Middle-East Journal of Scientific Research 20 (12): 1715-1720, 2014
ISSN 1990-9233
© IDOSI Publications, 2014
DOI: 10.5829/idosi.mejsr.2014.20.12.21085

Modeling Employments Rate Data Using Ordinal Logistic Regression

[1-3]Anwar Fitrianto and [1]Maziatul Izati Ab. Ghazab

[1]Department of Mathematics, Faculty of Science,
Universiti Putra Malaysia, Malaysia
[2]Laboratory of Applied and Computational Statistics,
Institute for Mathematical Research, Universiti Putra Malaysia
[3]Department of Statistics, Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University (IPB), Indonesia

Abstract: In 2007-2008, there are huge recession occur in employment rate. It is corresponding to a global economic crisis on that had a huge effect on the number of employment people around the world. According to International Labour Organization (ILO), the number of person who has loss their job increased from 178 million in 2007 to 197 million in 2012, with a peak of 212 million reached in 2009. In our study, we are interested to use logistic regression analysis to determine the factors which are considered to be a significant contributor to the employment. The logistic regression model was used to build models for nine independent variables which are urban population, inflation rate, literacy rates, health expenditure, spending on education, labour participation rate, agriculture expenditure, health index and human development index.

Key words: Ordinal Logistic Regression • Employment rate • Logit • Odd ratio

INTRODUCTION

According to Department of Statistics Malaysia [1], employed workers refer to the people who at any time during the reference week worked at least one hour for pay, profit or family gain (as an employer, employee, own account worker or unpaid family worker). People who do not work because of illness, injury, disability, bad weather, leave, labour dispute and social or religious reason but had a job, farm or other family enterprise to return are considered as employed. The people who are temporary lay-off with pay that would definitely be called back to work also included as employed.

Whereas, employment rate is a measurement of the proportion of the working age population (age 15 and above) that is employed [1]. These also include the people who have stopped looking for job. Employment rate is an important indicator of the state of the wider economy. According to [2], employment rate can be calculated as the following equation:

$$\text{Employment rate} = \frac{\text{Total employed people}}{\text{working age population}} \times 100$$

From the above formulation, it is clear that employment rate is a continuous variable with measurement scale of ratio. In this study, we build an ordinal logistic regression in order to know factors that have relationship with employment rate. Ordinal logistic regression is used when response categories are ordered in which the logit can utilize the ordering. Ordinal logistic regression models have been applied over the last few years for analyzing data, the response or outcome of which is presented in ordered categories. Abreu *et al*. [3] mentioned that ordered information in score-form has been increasingly used in epidemiological studies, such as quality of life in interval scales, health condition indicators and even for indicating the seriousness of illnesses. Depending on the study's purpose, these models also allow the odds ratio (OR) statistic or the probability of the occurrence of an event to be calculated.

Corresponding Author: Anwar Fitrianto, Department of Mathematics, Faculty of Science,
Universiti Putra Malaysia, Malaysia.

Some Logistic Regression Models in Applied Sciences: Logistic regression applies maximum likelihood estimation after transforming the dependent variable into a logit function (the natural log of the odds of the dependent variable occurring or not). Logistic regression estimates the probability of a certain event occurring. The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most appropriate model. A model is created that includes all predictor variables that are useful in predicting the response variable. Logistic regression then will test the fit of the model after each coefficient is added or deleted.

There are a lot of research have been conducted on logistic regression models. Bakar *et al.* [4] examines the labour force participation of women in Malaysia. They used logistic regression with probit link function to model the data. The variables used in their study are hourly cost of child care, age, years of education, working experience, husband income, number of children and dummy variable that sow marital status, urban or rural areas, health and place of birth.

In a study, [5] discussed that discriminant analysis and logistic regression were both suitable ways to model the outcome of binary dependent variable. In his research he found that logistic regression was preferable since it was more capable of handling several dummy variables simultaneously and did not assume normality. Meanwhile, [6] studied long term electric consumption forecasting model. Different regression models were developed, using historical electricity consumption, gross domestic product (GDP), gross domestic product per capita and population. Subramaniam *et al.* [7] has examined the Malaysian women's labour force participation. She used the correlation and multiple correlation analysis to analyze the data. The data was taken from field survey among 319 female employees in selected services organizations. The socio economic factor selected for this analysis are age, marital status, highest education achieved, occupation level, ethnicity, place of birth, personal income, average family expenditure and family responsibilities.

