ELSEVIER

Computational Statistics & Data Analysis 34 (2000) 461–472

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Adjustments for R^2-measures for Poisson regression models

M. Mittlböck[a,*], T. Waldhör[b]

[a] *Department of Medical Computer Sciences, Section of Clinical Biometrics, University of Vienna, Spitalgasse 23, 1090 Vienna, Austria*

[b] *Department of Epidemiology, Institute of Cancer Research, University of Vienna, Austria*

Received 1 October 1998; received in revised form 1 August 1999; accepted 26 November 1999

Abstract

In regression models not only the parameter estimates and significances of explanatory variables are of interest, but also the degree to which variation in the dependent variable can be explained by covariates. In recent publications an R^2-measure based on deviance was recommended for Poisson regression models, one of the most frequently used modelling tools in epidemiological studies. However, when sample size is small relative to the number of covariates in the model, simple R^2-measures may be seriously inflated and may need to be adjusted according to the number of covariates in the model. Two new adjustments for the R^2-measure in Poisson regression models based on deviance residuals are presented and compared by simulation with population values. The proposed measures are also applied to real data sets.

Keywords: Adjusted R^2; Poisson regression; Deviance residuals; Likelihood ratio; Degrees of freedom

1. Introduction

R^2-measures are frequently used in linear regressions and are also becoming more familiar in generalized linear models. Although R^2-measures are also called measures of the proportion of 'explained variation', for generalized linear models the term variation is not as clear as in linear regression, which is based on least-squares. Generalized linear models are usually fitted by maximum likelihood methods and many authors therefore prefer R^2-measures based on the proportion of the reduction

* Corresponding author. Tel.: +43-1-40400-2276; fax: +43-1-40400-2278.
E-mail address: martina.mittlboeck@akh-wien.ac.at (M. Mittlböck).

0167-9473/00/$ - see front matter
PII: S 0167-9473(99)00113-9

in maximized log-likelihood. All R^2-measures try to give a measure with values between zero and one, indicating no prognostic value of the covariates and perfect prediction, respectively.

In recent years, several papers have dealt with R^2-measures for Poisson regression models (e.g. Cameron and Windmeijer, 1996; Waldhör et al., 1998). The main interest has been in the behaviour of R^2 without considering the number of fitted parameters. Similar to regression models in clinical studies, Poisson regression models are often used to screen for prognostic factors in epidemiological studies which have only small to moderate sample size and many covariates. In such situations unadjusted R^2-measures may give substantially inflated values, jeopardizing the ability to draw valid interpretations. R^2-values of 30% or higher can easily be reached, even when no association between independent and dependent variables exists at all.

In linear regression models the use of an adjusted R^2-measure (R^2_{adj}) is well established and based on strong theoretical arguments, whereas in Poisson regression models only ad hoc corrections have been proposed. To address this problem, we suggest two new adjustments for an R^2-measure based on deviance residuals (R^2_{DEV}), which was recommended by Cameron and Windmeijer (1996) and Waldhör et al. (1998). We compare the behaviour of the new measures in a simulation study with the unadjusted R^2-measure (R^2_{DEV}) and an adjusted R^2-measure suggested by Waldhör et al. ($R^2_{\text{DEV,df}}$). Finally, we present real-data examples and discuss the advantages and disadvantages of the suggested measures.

2. Adjusted R^2 measures

Waldhör et al. (1998) reviewed four R^2-measures for Poisson regression models of the form $\log(\mu_i) = \boldsymbol{\beta}\boldsymbol{x}_i$, where μ_i is the expectation of a Poisson distributed variable, $\boldsymbol{x}_i$ is the vector of covariates for the ith observation and $\boldsymbol{\beta}$ is the parameter vector to be estimated with β^0 as the intercept and $\beta^1, \ldots, \beta^k$ as the parameters for the k covariates. They described and compared their properties in detail and recommended using the R^2-measure based on deviance residuals, which can also be expressed in terms of the log-likelihood:

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

where y_i is the observed value of the dependent variable, $\hat{\mu}_i$ the predicted value of the ith observation, $\bar{y}$ the mean of the dependent variable, $\log L(y)$ the log-likelihood of the saturated model, $\log L(\hat{\mu})$ the log-likelihood when all covariates are fitted and $\log L(\bar{y})$ the log-likelihood when only the intercept is fitted.

