Adv. & Appl. in Stat. **4**(3) (2004), 253-264

EXTRA-VARIATION AND BAYESIAN BINOMIAL REGRESSION MODELS FOR ESTIMATING EARNINGS PROBABILITIES USING A THREE-WAY LAYOUT INCOME DATA

IBRAHIM M. ABDALLA

Department of Statistics, College of Business and Economics
United Arab Emirates University, Al-Ain, P. O. Box 17555, U. A. E.
e-mail: i.abdalla@uaeu.ac.ae

Abstract

This paper attempts to obtain improved estimates of probabilities in a three-way cross-classification of categorical variables; namely, education level, labor market sector and workers' nationality. Analysis is based on data made available from Abu-Dhabi Emirate Family Expenditure Survey, 1997, to estimate probability of low earnings. The sample proportion, the maximum likelihood estimate, is the classical estimator for this parameter. However, the estimate may possess undesirable features, particularly, when the data are sparse. Alternatively, a main effects model could be fitted to the data, seeking relationships between the proportion at each combination and certain levels of the categorical variables involved. This could produce smaller standard errors, but does not take into account uncertainty associated with model parameters and may not account for extra sampling variance in the data. An expanded model that accounts for over-dispersion in the data could improve the fit. However, a random effect model through Bayesian estimation, based on Markov Chain Monte Carlo (MCMC) simulation implemented using WinBUGS software, provides better estimates that compromise between the two classical and main effects approaches.

2000 Mathematics Subject Classification: 62-XX.

Key words and phrases: logistic regression, extra-variation, earnings probabilities, Bayesian methods.

Received April 28, 2004

1. Introduction

Measurement of inequality in the distribution of earnings reveals wealth concentration and shows whether it is in line with generally accepted social values [4]. Monitoring inequality is necessary for effective administration of government policies such as social welfare programs and policy reforms aimed at economic liberalization and employment assignments into different labor market sectors. Human capital theory predicts that, in unregulated labor markets, workers' remuneration depends on their individual productivity, which is a function of experience, education and skill levels. Inequality in workers' pay not attributable to perceived productivity and income generating characteristics might reflect differences in employers' treatment and could be regarded as discrimination [1].

Mincer [10] advocates an earnings linear regression model based on employing a continuous response variable, viz., earnings or logarithm of earnings, in order to capture the effect of general human capital represented by years of schooling and labor market experience. Motivated by this, the question of earnings distribution discussed in this article is based on utilizing a methodology using data available on categorical scales [3]. Of particular interest is the Logit modeling technique for categorical data which shares many features with linear models for continuous data.

Given three categorical variables with I, J and K levels, let r_{ijk} represent the number of units identified with a certain characteristic and associated with a sample size n_{ijk} for each possible combination (i, j, k) of levels. Consider r_{ijk} as realizations of independent binomial random variables, each with a parameter p_{ijk} and index n_{ijk}. A typical scenario addresses the question of classifying labor market employees by labor market sectors, nationality and educational levels, according to their earnings or income. In this study, the p_{ijk} describes the proportion of low income earners, who earn less than three thousand U.A.E. Dirhams per month (approximately 800 US dollars), in education level i, labor market sector j and nationality group k. The U.A.E. has no official poverty line, therefore the definition of low income earnings of below three thousand U.A.E. Dirhams per month is arbitrary and has no basis in absolute and relative terms. However, employees receiving monthly pay below this threshold are regarded as disadvantaged as they are faced with growing economic costs, including housing, health and child education.

The problem considered in this paper relates to getting improved estimates of the p_{ijk}, probabilities of low earnings, and to comparing the performance of some proposed models of estimation. These estimates can be used to predict future proportions of low income earners in a labor market. This would have wide implications from a policy perspective. Particular interest may be focused on assessing how a labor market functions and what factors affect earnings. This is a pre-requisite for proactive measures that seek to reduce labor market discrimination.

