STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2867–2872
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2168

Careful use of pseudo *R*-squared measures in epidemiological studies

Harald Heinzl[1], Thomas Waldhör[2] and Martina Mittlböck[1,*,†]

[1]*Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Vienna, Austria*
[2]*Department of Epidemiology, Center of Public Health, Medical University of Vienna, Vienna, Austria*

SUMMARY

Many epidemiological research problems deal with large numbers of exposed subjects of whom only a small number actually suffers the adverse event of interest. Such rare events data can be analysed by employing an approximate Poisson model. The objective of this study is to challenge the interpretability of the corresponding Poisson pseudo *R*-squared measure. It will lack sensible interpretation whenever the approximate Poisson outcome is generated by counting the number of events within covariate patterns formed by cross-tabulating categorical covariates. The failure is caused by the immanent arbitrariness in the definition of the covariate patterns, that is, independent Bernoulli events, $B(1,\pi)$, are arbitrarily combined into binomially distributed ones, $B(n,\pi)$, which are then approximated by the Poisson model.

KEY WORDS: Poisson regression model; logistic regression model; pseudo *R*-squared measure; predictive accuracy; binomial outcome variable; Bernoulli outcome variable

1. INTRODUCTION

The Poisson regression model is often used as approximation for the logistic regression model in analysing epidemiological data sets. Mittlböck and Heinzl [1] have shown that even though both models yield nearly identical regression coefficients, *p*-values and confidence intervals, their corresponding pseudo *R*-squared values can differ enormously. They stated that this is due to the fact that in the logistic model the *R*-squared measure (or measure of predictive accuracy) quantifies the predictability of single events, whereas in the Poisson model the *R*-squared measure quantifies the predictability of event rates. Here the meaning of the latter statement will be challenged as it contains a considerable amount of latent arbitrariness which in the end considerably limits the sensible use of the Poisson *R*-squared measure in epidemiological settings. That is, the Poisson model is often used to approximate an artificial binomial outcome,

*Correspondence to: Martina Mittlböck, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, Vienna A-1090, Austria.
†E-mail: martina.mittlboeck@meduniwien.ac.at

Received 30 July 2004
Accepted 30 November 2004

$B(n,\pi)$, although the actual outcome are Bernoulli events, $B(1,\pi)$. The inherent potential pitfall with regard to pseudo *R*-squared measures will be exemplified in Section 2 with data on Austrian stillbirths for the period between 1984 and 1998. Section 3 contains a brief formal presentation of pseudo *R*-squared measures for Poisson, logistic binomial and logistic Bernoulli models [1–4]. Section 4 presents a discussion and some practical recommendations on how to use pseudo *R*-squared measures in epidemiological research.

All reported numerical results of generalized linear regression models were obtained by applying the SAS procedure GENMOD (SAS, Version 8, Cary: SAS Institute Inc., 2001). The values for the deviance-based pseudo *R*-squared measures are computed according to Mittlböck and Schemper [5] and Mittlböck [6].

2. EXAMPLE

In Austria, 5255 stillbirths among a total of 1 342 993 births (0.391 per cent) have been observed for the period between 1984 and 1998. Besides the dichotomous outcome for stillbirth/livebirth the data set contains three covariates of interest, these are sex of the infant (dichotomous), age of the mother (6 levels: '19 years or younger', '20–24 years', '25–29 years', '30–34 years', '35–39 years', '40 years or older') and calendar time (five three-year-intervals from '1984–1986' to '1996–1998'). The data set is available on request from the corresponding author. Note that the Austrian stillbirths data are used here for plain exemplification purpose, readers interested in more exhaustive data descriptions and statistical analyses are referred to Waldhör *et al.* [7].