The inflation of R^2-measures can be considered when the number of covariates is large relative to a given sample size. In linear models, R^2-measures are therefore adjusted by their degrees of freedom, so that the expectation of the adjusted R^2 (based on sums-of-squares) equals zero for $R^2 = 0$ in the underlying population:

$$R^2_{\text{adj}} = 1 - \frac{(n-k-1)^{-1} \sum_i (y_i - \hat{\mu}_i)^2}{(n-1)^{-1} \sum_i (y_i - \bar{y})^2},$$

where n denotes the sample size and k the number of estimated covariates without intercept.

Waldhör et al. suggested using the same correction for degrees of freedom in the R^2_{DEV}-measure as for the adjusted R^2-measure (R^2_{adj}) in linear regression models:

$$R^2_{\mathrm{DEV,df}} = 1 - \frac{(n-k-1)^{-1}[\log L(y) - \log L(\hat{\mu})]}{(n-1)^{-1}[\log L(y) - \log L(\bar{y})]}.$$

In a simulation study they demonstrated that this correction is adequate with $\beta^1,\ldots,\beta^k = 0$ and a large μ. However, for rare events (i.e. small values of μ) and small sample size, $R^2_{\mathrm{DEV,df}}$ gives values which are much too small. Similarly, the behaviour of $R^2_{\mathrm{DEV,df}}$ has not been investigated in situations where $\beta^1,\ldots,\beta^k \neq 0$. Furthermore, although this adjustment is appropriate in linear regression models when using the sum-of-squares approach, it may only be approximately valid in Poisson regressions using deviance residuals. We suggest two new correction terms based on the log-likelihood function and investigate them in comparison to the unadjusted R^2_{DEV} and with $R^2_{\mathrm{DEV,df}}$.

The likelihood-ratio statistic for testing all k explanatory covariates in regression models is $2[\log L(\hat{\mu}) - \log L(\bar{y})]$ which follows approximately a $\chi^2\sim$distribution with k degrees of freedom under the null hypothesis H_0: $\beta^1,\ldots,\beta^k = 0$. The expectation of the likelihood-ratio statistic under H_0 is therefore k and the bias of $\log L(\hat{\mu})$ is $k/2$ due to the estimation of the effect of k non-informative covariates. Our first proposal for a new correction of the R^2-measure is

$$R^2_{\mathrm{DEV,adj,1}} = 1 - \frac{\log L(y) - [\log L(\hat{\mu}) - k/2]}{\log L(y) - \log L(\bar{y})} = 1 - \frac{\log L(y) - \log L(\hat{\mu}) + k/2}{\log L(y) - \log L(\bar{y})}.$$

Mittlböck and Schemper (1996) used a similar measure for logistic regression which can also be applied to Poisson regression models. Namely, the likelihood-ratio statistic $2[\log L(\hat{\mu}) - \log L(\beta^0)]$, where β^0 is the true intercept parameter, follows approximately an $\chi^2\sim$distribution with $k+1$ degrees of freedom under the null hypothesis H_0: $\beta^1,\ldots,\beta^k = 0$. Therefore the bias of $\log L(\hat{\mu})$ is $(k+1)/2$. Similarly, the bias of $\log L(\bar{y})$ is $\frac{1}{2}$. The second corrected R^2-measure we propose is therefore

$$R^2_{\mathrm{DEV,adj,2}} = 1 - \frac{\log L(y) - [\log L(\hat{\mu}) - (k+1)/2]}{\log L(y) - [\log L(\bar{y}) - 1/2]}$$

$$= 1 - \frac{\log L(y) - \log L(\hat{\mu}) + (k+1)/2}{\log L(y) - \log L(\bar{y}) + 1/2}.$$

As can be easily seen $R^2_{\mathrm{DEV,adj,2}}$ always gives values closer to zero than $R^2_{\mathrm{DEV,adj,1}}$.

