European Journal of Business and Management
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.8, No.19, 2016

Modeling Semi-continuous Response by Proportional Odds Models: Investigating the Effect of Job Training on Future Income

Muhammad Abu Shadeque Mullah[1], Nabila Parveen[1] and Mohammad Ahshanullah[2]

[1]Department of Statistics, Biostatistics and Informatics, University of Dhaka

[2]School of Business and Economics, United International University

Abstract

Continuous outcome with numerous zero values is often known as semi-continuous response. Analyzing such data by using typical continuous distribution models may sometimes be misleading. With the aim of investigating the effect of job training program on the future earnings, we run into a response variable (future earnings) which is zero-inflated (semi-continuous). We convert the semi-continuous response into ordinal data by appropriately categorizing it. We then apply proportional odds models to study the association between job training and future income. We also evaluate the effect of job training on the difference in earnings before and after the training program. The association between pre- and post- training earnings has also been examined. It is found that job training has statistically significant impact on the post training earnings as well as on the change in earnings.

Keywords*: Zero-inflated Variables*; Proportional Odds Models; Job Traning

1. Introduction

The problem of unemployment is a world-wide reality. Although it is more pronounced in the developing countries, developed countries also suffer from this virtual reality. It is one of the major causes of poverty, backwardness, crimes and frustration among people. During recessions, an economy usually experiences a relatively high unemployment rate [1]. According to International Labor Organization report, more than 197 million people globally are out of work. About 6% of the world's workforces were without a job in 2012 [2]. Even within the workforce, many people are not satisfied with their earnings and always trying to get job with higher income.

To alleviate the problem of unemployment, many countries around the world provide funding and support for skills and employment training through different programs organized by governmental and non-governmental agencies. These training programs are usually designed to help unemployed and underemployed people to get job or better job. However, factors such as age, sex, education, race, family size, etc. are also important determinants for employment and future earnings. This study intends to evaluate the effect of job training program on future income using a dataset from a job training program conducted between 1995 and 1997 in the USA.

The data, however, contained an outcome variable (real earnings after job training) which is non-negative and continues but with many zero values. The outcome, therefore, has a continuous distribution except for a probability mass at zero. We term this response as semi-continuous. Unlike left-censored data, semi-continuous data include plenty of zero values which represent actual responses rather than censored. The existence of a probability mass at zero makes common response distributions such as log-normal or gamma inappropriate for statistical modeling.

While several competing methods are available to analyze such data, we adopt proportional odds models after converting the semi-continuous response to the ordered data by appropriately categorizing the response.

2. Objectives of the Study

The specific objectives of the present study are to

1. Estimate the effect of job training on income after adjusting for potential confounding variables.
2. Estimate the effect of job training on the difference in earnings between 1995 and 1998 (before and after the training).
3. Estimate the effect of pre- training earnings (in 1994 and 1995) on post- training earnings in 1998.

3. Research Methods

3.1 Study design and participants

This study is based on a dataset from a job training program which took place in 1995-97 in the USA. The training program was to encourage people to get to work, and to take better jobs. A total of 445 subjects aged between 17 and 55 were followed up till 1998. Of them 185 subjects participated in the training program and 260 did not. Most of the subjects (83%) in the study were black. It should be worth whiling to note that the training program was not randomly assigned to the participants.

3.2 Variables

In this follow up study, the treatment status (participating in the training program or not) and real earnings in 1998 were measured for each individual. In addition to these two variables, several other baseline measures such as age (in years), education (in years

of schooling), race (black or not), ethnicity (Hispanic or not), marital status, high school diploma (yes or no), earnings in 1995, and earning in 1994 were also collected to adjust for potential confounding and presence of effect modification.

3.3 Exposures and Outcomes

Real earnings in 1998 (in dollar) have been considered as outcome variable to address objectives 1 and 3. The difference between real earnings in 1998 and 1995 has been taken as outcome variable to asses the objective 2. Participation status in the training program (1 = yes, 0 = no) has been taken to be the only exposure to address objective 1 and 2. For evaluating objective 3, earnings indicators of being zero in 1994 and 1995 have been accounted as exposures.

