Count Data Models Applied to the Number of Household Members

Juan Pablo Díaz[1]

Resumen: Este documento de trabajo trata de modelar el número de miembros que viven en un hogar mediante la aplicación de modelos de datos de recuento, utilizando distribuciones Poisson y binomial negativa, básicamente. La idea principal es modelar el valor esperado del número de miembros de un determinado hogar en función de un conjunto de variables explicativas. La muestra procede de Ecuador, por lo que los resultados solo se aplican para este pais especifico. Sin embargo, este documento da algunas ideas sobre que factores afectan al numero de miembros en un hogar.

Palabras claves: Distribución poisson, distribución binonial negativa, miembros de hogar, Ecuador

Abstract: This working paper tries to model the number of members living in a household by applying Count Data Models, Poisson and Negative Binomial distributions basically. The main idea is to model the expected value of number of a certain household depending on a set of explanatory variables. The sample is from Ecuador, so the results only apply for this specific country. However, this paper gives some ideas about what factors affect the number of members in a household.

Key words: Poisson distribution, Negative Binomial distribution, household members, Ecuador

CÓDIGO JEL: C13, C16

1. Introduction

In this working paper, we try to determine what are the factors associated to the number of members in an Ecuadorean household. Understanding the size of the household is basic to understand a society as a whole. Not only because the family unit is one of the fundamental socioeconomic institution in human society as Bongaarts (2001) suggests, but also because the size of this household implies other social characteristics as the poverty, for example, especially in developing countries. In such

1 Student of the Graduate Program in Economics at University of Barcelona, Spain; moreover, he currently does an internship in the Department of Econometrics, Statistics, and Spanish Economy in such institution. He got his Master's degree in Economics and Business Administration at FLACSO-Ecuador and a Superior Diploma in Small and Medium Business Enterprises also.

a sense, the number of members in a certain household will be fundamental at the time to design sanity, educational, or security policies, just to claim several examples.

Moreover, the role of the household and residential family is also central in economic analyses, because these units are usually the locus of joint decisions regarding consumption, production, labor force participation, savings, and capital formation (Kuznets 1978; Becker 1991 quoted in Bonggarts 2001).

In addition, it is important to remark that the terms household and family are not always used consistently in the literature. As Bonggats (2001) says, a household is usually defined as a group of persons (or one person) who make common provision for food, shelter, and other essentials for living, but practices vary significantly among countries. On the other hand, the term family is used even less consistently. In the social science literature and in common usage "family" refers generally to a group of kin-persons related by blood, marriage, or adoption (Burch 1979, quoted in Bonggarts 2001). In contrast, demographers and economists usually follow the recommendations of the United Nations (1980) and define a family as the members of a household who are related through marriage, blood, or adoption. That is, they focus on the residential family. In any case, in this study, we are going to refer as household or family indistinctly.

Said so, the literature proposes several options to get our goal. In that sense, since the variable that we are studying is a natural number that takes relatively low values, we are going to use two types of count data models: Poisson and negative Binomial distributions. Then, we are going to compare between both regressions and select our preferred according to the one that shows better results.

2. Previous Studies

There are a great number of published papers that contributes to understand the size of the household. For example, Bongaarts (2001) published a study that uses data from household surveys in 43 developing countries to describe the main dimensions of household size and composition in the developing world. He found that the average household size varies only modestly among regions, ranging from 5.6 in the Near East/North Africa to 4.8 in Latin America. These averages are similar to levels observed in the second half of the nineteenth century in Europe and North America. Moreover, he observed that about four out of five members of the household are part of the nuclear family of the head of the household. In addition, Bongaarts suggests that household size is found to be positively associated with the level of fertility and the mean age at marriage, and inversely associated with the level of marital disruption. An analysis of trends and differentials in household size suggests that convergence to smaller and predominantly nuclear households is proceeding slowly in contemporary developing countries.

