J. Res. Sci. Mgt Vol 6 No. 1 Dec (2008), Vol 7 No. 1 Dec (2009)
© A Publication of the School of Sciences, Federal University of Technology, Akure, Nigeria.

Journal of Research in Science and Management

AN ALTERNATIVE TECHNIQUE TO ORDINAL LOGISTIC REGRESSION MODEL UNDER FAILED PARALLELISM ASSUMPTION

K.A ADELEKE[1] and A.A.ADEPOJU[2]

[1]Department of Statistics, University of Ibadan, Ibadan
[2]University of Mines and Technology, Tarkwa, Ghana

Abstract
Maternal health status is often measured in medical studies on an ordinal scale, but data of this type are generally reduced for analysis to a single dichotomy. Several statistical models have been developed to make full use of information in ordinal response data, but have not been much used in analyzing pregnancy outcomes. The authors discussed two of these statistical models the ordinal logistic regression model and the multinomial logistic regression model. Logistic regression models are used to analyze the dependent variable with multiple outcomes which can either be ranked or not. In this study, we described two logistic regression models for analyzing the categorical response variable. The first model uses the proportional odds model while the second uses the multinomial logistic regression model. The fits of these models using data on delivery from a Nigerian state hospital record/database were illustrated and compared to study the pregnancy outcomes. Analyses based on these models were carried out using STATA statistical package. The Multinomial logistic regression was found to be an important alternative to the ordinal regression technique when proportional odds assumption failed. The weight of the baby and the mother's history of disease (treated or not treated) were found to be important in determining the pregnancy outcome.

Keywords: Likelihood function, Multinomial regression, Ordinal regression, Parallelism, Response variable.

Introduction

Ordinal logistic regression has the same goal as ordinary least squares (OLS) regression in which we may wish to model a dependent variable in terms of one or more independent variables. However, OLS regression is for continuous (or nearly continuous) dependent variables; logistic regression is for dependent variables that are categorical. The dependent variable may have two categories (e.g., alive/dead; male/female; Republican/ Democrat) or more than two categories. If it has more than two categories they may be ordered (e.g. Live birth/Stillbirth/Abortion) or unordered (e.g. Married/single/divorced/widowed/other). Logistic regression deals with these issues by transforming the dependent variable. Rather than using the categorical responses, it uses the log of the odds ratio of being in a particular category for each combination of values of the independent variables. The odds is the same as in gambling, for example, 3-1 indicates that the event is three times more likely to occur than not. The ratio of the odds is taken in order to allow us to consider the effect of the independent variables. The log of the ratio is then taken so that the final number goes from -8 to +8, where 0 indicates no effect, and the result is symmetric around 0, rather than 1. Adepoju and Adegbite (2009) used ordinal logistic regression method to examine the relationship between the outcome variable, different levels of staff status in the Lagos State Civil Service of Nigeria; the explanatory variables are Gender, Indigenous status, Educational Qualification, Previous Experience and Age. The study revealed that two explanatory variables namely, Education Qualification and Previous Working Experience significantly predicted the probability of an individual staff being a member of any of the three levels of staff status.

Several works have been done using Ordinal Logistic Regression model. The works cut across medical, social and economical phenomena. Plank, Stephen B. and Jordan, Will J. (1997) used the logit model to predict college going behavior and found out that the probability of campus residency increased with the percentage of students living on campus in the absence of monetary constraints. The concepts of logistic regression are discussed by Agresti (1996), Hosmer and Lemeshow (2000). This paper deals with modeling pregnancy outcome as dependent variable.

The multinomial logit on the other hand, is a generalization of the logistic regression model to the case where there exists more than two outcomes, and where the outcomes are not ordered. For instance, the case of pregnancy outcome can be treated as unordered.

There are several reasons why the use of OLS with categorical dependent variables is a 'bad idea:

- Firstly, the residuals cannot be normally distributed (as the OLS model assumes), since they can only take on one of several values for each combination of level of the independent variables.
- Secondly, the OLS model makes nonsensical predictions, since the dependent variable is not continuous - e.g., it may predict that someone does something more than 'all the time'.
- Finally, for nominal dependent variables, the coding is completely arbitrary, and for ordinal dependent variables it is (at least supposedly) arbitrary up to a monotonic transformation. Yet recoding the dependent variable will give very different results.

Methodology

Ordinal logistic regression refers to the case where the dependent variable has an order. The most common ordinal logistic model is the proportional odds model. The model was originally proposed by Walker and Duncan (1967) called Cumulative Logit model and now called proportional model by McCullagh (1980).

