Professor: Veronica Grembi
March, 2021

Applying best subset selection method to model energy demand in residential sector

Mohammadhossein Jafari [1,2] and Quynh Anh Nguyen[1,3]

[1] *Master's student in Data Science and Economics, University of Milan*
[2] *Student ID: 964548 (mohammadhossein.jafari@studenti.unimi.it)*
[3] *Student ID: 947097 (quynh.nguyen@studenti.unimi.it)*

The research aims to explain the factors mostly influence residential sector energy demand in Italy. The data is collected from different sources including ISTAT, and EUROSTAT considering six explanatory variables including resident population, purchasing power parities, household size, median household income, residential electricity cost, and the cost of residential natural gas for the period of eleven years from 2009 to 2019. Best subset selection method is applied for the sake of choosing the optimal number of regressors by comparing different criteria including R^2, RSS, Mallows's C_p, AIC, and BIC. After that, an ordinary least squares regression model is implemented on selected independent variables in order to investigate how they can explain household energy consumption in Italy. The research suggest that an increase in resident population density in Italy would in fact lead to a decrease in overall energy consumption. Thus, in order to design more a comprehensive energy plan, and to improve the accuracy of energy consumption forecasts, resident population should be fully considered. Besides, we also find several policy implications that could be introduced to ensure Italy's energy independence.

Keywords: Residential energy demand; best subset selection method; OLS method, microeconomics

1. INTRODUCTION

Allocating sufficient amount of resources to each part of economy is one of the most important activities that has to be done accurately by policy makers. Among all of the resources, energy ones have a much more important role and countries have to make their best effort to $forecast$ sufficiently energy consumption of each sector of their economy. Most countries require to import energy resources for two main categories. Firstly, they need energy resources to guarantee decent services for the well-being of their society, needing energy for household usage and transportation fall into this category. Secondly, some countries have invested a large amount of financial resources to run industrial companies, where they are highly dependent on energy resources such as crude oil and natural gas as their inputs. For example, refining crude oil to build high-tech products is in this category. When a country is highly dependent on importing energy for both categories, the importance of having a strategic plan for predicting accurately both demand and supply of energy resources becomes a crucial subject. Therefore, it is highly critical for researchers to $forecast$ the proportion of energy which is demanded by each sector.

We decided to focus on Italy since it is one of the industrial countries whose economy is highly dependent on importing fossil fuels and exporting refined products. Italy is one of the main refining centers of petroleum in Europe, its rank is second after Germany (EIA, 2017). In 2016, it imported more than 1.2 million barrels per day and exported about 0.6 million barrels per day of refined products in the same year. It is the third country in consuming natural gas in Europe after Germany and the United Kingdom (EIA, 2017). Transport sector was the largest energy consumer in 2019, by allocating more than 32 percent of energy consumption to itself. Energy

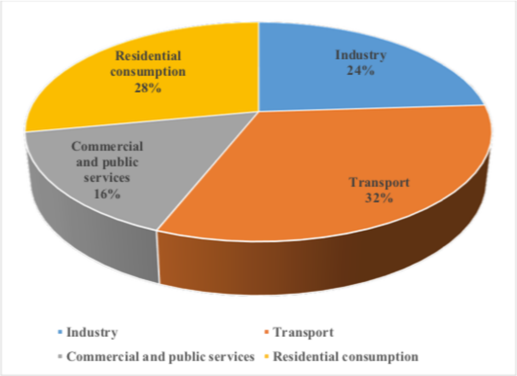


Figure 1. The proportion of energy consumption by each sector in 2019.

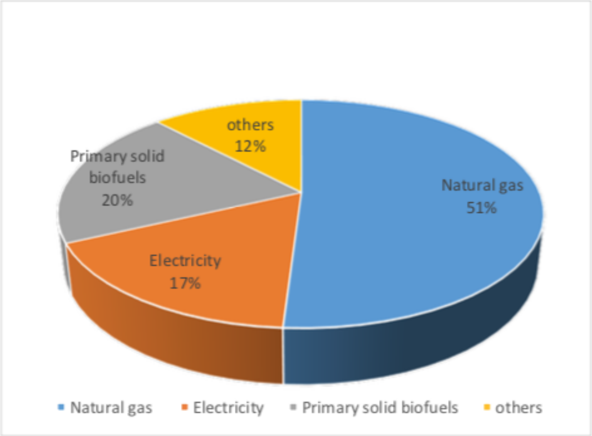


Figure 2. The proportion of types of fuel in final energy consumption of households in 2019.

consumption by sector in Italy in 2019 has been shown by figure 1 (Eurostat, 2021).

Italy has run strategic plans to decrease its independence to fossil fuels in the last 30 years. It became successful to decrease proportionally the role of fossil fuels in its energy consumption basket and add new methods such as using renewable energy sources to provide its energy needs, even though oil products and natural gas are still the main energy resources in its energy consumption basket. The figure 3 illustrates the decreasing trend of using fossil fuel category- which consists of oil products, natural gas, and coal- and the increasing trend of renewable energy sources- which includes bio-fuels, waste, wind, solar energy, and others in Italy from 1990 to 2018, and measured by toe.

At the same time, it is also important to consider that each sector has its own energy consumption basket. We limit our research to household final energy consumption because it has been the second major consumer of energy in several years. Moreover, having an accurate prediction of energy consumption of this sector is socially and politically critical. In other words, if there is not an accurate estimation of energy consumption in this field, it can decrease highly the quality of life of the society and increase greatly dissatisfaction with the government, which can cause crisis in the country. In 2019, natural gas accounted for 51 percent of energy which was consumed by residential sector. The figure 2 (Eurostat, 2021) shows the proportion of types of fuel in final energy consumption of households in 2019.