The maximum likelihood (ML) estimates of p_{ijk}, the sample proportion r_{ijk}/n_{ijk}, provide the classical unbiased estimates which depend on the data only in the specific level combination (i, j, k) and make no attempt to take advantage of global patterns in the data. ML estimates may possess other undesirable features such as large standard errors, particularly when the data are sparse [3]. Another alternative for estimating p_{ijk}, is to fit a main effects model that pools information across possible combinations [2, 5]. Such model-based estimators might suffer from extra sampling variation which may inflate the significance of model parameters.

Attempts to deal with previous defects include expanding the main effects model to incorporate a component of extra random variation between the model observations [12]. Recent attempts are mainly based on utilizing Bayesian approaches to estimating p_{ijk} by incorporating random effects model-based inference (see for example [5, 7, 8, 9]). These approaches might result in improved model fit and allow the data to determine a compromise between classical and main effects models estimates.

The objective of this paper is outlined as follows. The paper attempts to account for extra-binomial random variation between observed counts of low income earners, using extra-variation binomial regression as well as employing a full Bayesian framework in order to estimate p_{ijk}, probability of low income earnings. The techniques are demonstrated empirically by analyzing data on income distribution in the Emirate of Abu-Dhabi. Empirical analysis is based on the 1997 survey of family expenditure in the Emirate. The problem can be set up mathematically as shown in Section 2. Brief description of the data used to demonstrate the techniques is presented in Section 3. Section 4 presents the tools employed in the analysis. Analysis results and comparisons between different models are briefly discussed in Section 5.

2. The Models

Let labor market employees in education level i $(i = 1, \dots, I)$, employment sector j $(j = 1, \dots, J)$ and nationality group k $(k = 1, \dots, K)$ be classified according to their earnings. It is assumed that in education level i, employment sector j and nationality group k, the observed number of low income earners r_{ijk}, follows a binomial law with sample size n_{ijk} and true rate p_{ijk},

$$f(r_{ijk} \mid n_{ijk}, p_{ijk}) = \binom{n_{ijk}}{r_{ijk}} p_{ijk}^{r_{ijk}} (1 - p_{ijk})^{n_{ijk} - r_{ijk}}, \quad 0 \le r_{ijk} \le n_{ijk}. \tag{1}$$

A common approach is to estimate the p_{ijk} by fitting a main effects analysis of variance model utilizing the Logit link, that is,

$$\text{Logit}(p_{ijk}) = \eta_{ijk},$$

$$\eta_{ijk} = \mu + \alpha_i + \beta_j + \lambda_k, \quad i = 1, \dots, I, \; j = 1, \dots, J, \; k = 1, \dots, K.$$

2.1. Extra-variation model

The logistic regression model described above is a well known model that employs binary responses. In certain situations, however, residual variation may be greater than can be attributed to the binomial sampling variation assumed by the model. This may arise due to correlated cells' counts and that probabilities p_{ijk} are not constant among different cells. Such violations might lead to under inflating sampling variance.

Suppose that the probability of low income earnings p_{ijk} in each cell is a random variable and that it follows a beta distribution [6, 12] with mean and variance given by

$$E(p_{ijk}) = \frac{\gamma}{\gamma + \delta} = \theta$$

$$\mathrm{Var}(p_{ijk}) = \frac{\theta(1 - \theta)}{\gamma + \delta - 1} = \varphi\theta(1 - \theta).$$

For $\gamma > 1$ and $\delta > 1$, the beta density is equal to zero at both zero and one, that is, $0 < \varphi \leq 1/3$. Based on this, the unconditional mean and variance of r_{ijk} are given by

$$E(r_{ijk}) = n_{ijk}\theta$$

$$\mathrm{Var}(r_{ijk}) = n_{ijk}\theta(1 - \theta)(1 + (n_{ijk} - 1)\varphi).$$

For all values different from $n_{ijk} = 1$ or $\varphi = 0$, the unconditional variance of r_{ijk} is larger than the binomial variance (over-dispersed binomials).

