is always 0·5. In a preliminary investigation the measure $r^2(y, \hat{p})$ had performed very well and therefore samples were generated with approximate $R^2 = r^2(y, \hat{p})$ of 0·0, 0·25, 0·5, 0·75 and 1·0.

As comparisons of the investigated measures with the well established multiple R^2 of the general linear model were of interest, further data sets were generated for which analysis by the logistic and by the general linear model would be equally adequate.

For these simulations lower and upper limits for the explanatory variable x had to be set such that $0{\cdot}2 < E(\hat{p}) < 0{\cdot}8$ while b_1 was fixed at a constant value. It can be shown[2] that the logistic function is almost linear within these constraints and therefore analyses by both the linear and the logistic model are adequate for such data. It is obvious that by severely restricting the range for $\hat{p}$ and consequently for x, only small to moderate values of R^2 are obtainable. Thus results for data sets with continuous x and $R^2 = 0{\cdot}12$ and for dichotomous x and $R^2 = 0{\cdot}18$ and 0·36 are reported, where 0·12 and 0·36, respectively, are the maximum R^2 obtainable under circumstances where a linear model could be fitted to samples with dichotomous y.

Results are found in Table I. The desirable property of fair agreement with the standard R^2 of the general linear model, R^2_{GLM}, is only observed with r^2, r^2_s, R^2_G and R^2_{SS}. It can be seen that for high values of explained variation r^2, R^2_{SS} and also R^2_G continue to lead to identical results while with continuous x and r^2 increasing r^2_s, τ^2_a and τ^2_b show increasingly lower values than r^2. For perfect explained variation according to r^2 and R^2_{SS}, that is y and $\hat{p}$ almost identical, r^2_s, τ^2_b and τ^2_a give values of only 0·77, 0·54 and 0·25, respectively. This amazing property of these non-parametric measures of association is caused by forcing different ranks on the $\hat{p}$ which for high R^2 are almost identical in the tail areas of the logistic model. As discussed in Section 2.1, τ-measures of association tend to give lower values than Pearson or Spearman correlations and already therefore would not be suitable if comparability with R^2_{GLM} is considered an important criterion. From Table I it is particularly hard to understand why the standard logistic regression program of SAS,[12] PROC LOGISTIC, reports τ_a. Somers' $D_{\hat{p}y}$ squared and the Goodman–Kruskal γ-squared give much higher values than r^2 for continuous x. The entropy-based R^2_{E} and the classification-error-based R^2_{CER} are different in construction and thus cannot be assumed to agree with R^2_{GLM}; in Table I R^2_{CER} and R^2_{E} tend to be greater and smaller than R^2_{GLM}, respectively. We also direct attention to the R^2_{CER} value of 30 in the second row of Table I. The restriction on the range of the p_i's for the comparative simulations with R^2_{GLM}, $0{\cdot}2 < E(\hat{p}) < 0{\cdot}8$, 'artificially' increased the value of R^2_{CER}, which then, for higher values in all other measures, declined to $R^2_{\mathrm{CER}} = 25$ in the following row. This clearly is an undesirable property of R^2_{CER}. The measure based on model likelihoods, R^2_{LR}, is too low and cannot even obtain a value of 1 with 100 per cent predictability. This has already been criticized in Section 2.3 where the appropriateness of the modified measure R^2_{CU} was also questioned.

The results of those measures for which probability limits have been given in Section 2 can be obtained without simulation. For instance, the choice of $b_1 = 3{\cdot}53$ ($5{\cdot}32$) used for the simulation results given in lines 11 and 12 of Table I, assuming $\bar{p} = 0{\cdot}5$ leads to response probabilities of $p_1 = 0{\cdot}146$ ($0{\cdot}065$) and $p_2 = 0{\cdot}854$ ($0{\cdot}935$) for the levels of a dichotomous covariate. From this follows $r^2 = 0{\cdot}50$ ($0{\cdot}76$), $R^2_{\mathrm{E}} = 0{\cdot}40$ ($0{\cdot}65$) and $R^2_{\mathrm{CER}} = 0{\cdot}71$ ($0{\cdot}87$). The simulation study, however, is still needed for comparisons with rank-based measures.