In this research, we analyse the main factors which potentially influence the consumption demand. While some previous researches have tried to analyze some limited part of residential energy consumption such as electricity in Italy, we try to widen the research area and provide a model which explains the total energy consumption of residential sector. Kialashaki and Reisel made a similar research question in the United States (Kialashaki and Reisel, 2013). It is important to consider that we are running the model in Italy, which is suffering desperately from lacking enough energy resources and is greatly dependent on importing them from other countries, while the United States intrinsically owns a large amount of energy resources.

2. THEORETICAL FRAMEWORK

All those who benefit from the energy market, including consumers, producers, and intermediaries, desperately need models which can explain accurately different types of energy resources which is demanded by each economic sector. A review of different techniques for modeling the residential sector energy consumption has been done by Swan and Ugursal (Swan and Uqursal, 2009). In their research, they define two different approaches: bottom-up and top-down.

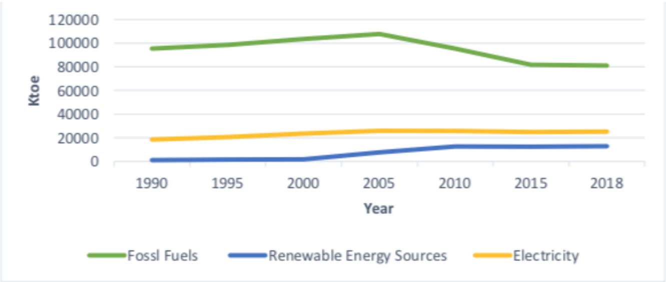


Figure 3. Total final energy consumption by source.

Each technique needs various levels of inputs, simulation and calculation methods, and the outputs of the techniques can be applied in different conditions. They also review precisely the advantages, purposes and negative points of each techniques when they run different models in their research.

The regression model which is run in our research is a part of the bottom-up approach. Models which are included in the bottom-up approach can explain energy demand of end-users, individual houses, and groups of houses. It has been intensively researched by scholars how to use statistical techniques and information to regress energy consumption on house characteristics. It is also critical to include all possible factors that can influence the output interested variable.

How economic factors and energy consumption are related to each other has been studied intensively by Min et al. (Min, Hausfather and Lin, 2010), Geem and Roper (Geem and Roper, 2009), Jin-ming and Xin-heng (Jin-ming and Xin-heng, 2009), Cayla et al. (Cayla, Maizi and Marchand, 2011), and Swan and Ugursal, among others (Swan and Uqursal, 2009). Many researchers provided models which clarify sectoral or total energy consumptions. Geem (Geem, 2011) run models to forecast the energy demand for transportation by consisting different predictors such as population, fossil fuel price, GDP, the figures of vehicle registrations, and passenger transport number. In his research, he compares the results of multiple linear regression models with other models by considering the RMSE and R2. Murat and Ceylan offer an energy consumption model which forecast energy demand for the next 20 years in Turkey (Murat and Ceylan, 2006).

Narrow our focus on the residential sector, Gilland (Gilland, 1988) forecast the global energy demand for a 20-year period by 2020 by having suitable assumptions regarding economic growth, population growth, and the relationship between GDP per capita and elasticity of energy demand by world region. A new method to model the energy demand of residential part of United States by considering both end use and fuel type has been done sufficiently by Min et al. (Min, Hausfather and Lin, 2010). In their research, they offer a detailed analysis about how energy is consumed by residential users in different part of the country and the differences between the way the energy is used in the members of a region and also across the regions. The impact of income changes on household energy usage in the residential and transport part of France has been provided in an in depth research by Cayla et al. (Cayla, Maizi and Marchand, 2011). Their research shows the poorest households are highly limited since they allocate a large amount of their budget to energy services. Song et al. (Song N., Aguilar F.X., Shifley S.R., Goerndt ME, 2012) had made researches about alternative fuel such as wood, which is used in residential sector. They discover that the non-wood energy price is positively related to US household energy usage in the long run when its elasticity is 1.82. At the same time, their research found that the salary rate is negatively related to wood energy demand both in short term and long- term. They also offer that the estimated trend for household wood energy usage is highly negative approximately -3 percent per year.

3. DATA

3.1. Data acquisition

According to mentioned literature reviews, various independent variables including resident population, purchasing power parity, median household size, median household income, cost of residential electricity and cost of residential gas are taken into consideration in order to build a residential energy demand regression model of Italy. The data of considered indicators are collected from two main sources which are Istat, Italian National Institute of Statistics and Eurostat, the statistical office of the European Union for the period of eleven years from 2009 to 2019. Table 1 shows the summary of independent variables and the estimation of energy consuming in residential sector as an dependent variable.

Resident population The annual resident population of Italy is retrieved from Demographic balance and crude rates at national level dataset provided by *Eurostat*. The data provided 58 demographic indicators for all European countries including Italian resident population by counting the number of person whose residence in Italy on 1st January each year.

Purchasing power parity Purchasing power parity (PPP) is a indicator demonstrating how many currency units a given quantity of goods and services costs in different countries. PPPs are calculated for each of European countries and there is no regional breakdown. PPP is to convert national accounts aggregates, i.e. the Gross Domestic Product (GDP) of a country, into comparable volume aggregates. The data is acquitted from Purchasing power parities (PPPs) dataset of Eurostat for the period from 2009 to 2019.

