Simple Regression Model for Energy Demand: Case Study in Rural Areas of Nepal

Anna-Kaarina Seppälä
Chair of Renewable and Sustainable Energy Systems
Technical University of Munich
Munich, Germany
anna.seppaelae@tum.de

Sudhir Jha
Chair of Renewable and Sustainable Energy Systems
Technical University of Munich
Munich, Germany

Stephan Baur
Chair of Renewable and Sustainable Energy Systems
Technical University of Munich
Munich, Germany
stephan.baur@tum.de

Emmanuel Benjamin
School of Life Sciences Weihenstephan
Technical University of Munich
Munich, Germany

***Abstract*—A significant share of Nepal's rural population still live without access to electricity, many in places where grid extension is not possible. To enable off-grid energy projects, this study analyses survey data from three locations powered by off-grid hydro systems, and determines current and future trends in energy consumption.**
A simple linear regression model to predict electrical energy demand in rural areas is also presented, based on empirical results on household energy use, including details on commonly used appliances and socio-economic data. The number of various electrical appliances and their use pattern were found to provide better indicators for energy demand than household income or the size of monthly electricity bills.

***Index Terms*—rural electrification, multiple linear regression, off-grid power systems, energy need, Nepal**

I. Introduction

Nepal has one of the lowest energy consumption figures per capita worldwide, and still ranks very low in many aspects of development [1]. Further, it has effectively no reserves of oil, coal, nor natural gas, and relies on expensive imports and biomass to meet its growing energy demand [2], [3]. The resulting lack of access to energy is one of the major obstacles to socio-economic development [4].

Over 80 percent of Nepal's nearly 29 million inhabitants live in rural areas [5], [6]. Due to the challenging geography of these regions, extending the national grid is strenuous and expensive, and until recently, the substantial hydropower potential in Nepal has remained underutilised [3], resulting in a low rural electrification rate, estimated at 80 percent by the World Bank [7]. Yet the government of Nepal ranks it at 64.9 percent [8], [9]. The chronic shortage of electricity leads to power-quality issues and frequent outages in the existing decentralised grids.

Building more small hydro off-grid systems could greatly improve the situation by offering an affordable and flexible solution to energy shortage in rural areas. An essential first step in the design of such systems is to analyse the trends of the current power consumption in Nepal, and to make reasonable projections for the near future. Accurate estimations provide answers to how much capacity needs to be added to solve existing power realibility problems. Furthermore, knowing the status quo and future trends will help currently unelectrified villages to avoid similar issues when they receive electricity.

The relationship between economic growth and energy consumption has been widely discussed [10], and the direction of causality is not always clear. This study took a simplistic approach based on number of household appliances and their use pattern, and compared it to more traditional indicators of household consumption, such as monthly electricity bills, education, and average household income. The causality relationship was determined as part of an energy needs survey.

Concretely, three districts with off-grid hydro power were surveyed, and the results were used to create a simple energy need model based on multiple linear regression. In the rest of the article, the terms energy and electrical energy will be used synonymously unless stated otherwise.

II. Method

The aim of this study was firstly the analysis of the current energy demand in specific rural areas of Nepal, and secondly the creation of a simple model to predict this demand in similar regions. To this end, a survey was conducted in three districts of the Pahad region, and the obtained data was then analysed to find reasons for Nepal's extremely low household electricity consumption, and to determine which factors most affect it. Also, the direction of causality was established by the analysis. Subsequently, a set of multiple linear regression models was created to predict energy need, and the models were validated with standard statistical tools. Lastly, based on the conducted survey, estimations were made on the future energy demand in a rural setting.

This research was partly funded by Maj and Tor Nessling Foundation.

A. Data Gathering

The survey was conducted in three districts of the Pahad region in Nepal: Bajhang, Kavrepalanchok, and Panchthar. The regions were chosen based on their similarities regarding geography and climate, household average income, and population size. Thus, homogeneous sampling was assured. 30 households per district were surveyed, resulting in a total sample size of 90. The households in each location were chosen by random sampling. Fig. 1 shows the visited districts on a map.

