

Designing a Pseudo R-Squared Goodness-of-Fit Measure in Generalized Linear Models

H. I. Mbachu

Dept. of Mathematics/Statistics, University of Port Harcourt, Port Harcourt

E. C. Nduka

Dept. of Mathematics/Statistics, University of Port Harcourt, Port Harcourt

M. E. Nja (Corresponding author)

Dept. of Mathematics/Statistics, Cross River University of Technology, Calabar

Received: December 19, 2011 Accepted: January 4, 2012 Published: April 1, 2012

doi:10.5539/jmr.v4n2p148 URL: <http://dx.doi.org/10.5539/jmr.v4n2p148>

Abstract

The coefficient of determination is a function of residuals in the General Linear Models. The deviance, logit, standardized and the studentized residuals were examined in generalized linear models in order to determine the behaviour of residuals in this class of models and thereby design a new pseudo R-squared goodness-of-fit measure. The Newton-Raphson estimation procedure was adopted. It was observed that these residuals exhibit patterns that are unique to the subpopulations defined by levels of categorical predictors. Residuals block on the basis of signs, where positive signs indicate success responses and negative signs failure responses. It was also observed that the deviance is a close approximation of the studentized residual. The logit residual is two times the size of the standardized residuals. Borrowing from the Nagelkerke’s improvement of Cox and Snell’s goodness-of-fit measure in generalized linear models and the coefficient of determination counterpart of the general linear model, a new pseudo R squared goodness-of-fit test which uses predicted probabilities and a monotonic link function is here proposed to serve both the linear and Generalized Linear Models.

Keywords: Deviance, Normalized residuals, Logit, Standardized residuals, Loglikelihood function, Response probability

1. Introduction

A generalized linear model is one in which each component of the response variable Y has a distribution in the exponential family, taking the form

$$f_y(y, \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\theta)} + c(y, \phi)\right\}$$

for some specific function $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ (McCullagh & Nelder, 1990). The functions a and c are such that $a(\phi) = \phi/w$ and $c = c(y, \phi/w)$, where w is a known weight for each observation. The model can be stated as

$$z_i = \sum_{j=1}^p x_{ij}\beta_j + e_i h'(\mu) = \sum_{j=1}^p x_{ij}\beta_j + e_i(y_i - \mu_i)h'(\mu), \quad i = 1, 2, 3, \dots, n \tag{1}$$

where z_i is the adjusted dependent variate, x_{ij} is the (i, j) th element of the design matrix, $h(\mu_i)$ is the link function and e_i is the residual error. The link between y_i and z_i is in the expression.

$$h_i = h(\mu_i) \tag{2}$$

Where y_i is a binomial random response variable.

From (1), a residual in generalized linear model can be defined as

$$e_i = \frac{z_i - x_{ij}\beta_j}{h'(\mu)} \tag{3}$$

e_i , so defined is called Pearson residual.

Standard theory for this type of distribution expresses the mean and variance of the response y as:

$$E(y) = b'(\theta) \text{ and } \text{var}(y) = \frac{b''(\theta)\phi}{w} = \frac{V(\mu)\phi}{w}$$

where V is the variance function.

The log-likelihood function, a goodness-of-fit measure is defined for the following exponential family models:

Generally, the log-likelihood function is of the form

$$L(y, \mu, \phi) = \sum_i \log(f(y_i, \mu_i, \phi))$$

with individual contribution for the binomial function as

$$l_i = [r_i \log(p_i) + (n_i - r_i) \log(1 - p_i)]$$

2. The Newton-Raphson Method

The Newton-Raphson estimation scheme is given as

$$\beta^{k+1} = \beta^k - H_g^{-1}$$

where H , the Hessian matrix is given as

$$H = \left\{ \frac{\partial^2 l}{\partial \beta_r \partial \beta_s} \right\}_{rs}$$

with

$$\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = \Sigma[(y - \mu) \frac{\partial}{\partial \beta_s} \{W \frac{d\eta}{d\mu} x_r\} + W \frac{d\eta}{d\mu} x_r \frac{\partial}{\partial \beta_s} (y - \mu)]$$

and

$$\frac{\partial^2 l}{\partial \beta_j^2} = \Sigma[(y - \mu) \frac{\partial}{\partial \beta_j} \{W \frac{d\eta}{d\mu} x_j\} + W \frac{d\eta}{d\mu} x_j \frac{\partial}{\partial \beta_j} (y - \mu)]$$

$$\frac{\partial l}{\partial \beta_j} = \left[\frac{y - \mu}{a(\phi)} \frac{1}{v} \frac{d\mu}{d\eta} x_j \right] = \frac{W}{a(\phi)} (y - \mu) \frac{d\eta}{d\mu} x_j$$

