

# Computing measures of explained variation for logistic regression models

Martina Mittlböck \*, Michael Schemper

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

Received 7 January 1998; accepted 15 April 1998

---

## Abstract

The proportion of explained variation ( $R^2$ ) is frequently used in the general linear model but in logistic regression no standard definition of  $R^2$  exists. We present a SAS macro which calculates two  $R^2$ -measures based on Pearson and on deviance residuals for logistic regression. Also, adjusted versions for both measures are given, which should prevent the inflation of  $R^2$  in small samples. © 1999 Elsevier Science Ireland Ltd. All rights reserved.

*Keywords:* Logistic regression; Explained variation; Sums-of-squares; Entropy; Adjusted  $R^2$ ; SAS-macro

---

## 1. Introduction

In classical regression analysis, the coefficient of determination  $R^2$  is routinely calculated by nearly every statistical software package.  $R^2$  has the desired interpretability as the proportion of variation of the dependent variable, which can be explained by the predictor variables of a given regression model.

For logistic regression, probably the most frequently used regression model following the general linear model, many different proposals have

been made to measure explained variation [1]. The PROC LOGISTIC of SAS [2] gives two so called  $R^2$ -measures, but these measures have major disadvantages so that their values can not be interpreted in a useful way. For the two mostly recommended  $R^2$ -measures, connected with the concepts of Pearson and deviance residuals, we have therefore written a SAS-Macro for routine use. The SAS-Macro also gives adjusted expressions for these measures, which can be used and interpreted analogous to adjusted- $R^2$  in a general linear model.

In Section 2, the two measures and their corresponding adjusted versions are described and in Section 3 two examples are given on how to use

---

\* Corresponding author. Tel.: + 43 1 404002276; fax: + 43 1 404002278; e-mail: Martina.Mittlboeck@AKH-Wien.AC.AT

the SAS-macro and how to interpret the results of the output. The program code of the SAS-macro is given in the appendix.

## 2. Description of measures

Given a sample of observations  $(y_i, x_i)$ ,  $i = 1, \dots, n$ ,  $y_i \in \{0, 1\}$  denotes the dependent variable and  $x_i$  is the corresponding covariate vector. The estimates from a logistic regression are  $\widehat{\text{Prob}}(y_i = 1|x_i) = \hat{p}_i = \exp(\hat{\beta}x_i)/(1 + \exp(\hat{\beta}x_i))$ , where  $\hat{\beta}$  is the estimated parameter vector. Furthermore  $\widehat{\text{Prob}}(y_{i=1}) = \bar{p} = \sum_i y_i/n$  is the proportion of one's in the sample.

The general form of  $R_2$ -measures of the proportion of explained variation (PEV) is

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

where  $D(y_i)$  and  $D(y_i|x_i)$  denote a measure of the distance of  $y_i$  from an unconditional and conditional (on a covariate vector  $x_i$ ) central location parameter, respectively.

Different specifications of  $D(y_i)$  and  $D(y_i|x_i)$  result in different  $R^2$ -measures. Two of them are recommended for logistic regression [1] and are defined as follows:

### 2.1. Sums-of-squares $R_2$ ( $R_{SS}^2$ )

For this measure  $D(y_i) = (y_i - \bar{p})^2$  and  $D(y_i|x_i) = (y_i - \hat{p}_i)^2$  denote the squared distance between observed ( $y_i$ ) and predicted ( $\bar{p}$  and  $\hat{p}_i$ ) outcomes under the null model (only with intercept) and under the full model (with covariates), respectively [3]:

$$R_{SS}^2 = 1 - \frac{\sum_i (y_i - \hat{p}_i)^2}{\sum_i (y_i - \bar{p})^2} = \frac{\sum_i (y_i - \hat{p}_i)^2}{n\bar{p}(1 - \bar{p})}$$

When the number of covariates  $k$  in a general linear model is large relative to a given sample size, then  $R_2$ -adjusted avoids the criticised property of inflation of  $R_2$  in small samples and  $E(R_{\text{adj}}^2) = 0$  for  $R_2 = 0$ . Although a thorough investigation of adjusted- $R_2$  in the context of logis-

tic regression is still missing up to now, Mittlböck and Schemper [1] suggest to calculate  $R_{SS,\text{adj}}^2$  analogous to the general linear model, which is preferable in any case to  $R_{SS}^2$  with small samples:

$$R_{SS,\text{adj}}^2 = 1 - \frac{\left[ \sum_i (y_i - \hat{p}_i)^2 \right] / (n - k - 1)}{[n\bar{p}(1 - \bar{p})] / (n - 1)}$$

