Presented at the Research Congress 2013
De La Salle University Manila
March 7-9, 2013

CLUSTERING AND MODEL FITTING OF OVERDISPERSED FOOD POVERTY DATA

Jacqueline Andan, Andrea Cortez & Shirlee Ocampo
Mathematics Department, De La Salle University
2401 Taft Avenue, Manila

ABSTRACT: Eradicating hunger and poverty is one of the major goals of many countries and local government units. Hunger profiles and food poverty models are needed to be generated from food poverty data in order to implement essential programs to the places where hunger incidence is high. In the Philippines, the National Statistics and Coordination Board (NSCB) uses the per capita income (PCI) criterion so as to classify whether a household is food poor or "hungry". If a household has a PCI below a food threshold set by NSCB for that locality, then it is considered as a food poor or "hungry" household. Model fitting of food poverty count data may encounter difficulties since there are times that food poverty count data may display high variability or overdispersion. Failure to model overdispersion may lead to serious underestimation of standard errors and misleading regression parameter estimates.

This research applied Poisson and negative binomial regression models to 2005 CBMS food poverty count data of Pasay City. Estimates of barangay level food poverty incidence were determined using Poisson and negative binomial (NB1, NB2) regression analyses, and then compared to actual proportions of food poor households in barangay level based on the per capita income (PCI) criterion. Cluster analysis was performed so as to characterize the barangays with high hunger incidence. Results show that the food poverty count data are overdispersed, and hence, negative binomial (NB2) regression model is preferred over Poisson regression model. Barangays with high food poverty incidence were clustered highlighting their common characteristics such as high proportion of households which are informal setters and under poverty threshold.

Key Words: food poverty; Poisson regression models; negative binomial regression models; overdispersed count data; cluster analysis

1. INTRODUCTION

The first of the United Nation's Millenium Development Goals (MDG) is to eradicate extreme poverty and hunger. Hunger, a form of deprivation is defined as the consumption of a diet inadequate to sustain good health and normal activity, growth and development (Millman and DeRose, 1998). Hunger profiles and food poverty models are needed to be generated from

food poverty data in order to implement essential programs to the places where hunger incidence is high. Food poverty data may display high variability or overdispersion, and this must be addressed in model fitting so as to avoid underestimation of standard errors and misleading parameter estimates.

This study fitted Poisson and negative binomial (NB) regression models to 2005 Community Based Monitoring System (CBMS) Pasay City food poverty data so as to obtain estimates of barangay level hunger incidence, and then to compare these generated estimates to actual proportions of food poor households based on the per capita income (PCI) criterion.

The PCI criterion is the hunger indicator officially used in generating hunger statistics in the Philippines. It involves the comparison between the diet actually consumed and what is required in terms of monetary values. A household will be classified food poor or "hungry" if it has a PCI below a food threshold set by National Statistics Coordinating Board (NSCB) for that locality. The food threshold is the minimum cost of the food items that will satisfy minimum nutritional requirements. The NSCB had set P11,199 per year as the food threshold in Pasay City in 2005.

The study utilized the 2005 CBMS complete enumeration data collected from Pasay city which is composed of 7 districts divided into 20 zones with a total of 201 barangays. Pasay city had a total of 65,117 households but with only 65,019 households in the database.

2. METHODOLOGY

From the 2005 CBMS Pasay City data, food poverty count (total number of food poor households) and hunger incidence were obtained in barangay level. The usual analysis for count data is by Poisson regression modelling using logarithm as its canonical link (log-link) function. The Poisson regression model is of the form

$$\log E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} \qquad \text{(Eq. 1)}$$

implying that

$$E[Y] = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}} \qquad \text{(Eq. 2)}$$

where Y is the response variable (food poverty count),
β is the p x 1 vector of regression parameters, and
X is the p x 1 vector of covariates.

This Poisson regression model was fitted to the CBMS food poverty data to estimate the food poverty count for each barangay. This is done by using the count regression procedure (proc countreg) in SAS which yields estimates for regression coefficients using the maximum likelihood estimation (MLE) method.The Poisson regression model has the property known as equi-dispersion which requires that the mean and variance of the data to be equal. However, in many real-life data, the variance is larger than the mean. Thus, overdispersion occurs when the

data exhibits high variability than allowed by the equality of mean and variance. Overdispersion can be checked through the use of either Pearson's chi-square statistic or Lagrange multiplier

statistic. Under Pearson's chi-square statistic, if its estimate divided by the degrees of freedom is greater than one, the data are overdispersed. Lagrange multiplier statistic which has a chi square distribution with degree of freedom equal to 1 is integrated in the count regression procedure in SAS through the dispersion parameter *alpha* = *r*. If the value of dispersion parameter *r* is significantly greater than zero, then there is overdispersion in the data. A more appropriate distribution to model count data with overdispersion is the negative binomial (NB). Its probability function is given by

