# Energy-Efficient Home Trends in NYS

Abdellah Ait Elmouden

December 13, 2021

**Abstract**

In most countries, energy efficiency is now considered a key factor in achieving long-term energy and climate goals. In this research project, we'll look at the New York State residential homes energy efficiency projects data to analyze how home and building owners and tenants of existing one- to four-family homes will benefit from implementing comprehensive energy efficiency-related improvements and technologies, and forecast the future energy consumption in New York State that is increasingly exposed to energy-efficient and renewal energy shift. The goal is to answer the following questions: Do energy-efficient homes save energy? What are the primary factors for adoption of energy-efficient program for energy and cost saving?

**keywords**: Energy-Efficient Home; home energy efficiency; home energy reduction; big data; Energy consumption; green energy saving; energy consumption prediction

## 1 Introduction

Global attempts to reduce greenhouse gas (GHG) emissions rely heavily on energy efficiency. According to the (International Energy Agency (2015) [**?**]), global planned climate mitigation strategies rely on energy efficiency to accomplish 42 percent of emissions reductions. Energy efficiency initiatives have a simple appeal: they may pay for themselves in the long run by cutting future energy bills. Lower energy use, on the other hand, reduces dependency on fossil fuel energy sources, resulting in the anticipated GHG reductions. Several government measures, like as efficiency standards, utility-sponsored rebate programs, and information disclosure regulations, aim to encourage increased energy efficiency investment.

Policymakers are likely drawn to energy efficiency because a number of analyses point to substantial unexploited opportunities for cost-effective investments (McKinsey and Company (2009); Tonn et al. (2014), Nadel and Ungar (2019)).

A more complete view of which energy efficiency opportunities are cost-effective, and investigating if energy-efficient homes save utility bills, and in what way. requires more evidence from a variety of settings.

A number of New York States counties have moved to change the way they get their energy. And the NYSERDA offers a range of residential programs designed to help New York State residents identify areas where their homes are driving up energy costs and can provide assistance in completing energy efficiency improvements for a healthier, more comfortable home. The Home Performance with ENERGY STAR Program encourages home and building owners and tenants of existing one- to four-family homes to implement comprehensive energy efficiency-related improvements and technologies by contractors accredited by the Building Performance Institute and participating in the HPwES program. Eligible measures include building shell measures, such as air sealing and insulation; appliances, such as ENERGY STAR refrigerators; heating measures, such as boilers and furnaces; cooling measures, such as ENERGY STAR room or central air conditioners, and certain renewable energy technologies. The HPwES program is designed to offer enhanced assistance to low- to moderate-income households. The "Assisted" component of the HPwES program is available to residents with up to 80 % of area median income, or 80 % of state median income, whichever is higher for the county.

In this research we'll look at the New York State residential homes energy efficiency projects to analyze how home and building owners and tenants of existing one- to four-family homes will benefit from implementing comprehensive energy efficiency-related improvements and technologies. and try to explore some factors related to home characteristics, project types, Incentives and programs that promote the installation of clean technologies will be used as predictors.

# 2 LITERATURE REVIEW

In this section we'll presents a brief summary of a literature review prepared as part of the Energy-Efficient Home Trends research project aimed at enabling a comparison between our project and available academic information. The literature reviewed was primarily selected for its relevance to energy conservation within households. Although quite comprehensive, the review is by no means a complete examination of all available literature within the field.

## 2.1 Energy-Efficient Home Trends

In this research project, we'll look at residential homes energy efficiency data provided by New York State Energy Research and Development Authority (NYSERDA) to analyze how home and building owners and tenants of existing one- to four-family homes will benefit from implementing comprehensive energy efficiency-related improvements and technologies. The primary purpose of this evaluation is to establish first year evaluated gross and evaluated net energy savings for program years' (PY) 2007 and 2008 participants. The primary vehicle for estimating evaluated gross savings was a billing analysis covering the pre- and post-installation periods. The purpose of this analysis is evaluate energy-efficient homes saving, and identify the primary factors that control the energy and cost saving? Other factors are not available in the dataset such socio-demographic, economic that can be used used as predictors. But we used home characteristics, project types, Incentives and programs that promote the installation of clean technologies.

## 2.2 Socio-Demographic correlation with energy usage

Previous research into the area of energy consumption has shown that socio-demographic variables can be highly related to household energy use ( [4] Gatersleben et al. 2002; [2] Abdo Abdullah Ahmed 2019; Scott Kelly 2011). Income, for example, influences purchase decisions while age increases the need for heating or cooling, thereby raising energy consumption ([1] Abrahamse and Steg 2009). An examination of these variables is relevant to our reseach project.

