British Journal of Economics, Management & Trade
14(3): 1-7, 2016, Article no.BJEMT.26043
ISSN: 2278-098X

SCIENCEDOMAIN *international*
www.sciencedomain.org

Tweedie Model for Analyzing Zero-Inflated Continuous Response: An Application to Job Training Data

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

[1]Department of Statistics, Biostatistics and Informatics, University of Dhaka, Bangladesh.
[2]School of Business and Economics, United International University, Bangladesh.

Authors' contributions

This work was carried out in collaboration between all authors. Author NP designed the study and wrote the first draft of the manuscript. Author MASM analyzed the data while author MA carried out the review of literature. All authors interpreted the results of data analysis, read and approved the final manuscript.

Article Information

DOI: 10.9734/BJEMT/2016/26043
Editor(s):
(1) O. Felix Ayadi, Interim Associate Dean and JP Morgan Chase Professor of Finance, Jesse H. Jones School of Business, Texas Southern University, TX, USA.
(2) Jie-hua Xie, NanChang Institute of Technology, Department of Science, China.
Reviewers:
(1) R. Gopalakrishnan, Anna University Regional Campus Madurai, Tamil Nadu, India.
(2) Mohsin Nadeem, Middlesex University, UK.
(3) H. C. Madubueze, Madumelu, Chukwuemeka Odumegwu Ojukwu University, Anambra State, Nigeria.
(4) Alaedin K. I. Alsayed, Al-Aqsa University, Palestine.
Complete Peer review History: http://sciencedomain.org/review-history/15344

Original Research Article

Received 30th March 2016
Accepted 23rd June 2016
Published 9th July 2016

ABSTRACT

Continuous data with substantial proportion of zero values arise often in many disciplines. Modeling such zero-inflated data is always challenging. We use Tweedie model to analyze zero-inflated continuous outcome with a view to evaluate the effect of job training on future earnings, and also on the difference between pre- and post-training earnings. We further assess the effect of pre-training earnings on the post-training earnings. We used data from a job training program in the USA where 445 subjects were followed for three years. Results suggest that job training has statistically significant impact (p-value <0.05) on future earnings, as well as on the change in pre- and post-training earnings. The effect of pre-training earnings on post-training earnings is however not found to be statistically significant. We found the Tweedie model to be particularly suitable for analyzing zero-inflated data to make valid statistical inference.

**Corresponding author: E-mail: ahshan@bus.uiu.ac.bd;*

Keywords: Zero-inflated data; Tweedie model; job training and future earning.

1. INTRODUCTION

Unemployment is one of the biggest problems all over the world especially in the developing countries. The world unemployment rate was estimated as 9.2% in 2012 [1]. The International Labor Organization (ILO) also projected that the world unemployment rate will be approximately 9% in 2016 [2]. It is one of the major causes which create poverty, crimes, family problems, social division and depressions among the people [3]. To alleviate this crucial problem, different types of job training programs are conducted by the government and non-government organizations in the developed countries. From these programs a person can take training to increase her/his knowledge and skills for doing a particular job so that s/he can elevate her/his income and lead a better life. The training has positive impact on economic returns at the individual level [4]. The human capital theory also proposes that workers with higher level of education trends to have higher wages [5]. Many statistical studies also show that workers who receive work-related training also earn higher wages [6]. For instance, one study indicated that 1% point increase in training is associated with an increase in hourly wages of about 0.3% [7]. Apart from job training, other factors that might influence the future earnings of an individual include her/his education, age, past job experience, gender, family income, race, family size, etc. [8].

In the present study, we intend to assess whether the job training affects the future earnings of an individual using a data set from a job training program, which took place in 1995-97 in the USA. The data, however, contained an outcome variable (real earnings after job training) which is zero inflated. That is, the response is a non-negative continuous variable with many zero values. Typical linear regression model to analyze such a data may yield seriously biased estimates and sometimes inference can be misleading [9].

While several competing methods are available to deal with such data, we adopt compound Poisson exponential dispersion model (i.e., Tweedie Model) that perhaps the most appealing statistical model under this situation. We evaluate the effect of job training on future earnings after adjusting for potential confounders. We also estimate the effect of job training on change in earnings before and after the job training. Moreover, we assess the effect of pre-training earnings on post-training earnings.

2. METHODS

2.1 Study Design and Participants

The data have been obtained from a job training program in the USA between 1995 and 1997. A total of 445 subjects aged between 17 and 55 were followed till 1998. Most of them (about 83%) were black. Among all followed up individuals, 42% (185 subjects) underwent the training program. The training program was not randomly assigned to the individuals.