3. Simulations

The performance of the unadjusted and the three adjusted R^2-measures was compared under various conditions for a Poisson regression model. Using the SAS

procedure FACTEX (SAS, 1996), a factorial design was produced for sample sizes of 16, 32, 64 and 16 384 [$=2^{14}$] with 1, 3 and 5 completely balanced dichotomous covariates with values 0 and 1. For simulations with five covariates and a sample size of 16 only a fractional factorial design was produced as there were too few observations for a factorial design with five covariates. With the SAS function RANPOI, Poisson distributed random variables were generated with mean $\mu_0 \times \exp[\beta^1 x_1]$ with $\mu_0 = \exp(\beta^0) = 1, 2, 5, 30$, respectively, and β^1 varying from 0 to 3. β^1 was chosen so that a wide range of R^2 was covered for all μ_0. Only the first covariate x_1 was assumed to influence μ. The prognostic effect of all other covariates was eliminated by setting $\beta^2, \ldots, \beta^k = 0$. The SAS procedure GENMOD was used to fit a Poisson regression to the data and a SAS macro was written to do all subsequent calculations. The number of repetitions was 50 000 for the smaller sample sizes of 16, 32 and 64 and 1000 for the large sample size of 16 384.

The mean of R^2_{DEV} for the 1000 repetitions with the large sample size (R^2_{large}) was taken to be the true value because the estimations of the unadjusted and the adjusted R^2-measures are of essentially the same value and the sample size is large enough that the number of investigated covariates (up to five covariates) has no influence on the resulting values.

The mean estimated R^2-values with varying mean (μ_0), number of covariates (k), sample size (n) and underlying R^2 (depending on β^1) are listed in Tables 1a–1c. For the sake of brevity, the results for $\mu_0 = 5$ are discussed but not shown in the tables. We see that under H_0: $\beta^1, \ldots, \beta^k = 0$, the unadjusted R^2_{DEV}-values increases dramatically with increasing number of covariates and with decreasing sample size. An increasing value of μ_0 also results in slightly increased R^2_{DEV} estimations. With $k = 5$, $n = 16$ and $\mu_0 = 1$, the unadjusted R^2_{DEV}-measure reaches nearly 30% where no dependence between μ and the covariates exists at all. This figure even rises to about 33% when μ_0 increases to a value of 30.

We also see that while the three adjusted measures, $R^2_{\mathrm{DEV,df}}$, $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$, correct this bias for the most part, they overcorrect in some situations so that the estimated mean of R^2 is lower than the large sample R^2. But this "small" bias seems to be negligible compared to the bias of the unadjusted measure. Therefore, all three adjusted measures represent an improvement over the unadjusted measure. However, because they generate values which are not always in agreement, the question remains as to which measure behaves the best when compared to the large sample estimate.

$R^2_{\mathrm{DEV,adj,2}}$ always gives values closer to zero than $R^2_{\mathrm{DEV,adj,1}}$, so it corrects more for positive R^2-values and fairly consistently it gives smaller values than the large-sample study. Therefore $R^2_{\mathrm{DEV,adj,1}}$ is almost always preferable to $R^2_{\mathrm{DEV,adj,2}}$, except when $R^2=0$, in which case not much attention is given to R^2-measures at all. $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,df}}$ also tend to provide smaller values than the large-sample study except for substantial β^1 and μ_0 ($\beta^1 \geq 1$ and $\mu_0 \geq 5$).

For all $R^2_{\mathrm{large}} > 65\%$, both $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,df}}$ give values very close to the large sample results. The maximum difference is less than 1%, compared to a maximum difference of about 11% for the unadjusted measure.

For $\beta^1, \ldots, \beta^k = 0$, all estimated values are negative, indicating that the corrections for the R^2-measure are too strong. $R^2_{\mathrm{DEV,adj,2}}$ is closest to the true value of zero for

$\mu_0 = 1$ and 2, but for larger μ_0, $R^2_{\mathrm{DEV,df}}$ is closest to zero. In the simulations we also allowed negative results, so that the estimated means would not be biased, but in practice one would choose $\max(0, R^2_{\mathrm{adj}})$. A negative R^2-value makes no sense in reality and would be equivalent to having no explained variation at all.

