
**Modelling Dichotomous Outcome Variable: A Modified Poisson
Regression Approach**

By

E. IOBISUE*Department of Statistics, Delta state
Polytechnic, Otefe-Oghara*

And

M. N. ONWUBUYA*Department of Statistics, Delta State
Polytechnic, Otefe-Oghara***Abstract**

Binary logistic regression model is popularly used in regression analysis when the outcome variable is dichotomous. However, it uses odds ratio as the risk measure in the analysis. In this study data on some characteristics of the borrower were collected from a bank in order to study loan default. The binary logistic and modified Poisson regression models that uses relative risk were used in the analysis. From the result, it is obvious that the relative risk of the modified Poisson regression gives a better understanding. Also, both models were able to capture the same factors namely age, occupation and asset-to-loan ratio as significant factors as far as loan default is concerned

When the outcome variable of interest is binary/dichotomous, the binary logistic regression is frequently used in analyzing the risk of exposure. It is fairly easy to run using many different statistical software packages. It's a non-linear model and one of the S-shaped curve models referred to as sigmoidal curves. It regresses against the logit of the dependent variable and not the dependent variable itself (i.e. log odds). It can be applied to many different study designs (cohort, case-control, cross-sectional). The binary logistic regression uses odds ratio as a risk measure. The odds ratio (OR) provides a good approximation of the Relative Risk when the outcome is rare (i.e. < 10) (Greenland, 1979). In situations when the emphasis on the importance of the rare

event assumption is not met, a conversion formula has been proposed (Zhang and Yu, 1998). However, despite advice on the "rare event rate assumption" consumers of health research literature often interpret the odds ratio as relative risk leading to its potential exaggeration (Schwartz, et al, 1999). Even with the conversion formula of (Zhang and Yu, 1998), it has been argued by McNutt et al, 2003, that it produces inconsistent estimates for the relative risk (i.e. as the sample size increases, the bias will not decrease.). Hence the relative risk is preferred over the odds ratio (Sinclair and Bracken,1994).

To estimate relative risk directly, the Poisson regression is usually recommended (McNutt et al, 2003). It was however criticized by the same author that it provides conservative results. The modified Poisson regression model uses relative risk as a risk measure. The purpose of the study is to demonstrate that the relative risk is better understood than the odds ratio and also to find out whether or not the modified Poisson regression model can capture the same significant variable(s) as the binary logistic regression model.

Methods and Materials

Table 2.1: 2x1 contingency table

	Default	Non default	
X = 1	a	b	n ₁ = a+b
			n = n ₁ + n ₀

Consider table 2.1

$$\text{Relative risk} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{Pr_1}{Pr_2} \tag{1}$$

Where P₁ is the probability of the outcome in group 1

P₂ is the probability of the outcome in group 2

$$\text{Odds ratio} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc} = \frac{\text{odds}_1}{\text{odds}_2} \tag{2}$$

$$\text{Where odds}_1 = \frac{Pr_1}{1-Pr_1} \tag{3}$$

$$\text{Odds}_2 = \frac{Pr_2}{1-Pr_2} \tag{4}$$

Binary Logistic Regression Model

$\pi(x)$ = success probability

$$\text{Logit} \left[\pi(x) = \ln \left(\frac{\pi(x_1)}{1-\pi(x_1)} \right) \right]$$

(5)

The above equation is called the logit. **i. e.**

$$\ln \left(\frac{\pi(x_1)}{1-\pi(x_1)} \right) = \alpha + \beta x$$

(6)

Modified Poisson Regression

When Poisson regression is applied to dichotomous outcome variable, the error for the estimated relative risk **will** be overestimated (Zhochei et al, **1995**). This problem **is** rectified by using the modified Poisson regression via the robust error variance procedure (Zou, 2004).

considering the table of 2.1, the procedure uses this estimator:

$$\widehat{var}(RR) = \frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_0}$$

(7)

as the variance **for** the Poisson regression model.

Methodology and Data Presentation

Data Presentation

Information on the class of borrowers was collected from UBA Oghara branch, Delta State. The researcher considered data from 2007 – 2015. The binary outcome variable *Y* is loan default and the covariates *X*, are: Age: Level of education; Amount of loan collected; Occupation of borrowers; duration of loan payment; Years of borrowers in business/job; Asset-to-loan ratio; Number of children of borrower; and the income of the borrower. the data is in Appendix I.