The Model I

Consider the following equation;

$$y_i^* = x_i\beta + \varepsilon i \qquad 2.0$$

Since the dependent variable is categorized, we use

$$c_k(x) = ln\frac{P(Y \le j|x)}{P(Y > j|x)}$$

and

$$In\left(\frac{\sum Pr(event)}{1-\sum Pr(event)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \cdots + \beta_kX_k \qquad 2.1$$

The model assumes a linear relationship for each logit and parallel regression lines.

Assumptions of Ordinal Models

- Relationship between probabilities and follows the assumed form (normal for probit, logistic for logit).
- Parallel regressions Coefficient is the same across all possible cut point of the outcome i.e equal slopes, (proportional odds for logistic models). If not, the generalized ordered logit is used.

Application and Discussion

Data used for model 1 was obtained from delivery records/database in a state hospital in Nigeria. The records for individual patient were systematically sampled over the course of the months in the last quarter of 2008. Summary of the data displayed in Table 1.0. The ordinal response variable 'pregnancy outcome' refers to the process of the end of delivery by which a fetus leaves the mother's womb. The outcome was categorized as live birth, stillbirth and abortion (Adeleke .K.A 2009). The response variables are coded as follows:

$$Y = \begin{cases} 0 & if \quad livebirth \\ 1 & if \quad stillbirth \\ 2 & if \quad abortion \end{cases}$$

Table 1.0

Case Processing Summary

		N	Marginal Percentage
Response	LIVEBIRTH	57	57.0%
	STILLBIRTH	32	32.0%
	ABORTION	11	11.0%
Age	35-50	15	15.0%
	25-34	55	55.0%
	15-24	30	30.0%
Antenatal	REGULAR	35	35.0%
	ONCE IN A WHILE	40	40.0%
	NOT AT ALL	25	25.0%
Diseases	YES BUT NOT TREATED	2	2.0%
	YES but treated	41	41.0%
	NO	57	57.0%
Parity	HIGH(>=6)	3	3.0%
	LOW(1<6)	54	54.0%
	NULLIPARA	43	43.0%
Weight	<2500	44	44.0%
	>=2500	56	56.0%
Valid		100	100.0%
Missing		0	
Total		**100**	

Table 2.0 shows the estimates of the parameters, standard errors of the estimates, Wald's statistics, and confidence intervals. The Wald's values showed that of all the factors Parity, Age and Antenatal showed no significant impact on the outcome of pregnancy. Others namely; Disease and Weight are statistically significant, that is, disease and weight are important factors in determining the outcome of birth. It can also be traced on table 2.0 at the top left hand corner where we have information for the overall model fit. The log-likelihood of -63.94 and likelihood ratio with 5d.f is 57.67 with probability of 0.0000 meaning that we reject the null hypothesis of zero coefficient of the overall estimate.

Table 2.0

Ordinal logistic regression	Number of obs = 100
	LR χ^2_5 = 57.67
	Prob > χ^2 = 0.0000
Log likelihood = - 63.94	Pseudo R^2 = 0.3108

ESTIMATES

Response	Coefficient	d.f	Std. Err.	Walds (z)	P > |z|	[Confidence interval]	
Age	0.3933	2	0.4522	0.87	0.384	-0.4931	1.2797
Antenatal	0.2657	2	0.3195	0.83	0.406	-.3605	0.8920
Diseases	-1.991	2	0.5057	-3.93	0.000	-2.9803	-0.9979
Parity	-0.0898	2	0.5254	-0.17	0.864	-1.1196	0.9399
Weight	2.8032	1	0.6262	4.48	0.000	1.5757	4.0307
/Cut 1	2.9159		1.4090			1.5419	5.6777
/Cut 2	5.6274		1.4880			2.7109	8.5439

In Table 3.0, the odds of a woman with history of baby's weight less than 2.5kg are 16.5 times more likely to have a live birth than women with history of babies 2.5kg. The odds could be as little as 4.8 times or as much as 56.3 times with 95% confidence. The factor, diseases revealed that women with history of diseases (whether treated or not) have the chance of having a live birth is 0.14, while women without history of diseases have 86% chance of having a live birth. The odds could be as little as 0.041 times as much as 0.37 times with 95% confidence.

