International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 14, Number 12 (2021), pp. 1206-1209

Using Multiple Linear Regression to Explain Income with Socio-Demographic Variables

Thomas Edison Guerrero Barbosa[1], Sir-Alexci Suarez Castrillon[2] and Isbelia Karina Rincon Parada[3]

[1] *Engineering Faculty, Civil Engineering Department. Research Group CERG, University Francisco of Paula Santander Ocaña, Colombia*

[2, 3] *Engineering Faculty, Department of Systems and Informatics. Research Group GRUCITE, University Francisco of Paula Santander Ocaña, Colombia*

ORCIDs: 0000-0003-3690-256X (Thomas); ORCIDs: 0000-0001-8010-0228 (Sir-Alexci)

ORCIDs: 0000-0002-6616-2147 (Isbelia)

Abstract

Multiple Linear Regression (MLR) models are estimated to explain the people income surveyed in different regions of Chile-based on demographic, education, health, housing and work variables. The database CASEN survey of 2011 was used. It is shown that variables associated with the gender of the person, age, marital status, educational level, job stability, possession of services such as the internet and condition of occupation of the home are explanatory of the individual's income. The ease and speed of estimations stand out from the estimates with RLM. However, the main challenge is to guarantee compliance with the assumptions of this approximation.

Keywords: Multiple Linear Regression (MLR), income, socio-demographic variables.

I. INTRODUCTION

Government agencies need decision tools to carry out formulation, preparation, planning, execution and monitoring of public policies that allow the economic and social development of any modern State. Through the National Socioeconomic Characterization Survey (CASEN), the Ministry of Social Development of Chile has a comprehensive and rich variety of information to characterize the situation of poverty and those groups defined as priorities by social policy, concerning demographic aspects, education, health, housing, work and income. This survey is applied every two or three years to households.

Multiple Linear Regression Models (MLR) are used to estimate models. They forecast a dependent variable associated with income (ytrabhaj, without considering pensions or subsidies) based on independent variables that characterize social, economic, and demographic conditions of the people who make up a given household. To carry out a sequence and fulfil the objectives of the work, it is proposed to carry out several activities, including calibrating RLM models and verifying the assumptions that support this approach.

The rest of the paper is structured as follows. Section 2 the database management and debugging are shown. Descriptive statistics of the explanatory variables is described in Section 3. Results and analysis of the estimated models are shown in Section 4. Finally, the conclusions are made in section 5.

II. DATABASE MANAGEMENT AND DEBUGGING

Initially, an exploration and knowledge phase were carried out with the type of information registered in the CASEN survey (2011). Once the relevant information was known, some variables (quantitative and qualitative) that intuitively for the modeller's consideration may present a direct or indirect relationship with the dependent variable ytrabhaj which refers to the individual's income registered in the survey, not considering pensions and subsidies within it. Likewise, the report of the prior state of the art made in national and international literature was also taken into account, which allowed identifying the nature and effect of independent variables on ytrabhaj. The selected variables correspond to different modules of the CASEN survey. Except for the variables age (age of individual) [1]–[5], age2 (age of individual squared), children (number of children born alive) Cohen and Huffman [6] and hrtrab (total hours worked last week at your main job) [6]–[9], the others are dummy variables.

Given that the RLM model assumes that each independent variable exerts a linear influence on the dependent variable, there are occasions when, given their nature (example: gender, age, marital status, socioeconomic status, among others), this principle is not fulfilled. To try to correct the effect of non-linearity and to be able to incorporate these variables in an adequate way to the model, one can go to the creation of dummy variables or dummies.

One aspect to consider within the treatment and inclusion of variables is the one referring to the age variable. For this study, the result associated with the square of the age variable (age2) was considered, which allows knowing the marginal effect of the independent variable on the dependent variable, differentiating whether it is increasing or decreasing [7]; On the other hand, it was considered in parallel to create a mute variable called age1559 (it is worth 1 if the age is between 15 to 59 years, 0 otherwise) given that according to the National Institute of Statistics [10] the non-dependent age ranges economically in Chile he is between 15 to 59 years old.

III. DESCRIPTIVE STATISTICS OF THE EXPLANATORY VARIABLES

The summary of the variables is shown in Table 1, as well as the minimum (Min), maximum (Max), mean (Mean) and standard deviation (Std. Dev.) values. The values of 0 and 1 correspond to the dummy variables. On the other hand, we see that the average age of the people surveyed corresponds to 44 years. Chilean workers work around 42.7 hours a week and have approximately two children.