1, the loglikelihood for a binary response variable can be written as

$$l = l(\beta; y) = \Sigma_i \Sigma_j y_i x_{ij} \beta_j - \Sigma m_i \log(1 + \exp \Sigma x_{ij} \beta_j)$$

$\eta = \beta_0 + \Sigma x_{ij} \beta_j$ is the linear predictor.

W , the weight matrix is given as $W = \text{diag}\{m_i (\frac{d\mu_i}{d\eta_i})^2 / \mu_i (1 - \mu_i)\}$.

m_i is row subtotal in the cross tabulation table. The gradient vector g is given as

$$g = \left(\frac{\partial l}{\partial \beta_0}, \frac{\partial l}{\partial \beta_1}, \dots, \frac{\partial l}{\partial \beta_n} \right) = \frac{\partial l}{\partial \beta_r} = \Sigma \frac{y_i - m_i \mu_i}{\mu_i (1 - \mu_i)} \frac{d\mu_i}{d\eta_i} = \Sigma (y_i - m_i \mu_i) x_{ir}$$

where the response or fitted probability μ_i is defined as

$$\mu_i = \frac{\exp \Sigma x_{ij} \beta_j}{1 + \exp \Sigma x_{ij} \beta_j}$$

An alternative estimation procedure is the Iterative Weighted Least Squares method which often adopted in order to avoid the computational tedium associated with the Hessian matrix.

3. Residuals in Generalized Linear Models

The coefficient of determination R^2 , is a function of the residual. It was originally developed for the normal-theory model. Cameron and Windmeijer (1996) designed an R^2 for the Poisson and related count data after observing that it was rarely used for count data. Nagelkerke (1991) generalized the definition of R^2 in what is called the generalized R^2 . The generalized R^2 is consistent with the classical R^2 and is also maximized by the maximum likelihood estimation of a model. The generalized coefficient of determination is given as follows:

$$R^2 = 1 - \left(\frac{L(0)}{L(\theta)} \right)^{\frac{2}{n}}$$

where $L(0)$ is the likelihood of the model with only intercept. $L(\theta)$ is the likelihood of the estimated model and n is the sample size. Residuals in a logistic model can be defined as the difference between y_i and the predicted probability θ for

y_i . We define the predicted probability in a cross-classified data as the probability that an object or a person selected from a subgroup is a success (Stroke *et al.*, 1997).

$$\theta = \frac{\exp\{\beta_0 + \sum \beta_i x_{ij}\}}{1 + \exp\{\beta_0 + \sum \beta_i x_{ij}\}}$$

The monotonic link function relates the predicted probability to the set of linear predictors. For the logistic regression where the underlying distribution is binomial, the link function is a logit. The deviance, Pearson χ^2 , standardized, logit and studentized residuals are the residuals normally associated with generalized linear models. The analysis of residuals made in this paper shows that the logit residual is approximately twice the size of standardized residuals. The standardized residual is approximately equal to the deviance residual. This can be seen in the appendix.

4. Goodness of Fit Measures in Generalized Linear Models

The deviance and the generalized Pearson χ^2 statistic are two measures of goodness of fit in generalized linear models. Both the deviance and the generalized Pearson χ^2 have exact χ^2 distributions for Normal-theory linear models if the models are true (McCullagh & Nelder, 1990). The deviance uses the log of the ratio of likelihoods. Cox and Snell R squared, another measure of goodness of fit in generalized linear models is a pseudo R squared and a modification of the deviance which configures the test interval to lie between 0 and 1 (excluding 1) such that a smaller ratio implies a greater improvement.