When the Poisson regression model approaches the linear regression (with large μ_0), $R^2_{\mathrm{DEV,df}}$ behaves best, whereas in Poisson regression situations with rare events $R^2_{\mathrm{DEV,adj,1}}$ provides lower bias than $R^2_{\mathrm{DEV,df}}$.

Table 1a
Simulation results with $\mu_0 = 1$

μ_0	β^1	R^2_{large}	k	n	R^2_{DEV}	$R^2_{\mathrm{DEV,df}}$	$R^2_{\mathrm{DEV,adj,1}}$	$R^2_{\mathrm{DEV,adj,2}}$
1	0.0	0.0	1	16	5.9	−0.8	−0.3	−0.3
				32	2.8	−0.4	−0.1	−0.1
				64	1.4	−0.2	0.0	0.0
			3	16	17.6	−3.1	−1.1	−0.9
				32	8.4	−1.4	−0.3	−0.3
				64	4.1	−0.7	−0.1	−0.1
			5	16	29.4	−5.9	−1.7	−1.4
				32	14.1	−2.5	−0.5	−0.4
				64	6.9	−1.1	−0.1	−0.1
1	0.5	6.5	1	16	11.7	5.4	5.8	5.6
				32	9.0	6.0	6.3	6.1
				64	7.7	6.2	6.4	6.3
			3	16	22.6	3.2	4.9	4.8
				32	14.3	5.1	6.1	5.9
				64	10.3	5.8	6.3	6.3
			5	16	33.6	0.3	4.1	4.1
				32	19.5	4.0	5.9	5.7
				64	12.9	5.4	6.3	6.2
1	1.0	26.8	1	16	30.8	25.8	26.1	25.0
				32	28.7	26.3	26.5	26.0
				64	27.7	26.6	26.7	26.4
			3	16	39.4	24.3	25.4	24.4
				32	32.9	25.7	26.4	25.8
				64	29.8	26.3	26.6	26.4
			5	16	48.2	22.3	24.8	23.9
				32	37.1	25.0	26.2	25.7
				64	31.8	26.0	26.6	26.3
1	2.0	71.5	1	16	73.9	72.0	72.1	70.9
				32	72.7	71.8	71.9	71.3
				64	72.1	71.7	71.7	71.4
			3	16	77.3	71.6	72.0	70.8
				32	74.3	71.6	71.8	71.2
				64	72.9	71.6	71.7	71.4
			5	16	80.6	71.0	71.9	70.6
				32	76.0	71.4	71.8	71.2
				64	73.8	71.5	71.7	71.4

Table 1b
Simulation results with $\mu_0 = 2$

μ_0	β^1	R^2_{large}	k	n	R^2_{DEV}	$R^2_{\text{DEV,df}}$	$R^2_{\text{DEV,adj,1}}$	$R^2_{\text{DEV,adj,2}}$
2	0.0	0.0	1	16	5.9	−0.8	−0.7	−0.6
				32	2.9	−0.4	−0.1	−0.1
				64	1.4	−0.2	−0.1	−0.1
			3	16	17.7	−2.8	−2.2	−1.8
				32	8.5	−1.3	−0.5	−0.4
				64	4.2	−0.6	−0.1	−0.1
			5	16	29.7	−5.5	−3.5	−3.0
				32	14.2	−2.3	−0.8	−0.7
				64	7.0	−1.1	−0.2	−0.2
2	0.4	8.0	1	16	13.3	7.1	7.1	6.8
				32	10.6	7.6	7.7	7.5
				64	9.3	7.8	8.0	7.9
			3	16	24.4	5.5	5.8	5.6
				32	15.9	6.9	7.5	7.3
				64	11.9	7.5	7.9	7.8
			5	16	35.5	3.3	4.4	4.5
				32	21.3	6.1	7.2	7.0
				64	14.5	7.2	7.8	7.7
2	0.8	30.3	1	16	34.3	29.6	29.6	28.4
				32	32.2	30.0	30.1	29.5
				64	31.3	30.2	30.3	30.0
			3	16	42.8	28.5	28.8	27.7
				32	36.3	29.5	29.9	29.3
				64	33.3	30.0	30.2	29.9
			5	16	51.4	27.1	28.1	27.0
				32	40.5	29.0	29.8	29.2
				64	35.3	29.8	30.2	29.9
2	1.5	67.8	1	16	70.4	68.3	68.4	67.0
				32	69.1	68.1	68.1	67.4
				64	68.5	67.9	68.0	67.6
			3	16	74.3	67.9	68.1	66.7
				32	71.0	67.8	68.0	67.3
				64	69.4	67.8	67.9	67.6
			5	16	78.1	67.2	67.8	66.5
				32	72.8	67.6	67.9	67.2
				64	70.3	67.7	67.9	67.6

