Modeling Electricity Consumption through Socioeconomic Indicators: A Multivariable Regression Approach

Gabriela Munguía Deras[1], *Alicia María* Reyes-Duke[1*], *Héctor* Villatoro Flores[2]

[1]Faculty of engineering, Universidad Tecnológica Centroamericana, (UNITEC), Honduras
[2]Application Engineer, e-Storage, Canadian Solar, Alberta Canada

Abstract. This study examines the relationship between electricity consumption and sociodemographic factors across 258 municipalities in Honduras using data from national statistical sources. Multiple regression models were simulated in RStudio to identify the main predictors of municipal electricity demand. The initial linear model achieved high explanatory power (R^2=0.85), but diagnostic tests revealed heteroskedasticity and non-normal residuals, limiting the model´s reliability. To correct these issues, a log-linear model using the natural logarithm of electricity consumption was developed. This specification eliminated heteroskedasticity (Breusch–Pagan p = 0.282) and produced a more stable fit (R^2 = 0.618; R^2 test = 0.81). Results indicate that urbanization, income per capita, and human development positively influence electricity consumption, while poverty level has a negative effect. This last model provides a tool to increase the decentralized power system planning since municipalities can better forecast their electricity demand and allocate resources (investments in distribution or generation) when and where they need them.

1 Introduction

Electricity consumption is a fundamental indicator of social and economic development. Many authors have studied the relation between human development and electricity consumption; Reza Torres [1] studied through a linear multiple regression a model of electricity consumption based on 31 independent sociodemographic variables through Minitab, Mazur [2] studied life expectancy with respect to energy and electricity consumption concluding that these two variables are essential for improving the wellbeing of people in less developed nations. Others have directly studied the Human Development Index (HDI) which represents the life expectancy index, GDP index and education index of a nation. Arto et al [3] studied the relationship between total primary energy demand and HDI whereas Leung & Meisen [4] studied the relationship between electricity demand and HDI in medium and low HDI countries, concluding that a logarithmic relationship exists in both cases. Mohamed, Z., & Bodger, P. [5] related electricity consumption to

* Corresponding author: aliciareyes@unitec.edu

sociodemographic variables like Gross Domestic Product, Population and Electricity Prices concluding that these three variables were significant for the model. In the United States of America, Haerer et al [6] studied the relationship between employment and the energy sector, concluding that significant changes in employment occurred in industry sectors that support the operation and management activities of U.S. electric power industry.

Like the previously mentioned studies, this study evaluates the relationship between sociodemographic and economic variables with respect to electricity consumption, with the novelty that this study will be based in Honduras and will be considering the consumption of each one of the municipalities (258 municipalities where data was available) in the territory. The analysis of electricity consumption related to sociodemographic variables of each one of the municipalities in Honduras provides support for the planning of decentralized energy systems given that the disaggregated data reveal patterns of demand and the model provide support by forecasting based on the independent variables. As noted by Quintero [7], energy generation that is locally distributed and based on regional resources helps diversify the energy mix and strengthens the self-sufficiency of each area. Similarly, Villatoro Flores [8] emphasizes that well-designed public policies promoting renewable, decentralized systems and encouraging private investment can transform a country's energy structure and expand access to electricity.

The rest of this article is divided into Methodology, Results and Conclusions.

2 Methodology

2.1 Data Collection

The first phase of this research was to obtain the data from legitimate sources such as statal organisms or companies associated to the electricity subsector of Honduras. Due to the goal of segmenting electrical demand, the data required for this study was the electricity consumption of each municipality in Honduras. Also, some demographic characteristics mentioned in Table 1 were required.

Table 1. National Sources to obtain dataset of dependent and independent variables.