The survey gathered data on socio-economic factors of the household and its energy needs, including the use pattern of common electrical appliances. Studied household details include household size, yearly income, educational level, occupation type of the head of house, and other economic activities. Household income was separated into two categories: 0 for less than 1,500 USD per year and 1 for more than 1,500 USD per year. Level of education was assessed on a scale from 0 to 7, from illiterate to Master's level or above.
The monthly household electricity need was determined based on the number of electrical appliances used within each household and the hours (per day) of use of said devices. The monthly energy costs, and amount of cooking fuel used per month were also provided. Additionally, questions were asked about the dwellers' satisfaction with power quality, and their desire for additional appliances.

B. Data Analysis

In order to find statistical significance in the empirical data gathered by means of the energy needs survey, a mathematical model was set up. A simple relationship between measured indicators and the resulting monthly energy consumption was assumed, and hence a linear model was chosen.

1) Linear Regression Models: The simplest linear represetation of the relationship between two variables is the linear regression model based on the *method of least squares* [12]. The model is of the form

$$\hat{y}_i = \beta_0 + \beta_1 x_i \tag{1}$$

where $\hat{y}_i$ is the prediction of the i$^{\text{th}}$ element in a set of values of the dependent (outcome) variable $\bar{y}$[1], and x_i is the corresponding value of the independent variable (feature). β_0 and β_1 are unknown constants. The predicted and real outcomes are assumed to differ from each other by a statistical error term e_i, or residual, with a zero mean and a normal distribution of unknown variance σ^2:

$$y_i = \hat{y}_i + e_i. \tag{2}$$

To determine β_0 and β_1, the model is iteratively 'trained' with a set of examplery values of feature $\bar{x}$ called a *training set*, and measured outcome $\bar{y}$. The objective is to minimise the squared sum of the residuals in the training set:

$$min\Big(\sum_{i=1}^{m} {e_i}^2\Big) \tag{3}$$

where m is the total number of elements of feature $\bar{x}$.

2) Multiple Linear Regression: If more than one independent variable is linearly correlated with the outcome $\bar{y}$, *multiple linear regression* may be used to study their combined effect. Then, the regression model takes the form

$$\hat{y}_i = \beta_0 + \beta_1 {x_i}^{(1)} + ... + \beta_n {x_i}^{(n-1)} \tag{4}$$

where $n-1$ is the total number of features $\bar{x}^{(k)}$. Please note that the superscript is just an index notation, not an exponent. Equation (4) may also be expressed in matrix form:

$$\hat{\bar{y}} = X\bar{\beta} \tag{5}$$

where $\hat{\bar{y}} = \begin{bmatrix} y_0 & y_1 & \dots & y_m \end{bmatrix}^T$ is the predicted outcome, and $\bar{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \dots & \beta_n \end{bmatrix}^T$. The feature matrix X is defined as follows:

$$X = \begin{bmatrix} 1 & {x_0}^{(1)} & \dots & {x_0}^{(n-1)} \\ 1 & {x_1}^{(1)} & \dots & {x_1}^{(n-1)} \\ 1 & {x_m}^{(1)} & \dots & {x_m}^{(n-1)} \end{bmatrix} \tag{6}$$

and has the dimensions $m \times n$. An artificial feature $\bar{x}^{(0)} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$ has been added to correspond to the coefficient β_0. In this form, the vector $\bar{\beta}$ may be solved analytically:

$$\bar{\beta} = (X^T X)^{-1} X^T \bar{y} \tag{7}$$

3) Validation of Regression Models: Even when variables $\bar{x}^{(k)}$ and $\bar{y}$ are unrelated, it is possible to fit a least-squares model to the measured values. Hence it is essential to validate the model with standard statistical tools to show its relevance. Before constructing the regression model, a study of the linear relationship between the outcome (monthly energy consumption) and all measured features was conducted using *Pearson's correlation coefficient* for sample data [15]:

$$r^{(k)} = \frac{(\sum_{i=1}^{m}(x_i{}^{(k)} - x_{ave})(y_i - y_{ave})}{\big[\sum_{i=1}^{m}(x_i{}^{(k)} - x_{ave})^2 \sum_{i=1}^{m}(y_i - y_{ave})^2\big]^{1/2}} \tag{8}$$

Features with a linear correlation of $r^{(k)} = 0.3$ or higher were deemed significant. For models with more than one feature, (8) was used to check the linear correlation between the independent variables. A value of 0.2 or less was assumed low enough to include both in the model without introducing

[1]In this articale, the bar symbol $^-$ is used for vectors.