3.4 Potential Confounders and Effect Modifier

In statistical analyses, to evaluate the effect of job training on real earnings in 1998, we include factors that appear to be potential confounders analytically and scientifically are: age at baseline, education, race, ethnicity, marital status, earnings indicator in 1994 (1 if earning = 0, otherwise 0), and earnings indicator in 1995 being zero. Except for earnings in 1994, all these factors are also considered as potential confounders to asses objective 2. However, for evaluating objective 3, we consider age, education, race, and ethnicity as potential confounders. Marital status has been omitted as it appears to be an effect modifier in assessing objective 3.

3.5 Model

The outcome real earning in 1998 is found to be a zero-inflated (i.e., semi-continuous) variable. Therefore, applying typical response distribution for modeling such data would be inappropriate. In this case, the most commonly used modeling approaches include Tobit censored regression model (Tobin, 1958) [3], Two-part model (Duan et al., 1983) [4], Sample selection models (Heckman, 1974) [5], Compound Poisson exponential dispersion model (Jørgensen, 1987) [6], Proportional odds model (Saei, Ward, and McGilchrist, 1996) [7], etc. However, proportional odds model (POM) is attractive and relatively easy to implement using available statistical software. The POM is the most popular model for analyzing categorical data when categories are ordered. It is also referred to as 'cumulative logit model' or 'ordinal threshold model'. After categorizing the semi-continuous data as: zero values in one category, and the remaining non-zero values into several ordered categories, we can apply the POM for modeling the dependence of an ordinal response on continuous or discrete covariates.

We therefore adopt the POM to study the effect of exposures on outcomes in addressing all the objectives after orderly categorizing the semi-continuous (real earnings). We primarily pick two models to consider in each case:

1. Including only the exposure (job training) with no adjustment, and
2. Adjusting for additional potential confounders mentioned earlier.

Categorical variables are modeled by using indicator variables. Continuous variables (age, education) are modeled by using natural (restricted) cubic splines with 4 knots.

3.6 Proportional Odds Model (POM)

Saei, Ward, and McGilchrist (1996)[7] suggested grouping the possible outcome values into k ordered categories and applying an ordinal response model. Let Y_g be the grouped response variable. The POM for an ordinal response posits an unobservable variable Z, such that one observes $Y_g = j$ (i.e., in category j) if Z is between $\beta_{(0,\,j-1)}$ and $\beta_{(0,\,j)}$. Suppose that Z has a cumulative distribution function G(z-η), where η is related to explanatory variables by

$$\eta = \mathbf{x}'\boldsymbol{\beta}$$

Then

$$P(Y_g \le j) = P(Z \le \beta_{(0,j)}) = G(\beta_{(0,j)} - \mathbf{x}'\boldsymbol{\beta})$$

The POM is then

$$G^{-1}\left[P(Y_g \le j \mid \mathbf{x})\right] = \beta_{(0,j)} - \mathbf{x}'\boldsymbol{\beta}$$

j=1, … , k-1. Assuming that G is logistic leads to a logit model for the cumulative probabilities, called a cumulative logit model which has the form

$$\log\left[\frac{P(Y_g \le j \mid \mathbf{x})}{1 - P(Y_g \le j \mid \mathbf{x})}\right] = \beta_{(0,j)} - \mathbf{x}'\boldsymbol{\beta}, \quad j=1, \ldots, k-1.$$

Here $\beta_{(0,1)} < \beta_{(0,2)} < \ldots < \beta_{(0,k-1)}$ to respect the ordered nature of the outcome. In application with semi-continuous data and a clump at 0, it is recommended to take the first category to be the 0 outcome, and then to select cut points on the positive outcome scale to define the other k−1 categories. It is also very essential to check the assumption that covariate effects are the same for each

cut point. This model has the simplicity of a single model to handle the clump at 0 and the positive outcomes, and it is simple to fit.

