

A note on R^2 measures for Poisson and logistic regression models when both models are applicable

Martina Mittlböck, Harald Heinzl*

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

Received 3 January 2000; received in revised form 23 May 2000; accepted 15 June 2000

Abstract

The aim of many epidemiological studies is the regression of a dichotomous outcome (e.g., death or affection by a certain disease) on prognostic covariables. Thereby the Poisson regression model is often used alternatively to the logistic regression model. Modelling the number of events and individual outcomes, respectively, both models lead to nearly the same results concerning the parameter estimates and their significances. However, when calculating the proportion of explained variation, quantified by an R^2 measure, a large difference between both models usually occurs. We illustrate this difference by an example and explain it with theoretical arguments. We conclude, the R^2 measure of the Poisson regression quantifies the predictability of event rates, but it is not adequate to quantify the predictability of the outcome of individual observations. © 2001 Elsevier Science Inc. All rights reserved.

Keywords: R^2 measure; Poisson regression; Logistic regression; Approximation; Deviance; Sums-of-squares

1. Introduction

The R^2 measure, also called coefficient of determination, is well established in classical linear regression analysis and also becomes more and more popular in generalized linear models, such as Poisson and logistic regression. R^2 is interpretable as the proportion of outcome variation, which can be explained by the predictor variables of a given regression model. It is mainly a measure for predictability of the dependent variable, which is the number of events in Poisson regression and the binary indicator for occurrence of an event in logistic regression. Quantifying predictability of the outcome measures the strength of a regression relationship, or more generally speaking, the amount of “knowledge” already attained in a specific research question. Therefore R^2 gives information additional to P-values and parameter estimates of prognostic factors. An R^2 measure of zero indicates complete lack of predictability of the outcome by the covariates fitted in the model, whereas a value of one indicates that the covariates can predict the outcome perfectly. An R^2 value of 0.4 would mean that 40% of the variation in the dependent variable can be explained by the covariates in the model whereas 60% are not explained and are due to chance. R^2 values increase if covariates are added

to a model, except when the parameter estimates are zero the R^2 value remains unchanged.

In many epidemiological studies an event of dichotomous nature (e.g., death from coronary artery disease or occurrence of cancer), is of interest. Such binomial data are often summarized for each distinct covariate pattern by event rates (number of persons affected relative to number of person-years lived) and analysed by regression models assuming an underlying Poisson distribution. Thus the Poisson regression, modelling the number of events, can be considered as an approximation to the logistic regression, for modelling dichotomous outcome. Both models can be used alternatively and lead to nearly the same results concerning the parameter estimates and their significances in case of a large sample size ($n \rightarrow \infty$) and rare events ($p \rightarrow 0$) [1,2]. A rule of thumb is that both distributions are approximately equivalent if $n \geq 20$ and $p \leq 0.05$ and very good if $n \geq 100$ and $p \leq 0.01$ (see [3,4] for an upper bound on the total error in the approximation). Accordingly we see from Fig. 1 that the logistic and the Poisson functions are approximately equivalent when the probability of an event (p) is small.

However, the R^2 measures of the logistic and Poisson regression may differ substantially. A goodness-of-fit test in logistic regression [5] evaluates how well the model can predict the observed probabilities. But to measure how well the individual outcome, usually coded as one for events and zero otherwise, can be predicted by a model, an R^2 -type measure has to be used (see [6]). In Fig. 2 an example from

* Corresponding author. Tel.: +43 1 40400 6686; fax: +43 1 40400 6687.

E-mail address: harald.heinzl@akh-wien.ac.at (Harald Heinzl)