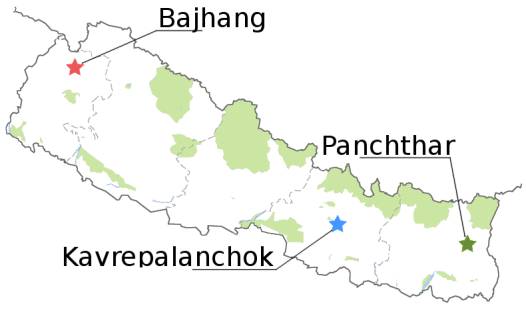

Fig. 1. Location of Bajhang, Kavrepalanchok, and Panchthar on the map. Own illustration, map data from Google [11].

errors due to cross-correlation.

To analyse the linear regression models, a variance analysis table corrected for the mean was used [13]. Table I presents the equations used for these statistics. Within, df stands for degrees of freedom, and ν is the number of elements in $\bar{\beta}$. The resgression is assumed significant if the regression sum of squares is large relative to the residual sum of squares [12].

Further indicators for the relevance of the overall regression model include the *F-test*, the *coefficient of multiple determination* (R^2), the *leave-one-out cross validation* (LOOCV), and the *root-mean-square error*[2]. For all validation tests, the confidence level $\alpha = 0.05$ was used. In this study, the LOOCV method was used on a datapoint level and a distric level. The method utilises a single observed datapoint of the original sample for validation, and the others form the training set. The procedure is repeated until each datapoint has been used once as validation data. On a district level, instead of leaving one single datapoint out of the training set, one full district sample (30 households) was used for validation, resulting in three iterations of the method. On the datapoint level, *predicted* R^2 was used as a validation statistic, and for the three districts, the relative error between the real and estimated total electricity consumption of each district sample was computed.

To test the contribution of a particular independent variable within the multiple regression model, the null hypothesis $H_0 : \beta_k = 0$ was taken, and its t-value and a corresponding probability (p-value) were computed. If the p-value is smaller than α, the null hypothesis is rejected and the tested feature deemed significant. Additionally, the standard error and confidence levels for each β_k were computed.

III. Results

A. Survey Outcome

In the survey, the average monthly household electricity consumption in the surveyed districts was found to be very low. For Bajhang, Kavrepalanchok, and Panchthar it was 10.7 kWh, 19.6 kWh and 15.6 kWh, respectively. A boxplot of the monthly energy consumption has been provided in Fig. 2. For reference, the average U.S. residential utility customer had a eletricity consumption of 897 kWh per month in 2016 [14]. Households typically use electricity for lighting and entertainment only, using conventional fuels like firewood and lpg for cooking. The most common household appliances were lights and and old CRT televisions. Bajhang faired worst in terms of access to electricity, with 10 percent of surveyed households without light. The affected families charge their mobile phones at their neighbours'. Bajhang was also found to have fewer literate adults, households of higher income and TV's than Kavrepalanchok and Panchthar. Fig. 3 shows the relative access to various lighting types for each district, while other appliances along with household income and adult literacy rate are plotted in Fig. 4.

Over 85 percent of all surveyed households pay a fixed minimum charge of 80 Rs. (0.78 USD) a month, irrespective of their consumption. Therefore, no real correlation between the two exists, and people have no economic incentive to control their power consumption. From the survey it is also clear that many households are unsatisfied with the grid quality, experiencing voltage fluctuations when plugging in devices. The number and type of appliances used is directly governed by the evident power capacity shortage in all surveyed districts. An interesting group of outliers are households with incadescent light bulbs. These are usually cheaper than the more modern LED lamps[3] or CFL's, and thus still popular in rural areas. Due to their high power consumption, however, some lower-income households with no other appliances still have spectacularly high monthly energy demand. Seeing that over 30 percent of surveyed households own incadescent lights (see Fig. 2), it is obvious that the current power shortage problems could be minimised by simply replacing all existing

[2]Due to space limitations, not all methods could be presented here in detail. The interested reader is referred to [12] and [13] for more information.