Household size The data is a subset of "Aspects of daily life" data which has executed by *Istat* annually throughout all regions in Italy. The purpose of this survey is to evaluate the integrated system of social by collecting fundamental information on individual and household daily life. The average household size is the average

Table 1. The summary of independent variables and dependent variable used in the analysis.

Denotation	Variable	Unit	Description	Reference
X1	Population	Thousands of people	The number of people having their usual residence in Italy on 1 January of the respective year.	(Eurostat, 2021)
X2	Purchasing power parities	Millions of Euros	Converting the Gross Domestic Product into comparable volume aggregates.	(Eurostat, 2020)
X3	Average household size	Individuals	The average number of person in a family in Italy.	(Istat, 2020)
X4	Median household income	Euro	Calculated by adding personal income of family members and income received at household level.	(Eurostat, 2020)
X5	Residential electricity cost	Euro/kWh	Measuring electricity prices for residential sector including all tax and VAT.	(Eurostat, 2019)
X6	Residential natural-gas cost	Euro/kWh	Measuring natural gas prices for residential sector end-users including all tax, levies and VAT.	(Eurostat, 2019)
Y	Consumption estimates	Ktoe	The energy consumption of households for heating, cooling, cooking, and electricity consumption by electrical appliances.	(Eurostat, 2019)

number of person in a family in Italy. It is identically the mean of the household size distribution over the period from 2009 to 2019.

Median household income This independent variable is measured using *Income and living condition* dataset from *Eurostat*. In this paper, we use annual median income of each family in Italy as an indicator. The household disposable income is calculated by adding the personal income received by family members plus income received at household level.

Cost of residential electricity The data provides information about electricity prices for household consumers in Italy for the period 2009-2019. The data is provided as an semesterly data, it is modified into annual data by obtaining mean of two semesters. The prices are in Euro and included all taxes and levies.

Cost of residential gas The data of this predictor is a subset of Gas prices for domestic and industrial consumers dataset which Eurostat measured natural gas prices for households and industrial end-users updated in 2019. Since the data is provided as an average half-yearly, we modified it into annual data by obtaining mean of two semesters. The prices are in Euro and including all taxes, levies, VAT.

Residential sector energy consumption estimates The data covers the energy consumption of households including individual dwellings, apartments, and so on for space heating, water heating, cooling, cooking as well as electricity consumption by various electrical appliances. The data is a subset of Final energy consumption in households by type of fuel dataset provided by the statistical office of Europe Commission *Eurostat* from 2008 to 2019. The variable is measured by thousand tonnes of oil equivalent (TOE) unit.

3.2. Descriptive statistics

Table 1 shows the summary of detailed information of all indicators. A two-dimensional panel data is used in our analysis. In this section, a brief statistic interpretation might help us have a overview of the data.

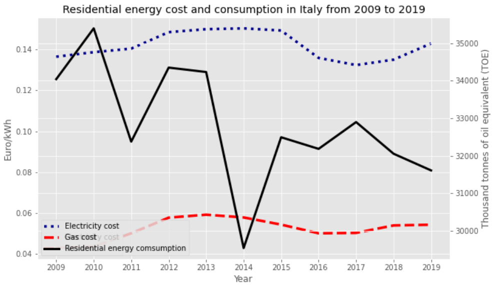


Figure 4. Residential energy cost and Estimated household energy consumption in Italy from 2009 to 2019

The figure 4 illustrates how residential energy consumption, electricity cost and gas cost changed from 2009 to 2019. Generally, household energy consumption had a downward trend from 2009 to 2019. It can be a sign that the plans of the government to encourage households to consume energy more efficiently have had sufficient results. The household energy consumption was 34000 ktoe in 2009 and it decreased sharply and reached to 30000 ktoe in 2014, which was its lowest amount from 2009 to 2019. During the next five years, the residential energy consumption increased by more than 2500 ktoe.

During this 10 years, electricity cost has been at least 0.1 Euro per kWh much more than gas cost. The electricity cost was well below 0.14 Euro per kWh in 2009. It experienced a rapidly increase to about 0.15 Euro per kWh from 2011 to 2012, and then remained stable for three years. The electricity cost started to decrease sharply from 2015 and reached to its smallest amount in this ten-year-period, which was 0.125 Euro per kWh, even though this downward trend did not continued and the cost of electricity grew again to well above 0.14 Euro per kWh in 2019. Gas cost had approximately the same trend from from 2009-2019. The gas cost went up by 20 percent from 2011 to 2012, and it remained stable for 2 years at 0.06 Euro per kWh. After that, it started to decrease moderately for the next two years, and then it went up to .055 Euro per kWh in 2019.

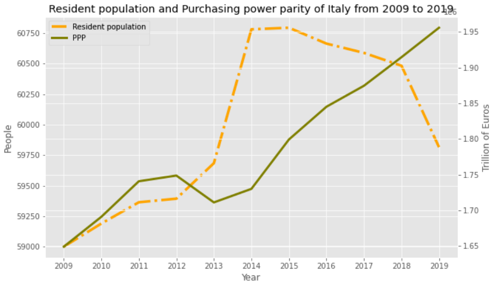


Figure 5. Resident population and Purchasing power parity of Italy from 2009 to 2019

The changes of the two vital macroeconomic factors have been shown in figure 5. The purchasing power parity had a moderately growth, by 6 percent from 2009 to 2010. It experienced a slight decrease during the next three years. Then, it grew sharply from 2013 to 2019, when it reached to its highest amount which was well above 1.95 trillion of Euro. The population had a gradual rise, more than half million people, from 2009 to 2013. After that, it increased swiftly to above 60.75 million people in the next year, even though it had also a sharp decreasing trend in the next five years, and reached finally to well above 59.750 million people in 2019. The figure 6 shows the two crucial household factors. Median household income increased to approximately 16000 Euros in 2010, and it remained stable for two years in this level. From 2012 to 2015, it had a small drop. Then, it started to go up sharply and reached to its highest amount, which was about 17200 Euros in 2019.