## 2.3 Altering Household Energy Consumption and energy efficiency upgrades

Mental changes and residents' adoption of energy-efficient devices was predicted by their attitude toward the environment and toward energy-efficient devices use, habitual energy-saving behavior, and their perception of the quality of energy-efficient device products, based on a three multiple regression analyses that were conducted ([6] Eunsil Lee, Nam-Kyu Park, Ju Hyoung Han, 2013).

Although recently built houses tend to include more energy saving features and appliances, they often have more appliances in general. In Ireland, [7] O'Doherty et al. (2008) suggest that "an increase of £100,000 in the market value of a home is likely to increase the number of energy-saving features by 3.4%, but is also likely to increase the number of energy-using appliances such that its potential energy use goes up by 5%".

There are many possible supervised machine learning methods that researchers could use to to study the effectiveness of energy efficiency upgrades ([3] Fiona Burlig, et al 2020). In another way the decision tree and neural network models appear to be viable alternatives understanding energy consumption patterns and predicting energy consumption levels, ( [8] Geoffrey K.F.TsoKelvin K.W.Yau 2007).

In order to better understand how consumers choose to adopt energy-efficient artificial lighting for their homes; a survey [5] was administered in four major urban areas: Chicago, Houston, New York, and San Francisco. Major insights from the survey indicate that if lighting becomes less expensive through adopting energy-efficient light sources, there is the potential for consumers to use considerably more. Regional factors such as lighting subsidies, taxation policies, laws, and educational information are also explored.

Another study showed that a Home energy management systems (HEMS) connect homes to a smart grid and may increase the overall use of renewable energy by directing energy demand to off-peak hours and increasing energy conservation [9], reduced the total consumption of electricity in the winter months by up to 30%, shifted the consumption to off-peak hours and decreased the number of high consumption hours. but in the study the correlation between each household's values (comfort, cost, sustainability, etc.) and consumption were only described in a qualitative way. they should be clearly demonstrated using quantitative indicators.

# 3 Methodology

## 3.1 The dataset

The model is based solely on publicly available data and comprises information available from one principle dataset: data from the New York State Energy Research and Development Authority (NYSERDA), hosted by the State of New York. The state has an open data platform found data.ny.gov and they update their information according the amount of data that is brought in.

The New York Residential Existing Homes (One to Four Units) dataset includes the following data points for projects completed during Green Jobs Green-NY, beginning November 15, 2010: Home Performance Project ID, Home Performance Site ID, Project County, Project City, Project Zip, Gas Utility, Electric Utility, Project Completion Date, Customer Type, Low-Rise or Home Performance Indicator, Total ProjectCost (USD), Total Incentives (USD), Type of Program Financing, Amount Financed Through Program (USD), Pre-Retrofit Home Heating Fuel Type, Year Home Built, Size of Home, Volume of Home, Number of Units, Measure Type, Annual kWh Savings, Annual MMBtu Savings, First Year Energy Savings $ Estimate (USD), Homeowner Received Green Jobs-Green NY Free/Reduced Cost Audit (Y/N).

Annual kWh Savings, Estimated Annual MMBtu Savings, and First Year Energy Savings Estimate represent contractor reported savings.

NYSEDRA dataset combined data from three primary sources:

- Program data on measures installed in each home.
- Billing records from the utilities and a full billing analysis of all participants with sufficient billing history
- Participant surveys

The Energy Change Survey: Telephone surveys were completed for a total of 699 participants.The survey was conducted from July 7, 2011 to August 29, 2011. The sample frame consisted of all participating homeowners included in the billing analysis testing (as described above). Thus, the sample frame was developed using the Program and billing data.

Contractor Surveys : The participating contractor survey was designed to obtain information about the timing of equipment replacement. The participating contractor data set was obtained from the HPwES program database.

## 3.2 Data-driven prediction techniques

### 3.2.1 ANOVA Analysis

In this phase we will explore our variables from several points of view to try to establish mathematical relationships between parameters and to predict response accuracy. The selection of the variables will be based on the correlation test. We'll focus on the relationship between First Year Energy saving (FYES) and the following variables: "Electric Utility", "Gas Utility", "Customer Type" (Assisted or Market): the program is designed to offer enhanced assistance to low- to moderate-income households. The "Assisted" component of the HPwES program is available to residents with up to 80 % of area median income, "Total Project Cost", "Total Incentives", "Homeowner Received Green Job (HRGJ)", "Year Home Built", "Measure Type": Eligible measures include building shell measures, such as air

### 3.2.2 Multiple linear regression

Multiple linear regression (MLR) is a statistical technique that allows finding a functional relationship (model or equation) among response Y and predictor variables $X_1, X_2, \dots X_n$. Given the quantitative variable Y, and set of predictor variables $n(X_1 1 \dots\dots X_n)$, the MLR model assumes that the average of Y identifies the values of predictor variables in a linear combination:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{1}$$

where $\beta$ are estimates of the regression coefficients.