### 2.2. Entropy-based $R^2$ ( $R_E^2$ )

The distance measure, based on deviance residuals, uses the entropy of the binomial distribution so that

$$D(y_i) = -[y_i \log \bar{p} + (1 - y_i) \log(1 - \bar{p})] \quad \text{and}$$

$$D(y_i|x_i) = -[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)],$$

leading to  $\sum_i D(y_i) = -\log L(\hat{\beta}_0)$  and

$$\sum_i D(y_i|x_i) = -\log L(\hat{\beta}),$$

respectively:

$$R_E^2 = 1 - \frac{\log L(\hat{\beta})}{\log L(\hat{\beta}_0)}$$

$L(\hat{\beta})$  and  $L(\hat{\beta}_0)$  denote the likelihoods of the fitted model with covariates and of the null model without covariates. Thus,  $R_2$  measures the reduction in maximized log-likelihood [4].

H. van Houwelingen (Leiden, The Netherlands, personal communication, see [1]) also suggested an adjusted version of  $R_2$ , which performs satisfactorily:

$$R_{E,\text{adj}}^2 = 1 - \frac{\log L(\hat{\beta}) - (k + 1)/2}{\log L(\hat{\beta}_0) - 1/2}$$

The corrections  $(k + 1)/2$  and  $1/2$  are the expected optimism of  $\log L(\hat{\beta})$  and of  $\log L(\hat{\beta}_0)$  under  $H_0: \beta_1 = \dots = \beta_k = 0$ , respectively. If these corrections are doubled, then the expression is identical to the Akaike information.

Two  $R_2$ -measures can be requested from SAS with the RSQ-option in the model statement of PROC LOGISTIC, but they do not have a useful interpretation. The measure RSquare [5] is defined

as  $R_{\text{RSquare}}^2 = 1 - [L(\hat{\beta}_0)/L(\hat{\beta})]^{2/n}$  and in logistic regression it can never reach a value of one even if the model predicts perfectly. Max-rescaled RSquare [6] is the measure RSquare corrected, so that the maximum of one can be achieved. This happens if RSquare is divided by the maximal achievable value in case of perfect prediction, which is  $1 - [L(\hat{\beta}_0)]^{2/n}$ . This seems to be a rather cosmetic correction as it forces Max-rescaled RSquare to 100% for complete agreement, however, there is no indication why the scaling of the intermediate values should be adequate. Max-rescaled RSquare nearly always gives higher values than both  $R_{\text{SS}}^2$  and  $R_E^2$ .

### 3. Program description and examples

A SAS macro has been written which gives the output of PROC LOGISTIC and unadjusted and adjusted measures of the proportion of explained variation of the model. It requires the existence of a SAS input data set containing the dependent variable, which must be coded with 0/1, and the independent covariates:

---

DATA:	name of the SAS-data set
TITLE:	title in the output listing
DEP:	name of the dependent variable (must be coded with 0/1)
VAR:	name of the independent variables, separated by blanks

---

The program code of the SAS-macro EVLOGIST is given in the appendix and it is also available via world wide web at <http://www.akh-wien.ac.at/imc/biometrie/evlogist.htm>

The following two examples serve to demonstrate the behaviour of the two measures with real data. They also help to elucidate the role of explained variation measures in addition to the standard description of logistic regression results by estimates of the relative risk associated with explanatory factors and by corresponding confidence intervals or  $p$ -values.

#### 3.1. Example 1: Physical characteristics of urine with and without crystals

This prognostic factor study [7] was conducted at Veteran's Administration Medical Center, Palo Alto and at the Stanford University School of Medicine, Stanford. From this study,  $n = 77$  completely documented urine specimens were analyzed to determine if certain physical characteristics of the urine might be related to the formation of calcium oxalate crystals. The six physical characteristics of the urine are: (1) specific gravity, the density of the urine relative to water; (2) pH, the negative logarithm of the hydrogen ion; (3) osmolarity (mOsm) is proportional to the concentration of molecules in solution; (4) conductivity (mMho) is proportional to the concentration of charged ions in solution; (5) urea concentration (UREA) in millimoles per litre; and (6) calcium concentration (CA) in millimoles/litre. Some of these characteristics are highly correlated. For computational reasons, specific gravity was multiplied by 1000 and all variables were standardized to a mean of zero.