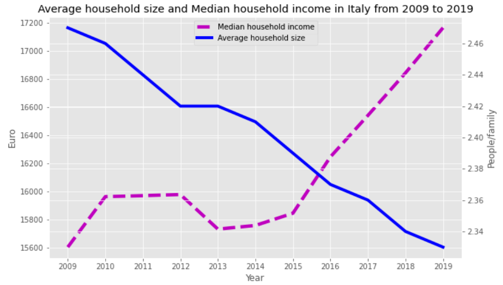


Figure 6. Average household size and Median household income in Italy from 2009 to 2019

Average household size had a totally different trend in the same period, it experienced a significant fall to 2.42 individuals per family from 2009 to 2012. It remained stable for one year at this level. It again started to decrease considerably to well below 2.34 people during the next five years.

4. MODELING METHOD

In our analysis, two methods will be applied subsequently on the pre-processed dataset. Initially, a best subset selection regression will be executed on all regressors in order to achieve a smaller subset of independent variables. There are five creteria including RSS, R^2, AIC, BIC and Mallows's C_p will be taken into account for selecting the final subset of independent variables. After that, we fit an ordinary least squares (OLS) regression model on the obtained subset from previous stage to analyze the casual effect of independent variables on the response variable which is energy consumption of residential sector.

4.1. Best Subset selection Regression

The model building provokes a common problem which is how to remove inefficient regressors within a set of all important candidates. In fact, not all of these candidate regressors are necessary for adequate modeling of the historical data, thus we are interested in screening the candidate variables to obtain the regression model that contains the best subset of regressor variables (Kialashaki

and Reisel, 2013). Moreover, a model with fewer regressor variables would be much easier to interpret and get the valuable insights.
For the sake of performing best selection, we fit separate regression models for each possible combination of the 6 predictors. That is we fit:

- All models that contain 1 predictors at the first step: $\binom{6}{1}$.
- All models that contain 2 predictors at the second step: $\binom{6}{2}$.
- All models that contain 3 predictors at the third step: $\binom{6}{3}$.
- All models that contain 4 predictors at the fourth step: $\binom{6}{4}$.
- All models that contain 5 predictors at the next step: $\binom{6}{5}$.
- Unique model contains all predictors at the last step: $\binom{6}{6}$.

By using combinatorial formula, the number of possible models results in 2^n possibilities. Since we have 6 candidate regressors, there are total $2^6 = 64$ models to be examined. All these models are built using the multiple linear regression method. After that, we evaluated all models using a group of statistical measures. The examined criteria include:

Residual sum of squares Denoted by RSS, it provides the estimation of errors by summing up the squares of residuals. The optimal subset of regressors are usually chosen so that RSS is as small as possible.

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

R-squared R^2 is a common used criterion represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model.

$$(R_p^2) = 1 - \frac{RSS}{TSS}$$

Using exclusively RSS and R^2 might not bring us to the ideal solution of choosing a regression model because the training set is generally an underestimate of the unknown data. This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training RSS is minimized. In particular, the training RSS decreases as we add more features to the model, but the test error may not (Xavier, 2018). Therefore, we can use other measures to examine the models such as $C - p$, AIC, BIC.

C_p Mallows's C_p, named for Colin Lingwood Mallows, is used to assess the fit of a regression model that has been estimated using ordinary least squares. The formula is defined as:

$$C_p = \frac{1}{m}(RSS + 2d\hat{\sigma}^2)$$

AIC This determines the relative information value of the model using the maximum likelihood estimate and the number of parameters in the model.

$$AIC = \frac{1}{m\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

BIC The Bayesian information criterion or also known as Schwarz criterion (SBC, SBIC) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function, and it is closely related to Akaike information criterion (AIC).

$$BIC = \frac{1}{m\hat{\sigma}^2}(RSS + \log(m)d\hat{\sigma}^2)$$

In order to lower the test error of our regression model, we will put a priority on the models with smaller values of C_p, AIC, and BIC.

4.2. OLS Regression

Formulated at the beginning of the 19th century by Legendre and Gauss, OLS method is one of the most popular statistical techniques used in the social sciences. It is used to predict values of a continuous response variable using one or more explanatory variables and can also identify the strength of the relationships.

In this analysis, we will run a ordinary least squares regression on the dataset to explore the relationship of dependent and independent variables after having determined the model with a optimal number of regressors from section 4.1. In other words, we are going to analyse how the residential energy consumption is effected by other indicators such as cost of gas, household size, population, etc.

5. RESULTS

5.1. Optimal regression model

The results of the regression models are presented in Appendix A. Table A1 contains the summary of the best models among all possible models. By comparing a benchmark of R^2 and RSS values, we choose the model in which its R^2 is maximum and RSS is minimum.

Figure 7 and 8 reveal that the optimal number of independent parameters for the regression model is four.

However, as having been discussed in the section 4.1, by choosing the optimal values of RSS and R^2 bring us a risk of having over-fitting model. Therefore, we will consider Mallows's C_p, AIC and BIC indices as well.