Several studies used MLR to predict energy consumption of individual buildings using outdoor environmental conditions and physical characteristics of building . MLR was also used to predict the annual building energy consumption of the banking sector in Spain by Aranda et al.

To develop our model, Stepwise regression will be used to reduce the number of parameters and only include the most effective parameters. Stepwise regression is an automatic method that is useful when the number of explanatory variables is large and it is not possible to fit all the possible models. R statistical analysis program will also be used to determine that a linear model is suitable for our dataset. otherwise we will try use Multilayer neural network technique.

#### 3.2.3 Multilayer neural network

The multilayer neural network is a technique that allows computational models with multiple processing layers to learn representations of data with multiple levels of abstractions [66,67]. A simplified neural network usually has three layers, i.e., one input layer, one hidden layer and one output layer, whereas multilayer neural network (MNN) has a number of hidden layers and each layer can have different activation functions, which performs non-linear transformations and convolution operations.

MNN has been successfully utilized to predict energy consumption of buildings. Rahman et al. also used deep recurrent neural networks (RNN) to predict the electricity consumption for commercial and residential buildings in the U.S. Singaravel et al. We split up our dataset into two parts training (%70) and testing datasets %(30).

### 3.3 Missing Data

Missing data in substantive research is common and can be problematic for structural equation modelling if not handled correctly. A well recognised statistical benchmark suggests datasets with missingness less than 5% is not a problem. Within this dataset we checked if our data variable had less than 5% missing data and therefore not be considered a problem to conduct further statistical tests. Histogram plot shows 14% missing data in volume of Home and Size of Home, and 30% missing data in Number of Units and Year Home Built. The distributions of these variables can therefore truncated. A better strategy would be to impute the missing values. So we Imputed the missing values using predictive mean matching method, that is a widely used statistical imputation method for missing values.

### 3.4 Data Preparation

#### 3.4.1 Data Cleaning

Data preparation and cleaning is a critically important step in any modeling project. So Before jumping to the sophisticated methods, there are some very basic data cleaning operations that was performed. yet are so critical that if skipped, models may break or report overly optimistic performance results. We removed column variables that only have a single or very few value, also unnecessary columns from our datasets (eg. Georeference, County code...etc)

#### 3.4.2 Variable Classification

All models do not understand strings use only numbers, Hence, we need to convert the input data into numeric before working with it. Most variables in our data are categorical string value, so we converted six categorical variable to numeric format either by binary encoding (1 or 0), and dummy variable ordinal encoding data us in some order (see Appendix).

# 4 Results

## 4.1 Descriptive statistics

As we can see from Table.1 the Standard error for small mean values is very small. The skewness coefficient is small for almost all variables, as confirmed by the below plots Histogram in the figure shows most data that are skewed to the right. The few larger values bring the mean upwards but don't really affect the median. So the mean is larger than the median.

Table 1: Variables Summary Statistics (N = 55830)

| | Mean | | Variance | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|
| | Statistic | Std. Error | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Gasutility | 8.16 | 0.022 | 26.602 | -0.399 | 0.01 | -1.433 | 0.021 |
| Elecutility | 5.43 | 0.008 | 3.915 | -0.273 | 0.01 | -0.732 | 0.021 |
| customertype | 1.61 | 0.002 | 0.238 | -0.453 | 0.01 | -1.795 | 0.021 |
| totalprojectcost | 8572.8 | 23.628 | 31169565.735 | 2.063 | 0.01 | 8.209 | 0.021 |
| totalincentive | 1687.31 | 7.219 | 2909413.163 | 0.961 | 0.01 | 0.262 | 0.021 |
| fueltype | 9.04 | 0.005 | 1.393 | -2.159 | 0.01 | 14.648 | 0.021 |
| yearhomebuilt | 1945.78 | 0.16 | 1436.601 | -0.932 | 0.01 | 1.004 | 0.021 |
| sizeofhome | 1909.51 | 2.889 | 466074.787 | 0.772 | 0.01 | 0.12 | 0.021 |
| masuretype | 1.01 | 0 | 0.01 | 10.382 | 0.01 | 110.132 | 0.021 |
| FYES | 596.21 | 2.893 | 467243.075 | 2.395 | 0.01 | 7.425 | 0.021 |
| HRGJ | 1.85 | 0.002 | 0.127 | -1.962 | 0.01 | 1.85 | 0.021 |