Characteristic [unit]	Description	Source	Name
Monthly Electricity Consumption [MWh]	This information is available exclusively for this study and corresponds to March 2022.	EEH 2022	Y
Urbanization Level [%]	Percentage of households that are in urban areas.	SGJD 2020 [9]	X_1
Human Development Index [-]	Measure of general humanitarian development for a region	SGJD 2020 [9]	X_2
Income per capita [L.]	Quotient of the municipality´s income divided by the total population of the municipality.	SGJD 2020 [9]	X_3
Industrial Occupation [Inhabitants]	Sum of all industrial and commercial level jobs.	INE 2013 [10]	X_4
Ethnic Population [%]	Percentage of ethnic population.	INE 2013 [10]	X_5
Poverty Level [-]	Measured by Unsatisfied Basic Needs (UBN) factor.	INE 2013 [10]	X_6

Characteristic [unit]	Description	Source	Name
Electrical coverage [%]	Represents which regions have the necessary infrastructure to transmit electrical energy.	ENEE 2019 [11]	X_7

2.2 Data filtering and definition of variables

Once the eight characteristics were identified in the national data sources for the achievable municipalities, the second phase started with the filtering and organization of the data. Although in Honduras there are 298 municipalities in total, the sample for this study is composed of 258 municipalities. This fact is due the national electric distribution utility only provides the resource to these regions; the remaining municipalities obtain electrical services from other two distribution companies: the Roatán Electric Company (RECO) in Islas de la Bahía and Inversiones de la Mosquitia en Honduras (INELEM).

2.3 Univariable and Multivariable Regression Model (Linear and Log Linear)

In this article RStudio is used to perform statistical analyses and regression modelling, the univariable regression model and multivariable regression model will both be achieved using the RStudio environment.

2.3.1 Univariable Regression Model

A study of the linear relationship between each independent variable (X) and the dependent variable (Y) was conducted to determine which variables presented higher Pearson correlation coefficients (r) and coefficients of determination (R^2). Given that it is not certain that the variables exhibit a strong linear relationship with respect to Y, an additional analysis using the natural logarithm of Y (lnY) against each X was proposed and performed. The strategy of using the logarithm of the dependable variable instead of their original form makes the effective relationship non-linear while preserving the linear regression model [12].

2.3.2 Multivariable Regression Model

A multiple linear regression approach was applied to quantify the combined effect of the explanatory variables (X_1-X_7) on electricity consumption (Y) in each municipality for the data set available to demonstrate a model that can predict electricity consumption based on sociodemographic factors. Two models were defined: one Lineal Regression Model using the dependent variable in its original form (Y), and another Log-linear Regression Model using its natural logarithm (lnY) given the observations made in the univariable regression model in which it was found that not all relationships were linear.

The steps followed to achieve the most fitted model for both the linear and the log-linear regression are described in Table 2.

Table 2. Steps to achieve the most fitted model for both regression models in RStudio.

1	All independent variables (X_1–X_7) were initially included in the regression.
2	A stepwise selection procedure based on the Akaike Information Criterion (AIC) was then applied to identify the most parsimonious set of predictors.
3	An ANOVA test was conducted to assess the overall significance of the selected model.
4	The statistical assumptions were verified using the Shapiro–Wilk test for residual normality, the Breusch–Pagan test for heteroscedasticity, the Variance Inflation Factor (VIF) for multicollinearity, and the Durbin–Watson test for autocorrelation.
5	If heteroscedasticity, robust (HC3) test
7	Finally, the predictive performance of the model was evaluated using a train–test split (80/20) approach.

3 Results

The results are structured based on the last section of the methodology (Section 2.3), first the univariable regression model and findings for linear and log-linear model are described and later the multivariable regression model is described for linear and log-linear model.

3.1 Univariable Regression Model

Pearson correlations (r) and coefficients of determination (R^2) were estimated between the dependent variable and each predictor (X_1–X_7), both on the original scale (Y) and on the logarithmic scale (ln Y).