To run the SAS program for this example, the following statements of the EVLOGIST-macro can be used:

```
%EVLOGIST (DATA=urine, TITLE=urine
example, DEP=crystals,
VAR=spgrav ph mosm mmho urea ca);
```

The results of an analysis by the logistic model are given in Fig. 1. Neither interactions nor quadratic or cubic effects could be detected.

From Fig. 1 we learn also that  $R_{\text{SS}}^2 = 0.52$  and  $R_E = 0.45$ . The observed explained variation of both measures of 0.45 and 0.52 is very high, as in our experience typical values of  $R_{\text{SS}}^2$  for such studies range between 0.15 and 0.45. Mittlböck and Schemper [1] have shown in a simulation study that  $R_E^2$  gives values smaller than  $R_{\text{SS}}^2$  in most cases. As it is well known for the general linear model that  $R_2$  values can be artificially inflated in samples where the ratio number of covariates to sample size is high, adjustments for  $R_2$ -measures should also be considered here. As we see from Fig. 1,  $R_{\text{SS,adj}}^2$  and  $R_{E,adj}^2$  are 4 and 6 percent points lower than the unadjusted measures, respectively. Nevertheless we learn from

## URINE EXAMPLE

## The LOGISTIC Procedure

Data Set: WORK.URINE  
 Response Variable: CRYSTALS  
 Response Levels: 2  
 Number of Observations: 77  
 Link Function: Logit

## Response Profile

Ordered Value	CRYSTALS	Count
1	1	33
2	0	44

## Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	107.168	71.560	.
SC	109.512	87.967	.
-2 LOG L Score	105.168	57.560	47.608 with 6 DF (p=0.0001)
			32.498 with 6 DF (p=0.0001)

RSquare = 0.4611

Max-rescaled RSquare = 0.6191

## Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	0.0686	0.3599	0.0364	0.8487	.	.
SPGRAV	1	0.3559	0.2221	2.5681	0.1090	1.435025	1.428
PH	1	-0.4957	0.5698	0.7569	0.3843	-0.197337	0.609
MOSM	1	0.0168	0.0178	0.8904	0.3454	2.219573	1.017
MMHO	1	-0.4328	0.2512	2.9679	0.0849	-1.909993	0.649
UREA	1	-0.0320	0.0161	3.9443	0.0470	-2.303055	0.968
CA	1	0.7837	0.2422	10.4717	0.0012	1.424500	2.190

(... output omitted ...)

NR OF OBSERVATIONS	NR OF COVARIATES WITHOUT INTERCEPT	SUMS OF SQUARES R2	ADJUSTED SUMS OF SQUARES R2	ENROPY-R2	ADJUSTED ENROPY-R2
77	6	0.51990	0.47875	0.45268	0.39191

Fig. 1. Output of the SAS-macro EVLOGIST from urine data example.

this example that nearly half of the variation in the occurrence of calcium oxalate crystals can be attributed to various physical characteristics of urine, an important message that is lost by just reporting the standard results of odds ratios and  $p$ -values.

### 3.2. Example 2: Dose-response example

The analysis of dose-response relationships is another frequent indication for the logistic model. In this example data [8] from 507 patients were analyzed which had received a drug with possible undesirable after-effects. It is thought that the chance of getting after-effects may depend on the dose level of the drug.