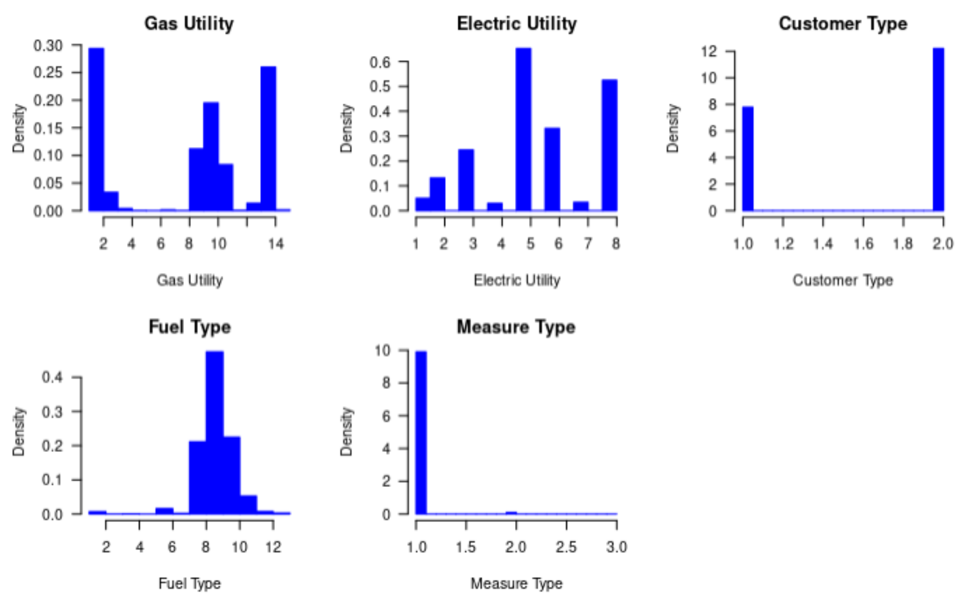



Figure 1: Categorical Variables Distribution

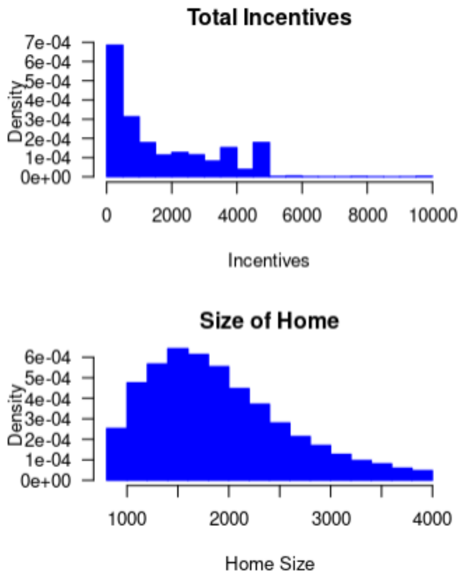



Figure 2: Numerical Variables Distribution

Some other graphical methods, maybe more helpful than the simple histogram. we used the fitdisrplus package in R to visualize the First Year Energy Saving (FYES) variable data together with some possible theoretical distributions in a skewness-kurtosis space:

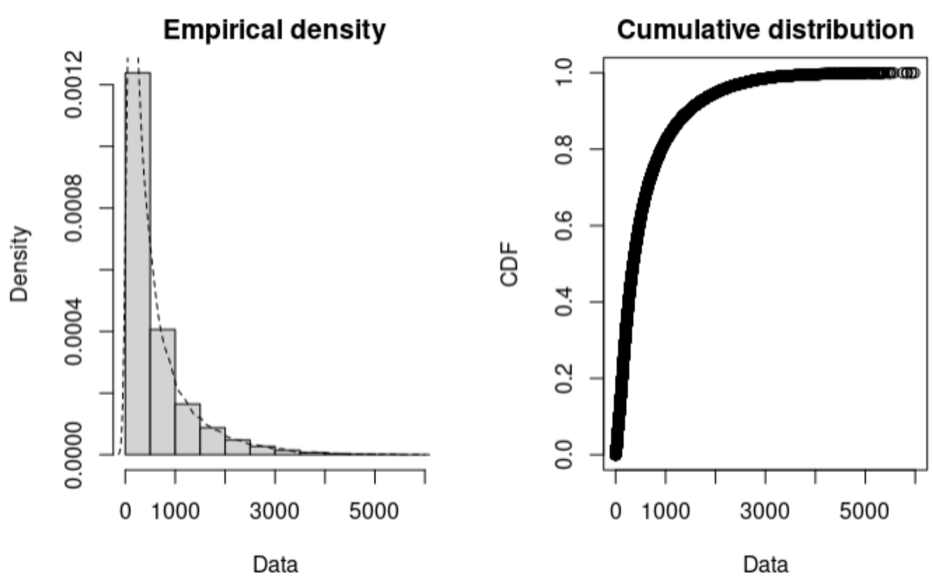



Figure 3: Empirical and theoretical distributions for data

From the empirical density above, our distribution is right skewed and appears to be an exponential type of distribution. The Cullen and Frey Graph below is a good way to exempt some distributions by the parameters of skewness and kurtosis using the descdist R function; The orange values around the blue (data) point are based on bootstrapping. From this Cullen and Frey Graph and the empirical graphs above, our choices for good fits would seem to be limited to the available distributions in the fitdistrplus package: Weibull, gamma, exponential