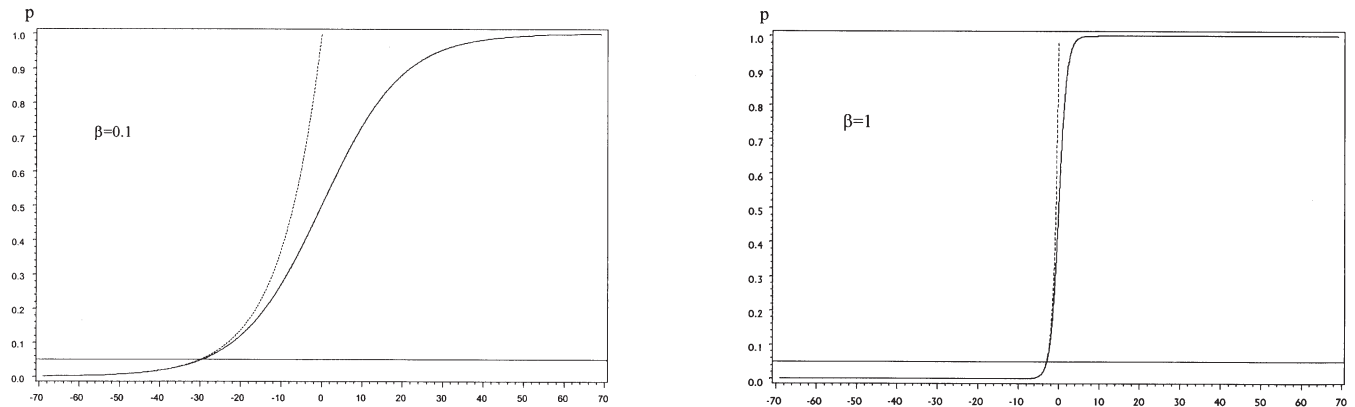


Fig. 1. Logistic function (solid line) and Poisson function (dotted line) with a single continuous covariate with range $(-70, 70)$ and parameters $\beta = 0.1$ and $\beta = 1$, respectively. The horizontal line at $p = 0.05$ denotes the maximum of p , where an approximation of the logistic regression by a Poisson regression is good.

a dose–response study shows that the probability of binary responses can be fitted very well for the various dose levels, and the goodness-of-fit is nearly perfect. But it is still difficult to predict which patient will respond as the covariate dose is not very effective in distinguishing outcomes [7]. Viewed on an individual-by-individual basis, the observations are all zeros and ones while for the various dose levels the predictions are between 0.30 and 0.82 in Fig. 2. Note, the closer the probability of an event is around 0.5 the more inaccurate is the prediction of an individual event.

The basic idea of Poisson regression is to estimate frequencies (number of responses) and not individual outcomes. This is similar to the estimation of probabilities for

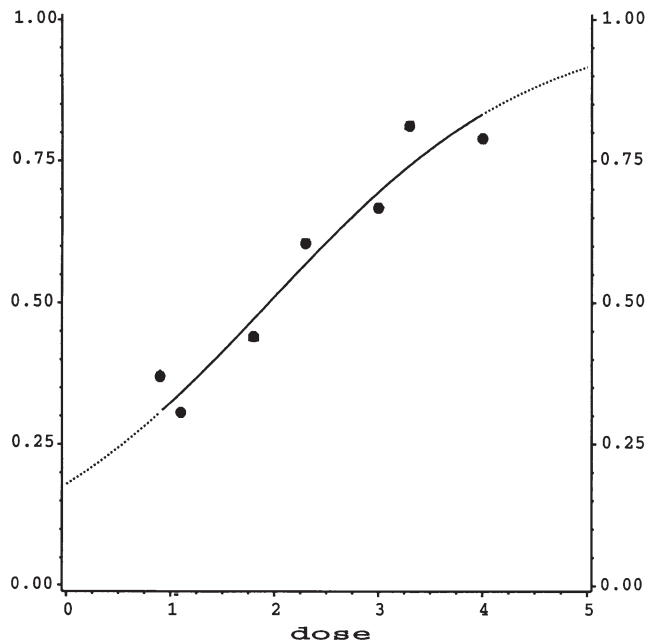


Fig. 2. Observed (dots) and by logistic regression estimated (line) probabilities of adverse events of a dose–response example.

binary events. Therefore there is no large difference between goodness-of-fit and R^2 measures in Poisson models [8, 9]. However, if a typical binomial situation with binary outcomes is fitted by Poisson regression, which is usually done in epidemiological studies for event rates, then model fit, parameter estimates and significances will be equivalent to the fit of a logistic regression, but the results of R^2 measures may be completely different and may also have different interpretations.

In the following section we describe R^2 measures for logistic and Poisson regression based on the sums-of-squares and on entropy concepts. Differences between R^2 measures of both models are exemplified using real data, and thereby stressing the differences in the interpretation of R^2 on the one hand and P-values and parameter estimates on the other hand.

2. Materials and methods

2.1. Data

We use data of a follow-up study [10,11], which investigates whether the covariates age and smoking can predict death from coronary artery disease of British male doctors. Age is given in decades, smoking is a dichotomous covariate which categorizes the study participants into smokers and non-smokers (Table 1). The numbers of death from coronary artery disease and person-years lived are given for each distinct covariate pattern formed by age-decades and smoking status.