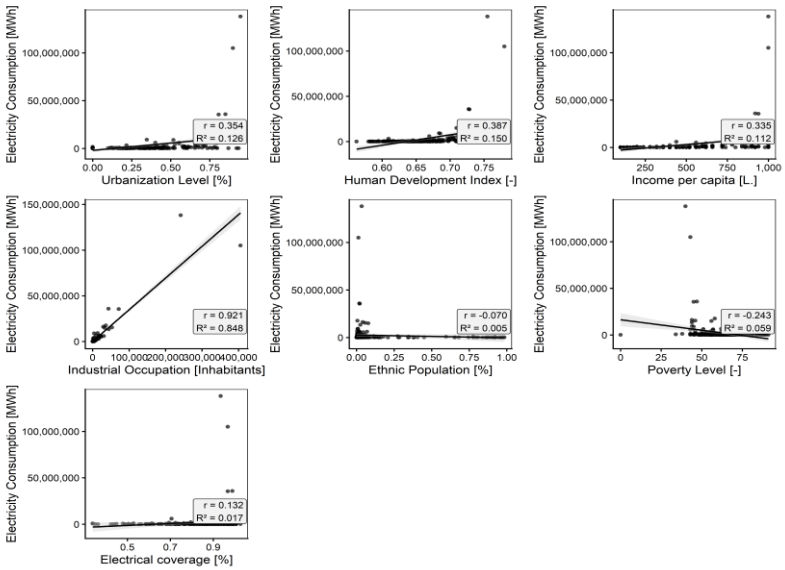

Fig. 1. Identification of linear correlation between electricity consumption and selected sociodemographic variables (X_1-X_7)

Based on Figure 1 it can be deducted that the strongest association was found with Industrial Occupation, which showed a very high and positive correlation with electricity consumption ($r = 0.921$; $R^2 = 0.848$). Moderate positive correlations were also observed with the Human Development Index; Income per capita and Urbanization Level (r between 0.39 and 0.34 while R^2 oscillates between 0.15 and 0.11), suggesting that more developed and urbanized municipalities have higher electricity consumption. In contrast, Ethnic Population ($r = -0.070$) and Poverty Level ($r = -0.243$) showed negative correlations. Nonetheless, Ethnic Population shows less explanatory power given a low coefficient of determination ($R^2 = 0.005$).

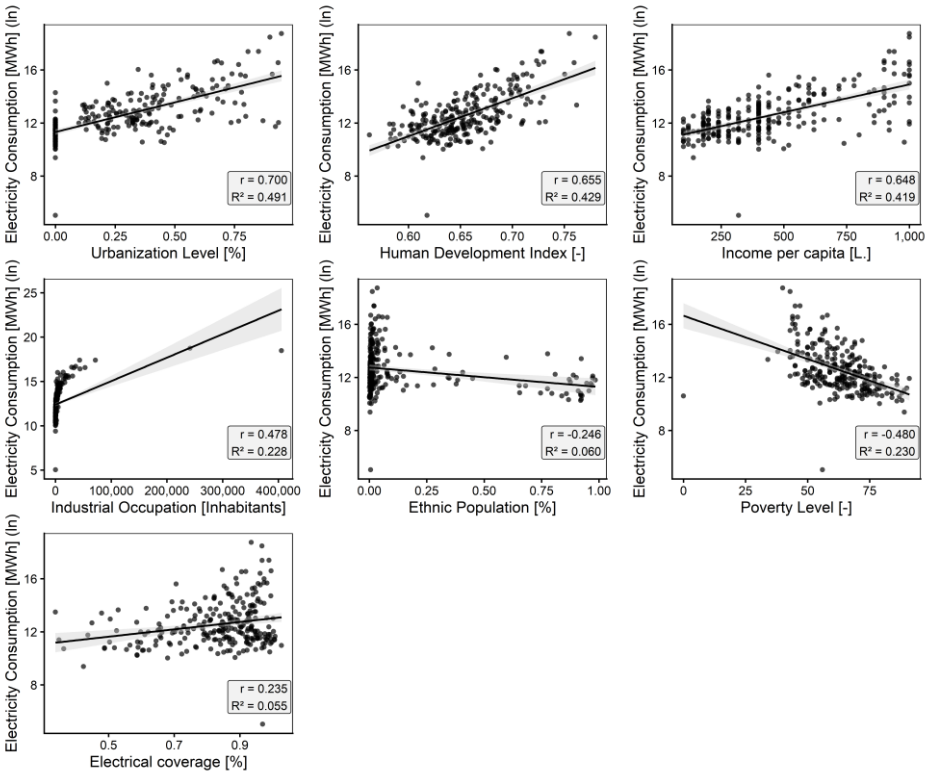

Fig. 2. Identification of log-linear correlation between electricity consumption and selected sociodemographic variables (X_1-X_7)