2.2 Variables

To assess whether the job training affects the future earnings, some of the base line characteristics such as age (in the year 1995), education, race (black/non-black), ethnicity (Hispanic or not), marital status (yes/no), pre-training earnings (in 1994 and 1995) of all individuals were measured so that we can adjust for potential confounder and take care of effect modifier.

2.3 Outcomes

We consider future earnings (i.e., earnings in 1998) as outcome to assess its dependency on job training, and also on pre-training earnings. Moreover, change in pre- and post- training earning (i.e., change in earnings between 1995 and 1998) has been taken as the outcome variable to evaluate the effect of job training on it.

2.4 Exposure

In this study we consider two main exposures, namely, job training (yes/no) and pre-training earnings (i.e., earnings in 1994 and 1995).

2.5 Potential Confounders and Mediators

To address all objectives, confounders have been selected based on literature and scientific consideration. For the first objective, that is to see the effect of job training on future earnings, potential confounders include age, education, race, ethnicity, marital status, earnings in 1994 and 1995. All of these covariates except for

earnings in 1994 have been included as confounders to evaluate the effect of job training on the income difference (pre- and post-training). Finally, to evaluate the third objective (i.e., to assess the impact of pre-earnings on post-training earnings), age, education, race and ethnicity have been included as potential confounders, while marital status has been considered as a mediator. We assume that pre-earnings affect marital status which in turn affects post-earnings, and as such, marital status has been regarded as a mediator.

2.6 Statistical Analysis

The outcome variable earnings in 1998 is a non-negative continuous variable and is also evident to be zero-inflated. For such zero-inflated (semi-continuous) response, where there is positive probability of a zero outcome, typical statistical model such as normal, gamma, log normal is inappropriate. There are several competing methods which can be applied to deal with such outcomes (which are continuous but are taking many zero values). Some of the approaches include

(1) Adding 1 to outcomes and then taking log so that it maps zero to zero and other values to log scale. Similar approach involves adding a small constant to outcomes and taking log. Sometimes this is problematic if there are more than a few zeros, because the constant is arbitrary and the answer may depend on it.
(2) Considering the zeros as censoring, i.e. values below a particular detection limit (see, [10]). This is sensible in some cases but fairly difficult to fit into the generalized linear model framework.
(3) Splitting the problem into two parts, i.e. analyze the zero/non-zero distribution as a binary variable and analyze the conditional (non-zero) data as Gamma or log-normal or whatever appropriate.
(4) Using the Tweedie distribution that has a point mass at zero plus a continuously distributed component. The model using Tweedie distribution is known as Compound Poisson exponential dispersion models. This is an appealing technique to deal with semi-continuous data as it makes possible to analyze data within a single model.

We therefore have adopted compound Poisson exponential dispersion models (Tweedie models) to assess all our objectives. In each case models have been fit after adjusting and not adjusting for potential confounders. The Tweedie model has been described in the following sub section. All categorical variables are modeled using indicator variables. Continuous variables (age, education) are modeled by using smoothing splines [11] with smoothness selection by generalized cross validation (GCV) (see, e.g., [12]).

2.6.1 Compound poisson exponential dispersion models (Tweedie models)

For modeling zero-inflated (semi-continuous) response, Jorgenson [13-14] proposed a type of compound Poisson distribution which belongs to the exponential dispersion family. The exponential dispersion models (EDMs) have density functions or probability mass functions of the form

$$f(y;\theta,\phi) = c(y,\phi)\exp\left[\frac{1}{a(\phi)}\{y\theta - b(\theta)\}\right],$$

for some known functions $a(.)$, $b(.)$ and $c(.)$. Here θ is the canonical parameter satisfying $b(\theta) < \infty$ and ϕ is the dispersion parameter. The mean of the distribution is $\mu = b'(\theta)$ and the variance is $\phi b'(\theta)$, where ' denotes the derivative with respect to θ. The mapping from θ to μ is invertible. So we may write $b''(\theta) = V(\mu)$ for a suitable function $V(\mu)$ known as variance function of the exponential dispersion model.

An important class of exponential dispersion model uses the variance function of the form $V(\mu) = \mu)^p$ for some p. These families are known as Tweedie models as the underlying linear exponential families were first systematically studies by Tweedie [15]. Jorgenson [13] showed that Tweedie exponential dispersion models exist for all values of p outside the interval $(0,1)$.