Table 3.1: Variable Description

Variable	Type	Variable description			
1. Age	Independent	Numeric			
2. Education	Independent	Categorical			
		Level	O' Level	ND, HND, B.Sc	Post-graduate
		Code	1	2	3
3. Amount of loan	Independent	Categorical			
		Level	≤500,000	<500,000≤1,000,000	>1,000,000
		Code	1	2	3
4. Occupation	Independent	Categorical			
		Level	Civil servant		Business
		Code	1		2
5. Duration of loan	Independent	Categorical			

Data Analysis and Findings Analysis of Data

Payment		Level	0-1 year	>1year≥2years	>2year	
		Code	1	2	3	
7. Years of business/job	Independent	Categorical				
		Level	Default	Non-default		
8. Asset to loan ration	Independent	Code	1	2		
9. Number of Children of borrower	Independent	Numeric				
		Categorical				
		Level	None	1-2	3-4	>4
10. Income of borrower	Independent	Code	0	1	2	3
		Numeric				

Table 4.1: Binary Logistic regression Model Parameter Estimation

Likelihood = - 118.81471 Pseudo R² 0.2742

Def	Coef.	St.Err.	Z	P> z	95% Conf. Interval	
Age	-.0824934	.0382047	-2.16	00.031	-.1573734	-.0076135
Education						
First degree	.8372516	.9560936	0.88	0.381	-1.036657	2.711161
Post graduate	1.347974	1.623677	0.83	0.06	-1.834375	4.530322
Amount of loan						
500000<1000000	.4525239	1.004675	0.45	0.652	-1.516602	2.42165
1000000 and above	-.70058	1.033745	-0.68	0.498	-2.726683	1.325522
Occupation						
Business	-1.962902	-1.962902	-3.03	0.002	-3.233318	-.6924857
Duration of loan payment						
2 years and above	-.1938527	.446333	-0.43	0.664	-1.068649	.6809438
Years in business/job						
>5≤10 years	.5387333	.4959936	1.09	0.277	-.4333963	1.510863
>10 years	.8733425	1.062694	0.82	0.411	-1.209499	2.956184
Asset to loan ratio	-.3103203	.0539832	-5.75	0.000	-.4161255	-.2045151
No. of children						
1-2 children	.91718337	.5555911	1.65	0.099	-.17175448	2.006122
3-4 children	.9920686	.6436105	1.54	0.123	-12693846	2.253522
>4 children	.7287565	.6809926	1.07	0.285	-.6059644	2.063477
Income of borrower	.0021912	.0050812	0.43	0.666	-.0077677	.012502
constant	2.208666	1.557918	1.42	0.156	-.8447968	5.262129

From table 4.1, Age, Occupation and Asset to loan ratio have significant effect on loan default with p-values <0.05. however, these three covariates have negative coefficients.

Modelling Dichotomous Outcome Variable: A Modified Poisson Regression Approach

This implies that older persons are less likely to default than younger persons. Business men are less likely to default than civil servants. Persons with higher asset to loan ratio are less likely to default than persons with lower asset to loan ratio.

Table 4.2: Binary Logistic Model Odds Ratio Estimation

Likelihood = 118.81471 Pseudo R² 0.2742

Def	Odds ratio	Std.Err.	Z	p> z	95% Conf. Interval	
Age	.9208175	.0351796	-2.16	0.031	.854385	.9924154
Education						
First degree	2.310009	2.208585	0.88	0.381	.3546381	15.04673
Postgraduate	3.849617	6.250534	0.83	0.406	.1597133	92.78845
Amount of loan						
500000<1000000	1.572275	1.579625	0.45	0.652	.2194563	11.26443
1000000 and above	.4962974	15130448	-0.68	0.498	.065436	3.764152
Occupation						
Business	.1404502	.0910375	-3.03	0.002	.0394264	.5003308
Duration of loan payment						
2 years and above	.8237792	.3676798	-0.43	0.664	0.3434721	1.97574'2
Years in business/job						
>5> 10 years	1.713835	.850051	1.09	0.277	.6483035	4.530638
>10 years	2.394902	2.545048	0.82	0.411	.2983466	19.22448
Asset to loan ratio	.7332121	.0395812	-5.75	0.000	.6595975	.8150424
No. of children						
1-2 children	2.502233	1.390219	1.65	0.099	.8421856	7.434432
3-4 children	2.696807	1.735694	1.54	0.123	.7638493	9.521212
>4 children	2.072502	1.411358	1.07	0.285	.545548	7.873301
Income of borrower	1.002194	.0050923	0.43	0.666	.9922623	1.012224
Constant	9.103565	14.18261	1.42	0.156	.4296447	192.8917

From the table 4.2, the odds of persons age (x + 1) years defaulting (having a loan default) is 0.92 times the odds of person at age x -years. The odds of a business person defaulting (having a Joan default) is 0.14 times the odds of a civil servant. The odds asset to loan ratio of an individual-with asset to ratio y+1 is 0.73 times higher that an individual with asset to loan ration y.