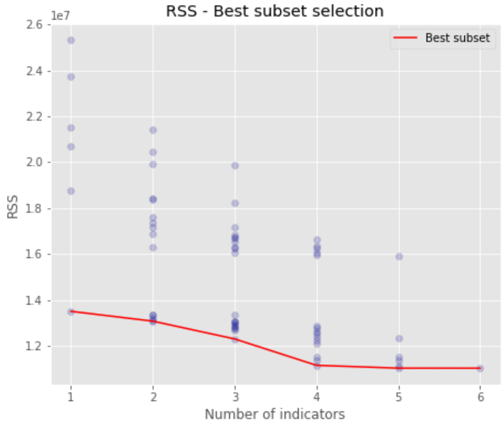


Figure 7. Best subset selection benchmark by RSS.

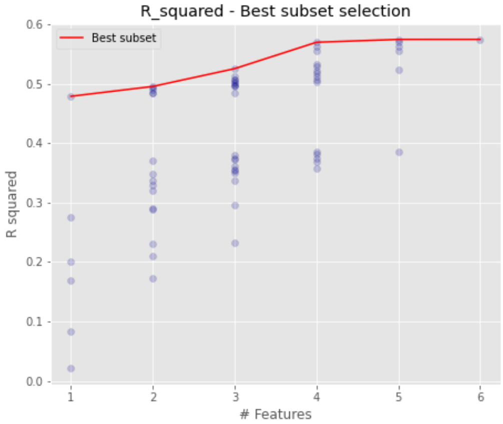


Figure 8. Best subset selection benchmark by R-squared.

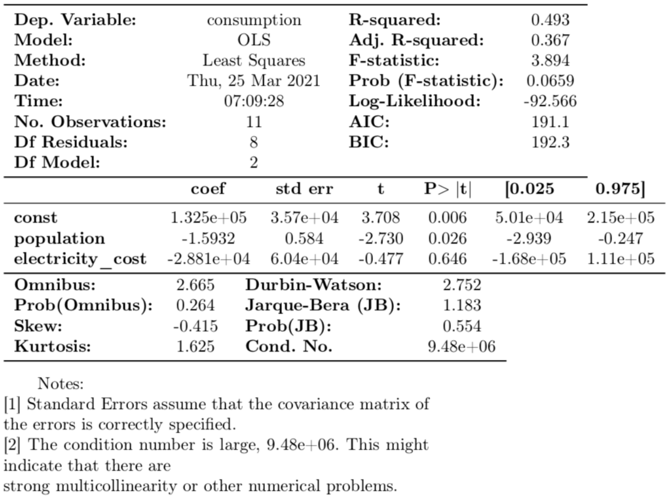
For the sake of selecting a set of appropriate independent variables for our regression model, all the possible models have been tested considering other indicators beyond R^2 and RSS. Table A2 in Appendix A shows the values of other indicators including Mallows's C_p, AIC and BIC when we run all 64 models. The results show that the model which give us the best values, i.e. smaller values, of these measures is regression model using two indicators, *population* and *the cost of electricity*. The following formula shows the regression equation for the proposed model:

$$Y_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

where Y_2 represents the residential energy consumption; β_0, β_1 and β_2 is the corresponding regression coefficients of the model; X_1 and X_2 are resident population and the cost of electricity in Italy over eleven years from 2009 to 2019.