The deviance for the set of distributions in generalized linear models is given as follows: for the normal distribution, it is stated as

$$D = \sum w_i (y_i - \mu_i)^2$$

For the poisson, binomial and gamma we have

$$2 \sum_i w_i [y_i \log(\frac{y_i}{\mu_i}) - (y_i - \mu_i)],$$

$$2 \sum_i w_i m_i [y_i \log(\frac{y_i}{\mu_i}) + (1 - y_i) (\log \frac{1 - y_i}{1 - \mu_i})]$$

and

$$2 \sum_i w_i [-\log(\frac{y_i}{\mu_i}) + \frac{y_i - \mu_i}{\mu_i}]$$

respectively. For the inverse-Gaussian, multinomial and negative binomial, we have

$$\sum_i \frac{w_i (y_i - \mu_i)^2}{\mu_i^2 y_i}$$

$$\sum_i \sum_j w_i y_{ij} \log(\frac{y_{ij}}{p_{ij} m_i})$$

and

$$2 \sum_i w_i [y \log(y/\mu) - (1 + 1/k) \log(\frac{y + 1/k}{\mu + 1/k})]$$

respectively. Cox and Snell R^2 is defined as

$$R^2 = 1 - \left\{ \frac{L(m_{int})}{L(m_{full})} \right\}^{2/N}$$

where $L(m_{int})$ is the conditional probability of the dependent variable for the intercept model. If $L(m_{full})$ is 1 then $R^2 < 1$. The Nagelkerke/Gragg & Uhler's modification is

$$R^2 = 1 - \left\{ \frac{L(m_{int})}{L(m_{full})} \right\}^{2/N} / 1 - L(m_{int})^{2/N}$$

In this paper a new goodness of fit test that makes use of fitted probabilities, a monotonic link function and the Nagelkerke range of possible values is proposed. The test is designed to serve both the general linear and the generalized linear models. It is given as follows:

$$R_{G\&G}^2 = 1 - \frac{[h'(\theta)]^{-1} \sum (y - \theta)^2}{\sum (y - h(\theta))}$$

$R_{G\&G}^2$, designed for the generalized linear models can be adapted for use as a goodness of fit measure in the general linear model by replacing the fitted probabilities and the link function values with fitted y values and the mean of y respectively. The value of $R_{G\&G}^2$ range from 0 to 1, with higher values implying better fits.

5. Illustrative Example

The hypothetical data below is used for the illustration of residual analysis in generalized linear models:

<Table 1>

The probability that a person from the i th sex level and the j th location status is infected with a certain virus.

The model

Let y_{ij} be a binomial random response variable corresponding to the i th sex status and the j th location which assumes the value 0 or 1. The probability θ_{ij} ; that a person of the i th sex and j th location is infected by the virus is modeled as

$$\theta_{ij} = \exp \frac{\exp[\beta_0 + \text{sex}(i) + \text{location}(j)]}{1 + \exp[\beta_0 + \text{sex}(i) + \text{location}(j)]}$$

where $i = 1, 2, j = 1, 2,$

β_0 = overall mean

$\text{sex}(i)$ = effect of i th sex level = β_1

$\text{location}(j)$ = effect of j th location status = β_2

e_{ij} = random error associated with observation. The Newton-Raphson estimates of the illustrative example are as follows:

Solution

$\beta_0 = 1.1568, \beta_1 = -1.2770$ is the effect of the i th sex level. $\beta_2 = -1.0545$ is the effect of the j th location status. The pseudo-R squared goodness of fit test reveals the following results:

Cox and Snell $R^2 = 0.140$

Nagelkerke/Gragg & Uhler's $R^2 = 0.187$

The proposed $R_{G\&G}^2 = 0.180$

The outlined residuals associated with this example are shown in the appendix. It is observed that residuals exhibit unique patterns in accordance with subpopulations defined by levels of the categorical variables. Residuals form blocks on the basis of signs, where positive signs indicate success and negative signs indicate failure responses. The deviance and the studentized residuals exhibit very close residual patterns.