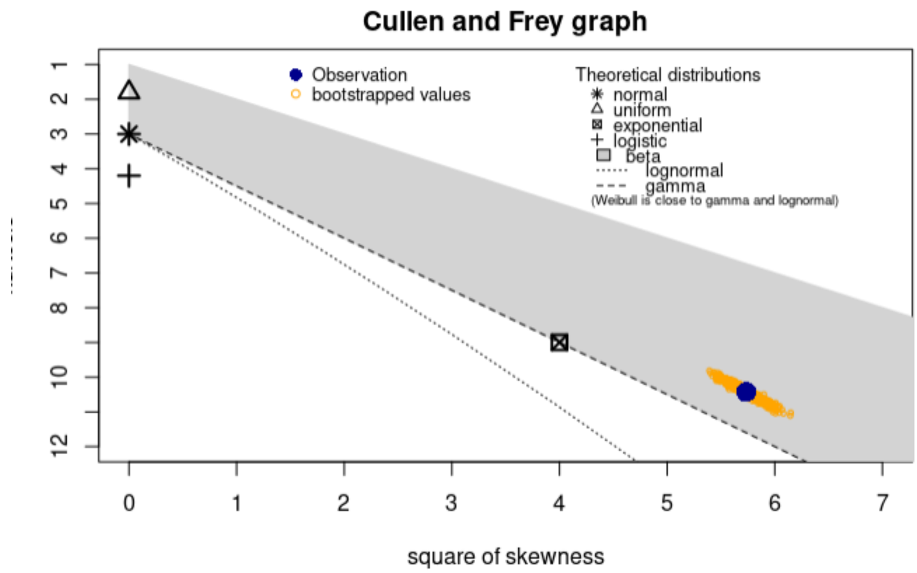



Figure 4: Cullen and Fry Graph

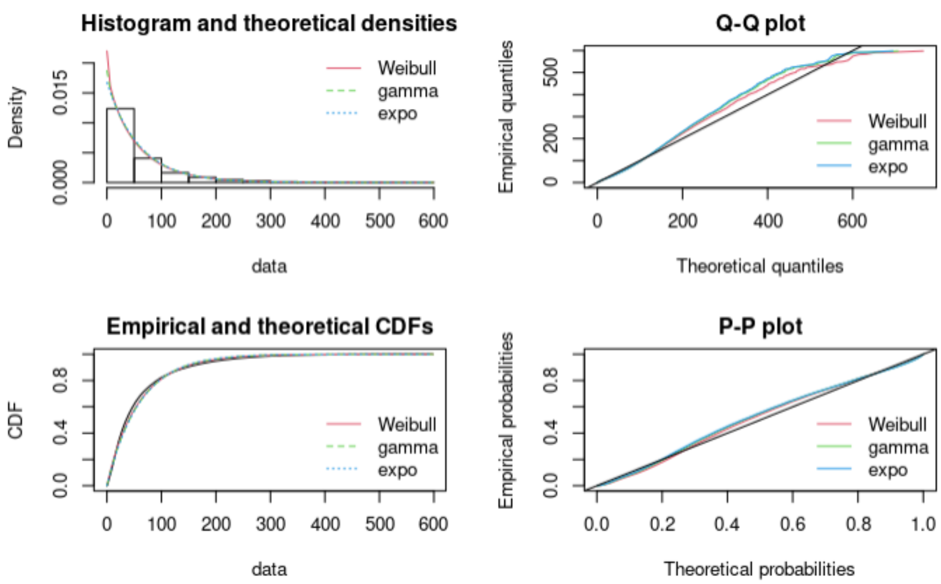



Figure 5: Distributions Plots

It seems that still both distribution fits this data the best. Let us confirm this against the Akaline's and Bayesian Information Criterion (AIC and BIC), which will give a sort of rank to the goodness of fit models as well as the Goodness-of-fit statistics, which give distances between the fits and the empirical data.