Regarding the associated effects of the size of the household, Lanjouw and Ravallion (1995) point out that there is considerable evidence of a strong negative correlation between household size and consumption (or income) per person in developing countries. It is often concluded that people living in larger and (generally) younger households are typically poorer. There has been much debate on which is the 'cause' and which is the 'effect' in this correlation. The position one takes in that debate can have implications for policy, including the role of population policy in development, and the scope for fighting poverty using demographically contingent transfers. In addition, they suggest that the existence of size economies in household consumption cautions against concluding that larger families tend to be poorer. The poor tend to devote a high share of their budget to rival goods such as food. But certain goods (water taps, cooking utensils, firewood, clothing, and housing) do allow possibilities for sharing or bulk purchase such that the cost per person of a given standard of living is lower when individuals live together than apart.

About the variables explaining the size of the households, we can found studies that use, for example, demographic variables. That is the case of Burch (1970) that investigates the influence of demographic variables (viz., mortality, fertility, age at marriage) on average household size under different family systems-nuclear, extended and stem. His study suggests that under all family systems, average household size is positively correlated with fertility, life expectancy, and average age at marriage. Households under nuclear and stem family systems never exceed 10 persons on average. By contrast, under extended family systems, when mortality is low and fertility is high, average household size reaches levels seldom if ever observed in reality, e.g., 25 persons per household. Large households under the extended family system also tend to be fairly complex, often containing 5 or more adults. A number of modifications in the model would make for greater fit between model and real family systems. A more fruitful approach would involve the simulation of household formation and development.

In any case, we would use variables that are more related to the composition of each household as it is possible to appreciate in the following pages. Basically we want to use variables that reflect the environment itself of the household (represented by a group of variables that capture the characteristics of the house) and a group of vectors that catch information about the economic situation of the family unit.

3. Methodology

Initially, we are going to consider that the number of members in the household can be model with a Poisson regression. The probability function of this distribution is given by:

$$P(x) = e^{-\lambda} \cdot \frac{\lambda^{x}}{x!} \qquad (1)$$

In which λ is a constant value that results from the multiplication between the probability that the event takes place (p), which is practically null, by infinite intents (n). Moreover, the Poisson distribution is characterized because its expectation and variance are equal to λ.

In addition, it is important to remark that the Poisson regression makes that the value of the parameter λ depends on the explicative variables:

$$\lambda_i = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_{ki}} \tag{2}$$

And if we apply logarithms to the equation:

$$\ln(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_{ki} \tag{3}$$

It is important to remember that by applying this model on our dependent variable, the number of members in the household, we are estimating the expected value of the variable, but not the probability as in the discrete choice models.

On the other hand, the Negative Binomial regression does consider that the expectation of the dependent variable is equal to λ, but incorporates an error term in the variance:

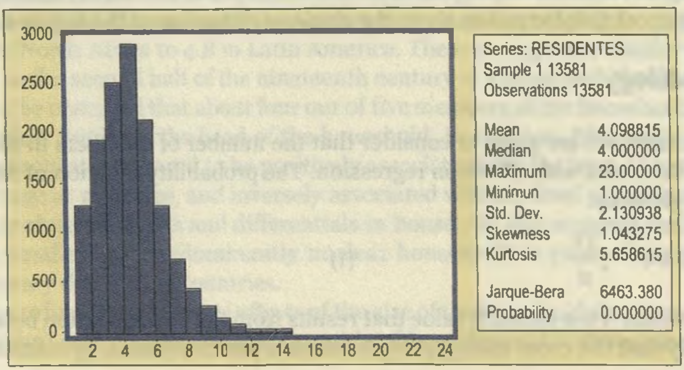
$$Var(y) = \lambda + \theta \tag{4}$$

Pay special attention to the fact that if λ equal to zero, then we are facing a Poisson distribution and this last would be our preferred functional form.

4. Empirical Analysis

First, it would be interesting to give a look to the distribution that the number of members of households has and its basic statistics as well:

Graph 1. Histogram and Basic Statistics of the Dependent Variable



In the previous graph, it is possible to appreciate that the majority of the data is concentrated in the left side of the distribution. Moreover, since there are not households with zero number of members, we do not face any problem of excess of zeroes. In addition, we can appreciate that the mean is 4, with a maximum value of 23 and a minimum of 1. Furthermore, the series presents a standard deviation of 2,13. It is clear that the shape of the distribution of the variable fits perfectly with the count data model distribution that we are going to apply in the following sections.