Table 4.3: Modified Poisson Regression Parameter Estimation

Log likelihood = -130.74479

Pseudo R² = 0.2280

Def	Coef.	Std. Err.	Z	P> z	95% Conf. Interval	
Age	-.0607152	.0264867	-2.29	0.022	-.1126281	-.0088023
Education						
First degree	.6428755	.6141196	1.05	0.295	-.5607768	1.846528
Post graduate	.9111033	1.177724	0.77	0.439	-1.397193	3.2194
Amount of loan						
500000<1000000	.248	.6898084	0.36	0.719	-1.104	1.6

1000000 and above	-.5270202	.721845	-0.73	0.465	-1.94181	.8877699
Occupation						
Business	-1.510491	.5305756	-2.85	0.004	2.5504	-.4705819
Duration of loan payment						
2 years and above	-.1691055	.274883	-0.62	0.538	-.7078664	.3696555
Years in business/job						
>5 ≥ 10 years	.330184	.3169972	1.04	0.298	-.291119	.330184
>10 years	.6102865	.654695	0.93	0.351	-.6728922	1.893465
Asset to loan ratio	-.2394762	.047112	-5.08	0.000	-.3318141	-.1471382
No. of children						
1-2 children	.6603691	.3769438	1.75	0.080	-.0784272	1.399165
3-4 children	.6809842	.4110853	1.66	0.098	-.1247283	1.486697
>4 children	.496987	.4355918	1.14	0.254	-.356242	1.350731
Income of borrower	.0019868	.0038832	0.51	0.609	-.0056242	.0095977
Constant	1.062805	1.135129	0.94	0.349	-1.162007	3.287617

From table 4.3 interestingly, there are three covariates that play significant roles in the loan default, namely; Age, Occupation and Asset to loan ratio. Like the Logistic regression model all the three variables have negative coefficients indicating that older person are less likely to default, businessmen are less likely to default compared to civil servants and the persons with Asset to loan ratio (x + 1) are likely to default than persons with Asset to loan ratio of x.=

Table 4.4: Modified Poisson Regression Risk Ratio Estimation

Log likelihood = -130.74479

Pseudo R² = 0.2280

Def	Robust RR	Std. Err.	Z	P> z	95% Conf. Interval	
Age	.9410912	.0249264	-2.29	0.022	.8934829	.9912364
Education						
First degree	1.901942	1.16802	1.05	0.295	.5707655	6.337776
Post graduate	2.487065	2.929076	0.77	0.439	.2472901	25.0131
Amount of loan						
500000<1000000	1.28146	.8839619	0.36	0.719	.3315424	4.95301
1000000 and above	.5903615	.4261495	-0.73	0.465	.143444	2.429705
Occupation						
Business	.2208016	.1171519	-2.85	0.004	.0780504	.62";.6387
Duration of loan payment						
2 years and above	.8444198	.2321167	-0.62	0.538	.4926943	1.447236
Years in business/job						
>5>= 10 years	1.391224	.4410141	1.04	0.298	.7474267	2.589558
>10 years	1.840959	1.205267	0.93	0.351	.5102308	6.642346
Asset to loan ratio	.78704	.0370791	-5.08	0.000	.7176027	.8631746
No. of children						
1-2 children	1.935507	.7295771	1.75	0.080	.9245694	4.051816

3-4 children	1.975821	.8122312	1.66	0.098	.8827368	4.422462
>4 children	1.643761	.7160089	1.14	0.254	.6999423	3.860248
Income of borrower	1.001989	.0038909	0.51	0.609	.9943916	1.009644
Constant	2.894478	3.285606	0.94	0.349	.3128577	26.77896

From table 4.4 persons with age (x+ 1) years have 0.94 times the risk of defaulting compared to persons with age x. businessmen have 0.22 times the risk of defaulting compared to civil servants. Persons with Asset to loan ratio (x+1) have 0.79 times the risk of defaulting compared to Asset to loan ratio x.

Conclusion/Recommendation

The findings from the study shows that the interpretation of the relative risk is far better understood than the odds ratio used in the binary logistic regression model. Also, since both models were able to capture the same three significant variables such as age, occupation and the Asset-to-loan ratio, the modified Poisson regression model is as good as the binary-logistic regression model.

References

- Greenland, S. (1979). Limitations of the logistic analysis of epidemiologic data. *Am J Epidemiol* 110:693-698
- McNutt, L. A., Wu, C. and Xue, X. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol* 157:940-943

- Schwartz, I. M., Woloshin, S., Welch, H. G. (1999). Misunderstanding about the efforts of race and sex on physicians' referrals for cardiac catheterization. *N Engl J Med.* 341
- Sinclair, J. C. and Bracken, M. B. (1994). Clinically useful measures of effect in binary analysis of randomized trials. *J Clin Epidemiol* 47:881-889
- Zhang, J. and Yu, K. F. (1998). What is the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998, 280 - 1690-1
- Zou, G. (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* 159:702-706