Table 3.0 Table of odds

RESPONSE	Odds Ratio	Std. Err.	Z	P>|z|	[95% Conf. Interval]	
AGE	1.481871	0.6701924	0.87	0.384	0.6107245	3.595635
ANTENATAL	1.304398	0.4168052	0.83	0.406	0.6972982	2.440066
DISEASE	0.136811	0.0691857	-3.93	0.000	0.0507769	0.3686173
PARITY	0.914053	0.4802457	-0.17	0.864	0.3263983	2.559734
WEIGHT	16.49827	10.33275	4.48	0.000	4.834336	56.30407
/Cut1	2.915949	1.409083			0.154197	5.6777
/Cut2	5.627454	1.48804			2.710949	8.54395

In table (4.0), Pearson and Deviance test showed that the p-values for the two tests in both logits are not significant, indicating good overall fit of the model.
Assumption of parallelism states that only the intercept is allowed to vary. Fig 1.0 interprets what we have from our estimates.

Table 4.0

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	85.732	85	.457
Deviance	74.197	85	.792

Link function: Logit.

Table (5.0) also showed that the assumption fails. The value of 22.42 at 9 d.f and with p-value of 0.008 compares with =16.92, is less hence we reject the hypothesis of equal slopes i.e the location parameters (slope coefficients) are not the same across response categories.

Table 5.0

Test of Parallel Lines(c)

Model	-2 Log Likelihood	Chi-Square	df	Sig.
NulHypothesis	94.133			
General	/1./0(a)	22.42/(b)	g	.008

Therefore the assumption fails and this led to an alternative technique, multinomial logistic Regression.

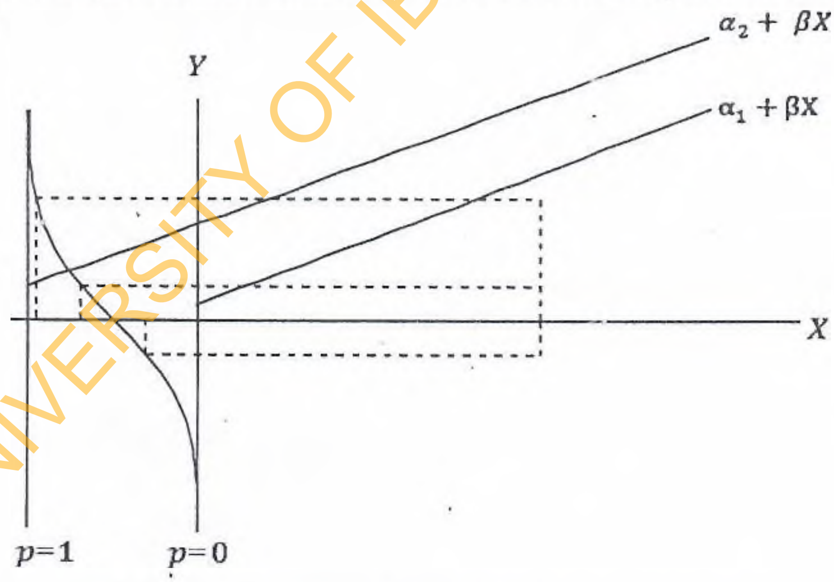

Fig 1.0: Parallel Regression with Different Intercepts and Cutpoints

Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) is an extension of binary logistic regressionto discrete/nominal outcome variables. This model was proposed by McFadden (1974), Amodification of the logistic model which is also called a discrete choice model. It was laterreferred to as polychotomous or Polytomous/Multinomial logistic regression model In health and life sciences (Hosmer Lemeshow (2000)). In the ordinal logistic model with the proportional odds assumption, the model included j-1 different intercept estimates (where j is the number of levels of the dependent variable but only one estimate of the parameter is associated with the independent variables. If the dependent variable is not ordered, however, this assumption makes no sense (i.e., because we could reorder the levels of the dependent variable arbitrarily). The multinomial model generates j-1 sets of parameter estimates, comparing different levels of the dependent variable to a base level. This makes the model Considerably more complex, but also much more flexible.

The Model II

Assuming we have k categories of outcome variables, Y, coded 0, 1, …, k-1. Using Y=0 as the baseline outcome or reference point and from a logit comparing Y=1, Y=2, …, Y=k-1 to it. To develop the model, assume we have p covariates and a constant term denoted by the Vector, of length p+1 here = 1, we have

$$f_1(x) = \ln\left[\frac{Y = 1/x}{Y = 0/x}\right]$$

$$= \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1p}X_p$$

$$= X'\beta_1 \qquad 3.0$$

And for k-1

$$f_{k-1}(x) = \ln\left[\frac{Y = (k-1)/x}{Y = 0/x}\right]$$

$$\beta_{(k-1)0} + \beta_{(k-1)1}X_1 + \beta_{(k-1)2}X_2 + \cdots + \beta_{(k-1)p}X_p$$

$$X'\beta_{k-1} \qquad 3.1$$

Hence, the conditional probabilities of each outcome categories given the covariate vector are;

$$Pr(y_i = 1|x_i) = \frac{1}{1+\sum_2^J \exp(x_i\beta_j)} \qquad \text{for } m = 1$$

$$Pr(y_i = m|x_i) = \frac{\exp(x_i\beta_m)}{1+\sum_2^J \exp(x_i\beta_j)} \qquad \text{for } m > 1 \qquad 3.2$$