Data

I am going to use data from the Survey of Conditions of Life (Encuesta de Condiciones de Vida) that provides us information of 13.581 households in Ecuador. The survey collected the information in 2006. Even though it has been 8 years since this information has been collected until the date I am writing this short essay, it does not suppose a serious problem mainly because changes in the structure of the households take, approximately, one generation or even more time, which in this type of econometric models suppose more than 25 years.

Now let's focus on the econometric part. What I want to do is to model the number of members in the household as a function of a set of variables that catch information about the characteristics of the dwelling and the head of the household, so given the functional form of the Poisson regression, our model will be as follows:

$$(\lambda_i) = e^{c + \beta_1 Bath + \beta_2 Bed + \beta_3 d\_rural + \beta_4 Father\_ed + \beta_5 Elec\_bill + \beta_6 d\_shower + \beta_7 Edadjefe + \beta_8 Edadjefeq} \quad (5)$$

Where:

Bath represents the number of bathrooms in the household;

Bed represents the number of bedrooms in the household;

d_rural is a dummy variable that takes the value of 1 when the household is located in the rural area or 0 otherwise;

Father_ed is an ordered variable that catches the level of education of the head of the household as follows: 0 = no education, 1= primary education, 2 = secondary education, 3 = post-secondary education (known in Ecuador as technical level), 4 = tertiary education. The idea of this variable is being a proxy of the income level of the family since it is well known that educational and income levels are highly correlated (previous studies have shown a correlation of 70 and 80% between income and education level);

Elec_bill is the last bill of electricity that the household paid in the last month. It is measured in dollars. This variable also works as a proxy of the expenditure level since I do not have information about the net expenditure of the household;

d_shower is a dummy variable that takes the value of 1 if there is a shower in the household and 0 otherwise;

edadjefe is the age of the head of household; and,
edadjefeq *is the age of head of the household squared.*

Results

We introduce the data into the software package and obtain the following results considering the Poisson and Negative Binomial functional forms:

Table 1. Results of Poisson Regression

Dependent Variable: NUM_MEMB
Method: ML/QML - Poisson Count (Quadratic hill climbing)
Date: 12/17/13 Time: 22:44
Sample (adjusted): 1 13577
Included observations: 9679 after adjustments
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.868441	0.051215	16.95678	0.0000
BATH	-0.061667	0.008726	-7.067388	0.0000
BED	0.149973	0.004879	30.73775	0.0000
D_RURAL	0.069637	0.011255	6.187113	0.0000
FATHER_ED	-0.066066	0.006633	-9.959443	0.0000
ELEC_BILL	0.002529	0.000348	7.274480	0.0000
D_SHOWER	-0.164713	0.011840	-13.91184	0.0000
EDADJEFE	0.023560	0.002091	11.26861	0.0000
EDADJEFEQ	-0.000292	2.03E-05	-14.41464	0.0000

R-squared	0.205327	Mean dependent var	4.112718
Adjusted R-squared	0.204670	S.D. dependent var	1.988290
S.E. of regression	1.773183	Akaike info criterion	3.921978
Sum squared resid	30404.21	Schwarz criterion	3.928652
Log likelihood	-18971.41	Hannan-Quinn criter.	3.924241
Restr. log likelihood	-19926.90	Avg. log likelihood	-1.960059
LR statistic (8 df)	1910.967	LR index (Pseudo-R2)	0.047949
Probability(LR stat)	0.000000		

Table 2. Results of Negative Binomial Regression

Dependent Variable: NUM_MEMB
Method: QML - Negative Binomial Count (Quadratic hill climbing)
Date: 12/17/13 Time: 23:18
Sample (adjusted): 1 13577
Included observations: 9679 after adjustments
QML parameter used in estimation: 1
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.911765	0.110999	8.214165	0.0000
BATH	-0.061393	0.019898	-3.085364	0.0020
BED	0.153490	0.011382	13.48492	0.0000
D_RURAL	0.067858	0.025630	2.647593	0.0081
FATHER_ED	-0.067747	0.014480	-4.678484	0.0000
ELEC_BILL	0.002616	0.000837	3.126275	0.0018
D_SHOWER	-0.167413	0.027021	-6.195583	0.0000
EDADJEFE	0.021703	0.004457	4.869183	0.0000
EDADJEFEQ	-0.000276	4.23E-05	-6.535432	0.0000