2.2. Statistical methods

The data are fitted by Poisson and logistic regression, so that the effect of age and smoking on death of coronary artery disease can be described by parameter estimates and P-values. The question, how good the individual outcomes can be predicted by these two factors is of further interest.

Table 1
Data from the study about death from coronary artery disease among British doctors [10]

Age	Smoke	Number of deaths	Person-years
40	0	2	18,790
50	0	12	10,673
60	0	28	5,710
70	0	28	2,585
80	0	31	1,462
40	1	32	52,407
50	1	104	43,248
60	1	206	28,612
70	1	186	12,663
80	1	102	5,317

Usually, this is answered by computing R^2 , which is also called the coefficient of determination or the proportion of explained variation.

The general form of R^2 measures is

$$R^2 = 1 - \frac{\sum_i D(y_i|x_i)}{\sum_i D(y_i)}$$

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

R^2 measures based on deviances make use of the log-likelihood of the saturated model ($\log L(y)$), when observed and estimated outcomes coincide, the log-likelihood of the full model ($\log L(\hat{\beta})$), when the effects of covariates ($\hat{\beta}$) are estimated, and the log-likelihood of the null model ($\log L(\bar{y})$), when only an intercept parameter ($\hat{\beta}_0$) is fitted. The conditional distance measure is the difference between the log-likelihoods of the saturated and the full model and the unconditional distance measure is the difference between the log-likelihood of the saturated and the null model. That is,

$$R^2_{DEV} = 1 - \frac{\log L(y) - \log L(\hat{\beta})}{\log L(y) - \log L(\bar{y})}$$

For Poisson regression the deviance-based R^2 measure is

$$R^2_{DEV,P} = 1 - \frac{\sum_i [(y_i^P \log(y_i^P) - y_i^P) - (y_i^P \log(\hat{\mu}_i) - \hat{\mu}_i)]}{\sum_i [(y_i^P \log(y_i^P) - y_i^P) - (y_i^P \log(\bar{y}_i^P) - \bar{y}_i^P)]}$$

[8,9] with observed frequencies $y_i^P \geq 0$, estimated frequencies $\hat{\mu}_i = n_i \exp(\hat{\beta}x_i)$ and $\bar{y}_i^P = n_i \exp(\hat{\beta}_0) = n_i(\sum_j y_j^P / \sum_j n_j)$ under the full and null model, respectively; n_i denotes the total number of exposures to risk (or person-years lived) for a given covariate vector x_i . For logistic regression the deviance-based R^2 measure is

$$R^2_{DEV,L} = 1 - \frac{\log L(\hat{\beta})}{\log L(\bar{y})}$$

as $\log L(y)$ is equal to zero (note, by definition $0 \log 0 \equiv 0$). Thus

$$R^2_{DEV,L} = 1 - \frac{\sum_i [y_i^L \log \hat{p}_i + (1 - y_i^L) \log(1 - \hat{p}_i)]}{\sum_i [y_i^L \log \bar{y}^L + (1 - y_i^L) \log(1 - \bar{y}^L)]}$$

[6,12,13] with observed outcome $y_i^L \in \{0,1\}$ and estimated probabilities

$$\hat{p}_i = \frac{\exp(\hat{\beta}x_i)}{1 + \exp(\hat{\beta}x_i)}$$

and

$$\bar{y}^L = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)} = \frac{1}{n} \sum_i y_i^L,$$

under the full and null model, respectively, and the number of observations $n = \sum_i n_i$.

The R^2 measure for Poisson regression models, based on sums-of-squares, uses as conditional distance measure the squared difference between observed and estimated frequencies $(y_i^P - \hat{\mu}_i)^2$. For the unconditional distance measure $\hat{\mu}_i$ is replaced by \bar{y}_i^P . The resulting R^2 measure gives

$$R^2_{SS,P} = 1 - \frac{\sum_i (y_i^P - \hat{\mu}_i)^2}{\sum_i (y_i^P - \bar{y}_i^P)^2}$$

[8,9]. Whereas the R^2 measure in logistic regression models, based on sums-of-squares, quantifies the difference between the binary observed outcomes and estimated probabilities, so that $D(y_i|x_i) = (y_i^L - \hat{p}_i)^2$. For the unconditional distance measure \hat{p}_i is replaced by \bar{y}^L , so that eventually

$$R^2_{SS,L} = 1 - \frac{\sum_i (y_i^L - \hat{p}_i)^2}{\sum_i (y_i^L - \bar{y}^L)^2}$$

[6,14].