The likelihood function for independent observation is

$$l(\beta) = \prod_{i=1}^{n}[\pi_0(x_i)^{y_{0i}}\,\pi_1(x_i)^{y_{1i}}\,\pi_2(x_i)^{y_{2i}} \ldots \pi_{k-1}(x_i)^{y_{k-1i}}]$$

Taking the first difference, and equating to zero gives the likelihood estimator

Application

For the multinomial model, one way to check model fit is to check each of the binomial models, separately. An observation with a residual that is far from 0 (both sides) is poorly fit by the model, A point with high/leverage has huge influence on the parameter estimates. Several measures have been proposed for analyzing residuals; influential points and high level points (Hosmer and Lemeshow (2000)). This method is applied using the same data on the pregnancy outcomes from a state Hospital (table 1.0).
Output from the main effect model is explained below.
From the table 1.0, the response variable is the outcome which as coded above is: live birth (0), Stillbirth (1), and Abortion (2). Abortion being the worst case is used as reference point or baseline

.Interpretation and Assessment of Fit and Diagnostic for Multinomial Logistic Regression

We obtained values for the estimates, standard error and Wald's as described in table (6.0). The essential thing to remember is that there are really two equations (one fewer than the number of categories). Eqn (i) compares women with live birth as outcome to abortion as outcome of pregnancies, and eqn (ii) compares those whose outcomes are stillbirth to abortion. Unlike ordinal logistic, the two equations have different intercept as well as slopes, the assumption of parallelism does not hold.
Examination of the Walds statistics in the table (6.0) showed that for live birth, presence or absence of diseases and weight are statistically influencing the chance of having live birth against abortion.

$$6.03 - 0.9959A - 0.309An + 2.3523D + 0.2740P - 3.6044W \quad \ldots (i)$$

$$f_0(x) = \ln\left[\frac{p(Y=0/x)}{P(Y=2/x)}\right]$$

$$f_1(x) = \ln\left[\frac{p(Y=1/x)}{P(Y=2/x)}\right]$$

$$= \quad 2.836 - 0.4189Ac + 0.2744An - 0.6047D - 0.3017P + 0.0636W \qquad (ii)$$

The odds ratio among the women with middle age and above having a live birth of 0.4 times as likely to have a live birth than are women of lower age group. The odds could be as little as 0.078 times or as much as 1.7 times larger with 95% confidence. While for stillbirth, middle aged women and above are 0.35 times more likely to have a stillbirth than women of lower group age and the odds could be as little as 0.16 times or as much as 2.7 times greater with 95 % confidence.
Presence or absence of diseases in live birth gives the odds of women with history of diseases treated or not treated having a live birth is 10.5 greater than women without history of diseases. The odds could be as little as 1.9 times or as much as 56.42 times larger with 95% confidence. While in stillbirth, the odds of women with history of diseases treated or not treated having a live birth is 0.05 i.e 5% lesser than women without history of diseases. The odds could be as little as 0.13 less times or as much as 2.28 times larger with 95% confidence.
The fit of the data to the model is as showed in table (8.0). The R^2 Pearson chi- square and Deviance statistics computed are 0.99 and 1.000 respectively compared with level of 0.05

showed that the model has a good fit. The pseudo of 0.3797 does not gives the actual interpretation as in OLS Hosmer and Lemeshow (2000). Generally in logistic are usually small.

Table 6.0 Multinomial Logistic Regression

Multinomial logistic regression — Number of obs = 100

$LR\ \chi^2_{10}$ = 70.46

Prob > χ^2 = 0.0000

Log likelihood = -57.551098 — Pseudo R^2 = 0.3797

RESPONSE	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
LIVEBIRTH						
AGE	-0.9958	0.7881	-1.26	0.206	-2.5406	0.5488
ANTENATAL	-0.3091	0.5301	-0.58	0.560	-1.3481	0.7299
DISEASE	2.3523	0.8574	2.74	0.006	0.6717	4.0329
PARITY	0.2739	0.9045	0.30	0.762	-1.4989	2.0469
WEIGHT	-3.6044	1.2378	- 2.91	0.004	-6.0301	-1.1783
_cons	6.0302	2.6293	2.29	0.022	0.8768	11.1836
STILLBIRTH						
AGE	-0.4189	0.7190	-0.58	0.560	-1.8282	0.9902
ANTENATAL	0.2744	0.4768	0.58	0.565	-0.6601	1.2090
DISEASE	-0.6046	0.7297	-0.83	0.407	-2.0350	0.8256
PARITY	-0.3017	0.8278	-0.36	0.715	-1.9241	1.3207
WEIGHT	0.0636	1.2423	0.05	0.959	-2.3713	2.4985
_cons	2.8359	2.6495	1.07	0.284	-2.3570	8.0288