Table 2: Goodness-of-fit-statistics

| Goodness-of-fit-statistics | | | |
|---|---|---|---|
| | 1-mle-weibull | 2-mle-gamma | 3-mle-exp |
| Kolmogorov-Smirnov-statistic | 0.042 | 0.051 | 0.055 |
| Cramer-von-Mises-statistic | 36.676 | 53.192 | 60.093 |
| Anderson-Darling-statistic | 224.160 | 281.895 | 307.472 |
| | | | |
| Goodness-of-fit-criteria | | | |
| | 1-mle-weibull | 2-mle-gamma | 3-mle-exp |
| Akaike's-Information-Criterion | 567029.00 | 567244.30 | 567262.60 |
| Bayesian-Information-Criterion | 567046.90 | 567262.10 | 567271.50 |

Since the weibull distribution has the min AIC, BIC, and minimum goodness-of-fit statistics, we will choose the weibull distribution.

## 4.2 Model development

Two basic steps occurred prior to model development: (1) an analysis of the relationship between Energy Saving and the explanatory factors including home characteristics, project types, Incentives and programs that promote the installation of clean technologies. we will check for correlation between the explanatory variable (2) a determination of the variation percentage, which can be interpreted by explanatory factors using generalized linear model (GLM) analysis, and multilayer neural network (MNN). The entire dataset was divided into training, validation and testing data with proportions of 60 %, 20 % and 20 % respectively. To predict energy consumption, development and training processes were carried out using RStudio package software and SPSS.

### 4.2.1 Correlation Analysis

First we try to put in as many features as we think can influence the output and progressively remove the least useful with the analysis. A Pearson correlation analysis was performed between all variables at first (see Table in Appendices), after that six variables were selected. We added significance test to the correlogram we shall compute the p-value. If the p-value is greater than 0.01 then it is an insignificant value for which the cells are either blank or crossed (Fig.6 and Fig.7).

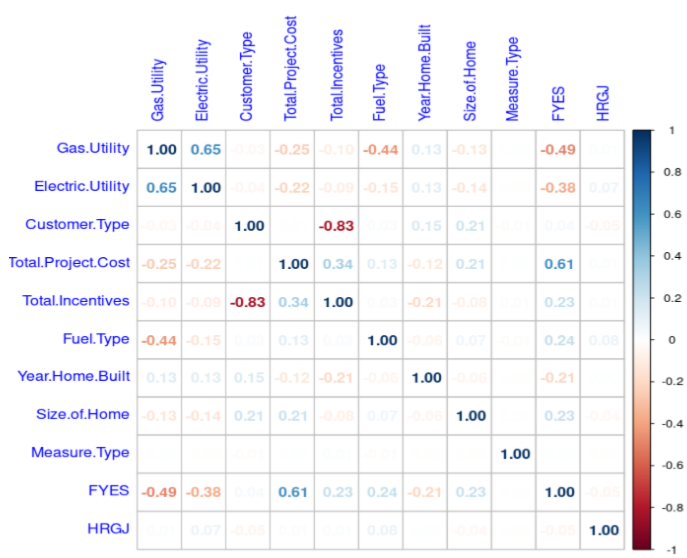



Figure 6: Full Correlation Matrix

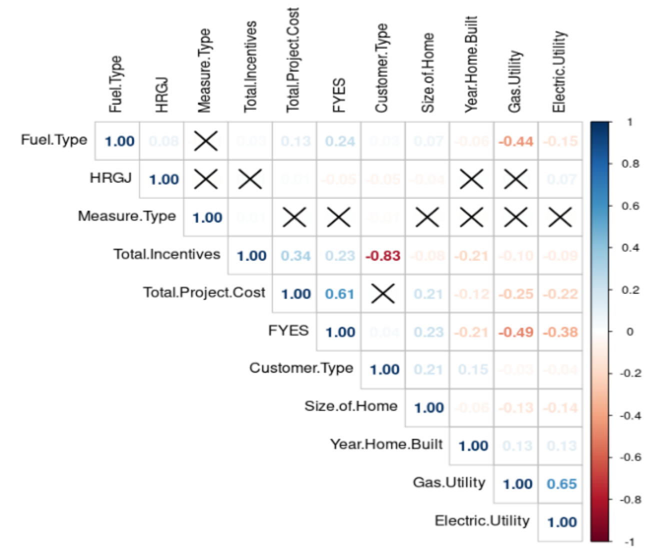



Figure 7: Upper triangular of the correlation matrix with Insignificant P-value

From the correlogram (Fig.6) we can see that our dependent variable (Energy Saving FYES) shows the best correlation with the "Total Project Cost" (0.61) and negatively high correlated to the "Gas Utility". The negative sign can be explained by having a larger number of Gas Utility categories(from 1..15) and the top one has more effect on Energy Saving (1: "Central Hudson Gas & Electric", 2: "Consolidated Edison", 3: "Corning Natural Gas", 4: "KeySpan Energy"...etc). "Electric Utility" shows a good correlation (-0.38), follow by Fuel Type (0.24), "Total Incentive" and "Size of Home" (0.23), Then "Year Home Built" which is the lowest. So it will be removed with "Measure Type", "Customer Type" and "Home Owner Received Green Job" (HRGJ) variables.