$$f_Y(y) = \frac{\Gamma(y + r^{-1}\mu^{2-p})}{\Gamma(r^{-1}\mu^{2-p})\Gamma(y+1)}\left(\frac{r^{-1}\mu^{2-p}}{\mu + r^{-1}\mu^{2-p}}\right)^{r^{-1}\mu^{2-p}}\left(1 - \frac{r^{-1}\mu^{2-p}}{\mu + r^{-1}\mu^{2-p}}\right)^{y}, y = 0, 1, 2, \ldots \quad \text{(Eq. 3)}$$

where $\mu_i = \exp\{X_i\beta\}$,

r is the dispersion parameter, and

$\Gamma(.)$ is the usual gamma function.

The NB distribution has a mean equal to μ_i and variance equal to $\mu_i + r\mu_i^p$ (Cameron and Travedi, 1986). There are two types of NB distribution, NB1 when p =1, and NB2 when p = 2. Similar to that of Poisson regression, NB regression models use the log-link function to calculate the estimated count of food poor in the barangay level. Count regression analyses using in SAS were employed for NB1 and NB2 regression models. To compare the Poisson, NB1 and NB2 regression models, goodness-of- fit tests such as Akaike Information Criterion (AIC), Schwarz Bayesian Criterion (SBC), and likelihood ratio test (LRT) were used. Actual and estimated hunger incidences generated from these models were compared by calculating the mean absolute deviation.

Hierarchical cluster analysis using Ward's method and non-hierarchical cluster analysis were performed on top 20 barangays with high hunger incidence (> 8%) using the significant correlates in the best fit loglinear model as clustering variables. Proportions of these count variables were obtained and used for clustering barangays with high hunger incidence so as to characterize and describe them.

3. RESULTS AND DISCUSSION

Pasay City had a hunger incidence of 4.08% based on the PCI criterion, that is, 2,650 out of 65,019 households had PCI below the food threshold P11,199. Table 1 shows that the ten food poorest barangays with the highest hunger incidence. Barangay 181 had the highest hunger incidence equal to 16.86%, followed by barangays 163 and 95. Barangay 143 was the poorest barangay with 42.09% poverty incidence, but it ranked as 6th food poorest barangay with 12.52% hunger incidence.

Table 1: Ten Food Poorest Barangays in Pasay City

Rank	Barangay	Hunger Incidence	Rank	Barangay	Hunger Incidence
1	181	16.86%	6	143	12.52%
2	163	14.97%	7	152	12.41%
3	95	14.07%	8	52	12.20%
4	5	12.95%	9	182	11.89%
5	34	12.79%	10	91	11.11%

Counts of food poor households were used for the Poisson and negative binomial (NB1 and NB2) regression analyses. The Poisson regression model is given by

$$\log\hat{\mu}_i = -3.340921 + 0.007830\text{Brgy\_totmem\_Male} - 0.005024\ \text{Brgy\_totmem\_Female} - 0.007797\text{Brgy\_Mem05\_Male} + 0.007483\text{Brgy-Mem05\_Female} - 0.010833\text{Brgy\_ntElem612\_male} - 0.022460\text{Brgy\_ntHs1316\_male} - 0.034576\text{Brgy\_ntLiter10\_male} + 0.027149\text{Brgy\_nt10Liter\_female} - 0.009501\text{Brgy\_LabforMale} + 0.003217\text{Brgy\_Labfor\_Female} + 0.003581\text{Brgy\_unemp115ab\_male} + 0.004535\text{Brgy\_ntSWS} + 0.002671\text{Brgy\_ntSTF} - 0.003945\text{Brgy\_infsets} + 0.001848\text{Brgy\_msh} + 0.016048\text{Brgy\_povp} \quad \text{(Eq. 4)}$$

where $\hat{\mu}_i$ is the estimated food poverty count, Brgy_totmem_Male is the total number of male members, Brgy_totmem_Female is the total number of female members, Brgy_Mem05_Male is the total number of male members 0 to 5 years old, Brgy_Mem05_Female is the total number of female members 0 to 5 years old, Brgy_ntElem612_male is the total number of male members 6 to 12 years old not attending elementary, Brgy_ntHs1316_male is the total number of male members 13 to 16 years old not attending high school, Brgy_ntLiter10_male is the total number of male members who are not literate, Brgy_ntLiter10female is the total number of female members who are not literate, Brgy_LabforMale is the total number of male members in the labor force, Brgy_Labfor_Female is the total number of female members in the labor force, Brgy_unemp115ab_male is the total number of unemployed male members 15 years old and above, Brgy_ntSWS is the total number of households without access to safe water supply, Brgy_ntSTF is the total number of households without sanitary toilet facilities , Bgy_infsets is the total number of households which are informal settlers, Brgy_msh is the total number of households living in makeshift housing, and Brgy_povp is the total number of households below the poverty threshold in i^{th} barangay.