Stat Computing (2011) gave three interpretations of R^2 as follows: (i) R^2 as explained variability: The denominator of the ratio indicates total variation in the dependent variable while the numerator is the variability in the dependent variable that is not predicted by the model. The ratio is the proportion of the total variability explained by the model which agrees with R^2 in Ordinary Linear Models (Koutsoyiannis, 1983). Thus a higher ratio implies a better model.

(ii) R^2 as improvement from null model to fitted model: A smaller ratio implies a greater improvement.

(iii) R^2 as the square of the correlation: correlation between predicted values and the actual values. A higher R^2 implies a greater improvement of fit.

It can be seen that the proposed R^2 goodness-of-fit measure compares favourably with the Nagelkerke/Gragg & Uhler's R^2 (0.180 against 0.187).

6. Conclusion

The Nagelkerke/Gragg & Uhler's Improvement of Cox and Snell R^2 is applicable in Generalized Linear models only. The existing R squared goodness of fit measure in General Linear models is not applicable in Generalized Linear model. This is because the model estimates from Generalized Linear models are maximum likelihood estimates which are obtained by iterative procedures. They are not calculated to minimize variance; so the Ordinary Least Squares approach to goodness of fit does not apply. To evaluate goodness of fit in generalized linear models a pseudo R^2 is required. This paper introduces a new pseudo R squared goodness of fit measure which has the advantage of assessing goodness of fit in both linear and generalized linear models. The result shows that the new pseudo-R squared method designed in this paper compares favourably with the existing Nagelkerke/Gragg & Uhler's design.

References

Cameron, A. C., & Windmeyer F. A. G. (1996). R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization. *Journal of Business and Economic Statistic*.

Koutsoyiannis, A. (1983). *Theory of Econometrics: An Introductory Exposition of Econometric Methods*. 2 Ed. The Macmillan Press Ltd, London.

McCullagh, P., & Nelder, J. A. (1990). *Generalized Linear Models*. Chapman and Hall. Madras.

Nagelkerke, N. (1991). A note on a General Definition of the Coefficient of Determination. *Biometrika*, 78 (3), pp. 691-692. <http://dx.doi.org/10.1093/biomet/78.3.691>

Nja, M. E., & Bamiduro, T. A. (2006). Relative performance of Optimization Methods In Solutions of Generalized Linear Models. An unpublished Ph. D thesis, University of Ibadan, Nigeria.

Stoke, M. E., Davis, C. S., & Koch, G. G. (1997). *Categorical Data Analysis using the SAS system*, SAS Institute Inc., Cary, NC, USA.

Table 1. Hypothetical data

I	Sex x_1	Location x_2	Infected y_i	Not infected	Total m_i
1	Female	Urban	4	11	15
2	Female	Rural	8	10	18
3	Male	Urban	9	9	18
4	Male	Rural	21	6	27

Appendix: Residuals

			PRE.1	COO.1	LEV.1	RES.1	LRE.1	SRE.1	ZRE.1	DEV.1	DFB0.1	DFB1.1	DFB2
1.00	.0	.0	.23601	.14361	.04248	.76399	4.23710	1.73663	1.79919	1.69934	-.05891	.12347	.12341
1.00	.0	.0	.23601	.14361	.04248	.76399	4.23710	1.73663	1.79919	1.69934	-.05891	.12347	.12341
1.00	.0	.0	.23601	.14361	.04248	.76399	4.23710	1.73663	1.79919	1.69934	-.05891	.12347	.12341
1.00	.0	.0	.23601	.14361	.04248	.76399	4.23710	1.73663	1.79919	1.69934	-.05891	.12347	.12341
1.00	.0	1.00	.46999	.05116	.04340	.53001	2.12770	1.25642	1.06193	1.22885	.02467	.07185	-.05169
1.00	.0	1.00	.46999	.05116	.04340	.53001	2.12770	1.25642	1.06193	1.22885	.02467	.07185	-.05169
1.00	.0	1.00	.46999	.05116	.04340	.53001	2.12770	1.25642	1.06193	1.22885	.02467	.07185	-.05169
1.00	.0	1.00	.46999	.05116	.04340	.53001	2.12770	1.25642	1.06193	1.22885	.02467	.07185	-.05169
1.00	.0	1.00	.46999	.05116	.04340	.53001	2.12770	1.25642	1.06193	1.22885	.02467	.07185	-.05169
1.00	.0	1.00	.46999	.05116	.04340	.53001	2.12770	1.25642	1.06193	1.22885	.02467	.07185	-.05169
1.00	.0	1.00	.46999	.05116	.04340	.53001	2.12770	1.25642	1.06193	1.22885	.02467	.07185	-.05169
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428
1.00	1.00	.0	.52555	.04097	.04341	.47445	1.90278	1.15975	.95015	1.13430	.02207	-.04625	.06428