For $p = 1$, the EDM has the Poisson distribution function. For Tweedie EDMs, the cumulant function $b(\theta)$ and mean can be obtained by equating $b''(\theta) = \mu^p$ and solving for μ and b. Setting the arbitrary constants of interaction to zero yields,

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p}, p \neq 1 \\ log\mu, p = 1 \end{cases} \text{ and } b(\theta) = \begin{cases} \frac{\mu^{2-p}}{1-p}, p \neq 2 \\ log\mu, p = 2 \end{cases}.$$

The remaining factor is the density, $c(y,\phi)$ is relatively difficult to derive. With link function g,

we can specify a model for the mean response as $g(\mu_i) = x_i'\beta$. Obtaining the ML estimator for β does not involve $c(y_i, \emptyset)$. When p is known, this model can be fitted with software for generalized linear models. In practice, however, p would itself be unknown and need to be estimated. To fit the compound Poisson exponential model, we use 'gam' function under 'mgcv' package in software 'R'. The maximum likelihood method has been used to estimate p.

3. RESULTS

Table 1 presents summaries of all characteristics with respect to the individual's participation status (yes/no) in the job training program. For quantitative variables, mean and confidence interval (CI) are reported whereas for factor variables, count and percentages are provided. The mean age, years of schooling, and proportion of married, black and Hispanic people are almost similar in two groups who participated in the training program and who did not. People with at least high school diploma participated more in the training program. Again, participants in the training program have higher average income than non-participants. Average baseline earning is also slightly higher for those who participated in the program than who did not.

Table 2 shows bivariate relationships between income in1998 and some important variables. It is observed that the amount of earnings does not depend on age, earnings before getting training, marital status, race, and ethnicity. Again, percentages of people with high school diploma are higher in the upper income group.

Table 3 shows the estimates and confidence intervals of the compound Poisson exponential dispersion (Tweedie) models parameters. In unadjusted model, the parameter estimate of the job training indicator has found to be statistically significant (odds ratio is different from 1; using t-test the p-value < 0.05) at 5 percent level. From the adjusted model fit, it is evident that (after adjusting for potential confounders) job training has statistically significant (t-test has p-value < 0.05) impact on earnings in 1998. Both the continuous variables age and year of schooling do not show non-linear effects (p-values are 0.21 and 0.06, respectively).

Table 4 presents the parameter estimates and confidence intervals of the compound Poisson models for assessing the effect of job training on change in earnings. From unadjusted model, the job training coefficient has not found have statistically significant (t-test has p-value = 0.31) at 5 percent level. However, the results from adjusted model suggest that (after adjusting for potential confounders) job training has statistically significant (odds ratio is different from 1; using t-test the p-value < 0.05) impact on the change in earning from 1995-1998. As before, the continuous variables age and year of schooling are not showing non-linear effects (p-values are 0.4 and 0.3, respectively).

Parameter estimates and confidence intervals of the Compound Poisson exponential dispersion (Tweedie) models for assessing the effect of pre-training earnings in 1994 and 1995 on post-training earnings in 1998 have been presented in Table 5. It is evident from both adjusted and unadjusted models that the coefficient of pre-training earnings is not statistically significant (odds ratio is not different from 1; using t-test the p-value = 0.23). While the year of schooling shows non-linear effect (p-value is 0.04), the age does not demonstrate any non-linear association with response variable (p-values is 0.22).

Table 1. Relation between exposure and important covariates, mean (95% CI) are provided; for factor variables, counts and percentages are reported

Characteristics	Participated in training program	Not participated in training program
Age in years	25.1 (24.2, 25.9)	25.8 (24.8, 26.9)
Education- years of schooling	10.1 (9.9,10.3)	10.3 (10.1 ,10.6)
Black	156 (84.3%)	215 (82.7%)
Hispanic	11 (5.9%)	28 (10.8%)
Married	35 (18.9%)	40 (15.4%)
High school diploma (Yes)	54 (29.2%)	43 (16.54%)
Real earnings in 1994	2095.6 (1386.8, 2804.4)	2107.0 (1412.4 , 2801.7)
Real earnings in 1995	1532.1 (1065.1, 1999.0)	1266.9 (888.0,1645.9)
Real earnings in 1998	6349.2 (5208.0, 7490.3)	4554.8 (3885.1, 5224.5)

Table 2. Relation between outcome (Income in 1998) and important covariates. For quantitative variables, mean (95%CI) are provided; for factor variables, counts and percentages are reported