The log-transformation (ln Y) of the dependent variable improved the linearity and interpretability of the relationships between electricity consumption and socioeconomic variables. When compared to the model that used the original dependable variable, the correlations became stronger and more consistent, this indicates a reduction of heteroscedasticity and the influence of extreme values. In Figure 2 it can be deducted that Urbanization Level, Human Development Index and Income per capita present strong correlation with the electricity consumption and the negative relations become more evident for Poverty Level and Ethnic Population. The relation with Industrial Occupation weakened compared with the model with the original value (Y).

3.2 Multivariable Regression Model

3.2.1 Linear regression model

In the first model, a linear regression was simulated based on the original dependable variable electricity consumption (Y), to assess which combination of independent variables are better predictors, In the first step the dataset was analyzed by RStudio, a model with all variables was created. Secondly, the stepwise AIC procedure concluded that only two out of seven were statistically relevant variables: Human Development Index (X_3) and income per capita (X_4). The resulting model is shown as Equation 1.

$$Y = -762{,}544.06 + 1{,}883.60x_3 + 341{,}70x_4 \quad (1)$$

Both predictors show a positive relationship with electricity consumption. The model achieved a high explanatory power, in the third step from Table 2, the ANOVA analysis shown that R^2 is equal to 0.85 and the F-statistic (832.8, $p < 0.001$) confirms that the overall regression model is statistically significant.

Following the fourth step, the tests were done and shown that two of the assumptions were violated. The residuals were strongly non-normal (Shapiro–Wilk) and heteroskedastic (Breusch–Pagan) was detected. Additionally, a mild positive autocorrelation was detected (Durbin–Watson).

Table 3. Diagnostic Test for OLS assumptions (multivariable linear regression)

Test	Ideal Range	Linear Model	Status
Shapiro–Wilk	W=1 $p \geq 0.05$	W = 0.25, $p < 0.001$	Unacceptable
Breusch–Pagan	$p \geq 0.05$	$p < 0.001$	Unacceptable
Durbin–Watson	DW ≈ 1.5 – 2.5 $p \geq 0.05$	DW = 1.75, p = 0.016	Acceptable
VIF	1- 5	1.10	Acceptable

Because heteroskedastic was detected in the model , the fifth step took place, applying robust (HC3) standard errors shown that the statistical significance of both predictors disappeared ($p > 0.10$), which suggests that the high R^2 value was influenced by heteroskedasticity and does not represent true correlation. To address these issues of not complying with the OLS assumptions a nonlinear model is proposed, a logarithmic transformation of the dependent variable that led to the simulation in RStudio of a log-linear model.

3.2.2 Log-linear regression model

As also done with the linear model, the stepwise AIC was used to identified the best predictors, this procedure identified four significant predictors: urbanization level (X_1), Human Development Index (X_3), income per capita (X_4), and Poverty level (X_6). The resulting model is shown as Equation 2.

$$\ln Y = 11.5216 + 2.6357x_1 + 0.0018489x_3 + 0.000012938x_4 - 0.0093212x_6 \quad (2)$$

In the ANOVA analysis it is shown that X_1, X_3 and X_4 were statistically significant at the 1% level ($p < 0.001$), while X_6 was not significant but improved the model based on AIC. The

model explains approximately **61.8%** of the variation in electricity consumption (R^2 = 0.618), this value is lower compared to the first multivariable linear regression model.

Table 4. Diagnostic Test for OLS assumptions (multivariable log-linear regression)

Test	Ideal Range	Linear Model	Status
Shapiro–Wilk	W=1 $p \geq 0.05$	W=0.94 $p < 0.001$	Unacceptable
Breusch–Pagan	$p \geq 0.05$	$p = 0.282$	Acceptable
Durbin–Watson	DW ≈ 1.5 – 2.5 $p \geq 0.05$	DW = = 1.76 $p = 0.023$	Acceptable (DW)
VIF	1- 5	1.10	Acceptable

The diagnostic tests confirmed a notable improvement in the model assumptions; results are shown in Table 4. The Breusch–Pagan test indicated no heteroskedasticity, and multicollinearity remained low (VIF < 2). The residuals were slightly non-normal (Shapiro–Wilk) and mildly autocorrelated (Durbin–Watson), the model satisfied homoscedasticity and provided more reliable model.