			<i>PRE</i> ₁	<i>COO</i> ₁	<i>LEV</i> ₁	<i>RES</i> ₁	<i>LRE</i> ₁	<i>SRE</i> ₁	<i>ZRE</i> ₁	<i>DEV</i> ₁	<i>DFB0</i> ₁	<i>DFB1</i> ₁	<i>DFB2</i>
.0	.0	1.00	.46999	.04023	.04340	-.46999	-1.88676	-1.15210	-.94168	-1.12682	-.02188	-.06371	.04584
.0	1.00	.0	.52555	.05027	.04341	-.52555	-2.10769	-1.24854	-1.05247	-1.22114	-.02444	.05123	-.07120
.0	1.00	.0	.52555	.05027	.04341	-.52555	-2.10769	-1.24854	-1.05247	-1.22114	-.02444	.05123	-.07120
.0	1.00	.0	.52555	.05027	.04341	-.52555	-2.10769	-1.24854	-1.05247	-1.22114	-.02444	.05123	-.07120
.0	1.00	.0	.52555	.05027	.04341	-.52555	-2.10769	-1.24854	-1.05247	-1.22114	-.02444	.05123	-.07120
.0	1.00	.0	.52555	.05027	.04341	-.52555	-2.10769	-1.24854	-1.05247	-1.22114	-.02444	.05123	-.07120
.0	1.00	.0	.52555	.05027	.04341	-.52555	-2.10769	-1.24854	-1.05247	-1.22114	-.02444	.05123	-.07120
.0	1.00	.0	.52555	.05027	.04341	-.52555	-2.10769	-1.24854	-1.05247	-1.22114	-.02444	.05123	-.07120
.0	1.00	.0	.52555	.05027	.04341	-.52555	-2.10769	-1.24854	-1.05247	-1.22114	-.02444	.05123	-.07120
.0	1.00	1.00	.76075	.09713	.02964	-.76075	-4.17967	-1.71693	-1.78316	-1.69129	-.12768	.09276	.09280
.0	1.00	1.00	.76075	.09713	.02964	-.76075	-4.17967	-1.71693	-1.78316	-1.69129	-.12768	.09276	.09280
.0	1.00	1.00	.76075	.09713	.02964	-.76075	-4.17967	-1.71693	-1.78316	-1.69129	-.12768	.09276	.09280
.0	1.00	1.00	.76075	.09713	.02964	-.76075	-4.17967	-1.71693	-1.78316	-1.69129	-.12768	.09276	.09280
.0	1.00	1.00	.76075	.09713	.02964	-.76075	-4.17967	-1.71693	-1.78316	-1.69129	-.12768	.09276	.09280
.0	1.00	1.00	.76075	.09713	.02964	-.76075	-4.17967	-1.71693	-1.78316	-1.69129	-.12768	.09276	.09280

Key:

<i>PRE</i> ₁	Predicted probability
<i>COO</i> ₁	Analog of Cook’s influence statistics
<i>LEV</i> ₁	Leverage value
<i>RES</i> ₁	Difference between observed and predicted probabilities
<i>LRE</i> ₁	Logit Residual
<i>SRE</i> ₁	Standard Residual
<i>ZRE</i> ₁	Normalized Residual
<i>DEV</i> ₁	Deviance value
<i>DFB0</i> ₁	DFBeta for constant
<i>DFB1</i> ₁	DFBeta for VAR00002(1)
<i>DFB2</i> ₁	DFBeta for VAR00003(1)