Williams [12] proposed a weighted logistic regression to allow for extra-binomial variation. The weights introduced are given by $w_{ijk} = (1 + (n_{ijk} - 1)\varphi)^{-1}$, where φ is estimated using an iterative procedure based on the expectation of χ^2.

2.2. Bayes model

The Bayes approach advocates a random effect model that allows for over-dispersion caused by unobserved factors. The model is based on the assumption that the parameters p_{ijk} of different binomial distributions are related. The Bayes estimates provide a compromise between the estimates r_{ijk}/n_{ijk}, sample proportion, and estimates based on a main effects logistic regression model, described earlier, under the assumption of independence of the p_{ijk}. Bayes estimates of p_{ijk} based on small sample sizes n_{ijk} are adjusted towards the global patterns suggested by the data as a whole, but estimates based on large sample sizes change slightly. The Bayes model can be described as follows. Assume that η_{ijk} follows a normal prior,

$$\eta_{ijk} \sim N(\mu + \alpha_i + \beta_j + \lambda_k, \tau).$$

The regression hyper-parameters μ, α_i, β_j, and λ_k are assigned normal priors with common mean zero and precision v [7]. The precision τ is assumed to follow a Gamma distribution with parameters v_1 and v_2. Consequently, a Bayesian model specification is completed. Given the data D, the posterior density of $\boldsymbol{\eta} = (\eta_{ijk})^{\tau}$, $\boldsymbol{\alpha} = (\alpha_i)^{\tau}$, $\boldsymbol{\beta} = (\beta_j)^{\tau}$ and $\boldsymbol{\lambda} = (\lambda_k)^{\tau}$ can be written as

$$\prod(\boldsymbol{\eta} \mid \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau) \propto \ell(D \mid \boldsymbol{\eta}) \prod(\boldsymbol{\eta} \mid \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau) \prod(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau), \quad (2)$$

where

$$\ell(D \mid \boldsymbol{\eta}) = A \exp\left\{\sum_i \sum_j \sum_k r_{ijk}(\eta_{ijk})\right\} \prod_i \prod_j \prod_k (1 + \exp(\eta_{ijk}))^{-n_{ijk}}, \quad (3)$$

$$A = \prod_i \prod_j \prod_k \frac{n_{ijk}!}{(n_{ijk} - r_{ijk})!\, r_{ijk}!},$$

$$\prod(\boldsymbol{\eta} \mid \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau) \propto \exp\left\{\sum_i \sum_j \sum_k -\frac{\tau}{2}(\eta_{ijk} - \mu - \alpha_i - \beta_j - \lambda_k)^2\right\}, \quad (4)$$

$$\prod(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau) \propto$$

$$\tau^{IJK(v_1 - 1)} \exp\left\{-\frac{v}{2}\left(\sum_i \alpha_i^2 + \sum_j \beta_j^2 + \sum_k \lambda_k^2\right) - \sum_i \sum_j \sum_k v_2 \tau\right\}. \quad (5)$$

Integrating the joint posterior density with respect to relevant parameters, the full conditional distributions are given as

$$\prod(\boldsymbol{\eta} \mid D, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau) \propto \ell(D \mid \boldsymbol{\eta}) \prod(\boldsymbol{\eta} \mid \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau)$$

$$\prod(\mu \mid D, \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau) \propto \prod(\boldsymbol{\eta} \mid D, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau) \prod(\mu \mid v) \quad (6)$$

$$\prod(\boldsymbol{\alpha} \mid D, \boldsymbol{\eta}, \mu, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau) \propto \prod(\boldsymbol{\eta} \mid D, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau)\prod(\boldsymbol{\alpha} \mid v)$$

