Measuring Deviance Goodness-of-Fit For Poisson Regression Model with Application to Count Data in Medicine

Isaac O. Ajao[1]; Adebisi A. Ogunde[2]

Department of Mathematics and Statistics,
The Federal Polytechnic, Ado-Ekiti,
Ado-Ekiti, Nigeria.
e-mail: isaacoluwaseyiajao@gmail.com[1]; debiz95@yahoo.com[2]

***Abstract*—Health researchers all over the world are often concerned with rare or infrequently occurring, repeatable, health-related events such as number of auto accidents, twins, caesarean sections and so on. Cases of the occurrence of such discrete events take the form of non-negative integer or count data. Because the counts of rare events tend to be non-normally distributed and highly positively skewed, the use of ordinary least squares (OLS) regression with non-transformed data has many lapses. The Poisson regression is an appropriate alternative for analyzing these data. This research is focused at analyzing dataset on low birth weight with the aim of obtaining: summary statistics, a well fitted model, and deviance goodness of fit. The result obtained shows that *low birth weight* is highest in the year 2016, also *number of births* makes the smallest contribution ($\beta = 0.0206$). We conclude that the model fits reasonably well because the goodness-of-fit chi-squared test is not statistically significant with p-value of 0.8323 at 95% confidence interval.**

Keywords: *Poisson regression, deviance, low-birth weight, count data, medicine.*

I. Introduction

A great deal of the data collected by scientists, medical statisticians and economists, however, is in the form of *counts* (whole numbers or integers). The number of individuals that died, the number of firms going bankrupt, the number of days of frost, the number of red blood cells on a microscope slide, or the number of craters in a sector of lunar landscape are all potentially interesting variables for study. With count data, the number 0 often appears as a value of the response variable.

The first application of Poisson regression was given by [4] with wireworm counts from an agriculture experiment as the response variable and the regression function $E(yi) = \mu i = (xi\ \beta)^2$, where xi is the ith row of the model matrix for a Latin square design. This model is a GLM with a Poisson response and a "square root" link function [6] for additional details. [7] proposed Poisson regression with a linear rate function for use in consumer demand analyses and reliability. [10] described GLM for response variables in the regular exponential family, also [9] for Poisson log-linear models. [6] described Poisson regression methods for general models with an emphasis on intrinsically nonlinear models. [6] described the analysis of event rates for Poisson data and [5] considered applications of these methods in epidemiologic follow-up studies. [2] described the use of Poisson regression in occupational cohort studies and [7] gave a more complete review of Poisson regression methods and areas of applications. [3] discussed Poisson regression in econometric applications. [12] reviewed the use of Poisson regression in occupational and environmental cohort studies and considered problems that may occur when person-time and events are tabulated by levels of an exposure variable that was originally measured on a continuous scale and has been categorized for analysis.

Straightforward linear regression methods (constant variance, normal errors) are not appropriate for count data for five main reasons:

i. the linear model might lead to the prediction of negative counts
ii. the variance of the response variable is likely to increase with the mean
iii. the errors will not be normally distributed
iv. zeros are difficult to handle in transformations
v. some distributions (e.g. log-normal or gamma) don't allow zeros

II. Research Methodology

2.1 The Poisson distribution

The Poisson distribution is widely used for the description of count data that refer to cases where we know how many times something happened (e.g. lightning strikes, bomb hits), but we have no way of knowing how many times it did not happen. This is in contrast to the binomial distribution where we know how many times something did not happen as well as how often it did happen (e.g. if we got 6 heads out of 10 tosses of a coin, we must have got 4 tails).

In the case of binary regression the fact that probability lies between 0-1 imposes a constraint. The normality assumption of multiple linear regression is lost, and so also is the assumption of constant variance. Without these assumptions the *F* and *t* tests have no basis. The solution was to use the logistic transformation of the probability p or logit p, such that

$$\log_e(p/1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots \beta_n X_n.$$

The β coefficients could now be interpreted as increasing or decreasing the log odds of an event, and expβ (the odds multiplier) could be used as the odds ratio for a unit increase or decrease in the explanatory variable. In survival analysis we used the natural logarithm of the hazard ratio, that is

$$\log_e h(t)/h_0(t) = \beta_0 + \beta_1 X_1 + \ldots.. + \beta_n X_n$$

When the response variable is in the form of a count we face a yet different constraint. Counts are all positive integers and for rare events the Poisson distribution (rather than the Normal) is more appropriate since the Poisson mean > 0. So the logarithm of the response variable is linked to a linear function of explanatory variables such that $\log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$ etc. and so $Y = e^{\beta 0} \times e^{\beta 1 X 1} \times e^{\beta 2 X 2} \ldots$ etc.