TABLE I
VARIANCE ANALYSIS TABLE

Source	*Sum of squares*	*df*	*Mean square*
Regression	$\bar{\beta}X^T\bar{y} - m(y_{ave})^2$	$\nu - 1$	$\frac{\bar{\beta}X^T\bar{y} - m(y_{ave})^2}{\nu - 1}$
Residual	$\bar{y}^T\bar{y} - \bar{\beta}X^T\bar{y}$	$m - \nu$	$\frac{\bar{y}^T\bar{y} - \bar{\beta}X^T\bar{y}}{m - \nu}$
Total	$\bar{y}^T\bar{y} - m(y_{ave})^2$	$m - 1$	–

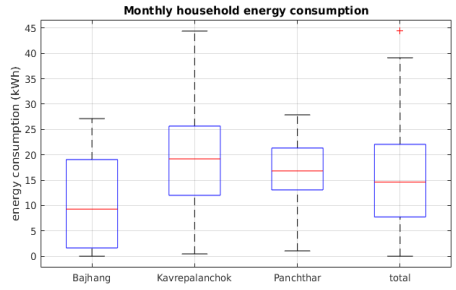

Fig. 2. Boxplot of the monthly energy consumption in Bajhang, Kavrepalanchok, and Panchthar.

[3]According to survey results, LED lamps in Nepal may be up to ten times more expensive than incadescent ones.

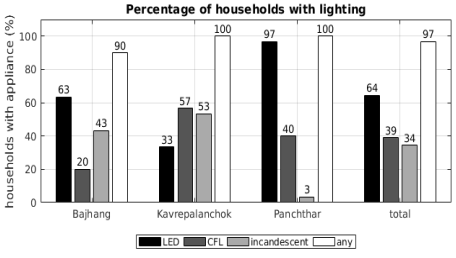

Fig. 3. Percentage of households with lights in Bajhang, Kavrepalanchok, and Panchthar. The label 'any' refers to households having any of the three light sources.

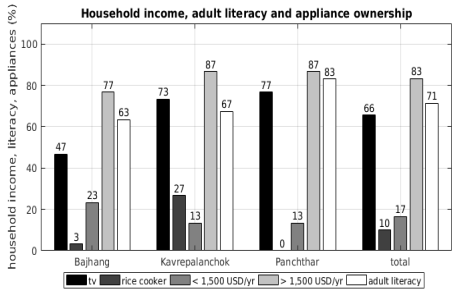

Fig. 4. Percentage of households with TV or rice cooker. Relative distribution of household income and adult literacy is also plotted.

incadescent bulbs (60W-100W) with LED lamps (5W).
Fig. 5 shows the relative popularity of various electrical devices within the surveyed sample. A TV is a desired appliance in households that do not own one, but electrical cooking devices obviously form a rising trend. Only 10 percent of sample households currently posess a rice cooker but 84 percent wish they had one. Seeing that over 80 percent of all surveyees belong to a higher income group, money clearly is not the limiting factor, but rather the lack of power.
Additionally, when asked what type of energy system ownership scheme they prefer, 87 percent of the families chose communal ownership over private-owned systems. Clearly, capital cost and maintenance of private systems are an issue.

B. Multiple Linear Regression Models

From household 90 samples, 85 datapoints were used to contruct the multiple linear regression models. Five datapoints were categorised as outliers due to significant deviations from the rest of the data. Three different methodologies were followed, each with its own model. Firstly, the number of electrical appliances was linked to monthly electricity needs. Secondly, the number of hours of use of each device was chosen as indication of consumption. And thirdly, average household education, along with lpg consumption and monthly electricity bill, was selected as the independent variable. Table II lists the correlation between monthly energy consumption (mec) and various exemplary features. The number of TV's and that of hours spent with the TV on have the best correlation with energy consumption. The average household income (hhi)

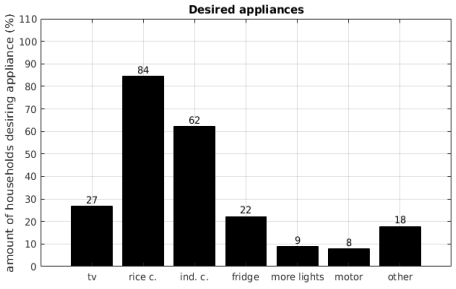

Fig. 5. Percentage of surveyed households wanting various appliances.