We further explored the performance of the reviewed measures under more special conditions in a similar way to that previously described for the empirical study. The small sample performance was checked for $n = 100$ based on 100 repetitions of the simulations. The effect of unbalanced dichotomous x (1 : 9) and of a skew distribution of x (exponential with $\lambda = 1$) was also investigated. Finally, the effect of misspecification in a logistic regression analysis was explored with y generated from continuous x using $P(y = 1|x) = \exp(b_0 + b_1x^2))/(1 + \exp(b_0 + b_1x^2))$. The analysis, however, assumed the usual model specified at the beginning of this section.

The message from these further experiments was simple. First, the small sample estimates are practically unbiased. Secondly, the differences between r^2 and the rank-based measures or R^2_{LR} can increase with unbalanced dichotomous x. In particular γ^2 gave a value of $0{\cdot}88$ when r^2 was $0{\cdot}25$ and τ^2_a achieved only $0{\cdot}03$ for $r^2 = 1{\cdot}00$. With unbalanced dichotomous x the measure R^2_{LR} decreased from $0{\cdot}75$ in Table I to $0{\cdot}48$. Finally, the measure R^2_{G} that was assumed sensitive to misspecification of the model failed to be so in our study and might be much more robust than expected.

4. ADAPTATION OF MEASURES OF EXPLAINED VARIATION FOR USE IN SMALL SAMPLES

When the number of covariates k in a general linear model is 'large' relative to a given sample size n, two variants of R^2 are often used. If inference is towards an underlying population then R^2-adjusted,

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}/(n-k-1)}{\mathrm{SST}/(n-1)} \quad \text{(see Myers,}^{28}\text{ p. 166),}$$

has the desirable property that $E(R^2_{\mathrm{adj}}) = 0$ for $R^2 = 0$ in the underlying population; this follows from equations (2) and (5) of Crocker.[29] Thus the criticized property of inflation of R^2 in small samples can be avoided. Though a thorough investigation of R^2_{adj} in the context of logistic regression ($R^2_{\mathrm{SS,adj}}$) is missing (in particular on the most appropriate correction of degrees of

Table II. Estimates of R^2_{SS}, $R^2_{\mathrm{SS,adj}}$, R^2_{E} and $R^2_{\mathrm{E,adj}}$ for underlying explained variation of $R^2_{\mathrm{SS}} = 0{\cdot}50$

Sample size	Number of covariates (k)											
	1				5				10			
	R^2_{SS}	$R^2_{\mathrm{SS,adj}}$	R^2_{E}	$R^2_{\mathrm{E,adj}}$	R^2_{SS}	$R^2_{\mathrm{SS,adj}}$	R^2_{E}	$R^2_{\mathrm{E,adj}}$	R^2_{SS}	$R^2_{\mathrm{SS,adj}}$	R^2_{E}	$R^2_{\mathrm{E,adj}}$
50	0·52	0·51	0·42	0·40	0·57	0·53	0·52	0·44	0·67	0·59	0·62	0·47
100	0·51	0·51	0·41	0·40	0·53	0·51	0·47	0·43	0·57	0·52	0·51	0·43
200	0·50	0·50	0·40	0·40	0·52	0·51	0·45	0·44	0·54	0·52	0·48	0·44
1000	0·50	0·50	0·40	0·40	0·51	0·50	0·44	0·44	0·51	0·51	0·45	0·44

All estimates are medians from 500 simulated trials.

Underlying explained variation of R^2_{E} as calculated from probability limits: 0·400, 0·434 and 0·438 for $k = 1$, 5 and 10, respectively

freedom for the error terms under heteroscedasticity), it is conjectured that the choice of $R^2_{\mathrm{SS,adj}}$ permits an analogous calculation of R^2_{adj} for logistic models.

This conjecture was supported by a simulation study comparing R^2_{SS} and $R^2_{\mathrm{SS,adj}}$ for samples of size $n = 50$, 100, 200 and 1000 with $k = 1$, 5 and 10 independent dichotomous covariates. Samples were generated as described for the study in the preceding section, but here with multiple covariates. The coefficients b_1 to b_k were identical and assumed values of 3·525, 2·268 and 1·662 for the simulations with $k = 1$, 5 and 10, respectively; they were obtained by setting the probability limit for R^2_{SS} to 0·50.