Assume that three researchers independently evaluate the prognostic effect of mothers' age on stillbirth outcome by applying three different models A, B and C. Both models A and B are simple Poisson regression models containing mothers' age as single prognostic factor. The results of models A and B show in unison that age has a statistically significant effect (p-value < 0.0001), and using the AGE category '40 years or older' as reference category for dummy coding, the estimated regression coefficients for '19 years or younger', '20–24 years', '25–29 years', '30–34 years' and '35–39 years' are $-0.91, -1.07, -1.11, -0.96$ and -0.53, respectively. Obviously, mothers aged 25–29 years have the lowest risk for stillbirth. The older the mother the more the risk increases. There seems to be a slight increase in stillbirth risk also for very young mothers. Model C is a simple logistic regression model containing mothers' age as single prognostic factor. The corresponding p-value is < 0.0001 and also the estimated regression coefficients are almost identical with that from the Poisson models A and B (the maximum absolute difference is < 0.01). Somewhat surprisingly, the reported pseudo *R*-squared values differ considerably with 100 per cent for model A, 68.7 per cent for model B and 0.38 per cent for model C, respectively.

The discrepancy in the *R*-squared values can be resolved by considering the number of observations which are 6, 60 and 1 342 993 in models A, B and C, respectively. The six observations of model A are constituted by the six age groups. That is, there have been 344 stillbirths among 81 399 births for mothers aged 19 years or younger, and 1424 stillbirths for 396 253 births for mothers aged 20–24 years, and so forth. The fitting of a Poisson model with six degrees of freedom (five for age and one for the intercept) and an appropriately specified offset (logarithm of birth numbers per group) will inevitably lead to a perfect prediction of the observed six stillbirth numbers. Consequently, the deviance for the full model will be

zero and the corresponding *R*-squared measure will yield the highest attainable value of 100 per cent (see Section 3 for formal details).

The 60 observations of model B are generated by cross-tabulating the categories of infants' sex, calendar time and mothers' age, i.e. $2 \times 5 \times 6 = 60$. That is, each observation of model A is split up into 10 new observations according to infants' sex and calendar time interval. Consequently, the *R*-squared value of 68.7 per cent reflects the ability of model B to predict these 60 observations with six fitted degrees of freedom.

The 1 342 993 observations of model C are the individual births. The quite low *R*-squared value of 0.38 per cent just reflects the inability of the model to individually predict stillbirth from mothers' age. However, this leads to the question what models A and B actually predict? Technically the study outcome (stillbirth *versus* livebirth) is a dichotomous Bernoulli event, $\mathrm{B}(1, \pi)$, and only model C is considering this fact. On the contrary, outcome variables generated by cross-tabulating arbitrarily chosen covariates will follow a binomial distribution, $\mathrm{B}(n, \pi)$. And there is the crucial difference: the Bernoulli outcome corresponds to single infants who represent well-defined and well-interpretable study objects, whereas the study objects of the binomial outcome correspond to the single cells of a cross-classified table, that is, they inevitably lack an authentic and inartificial interpretation. A Poisson approximation of the binomial outcome will inherit this problem. In other words, the Poisson models A and B may be useful for computing regression coefficients and *p*-values, but pseudo *R*-squared measures cannot be sensibly employed to assess the predictive accuracy of the models simply because both the outcome and predicted values are realizations of an arbitrary and artificial construct.

3. THEORY

Consider an outcome variable Y whose distribution belongs to the exponential family [8]. Consider the covariates $X_1 \ldots X_k$ which are either measured on categorized scales with $c_1 \ldots c_k$ levels, respectively, or can be sensibly transformed into such scales, e.g. in the Austrian stillbirths example the age of the mother has been categorized into five-year-intervals. In a generalized linear model the outcome expectation $\mathrm{E}(Y) = \mu$ is linked to a linear predictor *via* the so-called link function, $g(\mu) = \mathbf{Z}_{[1\ldots j]}\boldsymbol{\gamma}$, where $\mathbf{Z}_{[1\ldots j]} = (1, Z_1, \ldots, Z_p)$ is a $(p+1)$-dimensional row vector representing the j covariates of interest, $j \leqslant k$, and $\boldsymbol{\gamma}$ is the corresponding column vector of regression parameters, respectively [8]. The first element of $\mathbf{Z}_{[1\ldots j]}$ is set to a constant value of one to represent the intercept and the other elements are used to represent the j covariates of interest by dummy variables. Actually, any subset of j out of the k covariates can be chosen for modelling purpose, for the sake of notational simplicity it is assumed that only the first j covariates $X_1 \ldots X_j$ are represented in the model and the number of regression parameters is $p + 1 = (\sum_{i=1}^{j} c_i) - j + 1$. A deviance-based pseudo *R*-squared measure can be defined as relative reduction in deviance due to the covariates in the model by $R^2_{\mathrm{Dev}} = 1 - \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{D(\mathbf{y};\hat{\boldsymbol{\mu}}_{\mathbf{0}})}$, where $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ and $D(\mathbf{y};\hat{\boldsymbol{\mu}}_{\mathbf{0}})$ are the (scaled) deviances of the full model containing the covariates of interest, $X_1 \ldots X_j$, and the intercept-only model, respectively [1–4]. The vector $\mathbf{y} = (y_1, \ldots, y_n)'$ contains the observed values of the outcome variable Y; $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)'$ and $\hat{\boldsymbol{\mu}}_{\mathbf{0}} = (\hat{\mu}_{01}, \ldots, \hat{\mu}_{0n})'$ contain the corresponding predicted values for the full and the intercept-only models, respectively. Deviance-based *R*-squared measures are commonly preferred as they correspond to the model fitting criteria of maximum likelihood.