TABLE II
CORRELATION MATRIX FOR EXEMPLARY VARIABLES

	mec	***#tv's***	***hrs tv***	***hhi***	***meb***	***av.edu***	***#inc***
mec	1.000	–	–	–	–	–	–
#tv's	0.513	1.000	–	–	–	–	–
hrs tv	0.522	0.771	1.000	–	–	–	–
hhi	0.288	0.311	0.374	1.000	–	–	–
meb	0.179	0.201	0.112	0.197	1.000	–	–
av.edu	0.295	0.194	0.193	0.382	0.086	1.000	–
#inc	0.414	-0.200	-0.248	0.048	0.027	0.030	1.000
hrs inc	0.400	-0.105	-0.180	0.083	0.012	-0.117	0.778

falls somewhat below the threshold of significance chosen previously ($r^{(k)} = 0.3$), and the monthly electricity bill (meb) is far from being significant. Average education (av.edu) is very close to the chosen $r^{(k)}$ value and was thus chosen for the third model. Hosehold income could not be used as an additional feature due to its high cross-correlation with education.
Table III shows the created models and the percent error for each district when using the district-level LOOCV. The error for Bajhang is generally larger because of its slightly lower income level and lower electrification rate. Understandably, when the models are trained with similar district samples and compared with a dissimar one, bias is inevitable.
Based on the LOOCV validation, Model II seems to give the best overall results whilst Model III is extremely biased. A look at the regression power statistics in Tables IV, V and VI confirms that the regression of Model III is insignificant, and its prediction power minimal; the predicted R^2 value is 0.094, and the residual sum of squares is far greater than the regression sum of squares. Both appliance-based models have relatively good regression statistics, but Model II fairs slightly better in all aspects. It is thus safe to assume that the number of hours household appliances are used gives the best approximation of the actual energy consumption.

IV. DISCUSSION

An error of 20 percent is very near to the maximum obtainable accuracy of statistical models. Hence it is clear that tracking the number of household appliances and their hours of use provides a simple and effective way of predicting electrical energy consumption. When using this model, no profound surveys have to be conducted to get a feel for the needed power capacity, as a few simple, easy-to-answer questions provide acceptable results. Further, asking about income and

TABLE III
SUMMARY OF MODELS AND LOOCV RESULTS

Model	***Features***[a]			***LOOCV - % error***		
				B	K	P
Model I	#tv's	#inc	#rc	36.00	19.75	26.85
Model II	hrs tv	hrs inc	hrs rc	19.72	22.44	8.59
Model III	av.edu	lpg	meb	43.14	22.61	0.97

[a]inc = incadescent bulb, rc = rice cooker, lpg = liquid petroleum gas.

TABLE IV
REGRESSION POWER STATISTICS: MODEL I

Source	*SS*	*df*	*MS*
Regr.	5581.2	3	1860.4
Res.	2496.9	79	31.6
Total	8078.1	82	–

$F(3, 79)$	58.862
$P > F$	1.99e-20
adj. R^2	0.679
pred. R^2	0.646
RMSE	5.4848

mec	β_k	*std. err.*	*t*	*P*$>\|t\|$	*95% conf. interval*	
const	4.299	1.142	3.764	5.42e-04	2.3982	6.199
#tv's	11.037	1.282	8.610	1.27e-12	8.904	13.170
#inc	4.056	0.499	8.130	1.09e-11	3.2261	4.887
#rc	18.064	2.919	6.189	5.39e-08	13.208	22.921

other economic data is often met with mistrust, which the proposed method avoids.

Some substantial restrictions to the model still exist. Firstly, the energy consumption was not directly measured in most households but computed based on answers to the survey. Thus, the outcome of the models will still have to be tested in the field. Secondly, the model is limited to similar socio-economic groups and geography. Gathering more data is essential to expanding the model for other scenarios as well. And thirdly, the model does not consider migration effects. In rural areas, once a family reaches a certain income level, they tend to move to a more urban area. Not taking this and other effects into account may distort the outcome of the model.

Future work should concentrate on proper field validation of the models with real measurements for monthly energy consumption. Also, a Monte Carlo simulation shall be realised to extract predicted load curves for the modelled regions, an essential step in power system design. Incorporating other regions of Nepal and eventually other countries would also be of interest. Lastly, the model might be further improved by assuming non-liear correlation between some features and electricity consumption.

As the study has shown, it is evident that current demand has already outgrown the provided capacity in all surveyed districts of Nepal, and the single factor limiting the purchase of new household appliances is the scarcity of power. Due to their high power consumption compared to other devices, the introduction of rice and induction cookers into the districst will make the problem substantially more pronounced, increasing specifically the household peak load. Also the base load will be affected by the increasing number of television sets, bringing it from a few dozens of watts to over a hundred.