4. Examples

The following two examples are typical for epidemiological studies, where we have usually given the number of events (y_i), e.g. death, and person-years lived for each pattern of covariates. The number of events is then modelled by Poisson regression depending on the covariates and the offset, which is usually the logarithm of the person-years lived. A sequence of models is calculated for each example,

Table 1c
Simulation results with $\mu_0 = 30$

μ_0	β^1	R^2_{large}	k	n	R^2_{DEV}	$R^2_{DEV,df}$	$R^2_{DEV,adj,1}$	$R^2_{DEV,adj,2}$
30	0.0	0.0	1	16	6.6	0.0	−1.0	−0.9
				32	3.2	0.0	−0.2	−0.2
				64	1.6	0.0	−0.1	−0.1
			3	16	19.9	−0.1	−3.1	−2.6
				32	9.6	−0.1	−0.7	−0.6
				64	4.7	−0.1	−0.2	−0.2
			5	16	33.3	−0.1	−5.0	−4.3
				32	16.0	−0.1	−1.1	−1.1
				64	7.9	−0.1	−0.3	−0.3
30	0.1	7.2	1	16	13.2	7.0	6.1	5.8
				32	10.1	7.1	6.9	6.7
				64	8.7	7.2	7.1	7.0
			3	16	25.5	6.9	4.4	4.4
				32	16.0	7.0	6.5	6.3
				64	11.6	7.2	7.0	6.9
			5	16	37.8	6.8	2.6	2.9
				32	22.0	7.0	6.1	6.0
				64	14.5	7.1	6.9	6.8
30	0.2	24.0	1	16	28.5	23.4	22.9	21.8
				32	26.2	23.7	23.6	23.0
				64	25.1	23.9	23.9	23.6
			3	16	38.7	23.4	21.7	20.7
				32	31.0	23.7	23.3	22.7
				64	27.5	23.9	23.8	23.5
			5	16	48.8	23.3	20.4	19.7
				32	35.9	23.6	23.0	22.5
				64	29.9	23.8	23.7	23.4
30	0.5	70.4	1	16	72.9	71.0	70.9	69.5
				32	71.7	70.7	70.7	70.1
				64	71.1	70.6	70.6	70.3
			3	16	76.8	71.0	70.8	69.4
				32	73.6	70.7	70.7	70.0
				64	72.0	70.6	70.6	70.3
			5	16	80.6	70.9	70.6	69.2
				32	75.4	70.7	70.6	70.0
				64	72.9	70.6	70.6	70.2

beginning with a simple model with one covariate and continuing through a more complex model, fitting one additional effect on each step (type I). As the model fit of a Poisson regression depends on the covariates in the model, at least the last model of our examples shows an acceptable model fit without over- or underdispersion. Also non-significant covariates are left in the model to illustrate different situations.

Table 2a
Results of Poisson regression of Example 1 (suicides among academics in Denmark)

	Source	β	s.e.	df	Chi-square	p-value
(a)	Intercept	−7.55	0.156	1		
	Sex	−0.25	0.176	1	1.99	0.1583
(b)	Intercept	−7.89	0.227	1		
	Sex	−0.26	0.176	1	2.10	0.1472
	Age	0.07	0.032	1	4.73	0.0297
(c)	Intercept	−8.88	0.444	1		
	Sex	- 0.27	0.176	1	2.20	0.1379
	Age	0.53	0.174	1	10.10	0.0015
	Age-squared	−0.04	0.016	1	7.85	0.0051

Table 2b
Estimated R^2-measures in percent of Example 1 (suicides among academics in Denmark)