	Income in 1998 (annual income in dollars)				
	0	**1 – 3000**	**3001 – 6000**	**6001 - 9000**	**9001 +**
Age	25.5 (24.3, 26.6)	24.9 (23.3, 26.4)	24.2 (22.6, 25.9)	26.2 (23.9, 28.5)	26.0 (24.7, 27.3)
Education - years of schooling	10.2 (9.9, 10.5)	10.4 (9.9, 10.8)	9.8 (9.4, 10.2)	9.9 (9.3, 10.4)	10.6 (10.3, 10.9)
Black	127 (92.7%)	62 (88.6%)	63 (77.8%)	42 (73.7%)	77 (77.0%)
Hispanic	5 (3.7%)	5 (7.1%)	10 (12.4%)	10 (17.5%)	9 (9.0%)
Married	21 (15.3%)	11 (15.7%)	14 (17.3%)	10 (17.5%)	19 (19.0%)
High school diploma (Yes)	26 (19.0%)	17 (24.3%)	14 (17.3%)	10 (17.5%)	30 (30.0%)
Income in 1995 (in dollars)					
0	95 (69.3%)	45 (64.3%)	50 (61.7%)	39 (68.4%)	60 (60.0%)
1-3000	25 (18.3%)	16 (22.9%)	17 (21.0%)	9 (15.8%)	23 (23.0%)
3001-6000	10 (7.3%)	6 (8.6%)	6 (7.4%)	4 (7.0%)	8 (8.0%)
6001-9000	2 (1.5%)	2 (2.9%)	4 (4.9%)	3 (5.3%)	4 (4.0%)
9001+	5 (3.7%)	1 (1.4%)	4 (4.9%)	2 (3.5%)	5 (5.0%)
Income in 1994 (in dollars)					
0	106 (77.4%)	51 (72.9%)	58 (71.6%)	40 (70.2%)	71 (71.0%)
1-3000	9 (6.6%)	4 (5.7%)	12 (14.8%)	4 (7.0%)	10 (10.0%)
3001-6000	7 (5.1%)	5 (7.1%)	3 (3.7%)	4 (7.0%)	3 (3.0%)
6001-9000	5 (3.7%)	4 (5.7%)	3 (3.7%)	5 (8.8%)	5 (5.0%)
9001+	10 (7.3%)	6 (8.6%)	5 (6.2%)	4 (7.0%)	11 (11.0%)

Table 3. Results from compound poisson exponential dispersion (Tweedie) models fit for assessing the effect of job training on future earnings

Variables	**Unadjusted model Parameter estimates (e^{β}) (95% CI)**	**Adjusted model Parameter estimates (e^{β}) (95% CI)**
Job training	1.4 (1.1, 1.8)	1.3 (1.1, 1.7)
Age (in years)		*
Education in years		*
Black		0.7 (0.5, 1.1)
Hispanic		1.1 (0.6, 1.9)
Marital Status		1.0 (0.7, 1.4)
Earnings indicator in 1994		1.0 (0.7, 1.5)
Earnings in indicator in 1995		0.8 (0.6, 1.2)

** Year of schooling (education) and age are modeled using smoothing splines. A summary of odds ratio and CI are not provided by the method.*

For all model fits, the normalized residual plots have been considered to check the models adequacies and the plots raise no concerns about the models fits.

4. DISCUSSIONS

The aim of this study is to assess the effect of job training on the future earnings (in 1998) as well as on the difference in earnings before and after the job training program using a data set from the USA. We also intend to evaluate the effects of pre-training earnings (in 1994 and 1995) on future earnings (in 1998). The outcome variable future earning is zero inflated, and as such, an appropriate statistical tool which can deal with such semi-continuous response is required. To that end, we adopt the compound Poisson exponential dispersion (Tweedie) models to assess the effects of exposure on outcome. After adjusting for potential confounders, job training program has found to have statistically significant

Table 4. Compound poisson exponential dispersion (Tweedie) models fit for assessing the effect of job training on change in earnings from 1995-98

Items	Unadjusted model Parameter estimates (e^{β}) (95% CI)	Adjusted model Parameter estimates (e^{β}) (95% CI)
Job training	1.5 (0.9, 2.2)	1.5 (1.1, 2.3)
Age (in years)		*
Education in years		*
Black		0.4 (0.2, 0.9)
Hispanic		0.9 (0.3, 2.6)
Marital Status		0.9 (0.5, 1.6)
Earnings indicator in 1995		2.0 (1.3, 3.1)

** Year of schooling (education) and age are modeled using smoothing splines. A summary of odds ratio and CI are not provided by the method*