$$\prod(\boldsymbol{\beta} \mid D, \boldsymbol{\eta}, \mu, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau) \propto \prod(\boldsymbol{\eta} \mid D, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau)\prod(\boldsymbol{\beta} \mid v)$$

$$\prod(\boldsymbol{\lambda} \mid D, \boldsymbol{\eta}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau) \propto \prod(\boldsymbol{\eta} \mid D, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau)\prod(\boldsymbol{\lambda} \mid v)$$

$$\prod(\tau \mid D, \boldsymbol{\eta}, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) \propto \prod(\boldsymbol{\eta} \mid D, \mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \tau)\prod(\tau \mid v_1, v_2). \tag{7}$$

Generated samples of posterior distributions of model parameters are used to obtain parameter estimates via Gibbs sampling, implemented using WinBUGS package [11]. Gibbs sampling provides a Markov chain simulation of a random walk in the parameter space, which converges to a stationary distribution approximating the joint posterior distribution. The stationary distribution provides both estimates of parameter posterior mean, and variance.

3. The Data

The data set employed in the analysis is comprised of 2784 employees receiving market wages in Abu-Dhabi Emirate according to 1997 family expenditure survey. The data assess employees' earnings (low, high) as it relates to employment sectors (public and private, where the private sector is the reference level), nationality groups (Emiratis and expatriates, where the expatriate group is the reference group) and education levels (seven categories ranging from illiterate level to university level and above, where the illiterate education level is the reference level). The nationality variable is included in the analysis as a proxy variable in order to reflect government policies that encourage different labor market establishments, particularly in the private sector, to set efficient pay levels in order to attract nationals into their workforce. It can also suggest the ability of nationals to exploit other sources of secondary income that increase basic pay.

4. Models Fit

The maximum likelihood approach is used to estimate parameters of the main effects logistic regression. Parameter estimates are obtained using ARC software [6]. Estimates are in the directions anticipated. They are statistically significant at the 5% level, except two coefficients; namely, the read and write and the primary levels of workers' education.

To justify the use of extra-variation binomial model, it is important to have a measure of how much extra variation exists. The variance inflation factor $\widehat{c}$ can be used for this purpose. The measure is based on χ^2 statistic. When the model fits the data, the expected value of this statistic is equal to the number of degrees of freedom (df). Over-dispersion is estimated by how far the χ^2 statistic deviates from the expected value,

$$\widehat{c} = \frac{\chi^2}{\text{df}}.$$

That is, the expected value of $\widehat{c}$ is 1.0, with values of $\widehat{c} > 1.0$ indicating some degree of over-dispersion. Thus, based on the main effects logistic regression model fitted above, the obtained value for χ^2 is 38.90 and the model's degrees of freedom $\text{df} = 19$. This gives an estimated variance inflation factor $\widehat{c} = 2.05$. Based on this value, over-dispersion is detected, suggesting that the sampling variance must be inflated. Variance inflation is implemented by fitting an extra-variation binomial regression model using ARC software [6].

Gibbs sampling is performed with WinBUGS running for 50000 iterations with an initial burn-in of 10000 samples. A non-informative prior for each of the parameters μ, α_i, β_j and λ_k is approximated by a normal distribution with mean 0 and precision 10^{-6}. The precision parameters $\boldsymbol{\tau}$ follow a Gamma distribution with $\upsilon_1 = \upsilon_2 = 10^{-3}$ [11].

5. Results and Discussion

Table 1 displays parameter estimates for the main effects, extra-variation and Bayes binomial regression. The over-dispersion parameter associated with the extra-variation binomial model is estimated as $\widehat{\varphi} = 0.03$. Incorporation of the over-dispersion parameter into the model has inflated the variance (S.E.) associated with each parameter (see Table 1). As a result, some coefficients which were significant in the main effects model have turned insignificant; namely, the preparatory and the secondary levels of education, in addition to the read and write and the primary levels noted before. This suggests that the secondary education level and levels below have no significant influence on workers' low earnings compared to the illiteracy level. The rest of the coefficients have stayed significant at the 5% level. The Bayes estimates are closer to the estimates obtained from the main effects model than those associated with the extra-variation model (Table 1).