If the outcome for a single subject is a dichotomous event, then the logistic regression model will be commonly employed [8]. That is, a Bernoulli variable $Y \sim \mathrm{B}(1,\pi)$, $0 \leqslant \pi \leqslant 1$, with an expected value $\mathrm{E}(Y) = \mu = \pi$ is considered to depend on covariates and the logistic regression model is defined by $g(\mu) = \mathrm{logit}(\pi) = \log(\pi/(1-\pi)) = \mathbf{Z}_{[1\ldots j]}\boldsymbol{\gamma}$. Now a Bernoulli-specific pseudo R-squared measure, R^2_{Bern}, can be defined by simply inserting the Bernoulli deviance, $\mathrm{Dev}_{\mathrm{Bern}}(\mathbf{y};\boldsymbol{\mu}) = 2\sum_{i=1}^{n}[-y_i \log(\mu_i) - (1-y_i)\log(1-\mu_i)]$, into the generic R^2_{Dev}-formula. R^2_{Bern} allows the quantification of the predictive accuracy of a logistic regression model for a Bernoulli outcome variable [1, 2].

If a k-dimensional table is generated by cross-tabulating $X_1 \ldots X_k$, then the numbers of observed events and exposed subjects in the $m(k) = \prod_{i=1}^{k} c_i$ cells of the table will be generically denoted by $y_{[1\ldots k]}$ and $n_{[1\ldots k]}$, respectively, where $\sum_{[1\ldots k]} n_{[1\ldots k]} = n$. The resulting random outcome variable will follow a binomial distribution, $Y_{[1\ldots k]} \sim \mathrm{B}(n_{[1\ldots k]}, \pi_{[1\ldots k]})$, and its expectation $\mathrm{E}(Y_{[1\ldots k]}) = \mu_{[1\ldots k]} = n_{[1\ldots k]}\pi_{[1\ldots k]}$ can be modelled *via* the logistic regression model again, $\mathrm{logit}(\mu_{[1\ldots k]}/n_{[1\ldots k]}) = \mathbf{Z}_{[1\ldots j]}\boldsymbol{\gamma}$, $j \leqslant k$. Although the estimated regression coefficient vectors and their variance–covariance matrices of both the Bernoulli and the binomial approach will coincide, the corresponding pseudo R-squared measures R^2_{Bern} and R^2_{Bin} will differ simply because the corresponding deviances will differ, respectively [9]. The binomial deviance is

$$\begin{aligned}&\mathrm{Dev}_{\mathrm{Bin}}(\mathbf{y}_{[1\ldots k]};\boldsymbol{\mu}_{[1\ldots k]})\\ &\quad = 2\sum_{[1\ldots k]}\left\{y_{[1\ldots k]}\log\left(\frac{y_{[1\ldots k]}}{\mu_{[1\ldots k]}}\right) + (n_{[1\ldots k]} - y_{[1\ldots k]})\log\left(\frac{n_{[1\ldots k]} - y_{[1\ldots k]}}{n_{[1\ldots k]} - \mu_{[1\ldots k]}}\right)\right\}\end{aligned}$$