In the following, all computations for model fitting and calculations for the R^2 measures are made using the statistical software package SAS [15].

3. Results

Results of modelling coronary artery disease are given in Table 2 with the models: (a) smoking only, (b) age only, (c) smoking and age and (d) smoking, age and age-squared. We see that smoking and increasing age have a significant influence on death from coronary artery disease, but the increase in risk is larger for younger people than for older people (quadratic age effect). Model results of logistic and Poisson regression are nearly identical.

In Table 3, model (a), the estimated R^2 -values for fitting only the covariate smoking is shown. R^2_{DEV} gives 3.1% under Poisson regression and 0.3% under logistic regression. The difference between the deviance-based R^2 measures of Poisson and logistic regression becomes even more dramatic when age is modelled (model (b)): 90.9% under Poisson regression and 9.0% under logistic regression. The reason for this is simple: both measures are correct, but have different interpretation. In both models the underlying dependent variable is not the same and one has therefore to define what an R^2 measure should explain. The main intention of the Poisson regression is to explain the observed frequencies. Under model (d) the number of persons died with coro-

Table 2

Results of Poisson/logistic regression for the study about death from coronary artery disease among British doctors [10]

Model	β	S.E.	df	χ^2	P-value
(a) Intercept	-5.42/-5.42	0.040/0.040	1		
Smoke	0.54/0.54	0.107/0.107	1	25.59/25.69	0.0001/0.0001
(b) Intercept	-10.30/-10.33	0.189/0.190	1		
Age	0.08/0.08	0.003/0.003	1	840.63/841.34	0.0001/0.0001
(c) Intercept	-10.22/-10.25	0.191/0.192	1		
Smoke	0.41/0.41	0.107/0.107	1	14.37/14.49	0.0002/0.0001
Age	0.08/0.08	0.003/0.003	1	828.12/828.86	0.0001/0.0001
(d) Intercept	-17.51/-17.51	1.060/1.065	1		
Smoke	0.35/0.36	0.110/0.108	1	10.90/11.01	0.0010/0.0009
Age	0.33/0.33	0.030/0.034	1	90.60/89.70	0.0001/0.0001
Age-squared	-0.002/-0.002	0.0003/0.0003	1	51.25/50.42	0.0001/0.0001

nary artery disease can be predicted very well— R^2 is close to 100%. In contrast the dependent variable in logistic regression indicates on a year-by-year basis, if the person has died or is alive. Obviously it is much easier to predict the number of events, than to predict single events. Consequently, there will be a much higher R^2 -value for Poisson than for logistic regression.

Of course, the same is true if we consider an R^2 measure based on sums-of-squares (R_{SS}^2). As Poisson regression is usually fitted using maximum likelihood and not least-squares, R_{SS}^2 can become negative in some situations under Poisson regression (see [9]), as it is the case in Table 3, model (a). In models (b) to (d) the difference between R_{DEV}^2 and R_{SS}^2 under Poisson regression is minimal. But under logistic regression R_{SS}^2 is much smaller than R_{DEV}^2 , which is due to different construction principles—the improvement per observation in the likelihood of the full model relative to the null model is in our example larger than the improvement in the sums-of-squares of the full model relative to the null model.

From Table 2 we see that all variables (smoke, age and age-squared) have significant influence on the death from coronary artery disease. Despite the substantial differences in R^2 -values between Poisson and logistic regression, their ranking of the covariates remains unchanged. Under both models age explains most, whereas smoking, although significant, reduces uncertainty about death only slightly.

Table 3

Estimated R^2 measures in percent for the study about death from coronary artery disease among British doctors calculated under Poisson and logistic regression [10]

Model	Poisson regression		Logistic regression	
	$R_{DEV,P}^2$	$R_{SS,P}^2$	$R_{DEV,L}^2$	$R_{SS,L}^2$
(a) Smoke	3.1	-4.5	0.3	0.0
(b) Age	90.9	90.2	9.0	0.5
(c) Age + smoke	92.6	91.4	9.2	0.5
(d) Age + age-squared + smoke	98.7	99.6	9.7	0.6

4. Discussion

The example illustrates the structurally big difference in R^2 measures between Poisson and logistic regression. The likelihood, parameter estimates and P-values are approximately the same for both models, but one has to be cautious in interpreting R^2 measures. Do we want to explain the variability in the observed event rates or in the individual outcomes? In the example a major reduction in the uncertainty of the event rate prediction for a given covariate pattern can be achieved, however, the predictability of an individual outcome is still poor.