5.2. Regression output

We run an OLS test applied regression equation (1) with *population* and *electricity_ cost* regressors on the energy demand in residential sector. When running a regression, we try to discover whether the coefficients on *resident population* and *electricity cost* are really different from 0 or whether alternatively any apparent differences from 0 are just due to random chance.



```
Dep. Variable:        consumption          R-squared:              0.493
Model:                OLS                  Adj. R-squared:         0.367
Method:               Least Squares        F-statistic:            3.894
Date:                 Thu, 25 Mar 2021     Prob (F-statistic):     0.0659
Time:                 07:09:28             Log-Likelihood:         -92.566
No. Observations:     11                   AIC:                    191.1
Df Residuals:         8                    BIC:                    192.3
Df Model:             2

                      coef        std err     t        P> |t|    [0.025      0.975]
const                 1.325e+05   3.57e+04    3.708    0.006     5.01e+04    2.15e+05
population            -1.5932     0.584       -2.730   0.026     -2.939      -0.247
electricity_cost      -2.881e+04  6.04e+04    -0.477   0.646     -1.68e+05   1.11e+05

Omnibus:              2.665   Durbin-Watson:       2.752
Prob(Omnibus):        0.264   Jarque-Bera (JB):    1.183
Skew:                 -0.415  Prob(JB):            0.554
Kurtosis:             1.625   Cond. No.            9.48e+06

Notes:
[1] Standard Errors assume that the covariance matrix of
the errors is correctly specified.
[2] The condition number is large, 9.48e+06. This might
indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 9. Regression model output.

In other words, we would like to discover whether independent variables are having any effect on our dependent variable. Thus, the null hypothesis H_0 in our test is that *resident population* and *electricity cost* variables are having absolutely no effect, i.e. has a coefficient of 0, and we are seeking for a reason to reject this theory. The output of OLS regression is presented in the Figure 9.

The P value in our regression model as a whole is 0.0659 which is greater than 0.05. With a P value of 6.5 percent, there is only a 6.5 percent chance that H_0 would have come up in a random distribution, so we can say that a 93.5 percent probability of being correct that the variable is having some effect, assuming our model is specified correctly. Conventionally the 5 percent is set as the significance level, we fail to reject the null hypothesis in our case. In other words, we can not conclude whether there is any effect of *resident population* and *electricity cost* on residential energy consumption even though the p-value is pretty close the the significant level of 5 percent. In case, the significant level is more tolerant, e.i. the significant level is 7 percent instead of 5 percent, we can consider the coefficients of independent variables provided by proposed model. Apparently, *resident population* is the only one of which statistically significant in predicting (or estimating) *residential energy demand*; surprisingly, we see that as population increases by 1, energy demand in residential sector will decrease by 1.5932 ktoe. Besides, R-squared value is 0.493, meaning that this model explains 49.3 percent of the variance in our dependent variable. As we know, adding more variables to a regression model will make R-squared be higher while we have used only two variables for our model to avoid the over fitting problem when it comes to regression model, it is reasonable not to have a decent value of R-squared.

6. DISCUSSION

Although the general expectation is for the demand to continue to increase, the research presented to date suggest that an increase in resident population density in Italy would in fact lead to a decrease in overall energy consumption. Considering the effects of population on energy demand in Italy, we find several policy implications:

First, understanding energy consumption patterns must take precedence in order to implement effective demand management policies. As the population grows, the region will become increasingly dependent on external energy supplies to keep up with demand. Even though any current shortcomings in supply as mentioned might be counteracted by energy imports, in the long-term, population growth significantly diminishes the energy self-sufficiency of the area. Therefore, it is necessary that policy solutions are introduced to ensure Italy's energy independence by understanding factors that effect it.

Second, in order to design efficiently a comprehensive energy plan, and to improve the accuracy of energy consumption forecasts, population age distribution should be fully considered. The demographic makeup of populations might determine how and how much energy is being used. As such, ageing populations in countries like Italy lead to increased demand for small-scale household energy use, resulting in an overall lower consumption.

Limitation In the context of rising the attention of international community for sustainable development, the paper have not consider yet other factors that potentially correlate to energy consumption in residential sector. Thus, we might study additional indicators such as $CO2$ Emissions or the changing of temperature over years in Italy to have the better insights of the energy demand. On the other hand, the data in use is a panel data including multiple time series of 11 years. Therefore, multiple time series analysis for a longer period of time would be an appropriate approach for the purpose of forecasting residential energy consumption in the future periods. Last but not least, considering the diversity of regions of Italy in term of geographic and demographic features, a customized analysis for each regions would deepen the results regarding understanding residential energy demand.