R-squared	0.204766	Mean dependent var	4.112718
Adjusted R-squared	0.204108	S.D. dependent var	1.988290
S.E. of regression	1.773809	Akaike info criterion	5.016255
Sum squared resid	30425.67	Schwarz criterion	5.022929
Log likelihood	-24267.17	Hannan-Quinn criter.	5.018518
Restr. log likelihood	-24457.40	Avg. log likelihood -	2.507198
LR statistic (8 df)	380.4713	LR index (Pseudo-R2)	0.007778
Probability(LR stat)	0.000000		

Before analyzing the obtained results, we have to choose one specification of the both presented. In that sense, we compare the Likelihood Ratios of both regressions that in the case of the Poisson is -18971 and in the case of the Negative Binomial is -24617. Since LRPoisson > LRNB, we choose as our preferred specification the Poisson one.

Now, let's analyze the results. First, all variables are individually statistically significant at $\lambda = 1\%$. Regarding the variables that catch the physical characteristics of the place of residence there are variables that affect positively and negatively to the expected number of members in the house. In the first group are the number of bedrooms in the house and having a shower, meaning that the more bedrooms in the house, the more expected number of members in the household. In the same way, having a shower increases the expected number of members in the household. Regarding the house information variables that affect negatively, I have found that an increase in the number of bathrooms in the household leads to a decrease in the expected number of members in the house. In addition, if the household is settled in the rural area, the expected value of people living in there increases.

Regarding the variable that we used as proxy of the income level, level of education of the head of the household, we obtained a negative sign, meaning that the more educated the head of the household, the less the expected number of members in his home.

Regarding the variable that we used as a proxy of the expenditure level, the last electricity bill paid by the household, we have found that it has a positive effect, meaning that the more the expenditure in electricity, or net expenditure in general, the less the expected number of members in a household.

Combining the results obtained from the proxies of income and expenditure, it is possible to deduce that the richer the Ecuadorian household, the smaller the size of the household.

The contradictory signs obtained in the age and the squared age of the head of the household suggest that this variable has not a linear behavior.

5. Application

Once we have decided that the results obtained with the Poisson regression are our preferred, we can use them and run several types of simulations, for example. In that sense, now we are able to estimate the probability that an Ecuadorian household a certain number of members given the conditions we have used in our model such as having a shower or not at home, living in the rural area or not, having a certain number of bathrooms and bedrooms in the house and taking into account the characteristics of the head of the household such as his age, level of education, and the electricity bill finally.

6. Conclusion

We have constructed a Poisson regression that, in the framework of count data models, give us an idea about what are the variables that affect the expected value of the number of members in an Ecuadorean household. Since the sample used is representative, we consider that our results are consistent. In such sense, we can point out that it is possible to study the size of the household using characteristics of the head of the household and of the house itself as explanatory variables. The main contribution of this article is that give some hints about which factors are necessary to take into account at the moment of designing public policies that are address to the household such as health, reproductive or educational ones since now we know what factors affects its size.

References

Bongaarts, John. Household Size and Composition in the Developing World in the 1990s. Population Studies, Vol. 55, No. 3 (Nov., 2001), pp. 263-279.

Burch, Thomas. Some Demographic Determinants of Average Household Size: An Analytic Approach. Demography, Vol. 7, No. 1 (Feb., 1970), pp. 61-69.

Instituto Nacional de Estadística y Censos. Encuesta de Condiciones de Vida. www.inec.gob.ec. Visited in Decenber 10th, 2013.

Jiang, Leiwen and Brian C. O'Neill. Impacts of Demographic Trends on US Household Size and Structure. Population and Development Review, Vol. 33, No. 3 (Sep., 2007), pp. 567-591
Lanjouw, Peter and Martin Ravallion. Poverty and Household Size. The Economic Journal, Vol. 105, No. 433 (Nov., 1995), pp. 1415-1434.

Fiegehen, G. C. and P. S. Lansley. The Measurement of Poverty: A Note on Household Size and Income Units. Journal of the Royal Statistical Society. Series A (General), Vol. 139, No. 4 (1976), pp. 508-518.

Wooldridge, Jeffrey. Introductory Econometrics. Fourth Edition. Cengage Learning. 2009.