The results of Table II support our preference for $R^2_{\mathrm{SS,adj}}$ over R^2_{SS}, when the ratio k/n is 'high', though it appears that the discount in degrees of freedom for SSE in $R^2_{\mathrm{SS,adj}}$ is probably not large enough; this requires confirmation in larger simulation studies with more than a single setting of the parameter.

Also for R^2_{E} an 'adjusted' version has been suggested (H. van Houwelingen, Leiden, The Netherlands, personal communication):

$$R^2_{\mathrm{E,adj}} = 1 - \frac{\log L(\hat{\beta}) - (k+1)/2}{\log L(\beta_0) - 1/2}$$

which performs satisfactorily in Table II. The correction $(k + 1)/2$ is motivated by the χ^2_{k+1}-distribution of $2[\log L(\hat{\beta}) - \log L(\beta_0)]$ under H_0: $\beta_1, \ldots, \beta_k = 0$, where $\hat{\beta}$ is a $(k + 1)$-dimensional vector of maximum likelihood estimates and β_0 is the true intercept parameter. Therefore $E[\log L(\hat{\beta}) - \log L(\beta_0)] = (k + 1)/2$ and the expected optimism of $\log L(\hat{\beta})$ is $(k + 1)/2$.

No results are available for comparable adjustments of other measures of explained variation. If adjusted R^2-measures are used for a model obtained in a stepwise mode, then it is important to take k as the number of candidate factors rather than the number of factors included in the final model.[30,31]

If inference is towards a future sample then R^2-prediction, $R^2_{\mathrm{SS,pre}} = 1 - (\mathrm{SSE}_{\mathrm{PRESS}}/\mathrm{SST})$ (see Myers,[28] p. 171), is the suitable variant of R^2. In general linear models, $\mathrm{SSE}_{\mathrm{PRESS}}$ denotes the sum of squared Press[32] residuals for the fitted model. For a logistic model respective residuals could be obtained by cross-validation; in the distance function $D(y_i|x_i)$ of Section 2, $\hat{p}_i$ is then replaced by $\hat{p}_{i(-i)}$, which denotes a prediction obtained from a model fitted without the ith sampling unit. This approach mimics predicting the outcome for future individuals, to whose data a regression model cannot be fitted. Van Houwelingen and le Cessie[17] deal with cross-validation in logistic regression, in particular focusing on classification error and entropy.

Finally, $R^2_{\mathrm{E,pred}}$ is similar to $R^2_{\mathrm{E,adj}}$ but with the corrections of $R^2_{\mathrm{E,adj}}$ doubled; these are known as 'Akaike's corrections' (see Stone[33]).

5. CONCLUDING JUDGEMENTS OF MEASURES AND FURTHER REMARKS

To make a judgement on these measures we need criteria that a suitable measure should obey. We think that a 'good' measure of explained variation with logistic regression should possess the following properties (see also Kvalseth[6]):

1. intuitively clear interpretation;
2. consistency with the basic character of logistic regression, that is, non-linear monotonic transformations of explanatory variables should affect the measure, linear transformations should not;
3. the potential range of values of a measure should be [0, 1] with the endpoints corresponding to complete lack of predictability and perfect predictability, respectively;
4. if a data set could alternatively be analysed by a linear model, the resulting values of explained variation should be similar for both approaches. We term this property 'numerical consistency with classical multiple R^2'.

The compliance of the measures with the given criteria is shown in Table III. We think that the intuitive interpretability of the measures R^2_{E}, R^2_{LR} and R^2_{CU} is inferior to that of all other measures and that all nonparametric measures of correlation are not in agreement with the basic character of logistic regression. The R^2_{CER} measure has further unsatisfactory properties as mentioned in Sections 2 and 3. Numerical consistency with R^2_{GLM} where the general linear model could be applied was only observed for r^2, R^2_{SS}, R^2_{G} and r^2_{s}, while it was demonstrated that the measures r^2_{s}, τ^2_{a}, τ^2_{b} and R^2_{LR} do not always attain a value of one even when the outcome is completely determined by the predictor.