Results in Table 1 suggest that the strongest influence on the probability of low earnings is attributed to nationality factor, followed by the university/plus level of education, and then the sector of employment. It is evident from Table 1 that an increased number of nationals (Emiratis) in the Abu-Dhabi Emirate labor market decreases the probability of low earnings. This might be, as suggested earlier, a result of government policies to improve pay levels and rewards in establishments dominated by a national workforce. However, the fact that the nationality of workers is such an important factor influencing earnings might suggest that workers are not remunerated according to relative productivity.

The magnitude of low earnings is higher in the private sector and amongst expatriate workers compared to the public sector and the national workforce as depicted in Figure 1. Among both the public and the private sector employees, Figure 1 suggests that the scale of low earnings decreases with the increase in the level of education. This is particularly evident in the case of the private sector employees. This suggests that human capital in Abu-Dhabi Emirate, viz., educational attainments, positively influence earnings.

The mean square error (MSE) is used as a criteria for comparing predicted number of low income earners with the observed for each sector of employment. The advantage of the Bayesian approach is evident, Table 2, over the main effects and the extra-variation binomial regression models. The Bayesian model provides a better fit compared to the other two models. It is associated with a mean square error $(\text{MSE} = 4.05)$ followed by $\text{MSE} = 6.88$ for the main effects and $\text{MSE} = 21.65$ for the extra-variation binomial regression.

Table 1. Main effects, extra-variation and Bayes regression parameter estimates

	Main effects		Extra-variation		Bayes	
Variable	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Constant	0.5076	0.206	0.4322	0.351	0.4968	0.232
Public sector	–1.3602	0.107	–1.3612	0.244	–1.3600	0.139
Read/write	–0.1332	0.244	0.0157	0.428	–0.1261	0.276
Primary	–0.1623	0.262	–0.2210	0.452	–0.1421	0.290
Preparatory	–0.6007	0.243	–0.5461	0.442	–0.4982	0.277
Secondary	–0.8810	0.220	–0.7822	0.433	–0.8762	0.258
Diploma	–1.2587	0.230	–1.1784	0.462	–1.2660	0.273
University/plus	–1.9190	0.214	–1.6553	0.478	–1.8930	0.256
Emirati	–2.1026	0.185	–1.9655	0.347	–2.1270	0.203

Table 2. Observed and predicted numbers of low income earners under the main effects, extra-variation and Bayes regression models

		Public sector		Private sector	
Education	Predicted	Emirati	Expatriate	Emirati	Expatriate
Illiterate	Observed values	19.0	21.0	1.0	11.0
	Main effects	20.1	17.6	3.0	11.2
	Extra-variation	21.3	16.7	3.2	10.9
	Bayes	19.6	18.3	2.9	11.2
Read and write	Observed values	6.0	20.0	2.0	22.0
	Main effects	2.6	18.2	1.5	20.7
	Extra-variation	3.1	19.2	1.8	21.4
	Bayes	9.0	18.6	1.5	20.9
Primary	Observed values	11.0	14.0	0.0	16.0
	Main effects	5.6	18.1	0.3	17.0
	Extra-variation	5.6	16.4	0.3	16.0
	Bayes	6.3	17.5	0.3	16.9
Preparatory	Observed values	5.0	24.0	0.0	26.0
	Main effects	4.6	25.2	0.4	24.8
	Extra-variation	5.2	24.8	0.4	24.5
	Bayes	5.1	23.5	0.4	26.0
Secondary	Observed values	1.0	56.0	2.0	53.0
	Main effects	2.6	49.8	0.5	59.1
	Extra-variation	3.0	50.8	0.5	59.9
	Bayes	2.5	52.1	0.5	57.0