	R^2_{DEV}	$R^2_{\mathrm{DEV,df}}$	$R^2_{\mathrm{DEV,adj,1}}$	$R^2_{\mathrm{DEV,adj,2}}$
(a) Sex	7.4	1.7	3.5	3.4
(b) Sex + age	25.9	16.1	18.1	17.4
(c) Sex + age + age-squared	56.7	47.5	45.0	43.3

4.1. Suicides among academics in Denmark 1970–1980

Based on data from the Danish census of 9th November 1970 and official mortality statistics for the preceding 10 years, the mortality of the total Danish labour force aged 20–64 years was studied. Anderson et al. (1993, pp. 17, 18) give the number of suicides (y_i), a total of 193 death, and the number of person-years lived, at total 447.358 years, for academics specified by sex and age. Age is separated into nine age groups with an interval length of 5 years. Thus we have 18 observations (covariate patterns), nine age-groups for males and females. In our model for this data set we fit three covariates: sex, age and age-squared. There is no interaction between sex and age.

In Table 2a the results of the Poisson regressions are given, modelling (a) sex, (b) sex and age and (c) sex, age and age-squared, respectively. In Table 2b the corresponding estimated R^2 values are given. In model (a) sex has no significant influence. The values of $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ are about 3.5%, approximately half the value of R^2_{DEV}. $R^2_{\mathrm{DEV,df}}$ is only one fourth of R^2_{DEV}. When added to the model, both age and age-squared prove to be significant factors. Suicide increases with increasing age until 45–49 years and decreases afterwards. Table 2b shows that age-squared contributes most to the reduction of unexplained variation, measured by means of deviance. This could not be foreseen from parameter estimates or corresponding p-values alone. About 45% of the uncertainty can be explained by sex, age and age-squared.

Table 3a
Results of Poisson regression of Example 2 (death from coronary artery disease among doctors)

	Source	β	s.e.	df	Chi-square	p-value
(a)	Intercept	−5.42	0.040	1		
	Smoke	−0.54	0.107	1	25.59	0.0001
(b)	Intercept	−10.22	0.191	1		
	Smoke	−0.41	0.107	1	14.37	0.0002
	Age	0.08	0.003	1	828.12	0.0001
(c)	Intercept	−17.51	1.06	1		
	Smoke	−0.35	0.11	1	10.90	0.0010
	Age	0.33	0.03	1	90.60	0.0001
	Age-squared	−0.002	0.0003	1	51.25	0.0001

Table 3b
Estimated R^2-measures in percent of Example 2 (death from coronary artery disease among doctors)

	R^2_{DEV}	$R^2_{\mathrm{DEV,df}}$	$R^2_{\mathrm{DEV,adj,1}}$	$R^2_{\mathrm{DEV,adj,2}}$
(a) Smoke	3.1	−9.0	3.0	3.0
(b) Smoke + age	92.6	90.5	92.4	92.3
(c) Smoke + age + age-squared	98.7	98.0	98.4	98.3

4.2. Death from coronary artery disease among doctors

In 1961 Doll and Hill (1966) sent a questionaire to all male doctors on the British Medical Register enquiring about their smoking habits. Almost 70% of the doctors replied. Death certificates were obtained for medical practitioners and causes of death were assigned on the basis of these certificates. The data set of Breslow (1985) contains 10 observations (five age-groups and smoking status) with the corresponding person-years lived (a total of 181.467 years) and the number of deaths from coronary artery disease (a total of 731 deaths) accumulated during the first 10 years of study. In Table 3a the results of modelling the number of deaths from coronary artery disease (y_i) are given with the covariates (a) smoking, (b) smoking and age and (c) 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 bigger for younger people than for older people (quadratic age-effect).