REFERENCES

Cayla JM, Maizi N, Marchand C. The role of income in energy consumption behavior: evidence from French households' data. Energy Policy 2011; 39:7874–83.

EIA.International energy statistics. <https://www.eia.gov/international/analysis/country/ITA>.

EUROSTAT. Demographic balance and crude rates at national level. <https://ec.europa.eu/eurostat/databrowser/view/DEMO_GIND/default/table?lang=en>.

EUROSTAT. Purchasing power parities (PPPs), price level indices and real expenditures for ESA 2010 aggregates. <https://ec.europa.eu/eurostat/cache/ metadata/en/prc_ppp_esms.htm>.

EUROSTAT. Income and living conditions. <https://ec.europa.eu/eurostat/cache/metadata/en/ilc_esms.htm>.

EUROSTAT. Electricity prices for domestic and industrial consumers, price components. <https://ec.europa.eu/eurostat/cache/metadata/en/nrg_pc_204_esms.htm>..

EUROSTAT. Gas prices for domestic and industrial consumers. <https://ec.europa.eu/eurostat/cache/metadata/en/nrg_pc_202_esms.htm>.

EUROSTAT. Energy balances. <https://ec.europa.eu/ eurostat/databrowser/product/page/TEN00125>.

EUROSTAT. Final energy consumption by sector. <https://ec.europa.eu/eurostat/databrowser/view/ten00124/default/table?lang=en>.

EUROSTAT. Glossary. <https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Tonnes_of_oil_equivalent_(toe)>.

EUROSTAT. Final energy consumption in households by type of fuel. <https://ec.europa.eu/eurostat/databrowser/view/ten00125/default/table?lang=en>.

Geem Z.W., Roper W.E.. Energy demand estimation of South Korea using artificial neural network. Energy Policy 2009; 37(2009):4049–54.

Geem Z.W.. Transport energy demand modeling of South Korea using artificial neural network. Energy Policy 2011; 39(2011):4644–50.

Gilland B. Population, economic growth, and energy demand, 1985–2020. Popul Dev Rev 1988;14(2).

IEA. Total Energy consumption. <https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Tonnes_of_oil_equivalent_(toe)>.

ISTAT. Aspects of daily life. <http://dati.istat.it/ OECDStat_Metadata/ShowMetadata.ashx?Dataset=DCCV _AVQ_FAMIGLIE&ShowOnWeb=true&Lang=ennk>.

Jin-ming W., Xin-heng L.. The forecast of energy demand on artificial neural network. In: International conference on artificial intelligence and computational intelligence; 2009.

Kialashaki, A., Reisel, J.R.. Modeling of the energy demand of the residential sector in the United States using regression models and artificial neural networks. Applied Energy 108 (2013) 271–280.

Min J., Hausfather Z., Lin QF.. A high-resolution statistical model of residential energy end use characteristics for the United States. J Ind Ecol. 2010; 14(5):791–809.

Murat Y.S., Ceylan H.. Use of artificial neural networks for transport energy demand modeling. Energy Policy 2006; 34:3165–72.

Song N., Aguilar F.X., Shifley S.R., Goerndt M.E.. Analysis of U.S. residential wood energy consumption. Energy Econ; 2012.

Swan L.G., Uqursal V.I.. Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renew Sustain Energy Rev 2009; 13:1819–35.

Wasserman W., Neter J., Kutner M.H., Nachtsheim C.J.. Applied Linear Statistical Models, p289, 4th Edition.

Xavier B.S.. Choosing the optimal model: Subset selection. <https://xavierbourretsicotte.github.io/subset _selection.html>

APPENDIX

	Numb_features	RSS	R_squared	Features
0	1	13500600.6952	0.4789	(population,)
1	1	20702970.4468	0.2008	(PPP,)
2	1	18749956.0303	0.2762	(household_size,)
3	1	23737247.8917	0.0837	(median_income,)
4	1	25353261.2748	0.0213	(electricity_cost,)
5	1	21535506.2730	0.1687	(gas_cost,)
6	2	13342988.1119	0.4849	(population, PPP)
7	2	13355503.7710	0.4845	(population, household_size)
8	2	13213995.5448	0.4899	(population, median_income)
9	2	13127442.7892	0.4933	(population, electricity_cost)
10	2	13072380.9508	0.4954	(population, gas_cost)
11	2	16883485.6821	0.3483	(PPP, household_size)
12	2	17378222.5779	0.3292	(PPP, median_income)
13	2	18422548.7505	0.2889	(PPP, electricity_cost)
14	2	18364314.0633	0.2911	(PPP, gas_cost)
15	2	16286377.4331	0.3713	(household_size, median_income)
16	2	17184936.0108	0.3366	(household_size, electricity_cost)
17	2	17617569.2441	0.3199	(household_size, gas_cost)
18	2	21445938.1356	0.1722	(median_income, electricity_cost)
19	2	19920620.8748	0.2310	(median_income, gas_cost)
20	2	20453654.4716	0.2105	(electricity_cost, gas_cost)
21	3	13342119.3184	0.4850	(population, PPP, household_size)
22	3	12871836.9439	0.5031	(population, PPP, median_income)
23	3	12649825.2610	0.5117	(population, PPP, electricity_cost)
24	3	12938073.6970	0.5006	(population, PPP, gas_cost)
25	3	13047704.8406	0.4963	(population, household_size, median_income)
26	3	12793630.6510	0.5062	(population, household_size, electricity_cost)
27	3	13006568.0905	0.4979	(population, household_size, gas_cost)
28	3	12297619.0857	0.5253	(population, median_income, electricity_cost)
29	3	12772610.4849	0.5070	(population, median_income, gas_cost)
30	3	13042916.1955	0.4965	(population, electricity_cost, gas_cost)
31	3	16278410.2088	0.3716	(PPP, household_size, median_income)
32	3	16532732.8368	0.3618	(PPP, household_size, electricity_cost)
33	3	16828737.4706	0.3504	(PPP, household_size, gas_cost)
34	3	16684746.5151	0.3559	(PPP, median_income, electricity_cost)
35	3	16731467.7411	0.3541	(PPP, median_income, gas_cost)
36	3	18220570.5347	0.2967	(PPP, electricity_cost, gas_cost)
37	3	16040635.0435	0.3808	(household_size, median_income, electricity_cost)
38	3	16230447.4269	0.3735	(household_size, median_income, gas_cost)
39	3	17179547.5621	0.3368	(household_size, electricity_cost, gas_cost)
40	3	19879231.3921	0.2326	(median_income, electricity_cost, gas_cost)
41	4	12871596.4392	0.5031	(population, PPP, household_size, median_income)
42	4	12400796.8949	0.5213	(population, PPP, household_size, electricity_...
43	4	12511084.5469	0.5171	(population, PPP, household_size, gas_cost)
44	4	11537080.0284	0.5547	(population, PPP, median_income, electricity_c...