Ordinal Logistic Regression Model: There has been some work on ranking relations, which in the literature is often referred to as ordinal regression [8]. Since binary classification is much more studied than ordinal regression, a general framework to systematically reduce the latter to the former can introduce two immediate benefits. Well-tuned binary classification approaches can be readily transformed into good ordinal regression algorithms, which save immense efforts in design and implementation.

An ordinal regression problem was converted into nested binary classification problems that encode the ordering of the original ranks [9] and then the results of standard binary classifiers can be organized for prediction. Shashua [10] generalized the formulation of support vector machines to ordinal regression and the numerical results they presented showed a significant improvement on the performance. Next, new generalization bounds for ordinal regression can be easily derived from known bounds for binary classification, which saves tremendous efforts in theoretical analysis [11].

Currently, there are several approaches in logistic regression to deal with ordinal response, namely proportional odds model (POM), partial proportional odds model-without restrictions (PPOM-UR) and with restrictions (PPOM-R), continuous ratio model (CRM) and stereotype model (SM). The most common approach, which will be used in this article, is the proportional odds model [12].

Having *n* observations with ordinal response variable, *Y*, an ordinal logistic regression model relates the probability of an event occurs to predictor variables $\mathbf{x}' = (x_1, x_2,....,x_k)$ can be written as cumulative logits which is defined as follows:

$$\begin{aligned}\text{logit}\left[P(Y \le j) \mid \mathbf{x}\right] &= \log \frac{P(Y \le j \mid \mathbf{x})}{1\text{-}P(Y \le j \mid \mathbf{x})}\\ &= \log \frac{\pi_1(\mathbf{x}) + \pi_2(\mathbf{x}) + \ldots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \pi_{j+2}(\mathbf{x}) + \ldots + \pi_J(\mathbf{x})}\\ &= \alpha_j + \boldsymbol{\beta}'\mathbf{x}, \quad j = 1, 2, \ldots, J-1.\end{aligned}$$

In that model, each logit has its own α_j which is called as the threshold value and their values do not depend on the values of the independent variable for a particular case. Also, each cumulative logit uses all *J* response categories. The cumulative logit has its own intercept, but the model has the same effects $\boldsymbol{\beta}$ for each logit. According [8], the expression is called as a proportional odds model and satisfies the following expression:

$$\begin{aligned}&\text{logit}\left[P(Y \le j \mid x_1)\right] - \text{logit}\left[P(Y \le j \mid x_2)\right]\\ &\quad = \log \frac{P(Y \le j \mid x_1) / P(Y > j \mid x_2)}{P(Y \le j \mid x_2) / P(Y > j \mid x_1)}\end{aligned}$$

To fit the logistic regression model, we must estimate the **β**, the unknown parameters. The general method to estimates unknown parameters is maximum likelihood. The method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method we must first construct a function, called the likelihood function. This function expresses the probability of the observed data as a function of the unknown parameters.

After fitting the logistic model to a set of data, it is reasonable to determine how well the fitted values under the model compare with the observed values. In logistic regression, there are various possible measures to compare the overall differencebetween the observed and fitted values.Two of the most commonly used goodness of fit measures are the Deviance *D* and Pearson's chi-squared, x^2 goodness of fit test statistics.

In logistic regression, comparison of observed to predicted values is based on the log likelihood function. The comparison between observed and predicted values using the likelihood function is based on the following expression:

$$D = -2\ln\left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right].$$

This expression is called the likelihood ratio. A saturated model is one that contains as many parameters as there are data points. Hosmer and Lemeshow [13] explained that minus twice its log is necessary to obtain a quantity whose distribution is known and can therefore be used for hypothesis testing purposes. Such a test is called likelihood ratio test. The expression for this test is:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right],$$

where $\hat{\pi}_i = \hat{\pi}(x_i)$.

The statistic, *D*, in this expression is called the deviance and plays an important role in some approaches to assess goodness-of-fit. For the purpose of assessing the significance of an independent variable, we compare the value of *D* with and without the predictor variable in the equation. The change in *D* due to the inclusion of the independent variable in the model is obtained as:

$$G = -2\ln\left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right]$$

Table 1: Variables in the Employment Rate Data.