### 4.2.2 ANOVA Analysis

In addition to the correlation an ANOVA analysis has been performed to know the most significant factor effecting the Energy saving.

```
                      Df    Sum Sq  Mean Sq  F value   Pr(>F)    
Electric.Utility       1 3.778e+09 3.778e+09 16375.45  < 2e-16 ***
Gas.Utility            1 2.675e+09 2.675e+09 11595.86  < 2e-16 ***
Total.Project.Cost     1 6.609e+09 6.609e+09 28644.67  < 2e-16 ***
Total.Incentives       1 8.369e+06 8.369e+06    36.28 1.72e-09 ***
Size.of.Home           1 1.356e+08 1.356e+08   587.62  < 2e-16 ***
Residuals          55824 1.288e+10 2.307e+05                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8: ANOVA Analysis Summary

As we can see from the above table (Fig.8) all P values are almost zero and support our explanation of these variables effect on saving.

### 4.2.3 Multiple Regression

Since we are dealing with more response variables, the regression multivariate regression. For a multiple linear regression model of the form of Eq.(1),the Energy Saving is the response variable and "Gas Utility", "Electric Utility", ""Project Cost", "Total Incentives", "Fuel Type" and "Size of Home" are the predictor variables.

```
Call:
lm(formula = eng_imputed$FYES ~ eng_imputed$Size.of.Home + eng_imputed$Total.Incentives +
    Gas.Utility, data = eng_imputed, subset = Electric.Utility +
    Fuel.Type, weights = Total.Project.Cost)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-105513  -51072   -3101   17456   84658