3.7 Grouping for Outcomes

To use the proportional odds model logit link, the response variables earnings in 1998 and change in earnings from 1995-98 have been categorized as follows:

Original value	New category
Earnings in 1998	
$0	1
$1 - $4000	2
$4001 - $8000	3
$8000+	4
Change in earnings from 1995-98	
less than $0	1
$1 - $4000	2
$4001 - $8000	3
$0	4
$8000+	5

3.8 Sensitivity Analysis

In applying the proportional odds model, the way the positive scale is collapsed into ordered categories is very crucial as it might change the overall results. As part of a sensitivity analysis, we collapse the positive scale in several ways into different categories in addition to the categories reported above. With different categorical outcomes each objective has been evaluated with the goal of examining the effect of arbitrary categorization on the effect measure. We perform all statistical analyses using function 'lrm' function in software package 'R' (version 2.14.2).

4. Results

Table 1 shows the bivariate summaries of all the variables considered in the study with respect to the participation status in the job training program. Means and confidence intervals are reported for quantitative variables whereas for categorical variables counts and percentages are provided.

It is observed from the Table 1 that the participants in the job training program have higher average income ($6349.15) than the non-participants ($4554.80) after a year of training. Note that the average real earning at baseline is also a bit higher for those who participated in the program as compared to those who did not. People with at least high school diploma have higher participation in the training program. However, it is apparent that those who did participate and those who did not participate in the training program are almost similar with respect to the age, earnings before the year of training program, race, ethnicity, and marital status.

Table 1: Relation between exposure and other variables.

Characteristics	**Summary Measure**	
	Participated in training	Not participated in training
Continuous Variables (Mean(95% CI))		
Age (in years)	25.82 (24.78, 26.85)	25.05 (24.19, 25.92)
Education (schooling year)	10.35 (10.05, 10.64)	10.09 (9.89, 10.29)
Earnings in 1994	2095.57 (1386.75, 2804.40)	2107.03 (1412.41, 2801.65)
Earnings in 1995	1532.06 (1065.09, 1999.02)	1266.91 (887.97, 1645.85)
Earnings in 1998	6349.15 (5207.95, 7490.34)	4554.80 (3885.10, 5224.50)
Categorical Variables (Count (%))		
Black	156 (84.32%)	215 (82.69%)
Hispanic	11 (5.95%)	28 (10.77%)

Marital Status (Married)	35 (18.92%)	40 (15.38%)
High School Diploma (Yes)	54 (29.19%)	43 (16.54%)

Table 2 shows the bivariate summaries of important covariates by the post training earning level. It is evident from the Table 2 that the income level does not depend on age, earnings in the year and a year prior to the training program, race, ethnicity and marital status. However, percentage of people with no high school diploma is found to be higher in the higher income groups.

Table 2: Relation between outcome (earnings in 1998) and covariates.

	Earnings in 1998			
Covariate	**$0**	**$1 - $4000**	**$4001 - $8000**	**$8001+**
Continuous Variables Mean (95% CI)				
Age	25.48 (24.35, 26.62)	24.80 (23.48, 26.13)	25.48 (23.79, 27.16)	25.63 (24.34, 26.91)
Schooling Year	10.18 (9.91, 10.45)	10.02 (9.66, 10.39)	10.02 (9.64, 10.50)	10.46 (10.15, 10.77)
Categorical Variables Count (%)				
Black				
Black	127 (92.70%)	84 (86.60%)	72 (75.00%)	88 (76.52%)
Hispanic				
Yes	5 (3.65%)	9 (9.28%)	14 (14.58%)	11 (9.57%)
Marital Status				
Married	21 (15.33%)	16 (16.49%)	17 (17.71%)	21 (18.26%)
High School Diploma				
Yes	26 (18.98%)	20 (20.62%)	19 (19.79%)	32 (27.83%)
Earnings in 1994				
$0	106 (77.37%)	72 (74.23%)	64 (66.67%)	84 (73.04%)
$1-$4000	12 (8.76%)	8 (8.25%)	15 (15.63%)	12 (10.43%)
$4001-$8000	9 (6.57%)	6 (6.19%)	6 (6.25%)	7 (6.09%)
$8001+	10 (7.30%)	11 (11.46%)	11 (11.46%)	12 (10.43%)
Earnings in 1995				
$0	95 (69.34%)	65 (67.01%)	56 (58.33%)	73 (63.48%)
$1-$4000	29 (21.17%)	21 (21.65%)	26 (27.08%)	27 (23.47%)
$4001-$8000	8 (5.84%)	8 (8.25%)	8 (8.33%)	8 (6.96%)
$8001+	5 (3.65%)	3 (3.09%)	6 (6.25%)	7 (6.09%)