In the first row of Table 3b we see the estimated R^2-values for fitting smoking only. R^2_{DEV}, $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ give very similar values of around 3%, which is not unusual as we fitted only one covariate and the model-χ^2 is highly significant, but $R^2_{\mathrm{DEV,df}}$ becomes negative, which is not plausible as the effect of smoke is small but significant. In general, if the model fit is significant, a suitable measure for R^2_{adj} should be greater than zero. From the second row we see that age is a very important factor. The estimated R^2 increases from 3% to more than 90%. From the third row we see that the quadratic term of age increases R^2_{DEV}, $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ by about

6%. One has to be aware that we are explaining in this example the proportion of death, as it is the nature of a Poisson model, but we do not predict the event "death of a single individual". Therefore, it is not unusual to achieve an R^2-value of about 98% with 10 different covariate pattern and three fitted covariates.

5. Discussion

In linear regression models one should always use an adjusted R^2-measure. Otherwise the R^2-value increases monotonically as the number of covariates increases, even if they have no prognostic value at all. This rather undesirable property of R^2 can result in artificially high values and may discourage investigators from searching for further prognostic factors. Cameron and Windmeijer (1996) put forward major arguments for the use of R^2_{DEV} as a measure for explained variation and Waldhör et al. (1998) compared several candidates for an R^2-measure for Poisson regression. Whatever intrinsic qualities each measure might have, it is clear that the difference between the different measures is generally smaller than the difference between adjusted and unadjusted R^2-measures.

Waldhör et al. suggested the use of the same correction for R^2_{DEV} in Poisson regression models as for the R^2-measure based on sums-of-squares in linear regression models. This correction is based on the number of fitted degrees of freedom under the full model with covariates and under the null model, when only the intercept is fitted. This correction works well for larger μ, when the Poisson regression is approximated by linear regression. Cameron and Windmeijer (1996) stated that the concepts of deviance, maximum likelihood estimation and Kullback–Leibler distance are similar in function to the concept of residual sum of squares and least-squares estimation in linear models. Therefore, it is obvious that the same correction makes sense for R^2-measures based on sums-of-squares and for R^2-measures based on deviance residuals. However, in Poisson regression situations with rare events, where the normal approximation is not appropriate, this correction can underestimate the true values substantially and an adjustment based on the expected optimism of the log-likelihood under the null-hypothesis seems to be more appropriate. Such a correction based on the likelihood is also in accordance with the basic character of R^2_{DEV}.

Usually, the range of an adjusted measure of explained variation also includes negative values, as the correction is always based on expected values under the null hypothesis in which covariates exert no effect. But negative values for R^2_{adj} are not a problem, as in these situations the whole model is not significant and one is not interested in reporting and interpreting the proportion of explained variation. Normally R^2_{adj} is then assumed to be zero. But if the model fit is significant, a measure of R^2_{adj} should be positive. Also, if a significant factor is added to a model, then R^2_{adj} should increase. Neither are true for $R^2_{\mathrm{DEV,df}}$, as shown in the second example, whereas they are always true for $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$. For each degree of freedom due to the addition of covariates in the fitted model, a value of $\frac{1}{2}$ is subtracted from $\log L(\hat{\mu})$. Thus, $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ are only negative if the k covariates increase

$\log L(\hat{\mu})$ with less than $k/2$. But the model-χ^2 is only significant if the log-likelihood is increased by more than $k/2$. The same is true for adding covariates to a model. The correction term for $\log L(\hat{\mu})$ is $\frac{1}{2}$ higher for each added degree of freedom, but $\log L(\hat{\mu})$ has to increase more than $\frac{1}{2}$ for each added degree of freedom in order to be significant. If a covariate with one degree of freedom is added with a p-value of 0.3173, which corresponds with an χ^2_1 test statistic of 1 (this is twice the increase in $\log L(\hat{\mu})$), then $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ remain unchanged. For $p < 0.3173$, $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ increase and if $p > 0.3173$, $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ decrease.

The concepts of explained variation and of significant p-values are different. It is not possible nor desirable to draw conclusions from the magnitude of R^2-values to the significance of the model nor to the significance of single covariates. For instance, in Example 2 $R^2_{\mathrm{DEV}}, R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ are all around 3%. There is nearly no correction because the model-χ^2 is highly significant. Despite the relatively small p-value, the proportion of variation explained by smoking is very small, only 3%. In Example 1 R^2_{DEV} is about 7.4% and the model-χ^2 is not significant. Therefore, the correction is rather strong and $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ are around 3.5%, about half the value of $R^2_{\mathrm{DEV}}, R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ are even slightly higher than in the significant model in Example 2.