This is a typical example for studies with many observations, where contributions of the explanatory variables are highly significant and of substantial interest. However the R^2 measure of the logistic regression can be a useful reminder that the contribution of the variables may, in fact, explain only a small percentage of the variability in the individual response. Cox and Wermuth [16] stated that this type of interpretation is misleading in linear regressions with binary responses since low values of R^2 , roughly 0.1, are inevitable even if an important relation is present. But our example shows that low R^2 -values in logistic regression are a necessary hint that small P-values, impressive parameter estimates for prognostic factors and large R^2 -values calculated under Poisson regression may hide the necessity to search for further prognostic factors to improve the prediction of the individual outcome. It is even possible that the R^2 for Poisson regression achieves one, so that the event rates are perfectly predicted by the model, and in the same time the R^2 -value of the corresponding logistic regression may achieve only a value close to zero, indicating that it is still difficult to predict individual outcome.

R^2 measures usually increase monotonically with increasing number of covariates even if they have no prognostic value at all. Therefore the use of adjusted R^2 measures (as proposed by [6,9,17,18]) is advised especially in cases with small number of observations (that is distinctive covariate patterns in Poisson regression) relative to the number of covariates. However, the principal problem as described here remains the same.

References

- [1] Drolette ME. The vanishing 2×2 table: linking the hypergeometric, binomial and Poisson. *Am Stat* 1974;28:102–3.
- [2] Eisenberg H, Geoghagen R, Walsh J. A general use of the Poisson approximation for binomial events, with application to bacterial endocarditis data. *Biometrics* 1966;22:74–82.
- [3] Matsunawa T. Poisson distribution. In: Kotz S, Johnson NL, editors. *Encyclopedia of statistical sciences*, vol. 7. New York: Wiley, 1988. pp. 21–5.
- [4] Sheu S. The Poisson approximation to the binomial distribution. *Am Stat* 1984;38:206–7.
- [5] Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: Wiley, 1989.
- [6] Mittlböck M, Schemper M. Explained variation for logistic regression. *Stat Med* 1996;15:1987–97.
- [7] Korn EL, Simon R. Explained residual variation, explained risk, and goodness of fit. *Am Stat* 1991;45:201–6.
- [8] Cameron AC, Windmeijer FAG. R^2 measures for count data regression models with applications to health-care utilization. *J Bus Econ Stat* 1996;14:209–20.
- [9] Waldhör T, Haidinger G, Schober E. Comparison of R^2 measures for Poisson regression by simulation. *J Epidemiol Biostat* 1998;3:209–15.
- [10] Doll R, Hill AB. Mortality of British doctors in relation to smoking: observations on coronary thrombosis. *Nat Cancer Inst Monog* 1966; 19:205–68.
- [11] Breslow NE. Cohort analysis in epidemiology. In: Atkinson AC, Fienberg SE, editors. *A celebration of statistics*. New York: Springer Verlag, 1985. pp. 109–43.
- [12] Theil H. On the estimation of relationships involving qualitative variables. *Am J Sociol* 1970;76:103–54.
- [13] McFadden D. The measurement of urban travel demand. *J Public Econ* 1974;3:303–28.
- [14] Margolin BH, Light RJ. An analysis of variance for categorical data, II: small sample comparisons with chi-square and other competitors. *J Am Stat Assoc* 1974;69:755–64.
- [15] The SAS system for windows. SAS. 6.12. Cary, NC: SAS Institute Inc., 1996.
- [16] Cox DR, Wermuth N. A comment on the coefficient of determination for binary responses. *Am Stat* 1992;46:1–4.
- [17] Mittlböck M, Schemper M. Computing measures of explained variation for logistic regression models. *Comp Methods Prog Biomed* 1999;58:17–24.
- [18] Mittlböck M, Waldhör T. Adjustments for R^2 -measures for Poisson regression models. *Comp Stat Data Anal* 2000;34:461–72.