Table 3 summarizes the results obtain from the ordinal threshold models (proportional odds model) fit. It is clear from the second column of Table 3 that without adjusting for any potential confounder, job training program has found to have statistically significant impact on earnings after a year of ending the program. The odds ratio for higher income associated with participating in the training program is 1.62 (1.15, 2.27); that is, getting training is significantly associated with higher or increased income. Again, after adjusting for potential confounders (as shown in the third column of Table 3), the association still remains statistically significant. The continuous covariates age and education have not found to be statistically associated with outcome and they do not show nonlinear effects (p-values > 0.26). In order to check the proportionality assumption of the odds model (i.e., to test for whether the slopes of the covariates are equal across logit equations), a chi-squared test has been performed which compares the deviance of multinomial model with proportional odds model. The results are shown in Table 3 which suggests that for all the situations considered here, the proportionality assumption holds since the multinomial model assumption has not found to be significant. In addition, the plots of partial residuals are examined. The residual plots do not suggest any inadequacy of the model assumptions.

Table 3: Results from proportional odds models fit to evaluate the effect of job training on earnings in 1998.

Variables	**Model -1 (Unadjusted for confounders)**	**Model -2 (Adjusted for confounders)**
	Odds Ratio (95% CI)	Odds Ratio (95% CI)
Job Training	1.62 (1.15, 2.27)	1.54 (1.08, 2.18)
Age		*
Education		*
Black		0.35 (0.18, 0.67)
Hispanic		0.71 (0.31, 1.66)
Married		1.16 (0.73, 1.87)
Earnings indicator (1994)		0.88 (0.49, 1.56)
Earnings indicator (1995)		0.95 (0.55, 1.63)
Proportionality test	Chi-square statistic (p-value) 13.45 (0.64)	Chi-square statistic(p-value) 2.58 (0.28)

* Age and education have been modeled using natural cubic splines; the method does not give a summary odds ratio and CI per unit increase.

The results of proportional odds models fit are shown in Table 4. It is evident that without considering any confounder in the model, job training program does not have statistically significant impact on the change in earning before and after the program. However, after adjusting for potential confounders (as shown in the third column of Table 4), the association has found to be statistically significant. As before the continuous covariates age and education do not show nonlinear effects (p-values > 0.35). Significance association also holds for different types of grouping schemes for outcome variable (data not shown). For all the models considered here, the proportionality assumption holds which can be evident from the test results shown in Table 4.

Table 4: Results from proportional odds models fit to evaluate the effect of job training on change in earnings from 1995-98.

Variables	**Model -1 (Unadjusted for confounders)**	**Model -2 (Adjusted for confounders)**
	Odds Ratio (95% CI)	Odds Ratio (95% CI)
Job Training	1.34 (0.96, 1.87)	1.50 (1.06, 2.14)
Age		*
Education		*
Black		0.33 (0.17, 0.65)
Hispanic		0.54 (0.23, 1.30)
Married		0.90 (0.56, 1.45)
Earnings (1995)		3.84 (2.58, 5.71)
Proportionality test (based on deviance statistics)	Chi-square statistic (p-value) 4.6 (0.2)	Chi-square statistic (p-value) 25.1 (0.2)

* Age and education have been modeled using natural cubic splines; the method does not give a summary odds ratio and CI per unit increase.