It has been argued[34] that contrary to its use in the general linear model the R^2_{SS} of the logistic model is not optimized by the fitting process. Therefore some statisticians prefer likelihood-based measures because of their agreement with the maximum likelihood fitting procedure. Hosmer and Lemeshow[1] (p. 18) point out that the estimators in logistic regression can also be regarded as 'least-squares' estimators. The widely used BMDP-LR program,[35] for example, estimates the regression parameters via an iteratively reweighted Gauss–Newton least-squares fit of $\hat{p}$ to y, minimizing $\sum(y_i - \hat{p}_i)^2 w_i^{-1}$, where $w_i = \hat{p}_i(1 - \hat{p}_i)$. While this approach is optimal in the sense of efficiency, the very unequal weighting of residuals and thus of individuals harms the intuitive appeal of a related measure of explained variation. Also, Willett and Singer[36] argue against the use of weighted residuals and in favour of unweighted residuals for R^2-measures in weighted least-squares regression.

Another argument against R^2_{SS} when applied to logistic regression could be the lack of normality and homoscedasticity of residuals. However, in our understanding current likelihood-based measures do violate some of the desirable properties of an explained variation measure as summarized by Table III and therefore our personal preference is for R^2_{SS} or r^2. As a referee pointed out, some statisticians might prefer to cite the observed correlation r rather than use the 'proportion of explained variation interpretation' of r^2.

In this paper we have discussed explained variation mainly as a summary measure for a multiple logistic model. However, the proportion of variation explained by individual factors in

Table III. Summary of properties of measures of explained variation

Measures	Consistent with character of logistic regression	Intuitive interpretation	Range [0–1]	Numerical consistency with R^2_{GLM}
r^2	yes	yes	yes	yes
R^2_{SS}	yes	yes	yes	yes
R^2_{G}	yes	yes	yes	yes
R^2_{CER}	no	yes	yes	no
R^2_{E}	yes	no	yes	no
r^2_s	no	yes	no	yes
τ^2_a	no	yes	no	no
τ^2_b	no	yes	no	no
$D^2_{\hat{p}y}$	no	yes	yes	no
γ^2	no	yes	yes	no
R^2_{LR}	yes	no	no	no
R^2_{CU}	yes	no	yes	no

a marginal or partial sense and comparisons of the variation explained by different factors (their relative importance) may also be of medical interest. Respective analyses can be performed proceeding in a similar way as recently presented[37] for the Cox model.

In conclusion, we recommend routine evaluation of the proportion of explained variation when prognostic factor studies or dose-response studies are analysed by logistic regression. In any of these applications medical investigators are easily misled by highly significant p-values or impressive relative risk estimates for explanatory factors, while individual results are far from being determined. Quantification of the respective understanding of physiological processes is therefore particularly valuable for medical progress.

ACKNOWLEDGEMENTS

The authors appreciate several very useful suggestions by Hans van Houwelingen, University of Leiden, The Netherlands and by a referee.

REFERENCES

1. Hosmer, D. W. Jr. and Lemeshow, S. *Applied Logistic Regression*, Wiley, New York, 1989.
2. Cox, D. R. and Wermuth, N. 'A comment on the coefficient of determination for binary responses', *American Statistician*, **46**, 1–4 (1992).
3. Korn, E. L. and Simon, R. 'Explained residual variation, explained risk and goodness of fit', *American Statistician*, **45**, 201–206 (1991).
4. Cox, D. R. and Snell, E. J. *Analysis of Binary Data*, Chapman and Hall, London, 1989.
5. Van Houwelingen, J. C. and le Cessie, S. 'Logistic regression, a review', *Statistica Neerlandica*, **42**, 215–232 (1988).
6. Kvalseth, T. O. 'Cautionary note about R^2', *American Statistician*, **39**, 279–285 (1985).
7. Kendall, M. and Gibbons, J. D. *Rank Correlation Methods*, Edward Arnold, London, 1990.
8. Stuart, A. 'Kendall's Tau', in *Encyclopedia of Statistical Sciences*, Vol. 4, Wiley, New York, 1989, pp. 367–369.
9. Somers, R. H. 'A new asymmetric measure of association for ordinal variables', *American Sociological Review*, **27**, 799–811 (1962).
10. Goodman, L. A. and Kruskal, W. H. 'Measures of association for cross classifications', *Journal of the American Statistical Association*, **49**, 732–764 (1954).