In summary, $R^2_{\mathrm{DEV,adj,1}}$ behaves best in typical situations where a Poisson regression is based on a small sample and/or many covariates. Although its estimated value is too low in situations with no or nearly no explained variation, in such cases the model test would also not be significant. One would therefore not pay attention to the model, nor the R^2-measure. Furthermore, in small samples the likelihood ratio may not follow an χ^2-distribution, so the correction term is also not completely adequate even though the correction seems to fit well enough. Only when μ is large and the Poisson distribution is approximated by the normal distribution, $R^2_{\mathrm{DEV,df}}$ does give a better estimation of the underlying R^2-value.

All the measures presented here are easy to calculate, as most packages for Poisson regressions provide the log-likelihood of the fitted model. With most packages one can also fit Poisson models with intercept only. All corrections depend only on the number of fitted covariates. With these three values, $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$ can easily be calculated. To calculate $R^2_{\mathrm{DEV,df}}$, the sample size also is needed.

A major problem in Poisson regression may be overdispersion, when the dispersion of the data is greater than that predicted by the Poisson model, i.e. $\mathrm{var}(Y) > E(Y)$, which is most likely if the Pearson and deviance goodness-of-fit statistics indicate poor fit. So an apparent overdispersion could reflect missing covariates or it may be produced by a clustered Poisson process. If the precise mechanism that produces the overdispersion is known, specific methods may be used, e.g. a random effects model which are frequently based on the negative binomial likelihood. In the absence of such knowledge McCullagh and Nelder (1989) suggest to assume as an approximation that $\mathrm{var}(Y) = \phi\mu$ for some constant ϕ. The estimate of ϕ (which is also called dispersion parameter in generalized linear models) is usually the Pearson or deviance statistic divided by its degrees of freedom. With overdispersion present, the use of the Poisson maximum likelihood equations for estimating the regression parameters in the mean is still valid, however the variance structure may be misspecified.

Therefore, the predicted values and the deviance remain the same under overdispersion and R^2-measures are not changed. However, the adjustments of $R^2_{\mathrm{DEV,adj,1}}$ and $R^2_{\mathrm{DEV,adj,2}}$, which are based on log-likelihood theory assuming no overdispersion, may be too small. To find correct R^2-adjustments when overdispersion is handled with the simple method of McCullagh and Nelder is still a research area. If models based on negative binomial likelihood are used to deal with overdispersion a completely different model is fitted and R^2-measures as discussed in this paper may not be suitable.

In conclusion, we recommend routine evaluation of the adjusted proportion of explained variation in Poisson regression models. In any of these applications investigators may easily be misled by highly significant p-values or impressive parameter estimates for explanatory factors, while outcomes are far from being determined. R^2-measures offer a different and more accurate view. If investigators evaluate the effect of innumerable covariates, the unadjusted R^2-measure will increase with the number of covariates. Therefore, one should always use an adjusted R^2-measure which considers the number of evaluated covariates and gives a realistic view about the proportion of variation explained by covariates in the model.

References

Anderson, P.K., Borgan, O., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, New York.

Breslow, N.E., 1985. Cohort Analysis in Epidemiology. Atkinson, A.C. et al. (Ed.), A Celebration of Statistics. Springer, New York, pp. 109–143.

Cameron, A.C., Windmeijer, F.A.G., 1996. R^2 measures for count data regression models with applications to health-care utilization. J. Business Econom. Statist. 14, 209–220.

Doll, R., Hill, A.B., 1966. Mortality of british doctors in relation to smoking: observations on coronary thrombosis. Nat. Cancer Inst. Monogr. 19, 205–268.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, London.

Mittlböck, M., Schemper, M., 1996. Explained variation for logistic regression. Statist. Med. 15, 1987–1997.

SAS, 1996. The SAS System for Windows 6.12. SAS Institute Inc., Cary, NC.

Waldhör, T., Haidinger, G., Schober, E., 1998. Comparison of R^2 measures for Poisson regression by simulation. J. Epidemiol. Biostatist. 3, 209–215.