Table 5 summarizes the results obtain by fitting the proportional odds models. It is apparent from Table 5 that no matter whether adjusting or not adjusting for potential confounders, both the exposures (earnings indicator of being zero in 1994 and 1995) are found to have statistically no significant impact on the earnings in 1998. Moreover, the association remains statistically

insignificant while model is fitted after including unmeasured (generated) confounders as well. Although nonlinear effects are assumed for age and education, they do not appear to have nonlinear effects (p-values > 0.26). However, the proportionality assumptions of the odds models hold as can be observed from the test results in Table 5.

Table 5: Results from proportional odds models fit to evaluate the effect of earnings in 1994 and 1995 on the earnings in 1998.

	Model -1 (Unadjusted for confounders)	**Model -2 (Adjusted for confounders)**
Variables	Odds Ratio (95% CI)	Odds Ratio (95% CI)
Earnings (1994)	0.94 (0.54, 1.66)	0.90 (0.50, 1.60)
Earnings (1995)	0.82 (0.49, 1.40)	0.86 (0.50, 1.48)
Age		*
Education		*
Black		0.34 (0.18, 0.65)
Hispanic		0.68 (0.30, 1.56)
Proportionality test (deviance statistics)	Chi-square statistic(p-value) 2.16 (0.71)	Chi-square statistic (p-value) 10.15 (0.60)

* Age and education have been modeled using natural cubic splines; the method does not give a summary odds ratio and CI per unit increase.

5. Discussion

In this study we primarily intend to evaluate the effect of job training on the post training earnings (in 1998) and also on the change in earnings before (1995) and after the training program (1998). Because the responses are zero inflated non-negative variables, proportional odds model has been considered to evaluate the objectives. After adjusting for potential confounders, the job training has found to have statistically significant impact on future earnings as well as on the change in earnings before and after the program. Further, it is aimed to assess whether the pre training earnings (in 1994 and 1995) has had impact on post training earnings (in 1998). The association however has not been confirmed. Using the proportional odds model before and after adjusting for potential confounders, pre training earnings have found to have statistically no significant impact on post training earnings.

Two obvious concerns with the proportional odds model are that the way the positive scale is collapsed into categories is arbitrary, and by grouping the data one loses some information. It actually models grouped data instead of the original data. However, we consider collapsing positive scale in different ways. Fitting models on differently scaled categorical outcomes, the effects of exposures do not change. This ensures the robustness of the used model and findings of the study. Also, age, if treated as a categorical variable, might have an interaction effect with other factors. This analysis avoids examining any interaction effect in the sake of simple interpretation of the results.

Finally, a randomized control trial could have done with the goal of evaluating all the objectives discussed here. In that case with a large sample size, two groups of people are expected to be exchangeable with respect to all measured and unmeasured confounders. So the effect measure would be almost free from confounding if randomization is done properly.

6. References

1. The Saylor Foundation. "Unemployment Rate." pp. 1. Retrieved 20 June 2012
2. "Global employment trends 2013" (PDF). *International Labour Organization*. 21 January 2013.
3. Tobin, J. (1958), Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24–36.
4. Duan, N., Manning, W. G. Jr., Morris, C. N., and Newhouse, J. P. (1983), A comparison of alternative models for the demand for medical care (Corr: V2 P413). *Journal of Business and Economic Statistics*, 1, 115–126.
5. Heckman, J. (1974), Shadow prices, market wages, and labor supply. *Econometrica*, 42, 679–694.
6. Jørgensen, B. (1987), Exponential dispersion models. *Journal of the Royal Statistical Society*, Series B, Methodological, 49, 127–145
7. Saei, A., Ward, J. and McGilchrist, C. A. (1996), Threshold models in a methadone programme evaluation. *Statistics in Medicine*, 15, 2253–2260.