In other words, the typical Poisson regression model expresses the log outcome rate as a linear function of a set of predictors. The Poisson is a 1-parameter distribution, specified entirely by the mean. The variance is identical to the mean, so the variance/mean ratio is equal to one. In this paper we are studying the number of low birth weight babies (LBW) per month in a Teaching Hospital, and our data consist of the numbers of LBW babies per month (*x*). Some months have no LBW babies at all, but some may have as many as 5 or 6 cases. If the mean number of LBW babies per month is λ, then the probability of observing *x* LBW babies per month is given by:

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

$$P(x) = P(x-1)\frac{\lambda}{x} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

This means that if we start with the *zero term*

$$P(0) = e^{-\lambda} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

then each successive probability is obtained simply by multiplying by the mean and dividing by *x*. Log link function when Poisson errors are specified

$$y = \exp(\beta_i x_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

2.2 Overdispersion in Poisson

Overdispersion is often encountered when fitting very simple parametric models, such as those based on the Poisson distribution. The Poisson distribution has one free parameter and does not allow for the variance to be adjusted independently of the mean. The choice of a distribution from the Poisson family is often dictated by the nature of the empirical data. If overdispersion is a feature in Poisson regression, an alternative model with additional free parameters may provide a better fit. In the case of count data, a Poisson mixture model like the negative binomial distribution can be proposed instead, in which the mean of the Poisson distribution can itself be thought of as a random variable drawn – in this case – from the gamma distribution thereby introducing an additional free parameter, the resulting negative binomial distribution is completely characterized by two parameters. Overdispersion can be detected if residual deviance is much larger than the residual degrees of freedom.

Selection criteria: According to [1] Akaike information criterion (AIC) is defined as:

$$AIC = -2\ln L + 2k$$

where ln*L* is the maximized log-likelihood of the model and *k* is the number of parameters estimated. Some authors define the AIC as the expression above divided by the sample size.

Bayesian information criterion (BIC) [13] is another measure of fit defined as:

$$BIC = -2\ln L + k\ln N$$

where N is the sample size.

2.3 Deviance with Poisson errors

Up to this point, lack of fit has always been measured by SSE; the residual or error sum of squares

$$SSE = \sum (y - \hat{y})^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(5)$$

where $\hat{y}$ are the fitted values estimated by the model. With Generalized Linear models SSE is only the maximum likelihood estimate of lack of fit when the model has normal

errors and the identity link. Generally, we use **deviance** to measure lack of fit in a Generalized Linear model.
For Poisson errors the deviance is:

$$\text{Poisson deviance} = 2\sum O \ln\left(\frac{O}{E}\right) \quad \ldots\ldots\ldots\ldots(6)$$

where O is the Observed count, and E is the Expected count as predicted by the current model.

III. ANALYSIS

Table 1: Number of LBW by various maternal attributes

Month	Factor	MAB	Unbook mothers	Birth	LBW
Sep, 2014	A	31	0	3	4
Oct, 2014	A	31	15	31	5
Nov, 2014	A	32	2	14	1
Feb, 2015	B	29	3	16	2
Mar, 2015	B	31	7	27	4
Apr, 2015	B	32	9	38	5
May, 2015	B	31	5	14	6
Jun, 2015	B	29	1	9	1
Jul, 2015	B	33	1	3	0
Sep, 2015	B	29	0	3	0
Oct, 2015	B	32	5	25	0
Nov, 2015	B	31	4	24	6
Dec, 2015	B	31	5	27	4
Jan, 2016	C	31	6	27	4
Feb, 2016	C	32	9	31	10
Mar, 2016	C	31	6	33	8
Apr, 2016	C	32	5	19	4
Jun, 2016	C	32	3	30	5
Jul, 2016	C	31	4	31	5

MAB is the mean age of the mothers and LBW the number of low birth weight babies.

IV. RESULTS

From Table 1, it can be seen that year 2014 has 10 counts, year 2015, has 28 counts, while year 2016 has 36 counts. The total deviance (like sum of square total SST) is based on the whole sample of 19 numbers. The expected count is the overall mean which is 3.89**.** The total deviance can then be calculated thus using equation (6)

$$= 2\times\left[4\ln\left(\frac{4}{3.89}\right)+5\ln\left(\frac{5}{3.89}\right)+\ln\left(\frac{1}{3.89}\right)+2\ln\left(\frac{2}{3.89}\right).....+5\ln\left(\frac{5}{3.89}\right)\right]$$

$$= 43.7036$$

Now the three years means are $10/3$, $28/9$ and $36/6$ respectively. The residual deviance after fitting a 3–level factor for year should therefore be