Variable	Description
x_1	Urban population (% of total)
x_2	Inflation, GDP deflator (annual%)
x_3	Literacy rates, adult total (%)
x_4	Health expenditure, total (% of GDP)
x_5	Spending of education, total (% of GDP)
x_6	Labour participation rate, total (%)
x_7	Agriculture, value added (% of GDP)
x_8	Health index
x_9	Human development index

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}\left(1-\hat{\pi}_i\right)^{(1-y_i)}}\right]$$

MATERIALS AND METHODS

Data: We use employment data of year 2010 from 169 countries around the world to determine whether employment rate allocated based on predictor variables. This employment data is obtained from the 2013 World Bank Data, [14]. The data consist of several predictorssuch as urban population, health expenditure, education spending, labour participation rate, agriculture expenditure, inflation rate and literacy rates. Other variables such as health index and human development index are obtained from [15]. The original response variable for this study is employment rate. Detail of variables in the employment rate data is available in Table 1.

Methodology: In ordinal logistic regression, we need to have ordinal-scaled response variable. But, since the original response variable is employment rate which has ratio scale of measurement, we need to follow the following steps to conduct the data analysis:

Step 1: Convert the employment rate into four-scaled ordinal variable. This step will be done by building the original employment rate variable become four categories of employment rank, namely very low (employment rate below 25%), low (employment rate between 25% until 50%), medium (employment rate between 50% until 75%) and high (employment rate up from 75% until 100%). After the conversion, the values and their corresponding categories are very low employment (0), low employment (1), medium employment (2) and high employment rate (3).

Step 2: Fit the data with the ordinal logistic regression after converted the values of employment rate into ordinal scale. The value of ordinal logistic regression parameters will be obtained by using maximum likelihood method.

RESULTS AND DISCUSSION

Modelling the Employment Rate's Data: Table 2 shows an ordinal logistic regression analysis between employment category and corresponding predictor variables. The response variables is employment rank (category) which values are 0 (very low employment), 1 (low employment), 2 (medium employment) and 3 (high employment). From the result we can see that no countries have very low employment rank. In the analysis, we define the event of reference is 3. We can observe that total observations in this study are 137 countries and among them 32 countries are incomplete observations. There are 26, 95 and 16 countries have employment rate of category 0, 1, 2 and 3, respectively.

The above full model has log-likelihood, *G* statistics and *p* value of -23.262, 178.171 and 0.000, respectively which mean that at least one of the explanatory variables has significant influence on the probability of having very low, low, medium, or high employment rank.

In order to determine which variables have significant contribution into the model, we refer to the last column of Table 2, the *p*-value correspond to each predictor. At the 0.1 level of significant, we can conclude that, inflation rate, literacy rate, labour participation rate, agriculture expenditure and health index are significant predictors on employment rank.

The value of log-likelihood for the full model is -23.262 and its *G* statistics is 178.171 at degree of freedom of 9. *P*-value for the likelihood ratio test is 0.000 which highly significant at alpha level 0.1. We can conclude that there is at least one independent variable significantly contributes equal to zero to the employment rank. Since our objective is to obtain the best fitting model, the next step is to fit a reduced model containing only those variables that are significant. Result after selecting the important variables from the full model is displayed in Table 3. From the table we can see that all predictor variables have *p*-value less than 0.1.

The employment rank data has 3 different values in the response variable which are 1, 2 and 3. By having 3 values, we can construct 2 logistic regression equations. Referring to Table 3, first logistic regression equation can be is formed using the coefficient of constant, which is 57.29 and coefficients of x_2,x_3,x_6,x_7, and x_8,. The first fitted logistic regression equation is written as follows:

Table 2: Full Model of Ordinal Logistic Regression of the Employment Rate Data

Predictor	Coefficient	SE Coefficient	Z	*p*	Odds Ratio
Const(1)	56.172	12.590	4.46	0.000	
Const(2)	74.625	16.262	4.59	0.000	
x_1	0.034	0.032	1.09	0.277	1.03
x_2	-0.126	0.056	-2.25	0.024	0.88
x_3	0.076	0.045	1.70	0.090	1.08
x_4	0.232	0.186	1.24	0.213	1.26
x_5	-0.038	0.203	-0.19	0.852	0.96
x_6	-0.810	0.177	-4.57	0.000	0.44
x_7	-0.115	0.057	-2.02	0.043	0.89
x_8	-14.177	7.134	-1.99	0.047	0.00
x_9	-10.576	10.082	-1.05	0.294	0.00