Table 1. Descriptive statistics of the explanatory variables of the model

Variable	Mean	Std. Dev	Min	Max	Description
genero	0.46	0.50	0	1	Male or female
edad	44.2	18.0	18	105	Age of individual
edad2	2281.3	1761.7	324	11025	Age of individual squared
casado	0.39	0.49	0	1	Marital status
nivedu	0.15	0.35	0	1	Education level
hrtrab	42.7	15.5	0	120	Total hours worked last week at your main job
perman	0.71	0.45	0	1	Permanent job
solod	0.83	0.37	0	1	Work is daytime
solon	0.02	0.12	0	1	Work is night
secund	0.02	0.16	0	1	Secondary work
internet	0.36	0.48	0	1	Has internet service at home
propia	0.70	0.46	0	1	If the house is your own
arrendada	0.14	0.34	0	1	If the house is rented
nhijos	2.29	2.1	0	20	Number of children born alive

IV. RESULT

Many runs were carried out, varying the variables in each of them. The results of the estimates are presented in Table 2. RLM1 and RLM2 models are accompanied by the name of the variable, the estimated coefficient (Coef.), the test of statistical significance test t and the R^2 for each model, where some of the variables are common to both and not in other cases.

For the two estimated models, the independent variable corresponded to *lg_ytrabhaj*, which results from applying the logarithm to the variable *ytrabhaj*. These types of procedures are used as a mathematical transformation strategy to reduce the original spread of the data. The application of logarithm functions produces a reduction in dispersion concerning the original variable, a decrease in its value [11].

For the models presented in Table 2, the correspondence of the effect represented in the sign that accompanies the coefficient with the hypotheses formulated in the previous section can be evidenced. Additionally, it is worth highlighting the statistical significance of each of the variables considered in the model (since the t-test is more significant than 1.96 for a confidence level of 95%), which means that the variables considered explaining the modelled variable case, the income. Likewise, the adjustment of the RLM1 and RLM2 models corresponds to 0.3827 and 0.3489, respectively.

For the RLM1 model, the variable that explains income to a greater extent corresponds to *levedu*, understanding that the higher the educational level of the older person will be their income, aspect allows accepting the hypothesis proposed [7], [9], [12]. On the other hand, the gender variable showed a positive sign and statistical significance. It indicates that men

have a higher income than women [1], [2], [6], [7]. One aspect to consider is the internet variable since the positive sign of its estimation indicates that having this type of service indicates good income (aspect raised in our hypothesis). The variables that represent day or night working hours (*solod* and *solon*) present an inverse relationship concerning income, being significant from the modelled phenomenon. The variables remain, *married* [2], [9], *own*, *rented* [2], [13] and *age2* [1]–[3], [5], turned out to be significant and their effect is a direct relationship with people's income.

The RLM2 model presents the estimation with the same variables of the RLM1 model except the variables *gender*, *age2*, *solod* and *solon*; on the other hand, it contemplated the inclusion of the variables *nhijos* and *hrtrab*. In this model it is noteworthy that the standard variables in both models preserved the same effect reflected in their sign. Additionally, the variables *nchildren* and *hrtrab* present a direct and significant relationship with income, so that increases in the number of children and more remarkable Dedication of hours to work activities will have an impact on income increases, aspects that were expected according to the hypotheses raised [6].

Table 2. Estimation of the RLM1 and RLM2 models

	RLM1		RLM2	
lg_ytrabhaj	Coef.	t	Coef.	t
genero	.2203222	16.87		
casado	.1639452	13.80	.5337145	18.91
nivedu	.7536586	49.45	.7054456	23.68
perman	.3298993	26.24	.3520295	15.15
secund	.2843013	10.63	.2399733	5.11
internet	.4826854	42.99	.4857502	22.78
propia	.1434418	10.55	.1481514	5.65
arrendada	.2086276	12.84	.2037794	6.54
edad2	.0000131	2.56		
solod	-.1300493	-9.24		
solon	-.1650887	-3.87		
nhijos			.0128848	2.02
hrtrab			.0109742	18.26
_cons	12.20509	523.48	11.63532	313.42
R-squared	0.3827		0.3489	

V. CONCLUSION

It was possible to estimate and calibrate RLM models to estimate income (variable ytrabhaj) based on individual, demographic, education, health and housing characteristics reported in the CASEN survey. The estimated models obtained a good behaviour reflected in the R2 coefficients, the effect of the variables shown in the sign and the statistical significance (t test) for each of the variables considered. Verification of compliance with the RLM assumptions was carried out for each estimated model. When any of the assumptions were not fulfilled, tools and methodologies were used that made it possible to fulfil the associated belief.