11. Korn, E. L. and Simon, R. 'Measures of explained variation for survival data', *Statistics in Medicine*, **9**, 487–503 (1990).
12. 'The LOGISTIC Procedure' in *SAS/STAT User's Guide*, Version 6, 4th Edition, 1990, pp. 1071–1126.
13. Efron, B. 'Regression and ANOVA with zero-one data: measures of residual variation', *Journal of the American Statistical Association*, **73**, 113–121 (1978).
14. Agresti, A. 'Applying R^2-type measures to ordered categorical data', *Technometrics*, **28**, 133–138 (1986).
15. Margolin, B. H. and Light, R. J. 'An analysis of variance for categorical data, II: Small sample comparisons with chi square and other competitors', *Journal of the American Statistical Association*, **69**, 755–764 (1974).
16. Haberman, S. J. 'Analysis of dispersion of multinomial responses', *Journal of the American Statistical Association*, **77**, 568–580 (1982).
17. Van Houwelingen, J. C. and le Cessie, S. 'Predictive value of statistical models', *Statistics in Medicine*, **9**, 1303–1325 (1990).
18. Makuch, R. W., Rosenberg, P. S. and Scott, G. 'Goodman and Kruskal's λ: A new look at an old measure of association', *Statistics in Medicine*, **8**, 619–631 (1989).
19. Theil, H. 'On the estimation of relationships involving qualitative variables', *American Journal of Sociology*, **76**, 103–154 (1970).
20. McFadden, D. 'The measurement of urban travel demand', *Journal of Public Economics*, **3**, 303–328 (1974).
21. SPSS^x. *User's Guide*, 3rd Edition, SPSS Inc., Chicago, 1988.
22. Maddala, G. S. *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, 1983.
23. Magee, L. 'R^2-measures based on Wald and likelihood ratio joint significance tests', *American Statistician*, **44**, 250–253 (1990).
24. Cragg, J. G. and Uhler, R. 'The demand for automobiles', *Canadian Journal of Economics*, **3**, 386–406 (1970).
25. Nagelkerke, N. J. D. 'A note on a general definition of the coefficient of determination', *Biometrika*, **78**, 691–692 (1991).
26. GAUSS^{TM}. *Applications Manual*, Aptech Systems Inc., Maple Valley, U.S.A., 1988.
27. NAG. *Nag Fortran Library Manual-Mark 15.* Numerical Algorithms Group Ltd., Oxford, 1993.
28. Myers, R. H. *Classical and Modern Regression with Applications*, PWS-Kent, Boston, 1990.
29. Crocker, D. C. 'Some interpretations of the multiple correlation coefficient', *American Statistician*, **26**, 31–33 (1972).
30. Rencher, A. C. and Pun, F. C. 'Inflation of R^2 in best subset regression', *Technometrics*, **22**, 49–53 (1980).
31. Freedman, D. A. 'A note on screening regression equations', *American Statistician*, **37**, 152–155 (1983).
32. Allen, D. M. 'The relation between variable selection and data augmentation and a method for prediction', *Technometrics*, **16**, 125–127 (1974).
33. Stone, M. 'An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion', *Journal of the Royal Statistical Society, Series B*, **39**, 44–47 (1977).
34. Agresti, A. *Categorical Data Analysis*, Wiley, New York, 1990, p. 112.
35. Dixon, W. J. (ed.) *BMDP Statistical Software Manual*, Volume 2, University of California Press, Berkeley, 1990, p. 1304.
36. Willet, J. B. and Singer, J. D. 'Another cautionary note about R^2: Its use in weighted least-squares regression analysis', *American Statistician*, **42**, 236–238 (1988).
37. Schemper, M. 'The relative importance of prognostic factors in studies of survival', *Statistics in Medicine*, **12**, 2377–2382 (1993).