$$= 2\times\left[4\ln\left(\frac{4}{3.33}\right)+5\ln\left(\frac{5}{3.33}\right)+\ln\left(\frac{1}{3.33}\right)+2\ln\left(\frac{2}{2.8}\right).....+5\ln\left(\frac{5}{6}\right)\right]$$

$$= 28.2823$$

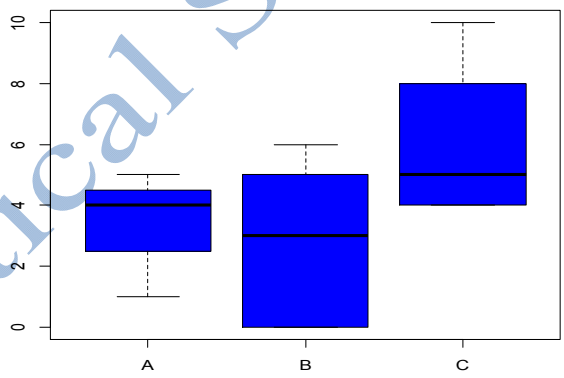

Fig 1: Boxplot showing the median of LBW in the factor levels

Table 2: Poisson estimates

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.20102	4.62038	0.260	0.795
FactorsB	-0.02628	0.45831	-0.057	0.954
FactorsC	0.49452	0.49710	0.995	0.320
MAB	-0.02907	0.15018	-0.194	0.847
Unbooked	0.05366	0.05489	0.978	0.328
No_of_births	0.02467	0.02212	1.115	0.265

Test for overdispersion

Null deviance: 43.702 on 18 degrees of freedom
Residual deviance: 24.759 on 13 degrees of freedom
AIC: 89.034

For model 1: The response variable, number of low birth weight (lbw) is regressed on one of the explanatory variables (mean age of mothers); then add other explanatory variables in the regression analysis

Table 2: Analysis of Deviance for model 1

	df	deviance	mean deviance	deviance ratio	AIC
Regression	1	1.5355	1.5355	1.54	5.1811
Residual	17	42.1666	2.4804		
Total	18	43.7022			

Table 3: Estimates of regression coefficients

	Estimates	s.e	z
Constant	-2.9927	3.5907	-0.83
Mean age of mothers	0.1395	0.1148	1.22

The regression equation may now be written as:
$\text{Log}_e(Y) = \beta_0 + \beta_1 X_1$
On substituting the values of Y and X, the equation can be written as:
$\text{Log}_e(\text{LBW}) = -2.9927 + 0.1395\ \text{MAB}$
Which leads to $\text{LBW} = (e)^{-2.9927} \times (e)^{0.1395\ \text{MAB}}$

Table 4: Analysis of Deviance for model 2

	df	deviance	mean deviance	deviance ratio	AIC
Regression	2	7.0691	3.5345	3.54	4.9952
Residual	16	36.6331	2.2895		
Total	18	43.7022			

Table 5: Estimates of regression coefficients

	estimates	s.e	z
Constant	-2.7324	3.5907	-0.75
Mean age of mothers	0.1008	0.1148	0.85
year	0.4194	0.1826	2.30

Table 6: Analysis of Deviance for model 3

	df	deviance	mean deviance	AIC
Regression	3	17.0977	5.6992	4.5726
Residual	15	26.6045	1.7736	
Total	18	43.7022		

Table 7: Estimates of regression coefficients

	estimates	s.e	z
Constant	-0.4544	4.2567	-0.11
Mean age of mothers	0.0032	0.1412	0.02
year	0.4902	0.1848	2.65
Unbooked mothers	0.1119	0.0347	3.22

Table 8: Analysis of Deviance for model 4

	df	deviance	mean deviance	deviance ratio	AIC
Regression	4	18.0486	4.5122	4.5122	4.6278
Residual	14	25.6536	1.8324		
Total	18	43.7022			

Table 9: Estimates of regression coefficients

	estimates	s.e	z
Constant	0.0520	4.4370	0.01
MAB	0.1470	-0.0900	0.93
year	0.2243	1.6400	0.10
Unbooked	0.0740	0.0526	1.41
number of births	0.0206	0.0213	0.97

This is the full model for which the regression equation may be written as:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

On substituting the values we have
LBW in 2014
= 1.0537 x 1.1526 x 1.0768 x 1.0208 = 1.3275
LBW in 2015
= 1.0537 x 1.1526 x 2.1536 x 1.0208= 2.6699
LBW in 2016
= 1.0537 x 1.1526 x 3.2304 x 1.0208= 4.0049