Table 5. Results from compound poisson exponential dispersion (Tweedie) models fit for assessing the effect of earnings in 1994 and 1995 on future earnings in 1998

Items	Unadjusted model Parameter estimates (e^{β}) (95% CI)	Adjusted model Parameter estimates (e^{β}) (95% CI)
Earnings indicator in 1994	1.08 (0.73, 1.60)	1.06 (0.72, 1.57)
Earnings in indicator in 1995	0.76 (0.53, 1.09)	0.79 (0.55, 1.13)
Age (in years)		*
Education in years		*
Black		0.79 (0.46, 1.04)
Hispanic		1.02 (0.58, 1.78)

** Year of schooling (education) and age are modeled using smoothing splines. A summary of odds ratio and CI are not provided by the method*

impact (odds ratio: 1.3; CI : (1.1, 1.7) ; using t-test the p-value < 0.05) on future earnings, as well as on the difference in pre- and post training earnings (odds ratio: 1.5; CI :(1.1, 2.3) ; using t-test the p-value <0.05). However, the pre-training earnings in 1994 and 1995 are not found statistically significant (t-test has p-value = 0.23) on affecting future earnings.

Due to unavailability of the recent data, we use historical data (1995-1998) and results from this study are expected to facilitate the use of appropriate statistical model to analyze similar data and also to contribute in the realm of contemporary research and knowledge.

For simplicity, we did not consider any interaction effects in the model which could have revealed some interesting facts. To evaluate all the objectives in this study, data from a randomized control trial (RCT) could be preferred because the RCT ensures that two groups of people are exchangeable for a reasonably large sample and hence no unmeasured confounder would be left to adjust for.

5. CONCLUSION AND RECOMMENDATION

We found that job training significantly affects the future earnings. Similar findings have been reported previously under different study designs and using different data. However, we found the future earnings are seriously zero-inflated which is not unexpected in general. We applied the most appropriate statistical model to analyze such data and our findings would be more accurate and interpretable. Based on our findings from this study, we strongly recommend using Tweedie model to analyze zero-inflated data to obtain valid statistical results.

Using recent data on job training program and analyzing such data by Tweedie model and then investigating how the model performance is affected if alternative statistical approaches are used would be a topic worth future research.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Saul DH. On-the-job training: Differences by race and sex. Research Summaries. Monthly Labor Review; 1981. Available:http://www.bls.gov/opub/mlr/1981/07/rpt3full.pdf
2. Sedhi A. Global unemployment forecast to hit 212m - country by country breakdown, The Gurdian; 2015. Available:http://www.theguardian.com/news/datablog/2015/jan/20/global-unemployment-forecast-to-hit-212m-country-by-country-breakdown
3. Graham W. The Poverty of conventional economic wisdom and the search for alternative economic and social policies. The drawing board: An Australian Review of Public Affairs. 2001;2(2):67-87.
4. Blundell R, Dearden L, Meghir C, Sianesi B. Human Capital Investment: The returns from education and training to the individual, The Firm and the Economy. Fiscal Studies. 1999;20(1):1-23.
5. Weiss A. Human capital vs. Signalling explanations of wages. The Journal of Economic Perspectives. 1995;9(4):133-154.
6. Arulampalam W, Booth A. Training and labor market flexibility: Is there a trade-off. British Journal of Industrial Relations. 1998;36(4):521-536.
7. Dearden L, Reed H, Van Reenen J. The impact of training on productivity and wages: Evidence from British panel data. Oxford Bulletin of Economics and Statistics. 2006;68:397-421.
8. Morgan J, David M. Education and Income. The Quarterly Journal of Economics. 1963;77(3):423-437.
9. Min Y, Agresti A. Modeling nonnegative data with clumping at zero: A survey. Journal of the Iranian Statistical Society. 2002;1:7-33.
10. Tobin J. Estimation of relationships for limited dependent variables. Econometrica. 1958;26(1):24–36.
11. Hastie T, Tibshirani R. Generalized additive models. Chapman and Hall: New York; 1990.
12. Wood S. Generalized additive models: An introduction with R. Chapman and Hall: New York; 2006.
13. Jorgensen B. Exponential dispersion models. Journal of Statistical Society. 1987;49:127-162.
14. Jorgensen B. Theory of dispersion models. Chapman & Hall, London; 1997.
15. Tweedie MCK. An index which distinguishes between some important exponential families. Statistics: Applications and new directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference (Eds. J. K. Ghosh and J. Roy), Calcutta: Indian Statistical Institute. 1984; 579-604.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/15344