Biometrika (1991), **78**, 3, *pp.* 691–2
Printed in Great Britain

Miscellanea

A note on a general definition of the coefficient of determination

BY N. J. D. NAGELKERKE
International Statistical Institute, 2270 AZ Voorburg, The Netherlands

SUMMARY

A generalization of the coefficient of determination R^2 to general regression models is discussed. A modification of an earlier definition to allow for discrete models is proposed.

Some key words: Discrete probability; Log likelihood; Multiple correlation coefficient; Regression model; Residual variation.

The use of R^2, the coefficient of determination, also called the multiple correlation coefficient, is well established in classical regression analysis (Rao, 1973). Its definition as the proportion of variance 'explained' by the regression model makes it useful as a measure of success of predicting the dependent variable from the independent variables.

It is desirable to generalize the definition of R^2 to more general models, for which the concept of residual variance cannot be easily defined, and maximum likelihood is the criterion of fit. The following generalization, but with misprint $1/n$ replaced by $2/n$ here in (1a) and (1b), was proposed by Cox & Snell (1989, pp. 208–9) and, apparently independently, by Magee (1990); but had been suggested earlier for binary response models by Maddala (1983),

$$-\log(1-R^2)=\frac{2}{n}\{l(\hat{\beta})-l(0)\} \tag{1a}$$

or

$$R^2=1-\exp\left[-\frac{2}{n}\{l(\hat{\beta})-l(0)\}\right]=1-\{L(0)/L(\hat{\beta})\}^{2/n}, \tag{1b}$$

where $l(\hat{\beta})=\log L(\hat{\beta})$ and $l(0)=\log L(0)$ denote the log likelihoods of the fitted and the 'null' model respectively.

It is easily found that this definition of R^2 has the following properties.

(i) It is consistent with classical R^2, that is the general definition applied to e.g. linear regression yields the classical R^2.

(ii) It is consistent with maximum likelihood as an estimation method, i.e. the maximum likelihood estimates of the model parameters maximize R^2.

(iii) It is asymptotically independent of the sample size n.

(iv) It has an interpretation as the proportion of explained 'variation', or rather, $1-R^2$ has the interpretation of the proportion of unexplained 'variation'. Variation should be construed very generally as any measure of the extent to which a distribution is not degenerate. To clarify, let M_1 be a model nested under M_2 which is nested under M_3, for example model M_1 contains only covariable x_1, for example a constant, while M_2 contains x_2 and x_1 and M_3 contains x_1, x_2 and

x_3 as covariables. Let $R^2_{2.1}$ denote the R^2 of M_2 relative to M_1, etc.; then

$$(1-R^2_{3.1})=(1-R^2_{3.2})(1-R^2_{2.1}). \tag{2}$$

In other words, the proportion of variation unexplained by model M_3 relative to model M_1 is the product of the proportion of variation unexplained by M_3 relative to M_2 and the proportion unexplained by M_2 relative to M_1.

(v) It is dimensionless, i.e. it does not depend on the units used.

(vi) Replacing the factor $2/n$ in (1a) and (1b) by k/n yields a generalization of the proportion of the kth central moment explained by the model.

(vii) Let y have a probability density $p(y\,|\,\beta x+\alpha)$, then using Taylor expansion, it can be shown that to a first order approximation, R^2 is the square of the Pearson correlation between x and the efficient score of the model $p(\,.\,)$, that is the derivative with respect to β of $\log\{p(y\,|\,\beta x+\alpha)\}$ at $\beta=0$.

However, R^2 thus defined achieves a maximum of less than 1 for discrete models, i.e. models whose likelihood is a product of probabilities, which have a maximum of 1, instead of densities, which can become infinite. This maximum equals

$$\max(R^2)=1-\exp\{2n^{-1}l(0)\}=1-L(0)^{2/n}.$$

For logistic regression, with 50% $y=1$ and 50% $y=0$ observations, this maximum equals 0·75. This maximum occurs when all observations are predicted with maximum probability, that is $\text{pr}\,(y=1)=1$ for the observations with $y=1$, and $\text{pr}\,(y=1)=0$ for the $y=0$ observations. This is clearly unacceptable for a R^2 coefficient. The same problem, but to a lesser degree, exists for Cox's model (Cox, 1972) with $l(\,.\,)$ being the logarithm of the partial likelihood (Cox, 1975).

We therefore propose to redefine R^2 as

$$\bar{R}^2=R^2/\max(R^2). \tag{3}$$

Properties (i), (ii), (iii), (v) and (vi) are automatically satisfied. Property (vii) reduces to first order proportionality, instead of equality, of R^2 and the Pearson correlation coefficient. Property (iv), that is (2), is more difficult to establish. However, from

$$\log\{1-\max(R^2_{2.1})\}-\log\{1-\max(R^2_{3.2})\}=\log(1-R^2_{2.1}) \tag{4}$$

criterion (iv) can also be established to hold for $\bar{R}^2$.

Acknowledgement

The author is grateful to an anonymous referee for many useful suggestions.

References

COX, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.
COX, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.
COX, D. R. & SNELL, E. J. (1989). *The Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.
MADDALA, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.
MAGEE, L. (1990). R^2 measures based on Wald and likelihood ratio joint significance tests. *Am. Statistician* **44**, 250–3.
RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

[*Received June* 1990. *Revised January* 1991]