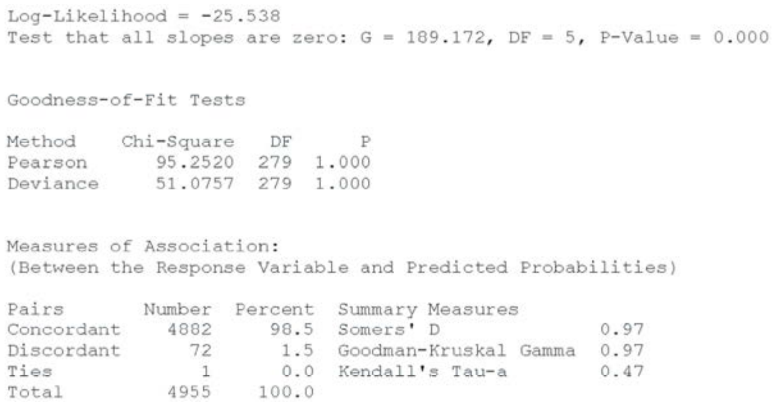
$$\operatorname{logit}\left[\pi\left(Y \le 1\right) \mid x_1, x_2, \ldots, x_9\right] = \ln\left(\frac{\pi\left(Y \le 1 \mid x_1, x_2, \ldots, x_9\right)}{\pi\left(Y > 1 \mid x_1, x_2, \ldots, x_9\right)}\right)$$
$$= 57.293 - 0.121x_2 + 0.059x_3 - 0.811x_6 - 0.101x_7 - 18.869x_8 \quad (1)$$

This logistic equation produces the probability that the employment rank is low for inflation, literacy rate, labour participation, agriculture expenditure and health index. From the Equation (1) we can see the coefficient value for literacy rates, x_3 is positive which is -0.059 rather than other variables that have negative values. It means that literacy rates have positive contribution to the employment rank.

Meanwhile, according Table 3, the second logistic regression equation is formed using the coefficient of constant, which is 75.8, x_2,x_3,x_6,x_7 and x_8. The ordinal logistic equation which is obtained by those coefficients will produce the probability that the employment is low or medium for inflation, literacy rate, labour participation, agriculture expenditure and health index. The fitted ordinal logistic regression model is written as Equation (2) which shows that no changes for the coefficient value for literacy rates, x_3,.

$$\operatorname{logit}\left[\pi\left(Y \le 2\right) \mid x_1, x_2, \ldots, x_9\right] = \ln\left(\frac{\pi\left(Y \le 2 \mid x_1, x_2, \ldots, x_9\right)}{\pi\left(Y > 2 \mid x_1, x_2, \ldots, x_9\right)}\right)$$
$$= 75.806 - 0.121x_2 + 0.059x_3 - 0.811x_6 - 0.101x_7 - 18.869x_8 \quad (2)$$

The other result to discuss in this research related to logistic regression is odd ratio. From the Table 3, the odd ratio of x_2 is 0.89. Its means a country which has one unit higher of inflation rate, the chance of the low or medium employment rank is reduced by a multiple of 0.89. Last two columns of Table 3 show confidence intervals for the odd ratios of the model. The confident interval of the odds ratio provides the range in which the odds ratio is



```
Log-Likelihood = -25.538
Test that all slopes are zero: G = 189.172, DF = 5, P-Value = 0.000


Goodness-of-Fit Tests

Method    Chi-Square   DF      P
Pearson      95.2520  279  1.000
Deviance     51.0757  279  1.000


Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant    4882     98.5  Somers' D              0.97
Discordant      72      1.5  Goodman-Kruskal Gamma  0.97
Ties             1      0.0  Kendall's Tau-a        0.47
Total         4955    100.0
```

Fig. 1: Other Statistics for Reduced Model of the Employment Rate Data

Table 3: Final Reduced Model of Ordinal Logistic Regression of the Employment Rate Data

						95% Confident Intervals for Odd Ratio	
Predictor	Coefficient	SE Coefficient	Z	p	Odds Ratio	Lower	Upper
Const(1)	57.294	12.381	4.63	0.000			
Const(2)	75.807	16.103	4.71	0.000			
x_2	-0.122	0.049	-2.51	0.012	0.89	0.81	0.97
x_3	0.059	0.035	1.69	0.090	1.06	0.99	1.14
x_6	-0.812	0.172	-4.73	0.000	0.44	0.32	0.62
x_7	-0.102	0.048	-2.12	0.034	0.90	0.82	0.99
x_8	-18.869	5.347	-3.53	0.000	0.00	0	0

expected to fall. For the x_2, we have 95% confidence that the odds ratio will be between 0.8 and 0.97. Notice the 95% confidence interval for the odd ratio does include 1 which means that countries which have different values of inflation rate tend to have different employment rank.