Table 10: Measures of Fit for Poisson of low birth weight

Model	current Poisson	saved Poisson	Difference
N	19	19	0.00
logLik intercept only	-46.857	-46.857	0.00
Log-Lik Full Model	-31.301	-31.293	-0.009
D	62.603	62.586	0.017
LR:	31.11	31.127	-0.017
Prob> LR:	0.00	0.00	0.00
Max. Likelihood R2	0.806	0.806	0.00
Cragg & Uhler's R2:	0.811	0.812	0.00
AIC:	3.926	4.031	-0.104
BIC':	-19.332	-16.405	-2.927

Difference of 2.927 in BIC' provides positive support for current model where *saved Poisson* is the full model, while *current Poisson* is without *number of births*. Since the effect

of *number of birth* in the model is low, it is therefore removed in order to increase the precision power of the fit.

Table 11: Summary statistics of the variables

Year	LBW	Mean age of mothers	Unbooked mothers	Births
2014	*0.3333	31.3333	2.0	10.3333
	**0.5774	2.0816	2.6457	12.7017
2015	3.2	31	5.6	21.5
	1.5491	1.1547	4.5509	11.0780
2016	6.6666	31.1667	4.6666	26.5
	1.9663	0.4082	1.2110	6.8920
Total	3.8421	31.1052	4.7368	21.3157
	2.6928	1.1002	3.6338	11.0254

where * represents the mean and ** is the standard deviation .

The deviance goodness of fit test

The null hypothesis is that our model is correctly specified, and we have strong evidence to reject that hypothesis. So we have strong evidence that our model fits badly.

```
Deviance goodness-of-fit     = 8.9800
Prob>chi2(13)           =  0.8323
```

V. Discussions and Conclusion

We conclude that the model fits reasonably well because the goodness-of-fit chi-squared test is not statistically significant. If the test had been statistically significant, it would indicate that the data do not fit the model well. In that situation, we may try to determine if there are omitted predictor variables, if our linearity assumption holds and if there is an issue of over-dispersion.

II. Discussions and Conclusion

Poisson regression is appropriate when the dependent variable is a count. The event is independent in the sense that the occurrence of one will not make another more or less likely, but the probability per unit time of events is understood to be related to covariates such as time of day.

We conclude that the model fits reasonably well because the goodness-of-fit chi-squared test is not statistically significant. If the test had been statistically significant, it would indicate that the data do not fit the model well. In that situation, we may try to determine if there are omitted predictor variables, if our linearity assumption holds and if there is an issue of over-dispersion.

References

[1] Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory, ed. B. N. Petrov and F. Csaki, 267–281. Budapest:Akailseoniai–Kiudo

[2] Breslow, N.E. and Day, N.E. (1987): *Statistical methods in cancer research, volume II: the design and analysis of cohort studies.* Number Scientific Publication 82. International Agency for Research on Cancer, Lyon.

[3] Cameron, A.C., &Trivedi, P.K. (1998): Regression analysis of count data. New York: Cambridge University Press.

[4] Cochran, W. G. (1940): The analysis of variance when experimental errors follow the Poisson or binomial laws. Annals.

[5] Frome E, and Checkoway H. (1985): Use of Poisson regression models in estimating incidence rates and ratios. *American Journal of Epidemiology* 121(2): 309-23.

[6] Frome E.L. (1984): Response to Nelder's reaction on Poisson rate analysis. *Biometrics* 40: 1160-62.

[7] Jorgenson, D. W., (1961): Multiple regression analysis of a poisson process. *Journal of the American Statistical Association*, 56, 235-245.

[8] Koch, G. G., Atkinson, S. S. and Stokes, M E., (1984): Poisson Regression, in: Encyclopedia of Statistical Sciences, edsKotz, S., Johnson, L. L. and Read, A, New York: J. Wiley & Sons, pages 32-40.

[9] Mccullagh, P., and Nelder, J. A., (1989): *Generalized Linear Model,* 2nd Edition. Chapman and Hall, London.

[10] Nelder J.A. and Wedderburn, R.W.M. (1972): Generalized linear models. *Journal of the Royal Statistical Society, Series A* 135(3): 370-84.

[11] R Development Core Team (2016): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, www.R-project.org.

[12] Richardson D.B. and Loomis D. (2004): The impact of exposure Categorisation for grouped analyses of cohort data. Occupational and Environmental Medicine 61(11): 930-35.

[13] Schwarz, G. (1978): Estimating the dimension of a model. Annals of Statistics 6: 461–464. Tong, H. 2010. Professor Hirotugu Akaike, 1927–2009. Journal of the Royal Statistical Society, Series A 173: 451–454.