The careful design of future energy systems is paramount in order to alleviate and avoid the problems now faced in Bajhang, Kavrepalanchok, and Panchthar. Only by providing access to sustainable, reliable power will the rural regions of Nepal be allowed to develop, both socially and economically, and their quality of life significantly improved.

REFERENCES

[1] M. Islar, S. Brogaard, M. Lemberg-Pedersen, "Feasibility of energy justice: Exploring national and local efforts for energy development in Nepal", Energy Policy, vol. 150(2017), pp. 668–676

[2] S. Pokharel, "An econometric analysis of energy consumption in Nepal", Energy Policy, vol. 35(2007), pp. 350–361

[3] B. B. Pradhan, B. Limmeechokchai, "Electric and Biogas Stoves as Options for Cooking in Nepal and Thailand", Energy Procedia, vol. 138(2017), pp. 470–475

[4] J. Sumanik-Leary, M. Delor, M. Little, M. Bellamy, A. Williams, S. Willimason, Engineering in Development - Energy, London: Engineers Without Borders UK, 2014.

[5] R. Nepal, "Roles and potentials of renewable energy in less-developed economies: The case of Nepal", Renewable and Sustainable Energy Reviews, vol. 16(2012), pp. 2200–2206

[6] The World Bank, Sustainable Energy for All (SE4ALL) database. Population, total. Retrieved 26.12.2017 from https://data.worldbank.org/indicator/SP.POP.TOTL?locations=NP

[7] The World Bank, Sustainable Energy for All (SE4ALL) database. Access to electricity, rural. Retrieved 26.12.2017 from https://data.worldbank.org/indicator/EG.ELC.ACCS.RU.ZS

[8] Government of Nepal - National Planning Commission, Sustainable Development Goals 2016-2030: National (Preliminary) Report, 2015.

[9] Government of Nepal - National Planning Comission, Annual Household Survey 2015-2016: Central Bureau of Statistics Report, 2016.

[10] R. Parajuli, P. A. stergaard, T. Dalgaard, G. R. Pokharel, "Energy consumption projection of Nepal: An econometric approach", Renewable Energy, vol. 63(2014), pp. 432–444

[11] Google, Scribble Maps: Nepal, Retrieved 22.12.2017 from https://www.scribblemaps.com/create/#lat=28.394857&lng=84.124008&z=7

[12] J. H. Pollard, A Handbook of Numerical and Statistical Techniques with Examples Mainly from the Life Sciences, 1st ed., Cambridge: Cambridge University Press, 1977.

[13] N. R. Draper, H. Smith, Applied Regression Analysis, 3rd ed., New York: John Wiley & sons, 1998.

[14] U.S. Energy Information Administration, "How much electricity does an American home use?". Retrieved 27.12.2017 from https://www.eia.gov/tools/faqs/faq.php?id=97&t=3

[15] G. L. Shevlyakov, H. Oja, Robust Correlation: Theory and Applications, 1st ed., Chichester: John Wiley & sons, 2016.

TABLE V
REGRESSION POWER STATISTICS: MODEL II

Source	*SS*	*df*	*MS*
Regr.	5773.6	3	1924.5
Res.	2304.5	79	29.2
Total	8078.1	82	–

$F(3, 79)$	65.974
$P > F$	7.88e-22
adj. R^2	0.704
pred. R^2	0.681
RMSE	5.269

mec	β_k	*std. err.*	*t*	*P*$>\|t\|$	*95% conf. interval*	
const	4.854	1.016	4.778	1.55e-05	3.164	6.5446
hrs tv	2.999	0.306	9.790	6.30e-15	2.489	3.509
hrs inc	3.440	0.461	7.458	2.19e-10	2.673	4.208
hrs rc	23.735	3.277	7.244	5.64e-10	18.282	29.187

TABLE VI
REGRESSION POWER STATISTICS: MODEL III

Source	*SS*	*df*	*MS*
Regr.	1352.4	3	450.8
Res.	6725.7	79	85.1
Total	8078.1	82	–

$F(3, 79)$	5.295
$P > F$	0.003
adj. R^2	0.136
pred. R^2	0.094
RMSE	9.002

mec	β_k	*std. err.*	*t*	*P*$>\|t\|$	*95% conf. interval*	
const	6.925	2.828	2.449	0.021	2.220	11.631
av.edu	2.286	1.023	2.234	0.034	0.583	3.988
lpg	5.328	2.296	2.321	0.028	1.508	9.148
meb	0.014	0.018	0.818	0.284	-0.015	0.044