In the case of large sample size and rare events the results of the Poisson regression model can be used as a close approximation for the logistic regression model, see e.g. Mittlböck and Heinzl [1] and references therein. If the Poisson approximation is employed, then the numbers of events in the $m(k)$ cells of the cross-classified table will be considered to follow a Poisson distribution with expected value $\mathrm{E}(Y_{[1\ldots k]}) = \mu_{[1\ldots k]}$, and the Poisson regression model is defined by $\log(\mu_{[1\ldots k]}) = \mathbf{Z}_{[1\ldots j]}\boldsymbol{\beta} + \log(n_{[1\ldots k]})$, where $\log(n_{[1\ldots k]})$ is usually called offset. If the Poisson approximation of the logistic regression model is justified, then the estimated regression coefficient vectors will be close, $\boldsymbol{\beta} \simeq \boldsymbol{\gamma}$, as well as their variance–covariance matrices. Naturally, R^2_{Poi} and R^2_{Bin} will then be close as well. The Poisson deviance is

$$\mathrm{Dev}_{\mathrm{Poi}}(\mathbf{y}_{[1\ldots k]};\boldsymbol{\mu}_{[1\ldots k]}) = 2\sum_{[1\ldots k]}\left\{y_{[1\ldots k]}\log\left(\frac{y_{[1\ldots k]}}{\mu_{[1\ldots k]}}\right) - y_{[1\ldots k]} + \mu_{[1\ldots k]}\right\}$$

If $X_1 \ldots X_h$ are used for cross-tabulation, $j \leqslant h \leqslant k$, the number of Poisson observations will be reduced to $m(h) = \prod_{i=1}^{h} c_i$, however, the estimates for $\boldsymbol{\beta}$ and their corresponding variance–covariance matrices will not change (a proof of this fact is available on request from the corresponding author). Contrariwise, the R^2_{Poi}-value will crucially depend on whether $X_1 \ldots X_h$ or $X_1 \ldots X_k$ have been used for cross-tabulation. A basically identical argument holds for the binomial approach as well.

4. DISCUSSION

A pseudo *R*-squared measure quantifies predictive accuracy (or prognostic separation) of a regression model. In other words, the strength of association between covariates and outcome variable is assessed within the framework of the chosen model. The reasonable use of a pseudo *R*-squared measure needs an accurate determination of the intrinsic nature of the outcome variable. In the case of the Austrian stillbirth data the outcome is a Bernoulli variable (i.e. the survival status of the individual newborn) and R^2_{Bern} can be given a valid predictive accuracy interpretation. The cross-tabulation of the data by categorized covariates leads to a binomial outcome or an approximate Poisson outcome, provided that the Poisson approximation of the binomial distribution is justified. However, predictive accuracy interpretations of R^2_{Bin} and R^2_{Poi} have to fail inevitably as they cannot be considered to measure 'groupwise' predictive accuracy, simply because the groups are arbitrarily defined and cannot be given a natural interpretation in terms of study objects.

R^2_{Bin} and R^2_{Poi} can only be used as predictive accuracy measures in the case of *genuine* binomial or Poisson observations, but this rather rarely happens for cross-classified tables. To give an example for genuine Poisson observations we could consider the number of car accidents in the administrative districts of, say, Austria. For some reason we could find it useful to study covariates like the amount of wine-production in the districts, the degree of urbanization, the length of the roads, the amount of truck traffic and the activity-level of the districts traffic police. Now the outcome and the covariates in the Poisson regression model would relate to the study objects 'administrative districts' and the computation of R^2_{Poi} could be a sensible task as its interpretation and prediction would relate to the number of car accidents on districts level (for a word of caution see below). Note that the number of car accidents could also be studied on a more aggregated level such as the federal states of Austria of which each of them consists of a cluster of districts. However, if we would do so then we would have changed the study objects from districts to federal states, and interpretation and prediction of R^2_{Poi} would be related to the number of car accidents on federal states level. Instead of car accidents we could study the number of stillbirths among all births in the administrative districts of Austria, which would provide an example for a genuine binomial outcome variable. The covariates of interest would be related to districts, e.g. the number of maternity clinics in the districts, the number of midwives and the existence of education programmes for mothers-to-be. That is, we would again consider the districts as study objects and R^2_{Bin} could have a sensible interpretation. Note that R^2_{Bern} as a measure of predictive accuracy will commonly result in quite low values compared to R^2_{Bin}, as it is a much easier task to build a model which predicts that, for example, four out of 1000 subjects within a district will suffer a loss than to build a model which can actually identify those four unhappy subjects.