Diploma	Observed values	0.0	34.0	0.0	46.0
	Main effects	0.8	36.5	0.1	42.6
	Extra-variation	0.9	36.6	0.1	42.8
	Bayes	0.8	35.8	0.1	43.4
University and plus	Observed values	1.0	39.0	2.0	116.0
	Main effects	0.8	41.5	0.2	115.5
	Extra-variation	1.1	49.4	0.3	133.9
	Bayes	0.8	41.1	0.2	115.9

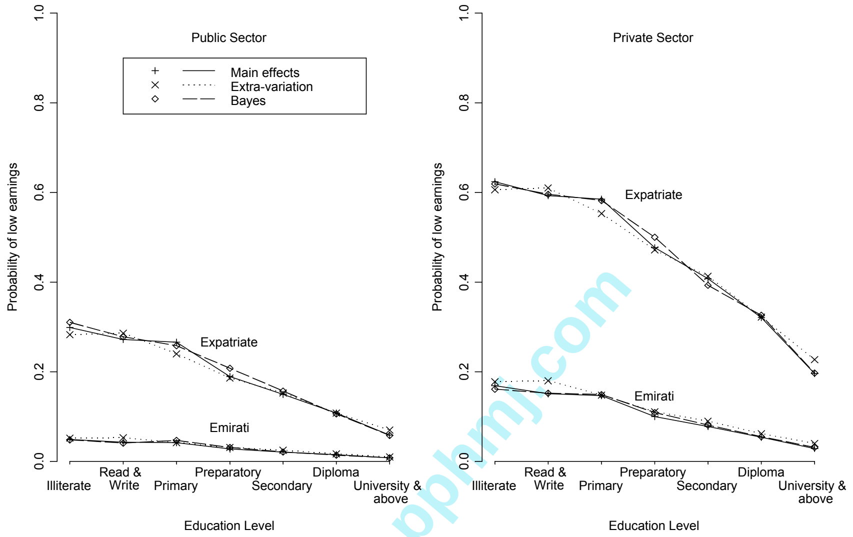

Figure 1. Probability of low income earnings based on main effects. Extra-variation and Bayes binomial regression estimates by labor market sector and nationality

References

[1] I. M. Abdalla, Sectoral assignment and earnings determinants and differentials in Abu-Dhabi Emirate, UAE, Arab J. Admin. Sci. 9(3) (2002), 457-475.

[2] A. Agresti, Categorical Data Analysis, John Wiley, New York, 1990.

[3] A. Agresti, An Introduction to Categorical Data Analysis, John Wiley, New York, 1996.

[4] J. Borland, Earnings inequality in Australia: Changes, causes and consequences, Economics Record 75 (1999), 177-202.

[5] P. Congdon, Bayesian Statistical Modelling, Wiley, New York, 2001.

[6] R. D. Cook and S. Weisberg, Applied Regression including Computing and Graphics, Wiley, New York, 1999.

[7] T. M. Laird, Empirical Bayes methods for two-way contingency tables, Biometrica 58 (1978), 581-590.

[8] B. MacGibbon and T. J. Tomberlin, Small area estimates of proportions via empirical Bayes techniques, Survey Methodology 15 (1989), 237-252.

[9] D. Malec, J. Sedransk, C. L. Moriarity and F. B. LeClere, Small area inference for binary variables in the National Health Interview Survey, J. Amer. Statist. Assoc. 92 (1997), 815-826.

[10] J. Mincer, Schooling, Experience and Earnings, NBER Press, New York, 1974.

[11] D. J. Spiegelhalter, A. Thomas, N. G. Best and D. Lunn, WinBugs User Manual, Version 1.4, Cambridge, England: MRC Biostatistics Unit, 2002.

[12] D. A. Williams, Extra-binomial variation in logistic linear models, Appl. Stat. 31(2) (1982), 144-148.

■