45	4	12091076.4135	0.5333	(population, PPP, median_income, gas_cost)

Table A1: Benchmark of all possible models with RSS and R-squared.

Continued on next page

46	4	12630587.6494	0.5124	(population, PPP, electricity_cost, gas_cost)
47	4	11147241.8625	0.5697	(population, household_size, median_income, el...
48	4	11349843.2315	0.5619	(population, household_size, median_income, ga...
49	4	12776707.2110	0.5068	(population, household_size, electricity_cost,...
50	4	12220016.6072	0.5283	(population, median_income, electricity_cost, ...
51	4	16039674.4113	0.3809	(PPP, household_size, median_income, electrici...
52	4	16230384.4687	0.3735	(PPP, household_size, median_income, gas_cost)
53	4	16341008.9123	0.3692	(PPP, household_size, electricity_cost, gas_cost)
54	4	16644198.4439	0.3575	(PPP, median_income, electricity_cost, gas_cost)
55	4	15943331.2556	0.3846	(household_size, median_income, electricity_co...
56	5	11119230.8984	0.5708	(population, PPP, household_size, median_incom...
57	5	11344653.0685	0.5621	(population, PPP, household_size, median_incom...
58	5	12357856.6263	0.5230	(population, PPP, household_size, electricity_...
59	5	11515031.8033	0.5555	(population, PPP, median_income, electricity_c...
60	5	11022991.3789	0.5745	(population, household_size, median_income, el...
61	5	15934078.5873	0.3849	(PPP, household_size, median_income, electrici...
62	6	11021574.1095	0.5746	(population, PPP, household_size, median_incom...

	Numb_features	Features	C_p	AIC	BIC
0	1	(population,)	1728307.9773	0.6272	0.6634
1	1	(PPP,)	2383068.8638	0.8649	0.9010
2	1	(household_size,)	2205522.0986	0.8004	0.8366
3	1	(median_income,)	2658912.2679	0.9650	1.0012
4	1	(electricity_cost,)	2805822.5754	1.0183	1.0545
5	1	(gas_cost,)	2458753.9389	0.8923	0.9285
6	2	(population, PPP)	2214960.2019	0.8039	0.8762
7	2	(population, household_size)	2216097.9891	0.8043	0.8766
8	2	(population, median_income)	2203233.6049	0.7996	0.8720
9	2	(population, electricity_cost)	2195365.1726	0.7968	0.8691
10	2	(population, gas_cost)	2190359.5509	0.7949	0.8673
11	2	(PPP, household_size)	2536823.6174	0.9207	0.9930
12	2	(PPP, median_income)	2581799.6989	0.9370	1.0093
13	2	(PPP, electricity_cost)	2676738.4418	0.9715	1.0438
14	2	(PPP, gas_cost)	2671444.3793	0.9695	1.0419
15	2	(household_size, median_income)	2482541.0493	0.9010	0.9733
16	2	(household_size, electricity_cost)	2564228.1928	0.9306	1.0030
17	2	(household_size, gas_cost)	2603558.4867	0.9449	1.0172
18	2	(median_income, electricity_cost)	2951592.0223	1.0712	1.1435
19	2	(median_income, gas_cost)	2812926.8168	1.0209	1.0932
20	2	(electricity_cost, gas_cost)	2861384.4165	1.0385	1.1108
21	3	(population, PPP, household_size)	2715861.8621	0.9857	1.0942
22	3	(population, PPP, median_income)	2673108.9189	0.9701	1.0787
23	3	(population, PPP, electricity_cost)	2652926.0387	0.9628	1.0713
24	3	(population, PPP, gas_cost)	2679130.4419	0.9723	1.0808
25	3	(population, household_size, median_income)	2689096.9095	0.9759	1.0845
26	3	(population, household_size, electricity_cost)	2665999.2559	0.9676	1.0761
27	3	(population, household_size, gas_cost)	2685357.2050	0.9746	1.0831

Table A2: Benchmark of all possible models with C$_p$, $AICandBIC$.

Continued on next page

28	3	(population, median_income, electricity_cost)	2620907.2955	0.9512	1.0597
29	3	(population, median_income, gas_cost)	2664088.3317	0.9669	1.0754
30	3	(population, electricity_cost, gas_cost)	2688661.5782	0.9758	1.0843
31	3	(PPP, household_size, median_income)	2982797.3976	1.0825	1.1910
32	3	(PPP, household_size, electricity_cost)	3005917.6365	1.0909	1.1994
33	3	(PPP, household_size, gas_cost)	3032827.1486	1.1007	1.2092
34	3	(PPP, median_income, electricity_cost)	3019737.0618	1.0959	1.2045
35	3	(PPP, median_income, gas_cost)	3023984.4459	1.0975	1.2060
36	3	(PPP, electricity_cost, gas_cost)	3159357.4272	1.1466	1.2551
37	3	(household_size, median_income, electricity_cost)	2961181.4734	1.0747	1.1832
38	3	(household_size, median_income, gas_cost)	2978437.1447	1.0809	1.1895
39	3	(household_size, electricity_cost, gas_cost)	3064718.9751	1.1123	1.2208
40	3	(median_income, electricity_cost, gas_cost)	3310144.7779	1.2013	1.3098
41	4	(population, PPP, household_size, median_income)	3174067.6962	1.1519	1.2966
42	4	(population, PPP, household_size, electricity_...	3131267.7376	1.1364	1.2811
43	4	(population, PPP, household_size, gas_cost)	3141293.8878	1.1401	1.2847
44	4	(population, PPP, median_income, electricity_c...	3052748.0225	1.1079	1.2526
45	4	(population, PPP, median_income, gas_cost)	3103111.3302	1.1262	1.2709
46	4	(population, PPP, electricity_cost, gas_cost)	3152157.8062	1.1440	1.2887
47	4	(population, household_size, median_income, el...	3017308.1892	1.0951	1.2397
48	4	(population, household_size, median_income, ga...	3035726.4955	1.1017	1.2464
49	4	(population, household_size, electricity_cost,...	3165441.4027	1.1488	1.2935
50	4	(population, median_income, electricity_cost, ...	3114833.1660	1.1304	1.2751
51	4	(PPP, household_size, median_income, electrici...	3462074.7846	1.2565	1.4012
52	4	(PPP, household_size, median_income, gas_cost)	3479412.0625	1.2628	1.4075
53	4	(PPP, household_size, electricity_cost, gas_cost)	3489468.8301	1.2664	1.4111
54	4	(PPP, median_income, electricity_cost, gas_cost)	3517031.5148	1.2764	1.4211
55	4	(household_size, median_income, electricity_co...	3453316.3159	1.2533	1.3980
56	5	(population, PPP, household_size, median_incom...	3515742.3793	1.2759	1.4568
57	5	(population, PPP, household_size, median_incom...	3536235.3038	1.2834	1.4642
58	5	(population, PPP, household_size, electricity_...	3628344.7182	1.3168	1.4977
59	5	(population, PPP, median_income, electricity_c...	3551724.2797	1.2890	1.4699
60	5	(population, household_size, median_income, el...	3506993.3321	1.2728	1.4536
61	5	(PPP, household_size, median_income, electrici...	3953455.8056	1.4348	1.6157
62	6	(population, PPP, household_size, median_incom...	4007845.1307	1.4545	1.6716