All the estimated models had similar behaviour in the sign according to what was stipulated in the initially contemplated hypotheses and statistical significance associated with each variable. It is easy to explain the prediction of income from variables such as gender, age, marital status, educational level, job stability, possession of services such as internet and condition of occupation of the home. In all the models, the variables turned out to be significant, and their sign is by econometric theory. The conversion of categorical variables to dummy variables allowed the linearization and proper use within the model of these variables.

The information obtained in the CASEN survey is wealthy since it reports data at the individual level in a household, characterizing an essential totality of factors to a large extent. Although obviously, this type of information is not free from fingering biases, denial in response, unknown response, among others, it constitutes an excellent raw material to carry out studies such as the one presented in this document and for other research purposes.

This paper aims to compare nine supervised algorithms' performance towards DDoS intrusion. DDoS attack will result

REFERENCES

[1] J. I. de la Maza Díaz, «Factores que Afectan los Niveles de Ingreso del Trabajador de Microempresa del Sector Comercio de la Provincia de Santiago», 2010, Accedido: dic. 05, 2021. [En línea]. Disponible en: https://repositorio.uchile.cl/handle/2250/103744

[2] M. Gentile y S. Marcińczak, «Housing inequalities in Bucharest: shallow changes in hesitant transition», *GeoJournal*, vol. 79, n.º 4, pp. 449-465, ago. 2014, doi: 10.1007/s10708-014-9530-5.

[3] Y. Huang y L. Jiang, «Housing Inequality in Transitional Beijing», *Int. J. Urban Reg. Res.*, vol. 33, n.º 4, pp. 936-956, 2009.

[4] S. Li, «Housing Inequality in Urban China : Guangzhou 1996 and 2005», *Espace-Popul.-Soc.*, vol. 2009, dic. 2009, doi: 10.4000/eps.3839.

[5] J. R. Logan, Y. J. Bian, y F. Q. Bian, «Housing inequality in urban China in the 1990s», *Int. J. Urban Reg. Res.*, pp. 7-25, 1999.

[6] P. N. Cohen y M. L. Huffman, «Working for the woman? Female managers and the gender wage gap», *Am. Sociol. Rev.*, vol. 72, n.º 5, pp. 681-704, 2007, doi: 10.1177/000312240707200502.

[7] A. Araújo Freitas, «La desigualdad salarial de género medida por regresión cuantílica: el impacto del capital humano, cultural y social», *Rev. Mex. Cienc. Políticas Soc.*, vol. 60, n.º 223, pp. 287-315, abr. 2015.

[8] L. McCall, «Explaining Levels of Within-Group Wage Inequality in U.S. Labor Markets», *Demography*, vol. 37, n.º 4, pp. 415-430, 2000, doi: 10.2307/2648069.

[9] H. Mandel y M. Semyonov, «Family Policies, Wage Structures, and Gender Gaps: Sources of Earnings Inequality in 20 Countries», *Am. Sociol. Rev.*, vol. 70, n.º 6, pp. 949-967, dic. 2005, doi: 10.1177/000312240507000604.

[10] «National Institute of Statistics. Población y Sociedad, aspectos demográficos.» Santiago – Chile 2008.

[11] J. de D. Ortúzar S. y L. G. Willumsen, *Modelling Transport*, Fourth edition. Chichester, West Sussex, United Kingdom: John Wiley & Sons, 2011.

[12] T. A. Diprete y C. Buchmann, «Gender-specific trends in the value of education and the emerging gender gap in college completion», *Demography*, vol. 43, n.º 1, pp. 1-24, feb. 2006, doi: 10.1353/dem.2006.0003.

[13] J. Kemeny, *The myth of home-ownership: private versus public choices in housing tenure*. London; Boston: Routledge & Kegan Paul, 1981. Accedido: dic. 05, 2021. [En línea]. Disponible en: https://archive.org/details/mythofhomeowners0000keme