Since the dispersion parameter r is significantly different from 0 ($p < 0.0001$), there is overdispersion in the data, and so, negative binomial (NB) regression model is more appropriate than Poisson regression model.

The negative binomial (p = 1) or NB1 regression model is given by

$\log\hat{\mu}_i$ = -3.348145 + 0.00453Brgy_totmem_Male – 0.002511Brgy_totmem_Female – 0.011548 Brgy_ntElem612_male + 0.018012Brgy_ntElem612_female – 0.019161Brgy_ntHs1316_male – 0.032553Brgy_ntLiter10_male + 0.031514Brgy_ntLiter10_female – 0.007968Brgy_LabforMale + 0.002547Brgy_Labfor_Female + 0.006172Brgy_unemp115ab_female + 0.013247Brgy_povp (Eq. 5)

where $\hat{\mu}_i$ is the estimated food poverty count, Brgy_totmem_Male is the total number of male members, Brgy_totmem_Female is the total number of female members, Brgy_ntElem612_male is the total number of male members 6 to 12 years old not attending elementary, Brgy_ntElem612_female is the total number of female members 6 to 12 years old not attending elementary, Brgy_ntHs1316_male is the total number of male members 13 to 16 years old not attending high school, Brgy_ntLiter10_female is the total number of female members who are not literate, Brgy_ntLiter10female is the total number of male members who are not literate, Brgy_LabforMale is the total number of male members in the labor force, Brgy_Labfor_Female is the total number of female members in the labor force, Brgy_unemp115ab_female is the total number of unemployed female members 15 years old and above, and Brgy_povp is the total number of households below the poverty threshold in i^{th} barangay.

The negative binomial (p=2)or NB2 regression model is given by

$\log\hat{\mu}_i$ = -3.3593621 + 0.006416Brgy_totmem_Male – 0.002471Brgy_totmem_Female – 0.017275 Brgy_Mem05_male – 0.012438Brgy_ntHs1316_male – 0.005437Brgy_LabforMale + 0.004765Brgy_ntSWS – 0.004969Bgry_infsets + 0.025915Brgy_povp (Eq. 6)

where $\hat{\mu}_i$ is the estimated food poverty count, Brgy_totmem_Male is the total number of male members, Brgy_totmem_Female is the total number of female members, Brgy_Mem05_Male is the total number of male members 0 to 5 years old, Brgy_ntHs1316_male is the total number of male members 13 to 16 years old not attending high school, Brgy_LabforMale is the total number of male members in the labor force, Brgy_ntSWS is the total number of households without access to safe water supply, Bgy_infsets is the total number of households which are informal settlers, and Brgy_povp is the total number of households below the poverty threshold in i^{th} barangay.

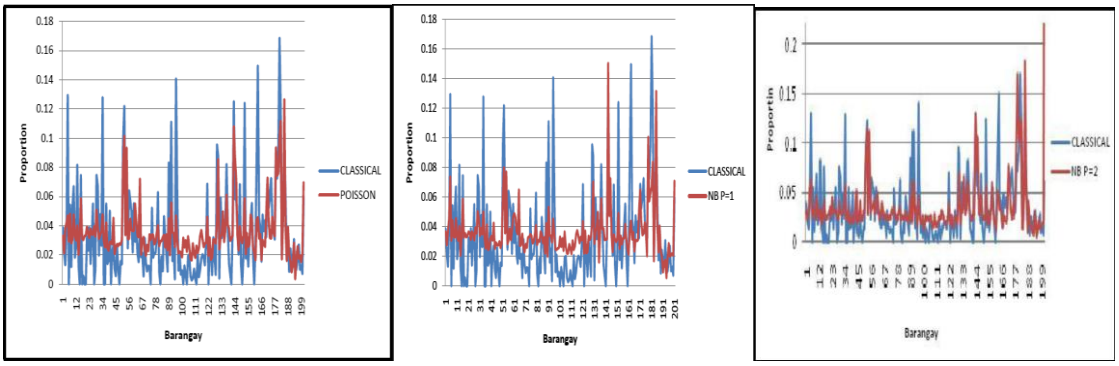
Good-of-fit tests show that NB2 has a lower AIC (1131) compared to that of NB1 (1143) and Poisson (1295). It has also a lower SBC (1164) compared to that of NB1(1186) and Poisson (1355). This indicates that NB2 regression model has a better fit compared to NB1 and Poisson regression models.