Meanwhile, the odd ratio for x_6 is 0.44 which indicates that for every one unit increase in labour participation, the chance of a country which has low or medium employment rank is reduced by a multiple of 0.44. The 95% confident interval for odds ratio of the variable x_6 lies between 0.32 and 0.62. The 95% confidence interval does include one, so we cannot conclude there is an association between employment value and labour participation on the odds ratio scale. Other odd ratios can be referred to the Table 3.

Test statistic for null hypothesis that all slopes are zero also shown in a Minitab output which is displayed in Figure 1. We can see that the log-likelihood for this model is -25.538 whereas G statistic of 189.172 at 5 degree of freedom. P value for this test statistics is 0.000, which indicates that there is sufficient evidence that coefficient is different from zero at 0.1 significant value.

The measure of association in the result shown above shows the relationship between the response variable and the predicted probabilities. From the result, 98.5% of the pairs were concordant, while only 1.5% of the pairs were discordant. Thus, there is a better chance for a pair to be concordant than discordant, which indicates that the predictive ability of the model is good.

CONCLUSION

In this study, nine predictor variables were used to analyse the data using ordinal logistic regression model which are urban population, inflation rate, literacy rate, health expenditure, spending on education, labour participation rate, agriculture expenditure, health index and human development index. We estimated the parameter and test the significant of the model. We tested the full model which all predicted variables are tested using the same model. As a conclusion, we obtain that the inflation rate, literacy rate, labour participation rate, agriculture expenditure and health index are having significant contribution to employment rate. This finding does not contradict to the previous study which was conducted by [16].

REFERENCES

1. Department of Statistics Malaysia, 2013 Labour Force Survey Report, 2012, www.statistics.gov.my (accessed 10 May 2013).

2. Ward-Warmedinger M., K. Masuch, R. Gómez-Slavador, N. Leiner-Killinger, R. Strauch, J. Turunen, J. De Mulder, H.Sthal, D. Nicolitsas, P. Cipollone, A. Lacuesta, A. Stigblauer, K. Stovicek, A. Balleer, K. McQuinn, P. Montanaro, A. Rosolia, E. Viviano and C. Duarte, 2008. Labour Supply and Employment in the Euro Area Countries: Developments and Challenges, ECB Occasional, pp: 87.
3. Abreu, M.N.S., A.L. Siqueira and W.T. Caiaffa, 2009. Ordinal logistic regression in epidemiological studies. Revista de Saúde Pública, 43(1): 183-194.
4. Bakar, A., N. Aznin and N. Abdullah, 2007. Labour force participation of women in Malaysia.International Economic Conference on Trade and Industry (IECTI) 2007, 3 - 5 December 2007, Bayview Hotel Georgetown, Penang. (Unpublished) http://repo.uum.edu.my/2469/ (accessed 2 June 2013).
5. Morrell, K.L.C.J.A.J.W.A., 2008. Mapping the decision to quit: A refinement andtestof the unfolding model of voluntary turnover. Applied Psychology, 57: 128-150.
6. Bianco, V., O. Manca and S. Nardini, 2009. Electricity consumption forecasting in Italy using linear regression model. Energy, 34(9): 1413-1421.
7. Subramaniam, G., B. Maniam and E. Ali, 2011.Can workplace flexibility have an effect onwomen's lifestyles and work-life balance? International Journal of Business Research, 11(4): 168-173.
8. McCullagh, P., 1980a. Generalized Linear Models: Chapman &Hall, London.
9. Frank, E. and M. Hall, 2001. A simple approach to ordinal classification, in: Proc. of European Conference on Machine Learning (L. De Raedt, P.A. Flach, Eds.), vol. 2167 of Lecture Notes in Artificial Intelligence, Springer, Freiburg, Germany, pp: 145-157.
10. Shashua, A. and A. Levin, 2003. Ranking with large margin principle: two approaches. Proceeding of Advances in Neural Information Processing Systems 15: 937-944.
11. Ling, L. and H. Lin, 2007. Ordinal regression by extended binary classification. Proceeding of Advances in Neural Information Processing Systems Vancouver, Canada, 19: 865-872.
12. McCullagh, P., 1980b. Regression models for ordinal data. J. Royal Stat Soc B., 42(2): 109-142.
13. Hosmer, D.W. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd ed. Wiley, New York.
14. World Bank Data, 2013. World Development Indicators. http://data.worldbank.org/ (accessed 10 May 2013).
15. International Human Development Indicator, Human Development Report, 2013. http://hdr.undp.org/en/ (accessed 20 May 2013).
16. Amjad, R. and A.R. Kemal, 1997. Macro Economic Policies and Their Impact on Poverty Alleviation in Pakistan. The Pakistan Development Review, 36(1): 39-68.