The statements for the SAS-macro are as follows:

```
%EVLOGIST (DATA=drug, TITLE=dose-
response example,
DEP=adverse, VAR=dose);
```

We can see from Fig. 2 that observed and fitted responses are close so that goodness-of-fit is satisfactory. Fig. 3 shows, that the relative risk, given the range of doses (0.9–4.0), is strong and highly significant. However the explained variation is

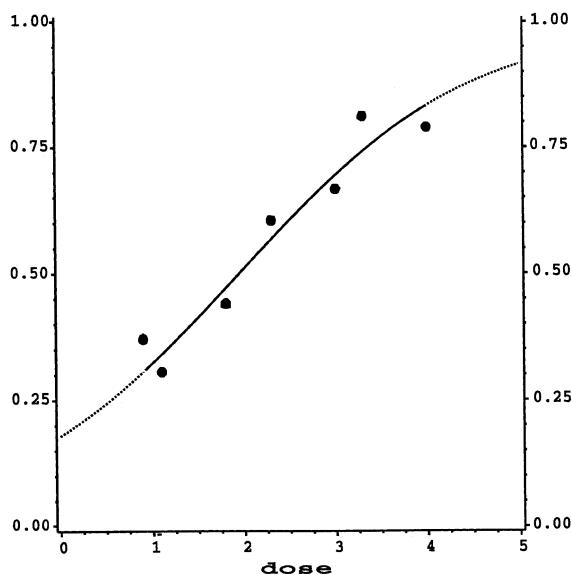


Fig. 2. Dose response relationship of example 2; observed and fitted responses are shown with dots and a line, respectively.

low with  $R_{SS}^2 = 0.11$  and  $R_E^2 = 0.08$ . These values are both very plausible as the individual result is far from being determined by the dose received. For such studies a value around 0.1 is not untypical, higher values of  $R_2$  being observable only for studies with more extreme doses investigated. As we investigate only one covariate and the sample size is large, there are only very small differences between adjusted and unadjusted values.

In any of these applications medical investigators are easily misled by highly significant  $p$ -values or impressing relative risk estimates for explanatory factors. This may discourage the search for further factors and therefore  $R_2$  should be evaluated on a routine basis.

## 4. Conclusions

Mittlböck and Schemper [1] preferred the use of  $R_{SS}^2$  to  $R_E^2$ , as  $R_{SS}^2$  is consistent with the results of  $R_2$  in the general linear model, when both models are applicable (if the fitted values ( $\hat{p}$ ) range between 0.2 and 0.8). Furthermore, the interpretation of  $R_{SS}^2$  is more intuitive to many people, as the use of squared residuals is a very basic idea in statistics whereas the use of the likelihood is not so clear in its interpretation. However, as the concept of entropy is also basic in logistic regression, both measures are routinely given by the macro.

The use of the adjusted- $R_2$  is recommended in any case but especially in case of small samples and/or many covariates, where the unadjusted  $R_2$  may give artificially inflated values. In our two examples, the adjusted measures only lead to moderate changes in  $R_2$ .

Summing up, we recommend to routinely evaluate explained variation in the analysis of factors possibly affecting a binary outcome. Impressive relative risk for explanatory factors and/or highly significant  $p$ -values may mislead investigators in quantifying the results, while individual results may be far from being determined. Quantification of the respective understanding of underlying processes is therefore valuable for any empirically based scientific progress.

## DOSE-RESPONSE EXAMPLE

## The LOGISTIC Procedure

Data Set: WORK.DRUG  
 Response Variable: ADVERSE  
 Response Levels: 2  
 Number of Observations: 507  
 Link Function: Logit

## Response Profile

Ordered Value	ADVERSE	Count
1	1	278
2	0	229

## Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	700.108	645.808	.
SC	704.337	654.265	.
-2 LOG L	698.108	641.808	56.300 with 1 DF (p=0.0001)
Score	.	.	53.874 with 1 DF (p=0.0001)

RSquare = 0.1051

Max-rescaled RSquare = 0.1406

## Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.5207	0.2566	35.1251	0.0001	.	.
DOSE	1	0.7806	0.1107	49.7667	0.0001	0.394471	2.183

(... output omitted ...)

NR OF OBSERVATIONS	NR OF COVARIATES WITHOUT INTERCEPT	SUMS OF SQUARES R2	ADJUSTED SUMS OF SQUARES R2	ENROPY-R2	ADJUSTED ENROPY-R2
507	1	0.10780	0.10604	0.080647	0.079101

Fig. 3. Output of the SAS-macro EVLOGIST from dose-response example.