Comparison of generated estimates of hunger incidence of Poisson and NB2 regression models are shown in Table 2 for the first 5 barangays.

Table 2: Comparison of Actual and Estimated Barangay Level Hunger Incidence

Barangay	Actual Hunger Incidence	Poisson Regression Model		NB2 Regression Model		NB2 Regression Model	
		Estimated Hunger Incidence	Absolute Difference	Estimated Hunger Incidence	Absolute Difference	Estimated Hunger Incidence	Absolute Difference
1	3.88%	3.31%	0.57%	3.68%	0.2%	3.22%	0.66%
2	2.09%	2.11%	0.02%	2.77%	0.68%	2.19%	0.10%
3	1.31%	3.82%	2.51%	3.71%	2.40%	3.42%%	2.11%
4	3.67%	4.21%	0.54%	4.56%	0.89%	5.23%	1.56%
5	12.95%	4.70%	8.25%	7.38%	5.57%	6.16%	6.79%

Graphs of estimated hunger incidence for 210 barangays in comparison with the actual (classical) shown in Figure 1 indicate that among the loglinear models, NB2 regression model yielded estimates of hunger incidence closest to the actual values.



(a) Poisson (b) NB1 (c) NB2

Figure 1: Graphs of Actual and Estimated Barangay Level Hunger Incidence

Using the significant correlates (expressed as proportions in barangay level) under the NB2 regression model, two clusters of first 20 barangays (top 10%) with high hunger incidence (> 8%) were generated using nonhierarchical and hierarchical cluster analyses (See Table 3).

Table 3: Clustering of Barangays with High Hunger Incidence in Pasay City

Nonhierachical Cluster Analysis							Hierarchical Cluster Analysis						
Cluster 1 Barangays			Cluster 2 Barangays				Cluster 1 Barangays						Cluster 2 Barangays
129	13	137	130	143	163	180	182	51	13	137	180	89	143
152	162	178	181	182	185	5	185	129	34	152	178	95	5
34	89	91	51	52	95		91	181	130	52	163	162	

Under nonhierarchical cluster analysis, cluster 1 is characterized by a significantly higher proportion of male students 13 to 16 years old not attending high school compared to cluster 1, and cluster 2 is characterized by a significantly higher proportion of informal settlers compared to cluster 2. Hierarchical cluster analysis using Ward's method resulted to cluster 2 with composed of two barangays only, namely, barangays 143 and 5, characterized by significantly higher proportions of households which informal settlers and under poverty threshold.

4. CONCLUSIONS

The paper yielded three loglinear models such as Poisson, NB1 and NB2 regression models in generating estimates of hunger incidence in Pasay City. Test for overdispersion showed that the food poverty data are overdispersed, and hence, NB regression models are preferable than Poisson model. Goodness-of-fit statistics indicated that NB2 regression model is better compared to NB1 model since it has lower AIC and SBC. The NB2 regression model yielded the closest estimates of barangay level hunger incidence compared to the actual proportions. Significant correlates of hunger in the NB2 model includes total number of male members, total number of female members, total number of male members 0 to 5 years old, total number of male members 13 to 16 years old not attending high school, total number of male members in the labor force, total number of households without access to safe water supply, total number of households which are informal settlers, and the total number of households below the poverty threshold. Cluster analysis using these correlates as clustering variables showed that barangays with significantly high proportion of households which are informal settlers and high poverty incidence are barangays 143 and 5 among the twenty food poorest barangays in Pasay City.

5. ACKNOWLEDGEMENTS

The researchers would like to express their sincere gratitude to Community Based Monitoring System (CBMS) and Pasay City government officials for sharing the needed data.

6. REFERENCES

Arcilla, R., Co, F., and Ocampo, S., Correlates of Poverty: Evidence from the Community-Based Monitoring System (CBMS) Data, *De La Salle University*, (2011).

Co, F., Arcilla, A. and Ocampo, S., Correlates of Hunger: Evidence from the Community-Based Monitoring System (CBMS) Data, *De La Salle University*, (2012).

The CBMS Core Indicator, *Poverty and Economic Policy (PEP),* (2009). Retrieved on June 23, 2011 from http://www.pepnet.org/fileadmin/medias/pdf/CBMS_country_proj_profiles/Philippines/CBMS_forms/definition_of_cbms_core_indicatorsrev11242009.pdf

Community-Based Monitoring System, *Poverty and Economic Policy (PEP),* (2011). Retrieved on June 23, 2011 from http://www.pep-net.org/programs/cbms/about-cbms/