There is a word of caution to be sounded. Both examples provided above do not necessarily avoid difficulties, e.g. overdispersion could be such a difficulty which then would have its own implications to the understanding of predictive accuracy [10]. However, overdispersion should only be addressed if there is a sound explanation for the phenomenon [8]. Another type of difficulty could be ecologic bias [11]. The crucial point is that there is no inherently right use of a pseudo *R*-squared measure, all uses need an appropriately justified regression model.

The counted event numbers within a cross-classified table, which has been generated by cross-tabulating covariates of interest, may be of particular importance for descriptive and

explorative purposes. This is analogous to the common linear regression model context where the prediction focusses on the individual outcome values, whereas groupwise mean values will be used for description, exploration and communication of results. Note that if the common linear model R-squared measure would be calculated for both individual and groupwise mean values, then analogous effects as mentioned here for R^2_{Bern} and R^2_{Bin} would be observed, respectively [9].

Instead of measuring predictive accuracy R^2_{Bin} and R^2_{Poi} could be employed to measure goodness of fit and model improvement within a specific cross-classified table. That is, the bundle of covariates forming the table would be fixed, and therefore the number of artificial binomial or Poisson observations would be fixed as well. Consequently, the absolute R-squared values would be irrelevant but the relationship between R-squared values of different models could be sensibly compared and assessed.

Note that for the cases of small sample size and/or many covariates R-squared measures should be adjusted for bias due to optimism. For deviance-based R-squared measures it has been recommended to use shrinkage-based bias-adjustments [6, 12]. Such adjustments also work well for R-squared measures for over-/underdispersed Poisson models [10].

Finally, note that the arguments which have been brought forward here for the case of deviance-based pseudo R-squared measures also apply to any measure which can be analogously defined for both the logistic and the Poisson regression model, e.g. sum-of-squares-based R-squared measures [1–4].

ACKNOWLEDGEMENTS

We would like to thank an anonymous reviewer for useful comments.

REFERENCES

1. Mittlböck M, Heinzl H. A note on R^2 measures for Poisson and logistic regression models when both models are applicable. *Journal of Clinical Epidemiology* 2001; **54**:99–103.
2. Mittlböck M, Schemper M. Explained variation for logistic regression. *Statistics in Medicine* 1996; **15**: 1987–1997.
3. Cameron AC, Windmeijer FAG. R^2 measures for count data regression models with applications to health-care utilization. *Journal of Business and Economic Statistics* 1996; **14**:209–220.
4. Waldhör T, Haidinger G, Schober E. Comparison of R^2 measures for Poisson regression by simulation. *Journal of Epidemiology and Biostatistics* 1998; **3**:209–215.
5. Mittlböck M, Schemper M. Computing measures of explained variation for logistic regression models. *Computer Methods and Programs in Biomedicine* 1999; **58**:17–24.
6. Mittlböck M. Calculating adjusted R^2 measures for Poisson regression models. *Computer Methods and Programs in Biomedicine* 2002; **68**:205–214.
7. Waldhör T, Haidinger G, Langgassner J, Tuomilehto J. The effect of maternal age and birth weight on the temporal trend in stillbirth rate in Austria during 1984–1993. *Wiener Klinische Wochenschrift* 1996; **108**: 643–648.
8. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
9. Agresti A. *Categorical Data Analysis*. Wiley: New York, 1990.
10. Heinzl H, Mittlböck M. Pseudo R-squared measures for Poisson regression models with over- or underdispersion. *Computational Statistics and Data Analysis* 2003; **44**:253–271.
11. Greenland S. Ecologic *versus* individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology* 2001; **30**:1343–1350.
12. Mittlböck M, Waldhör T. Adjustments for R^2-measures for Poisson regression models. *Computational Statistics and Data Analysis* 2000; **34**:461–472.