## Appendix A. SAS-macro

---

```

*****;
*** COMPUTATION OF EXPLAINED VARIATION MEASURES:
***
*** SUMS-OF SQUARES R-SQUARE
*** SUMS-OF SQUARES R-SQUARE ADJUSTED
*** ENTROPY R-SQUARE
*** ENTROPY R-SQUARE ADJUSTED
***
*** FOR LOGISTIC REGRESSION
*****;

```

---

```

MACRO EVLOGIST (DATA=, DEP=,
TITLE=' ', VAR=);

```

```

%LET COUNT=1;
%LET WORD=%NRBQUOTE(%SCAN
(&VAR,&COUNT,%STR()));
%DO%WHILE(&WORD^=);
%LET COUNT=%EVAL(&COUNT+1);
%LET WORD=%NRBQUOTE(%SCAN
(&VAR,&COUNT,%STR()));
%END;
%LET COUNT=%EVAL(&COUNT-1);

```

```

TITLE ``&TITLE``;
PROC LOGISTIC DATA=&DATA
DESCENDING OUTEST=TEST;
MODEL &DEP=&VAR/RL MAXITER=
500 RSQ;
OUTPUT OUT=D1 P=P_I XBETA=
SCORE;
RUN;

```

```

DATA TEST; SET TEST; MM=1;
KEEP INTERCEP MM;
RUN;

```

```

PROC SORT DATA=D1; BY &DEP;
DATA D1; SET D1; BY &DEP; MM=1;
IF P_I=. OR &DEP=. THEN DELETE;
KEEP &DEP P_I MM SCORE;
RUN;

```

```

PROC FREQ DATA=D1 NOPRINT;
TABLES &DEP/OUT=F;
DATA F; SET F; MM=1;
P_ROH=PERCENT/100;
IF _N_=2;
WERT=&DEP;
KEEP MM P_ROH WERT;
RUN;

```

```

DATA D1; MERGE D1 F TEST; BY MM;
RUN;

```

```

DATA D1; SET D1;
IF Y=WERT THEN Y=1; ELSE Y=0;
ERR_E=(&DEP-P_I)**2;
SSE+ERR_E;
ERR_T=(&DEP-P_ROH)**2;
SST+ERR_T;
RESP+&DEP;
ENTR_E=(&DEP*LOG(P_I)
+(1-&DEP)*LOG(1-P_I));
SS_ENE+ENTR_E;
ENTR_T=(&DEP*LOG(P_ROH)
+(1-&DEP)*LOG(1-P_ROH));
SS_ENT+ENTR_T;
DROP WERT;
RUN;

```

```

DATA D1;
LABEL N='NR OF OBSERVATIONS'
K='NR OF COVARIATES WITHOUT
INTERCEPT'
R2_SS='SUMS OF SQUARES R2'
R2_SS_AD='ADJUSTED SUMS OF
SQUARES R2'
R2_EN='ENROPY-R2'
R2_EN_AD='ADJUSTED ENROPY-R2';
SET D1; BY MM; DROP MM;
N=_N_;

```

```

IF LAST.MM;
R2_SS=1-SSE/SST;
R2_EN=1-SS_ENE/SS_ENT; K=
&COUNT;
R2_SS_AD=1-(SSE/(N-K-1))/
(SST/(N-1));
R2_EN_AD=1-(SS_ENE-(K+1)/2)/
(SS_ENT-1/2);
CALL SYMPUT('NOBS',N);
RUN;

PROC PRINT LABEL NOOBS;
VAR N K R2_SS R2_SS_AD R2_EN
R2_EN_AD;

%MEND;

```

## References

- [1] M. Mittlböck, M. Schemper, Explained variation for logistic regression, *Stat. Med.* 15 (1996) 1987–1997.
- [2] SAS, SAS/STAT User's Guide, version 6. SAS Institute, Cary, 1990
- [3] B.H. Margolin, R.J. Light, An analysis of variance for categorical data II: Small sample comparisons with chi square and other competitors, *J. Am. Stat. Assoc.* 69 (1974) 755–764.
- [4] H. Theil, On the estimation of relationships involving qualitative variables, *Am. J. Sociol.* 76 (1970) 103–154.
- [5] G.S. Maddala, *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, 1983.
- [6] J.G. Cragg, R. Uhler, The demand for automobiles, *Can. J. Econ.* 3 (1970) 386–406.
- [7] D.F. Andrews, A.M. Herzberg, *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Springer, New York, 1985.
- [8] C. Chatfield, *Problem Solving. A statistician's guide*, Chapman and Hall, London, 1988.