(RESPONSE ==ABORTION is the base outcome)

Conclusion

The choice between ordinal and multinomial are whether the more complex model offers either

(1) Greater insight into the substantive area, or
(2) Better fit or substantially different fitted values.

One substantive difference between the two models is that, in multinomial model, weight has negative effect on the odds ratio comparing live birth to abortion (OR = 0.03 i.e 3%per kg) but a large statistically significant effect on the ORs comparing stillbirth to abortion (OR = 1.06 per kg). While in ordinal logistic, the ORs for live birth or abortion are the same as ORs for stillbirth or abortion for any of the covariate.
One way to compare the fit of the two models is to compare the predicted outcomes with the actual outcomes. The fit for the ordinal model was shown in table (4.0), while the one for multinomial model is in table (8.0). The multinomial model actually fits than ordinal model. Hence, in summary, there is no reason to prefer the more complicated model.

i. The proportional odds assumption is not violated,
ii. The ordinal model fits the data slightly better than nominal model.
iii. The predicted pregnancy outcome is quite similar in the two models.

References

Adeleke K.A. (2009): "Ordinal Logistic Regression model an application to pregnancy outcome" unpublished project work. .

Adepoju A.A. and Adegbite M.(2009): "Application of Ordinal Logistic Regression Model to Occupation Data." *Journal of Scientific and Industrial Studies*. 7, 1. 39-49, Nigeria.

Agresti A. (1996): "An Introduction to Categorical Data Analysis", New York, Wiley.

Hosmer, D.W. and Lemeshow M. (2000): "Applied Logistic Regression" New York.

Kramer M.S (1987): "Intrauterine growth and gestational duration determinants." *Paediatrics* 80, 502-511.

Kramer M.S., Seguin L., Lydin J., Goulet L.(2000): "Socio-economic disparity in pregnancy outcome: why do the poor fare so poorly? Paediatrics. *Perinat Epidemiol* 14, 194-210

Logan S. (2003): "Research and equity in health. *Paediatrics*" 112, 759-762.

LALL R, Campbell M.J and Walters S.J.(2002): "A review of ordinal regression models applied on health quality of life assessments". *Stat. Methods Med Res* 11:49-67.

MacCulagh, P. and Nelder (1989): J.A. "Generalized Linear Models" 2nd edn. Chapman and Hall, New York.

McCullagh P. (1980): "Regression Models for ordinal Data (with discussion)", *Journal of the Royal Statistical Society* B42, 109-42.

Mitchell C. Dayton (1992): "Logistic Regression Analysis. University of Maryland". Minitab Inc 2007 "Meet Minitab 15"

Plank, Stephen B. and Jordan, Will J. (1997): Reducing Talent Loss. The Impact of *Information, Guidance, and Actions on Postsecondary Enrolment,* Report No. 9 Eric No: Ed405429

SPSS inc. 2005: "Ordinal Regression analysis", SPSS Advanced models 15 Chicago IL.

World Health Organization (1985): "Physical status: the use and interpretation of anthropometry. Report of a WHO expert committee". WHO *Technical Report Series* No. 854.Geneva

Yongquan Dong (2007): "Logistic regression models for ordinal response". Medicine-Texas Medical center.

APPENDIX

Table 7.0 **Model Fitting Information**

	Model Fitting Criteria			Likelihood Ratio Tests		
Model	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	159.090	164.300	155.090			
Final	100.090	152.194	60.090	94.999	18	.000

Table 8.0 **Goodness-of-Fit**

	Chi-Square	df	Sig.
Pearson	44.477	76	.999
Deviance	40.154	76	1.000

Table 9.0 **Classification**

	Predicted			
Observed	LIVEBIRTH	STILLBIRTH	ABORTION	Percent Correct
LIVEBIRTH	50	7	0	87.7%
STILLBIRTH	5	27	0	84.4%
ABORTION	2	8	1	9.1%
Overall Percentage	57.0%	42.0%	1.0%	78.0%