3.2.3 Comparison and interpretation

A comparison between both models (Table 5) highlights that the linear regression achieved a higher R^2 nevertheless considering Table 3 and 4 it is easily deductible that the strong heteroskedasticity and residual non-normality limits the model´s robustness. In contrast, the log-linear model offers greater statistical stability, it effectively reduces heteroskedasticity and residual non-normality, it complies with the base OLS assumptions.

Table 5. Comparison between the coefficient of determination (R^2) of the linear and log-linear regression models

Criterion	Linear model (Y)	Log-linear model (lnY)
R^2 (full sample)	0.850	0.618
R^2 (test set, Y)	0.985	0.808
Overall model suitability	Predictive, but affected by heteroskedasticity and inference issues	Statistically more robust and interpretable, with better residual behavior

4 Conclusion

This paper analyzed the impact of sociodemographic factors on municipal electricity consumption in Honduras through linear and log-linear regression models. Even though the linear model demonstrated remarkably high explanatory power in the 20/80 test (R^2=0.985) the assumptions of OLS were seriously violated due to heteroskedasticity and non-normality of residuals. The log-linear model eliminated heteroskedasticity, significantly improved residual behavior, and had a R^2 value of 0.81 in the original scale. The results suggest that the process of urbanization, human development, per capita income, and level of poverty are significant contributors to shaping the demand for municipal electricity. Overall, the log-linear specification is more statistically robust and interpretable for making forecasts of municipal electricity consumption and further supports better-informed decentralized energy planning in Honduras.

References

1. R . Torres, *Modelado del consumo de energía eléctrica residencial*, REI ITESO Repository, available at: https://rei.iteso.mx/handle/11117/3081 (2015)
2. A. Mazur, *Energy Policy* **39**, 2568–2572 (2011), https://doi.org/10.1016/j.enpol.2011.02.024
3. O. Arto, A. Capellán-Pérez, I. Arto, M. López-González, *Energy Sustain. Dev.* **33**, 1–13 (2016), https://doi.org/10.1016/j.esd.2016.04.001
4. G.C.K. Leung, P. Meisen, *Electricity consumption and human development*, Global Energy Network Institute (GENI), available at: http://www.geni.org/globalenergy/issues/global/qualityoflife/HDI-vs-Electricity-Consumption-2005-07-18.pdf (2005)
5. Z. Mohamed, P. Bodger, *Energy* **30**, 1833–1843 (2005), https://doi.org/10.1016/j.energy.2004.08.012
6. D. Haerer, L. Pratson, *Energy Policy* **82**, 85–98 (2015), https://doi.org/10.1016/j.enpol.2015.03.008
7. J.P. Quintero, *Generación distribuida: democratización de la energía eléctrica*, *Criterio Libre*, Universidad Libre, Bogotá (2008)
8. H.F. Villatoro Flores, *Clean Technol. Environ. Policy* **17**, 1975–1985 (2015)
9. Secretaría de Gobernación, Justicia y Descentralización (SGJD), *Categorización Municipal*, available at: https://www.sgjd.gob.hn/biblioteca-virtual/sgd/categorizacion-municipal (2020)
10. Empresa Nacional de Energía Eléctrica (ENEE), *Cobertura del servicio de energía eléctrica en Honduras* (2019)
11. INE, *Censo de Población y Vivienda de Honduras 2013*, available at: http://181.115.7.199/binhnd/RpWebEngine.exe/Portal?BASE=CPVHND2013NAC&lang=ESP (2013)
12. K. Benoit, *Linear regression models with logarithmic transformations*, London School of Economics, Methodology Institute, available at: https://kenbenoit.net/assets/courses/me104/logmodels2.pdf (2011)