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.309e+03  1.229e+01  -106.5   <2e-16 ***
eng_imputed$Size.of.Home     9.723e-01  4.655e-03   208.9   <2e-16 ***
eng_imputed$Total.Incentives -2.826e-02  2.125e-03  -13.3   <2e-16 ***
Gas.Utility                  2.842e+01  1.128e+00    25.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40940 on 55826 degrees of freedom
Multiple R-squared:  0.6168,    Adjusted R-squared:  0.6168
F-statistic: 2.996e+04 on 3 and 55826 DF,  p-value: < 2.2e-16
```

Figure 9: R MLR Output

Looking at the Residuals output above, it looks like our distribution is not quite symmetrical and is slightly right-skewed. This tells us that our model is not predicting as well at the higher Energy saving ranges as it does for the low ranges. Also the all p-value coefficient are significant. we can see that the p-values for the Intercept and points are extremely small. This leads us to conclude that there is strong evidence that the coefficients in this model are not zero. The Adjusted R-square is 0.61, but It's important to note that the R-square value (Multiple or Adjusted) is not fool-proof and shouldn't necessarily be used alone just by virtue of how the value is calculated.

#### 4.2.4 Multilayer Neural Network

The energy saving profile of a residential homes is inherently transient and non-linear in nature, so in this section we aims to develop reliable and accurate models from multilayer neural network to identify the primary factors for adoption of energy-efficient program and to predict energy saving for residential buildings. Both R and SPSS were used to develop the MNN equations.

Table 3: MNN Case Processing Summary

| | N | Percent |
|---|---|---|
| Training | 40040 | 71.70% |
| Testing | 9810 | 17.60% |
| Holdout | 5980 | 10.70% |
| Valid | 55830 | 100.00% |
| Excluded | 1 | |
| Total | 55831 | |

Table.3 shows our training, Testing and Holdout selection for our MNN process, and also shows the software better selection in the Percent column.

**Parameter Estimates**

| Predictor | | Predicted | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Hidden Layer 1 | | | | | | | Output Layer |
| | | H(1:1) | H(1:2) | H(1:3) | H(1:4) | H(1:5) | H(1:6) | H(1:7) | FYES |
| Input Layer | (Bias) | -.260 | 3.200 | 1.799 | 1.049 | .566 | -.005 | -.049 | |
| | Totalprojectcost | .254 | -.542 | .193 | .208 | .714 | .109 | -.235 | |
| | Totalincentive | .068 | -.321 | -.074 | -.319 | -.149 | -.392 | -.168 | |
| | sizeofhome | .166 | -.149 | -.162 | -.143 | -.228 | .030 | .281 | |
| | Gasutility | .298 | 1.019 | 2.070 | .321 | .112 | -.331 | .072 | |
| | Electricutility | -1.301 | .199 | -.824 | -.313 | .534 | -.209 | -.286 | |
| | Fueltype | -.337 | -.073 | -1.323 | .965 | .220 | -.009 | -.346 | |
| Hidden Layer 1 | (Bias) | | | | | | | | 2.239 |
| | H(1:1) | | | | | | | | .643 |
| | H(1:2) | | | | | | | | -2.000 |
| | H(1:3) | | | | | | | | -.607 |
| | H(1:4) | | | | | | | | -.405 |
| | H(1:5) | | | | | | | | .821 |
| | H(1:6) | | | | | | | | -.189 |
| | H(1:7) | | | | | | | | .347 |

Figure 10: Neural Network Parameter Estimates

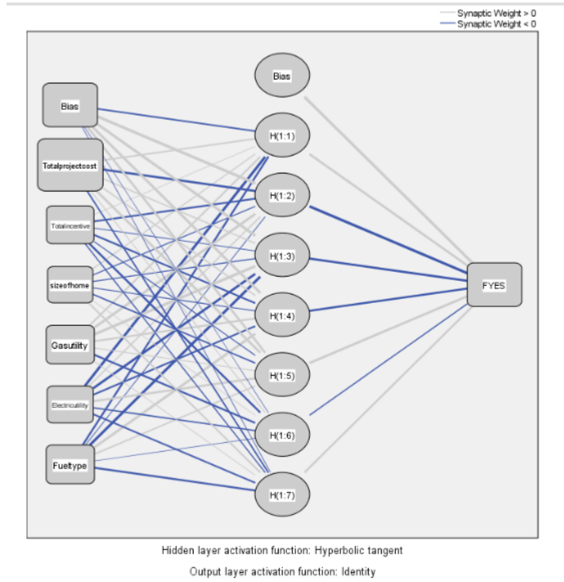



Figure 11: Neural Network Diagram

Fig.10 shows the MNN parameters, the first raw is the Bias vector $V0Bias$ of the hidden layer. Fig. 11 shows the MNN graphical representation, we have 6 input variables and 7 hidden layers. The blue lines are the positive weights. Also from the diagram we can see that the Total Project Cost has a bigger box, which means it is the most important independent variable contributor to the Energy saving.

Table 4: Independent Variable Importance

| | Importance | Normalized Importance |
|---|---|---|
| Total Proje Cost | 0.561 | 100.00% |
| Total Incentive | 0.095 | 17.00% |
| Size of Home | 0.044 | 7.80% |
| Gas Utility | 0.116 | 20.70% |
| Electric Utility | 0.065 | 11.70% |
| Fuel Type | 0.119 | 21.10% |

Table 4 show the Independent variable importance and again it confirm that The Total Project Cost is the most important follow by Fuel Type, Gas Utility then Total Incentive. So investing more money in energy saving project increase the saving.

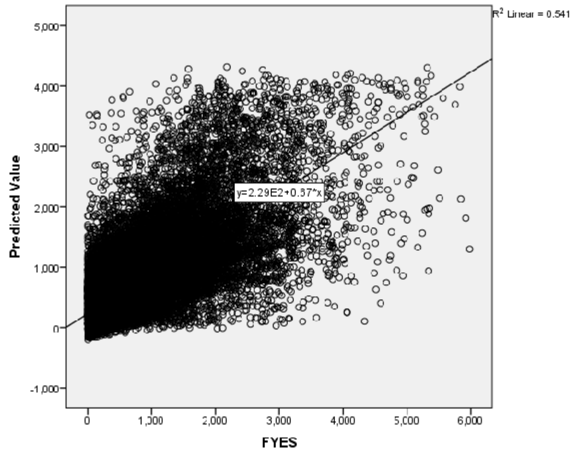



Figure 12: Predictive values of First Year Energy Saving (FYES)

Fig.12 shows the Predictive values of First Year Energy Saving (FYES) vs FYES. The standard error is 0.541, when we have a huge number of data it is difficult to predict high determination coefficient.

$$V_{0bias(7x1)} = \begin{bmatrix} -.26 \\ 3.2 \\ 1.8 \\ 1 \\ 0.57 \\ -0.005 \\ -0.049 \end{bmatrix} \text{ and } W_{bias(1x1)} = \begin{bmatrix} 2.24 \end{bmatrix} \text{ and } W_{6x1} = \begin{bmatrix} .643 \\ -2 \\ -.607 \\ -.405 \\ .821 \\ -.189 \end{bmatrix} \tag{2}$$

$$V_{(6x7)} = \begin{bmatrix} .254 & -.542 & .193 & .208 & .714 & .109 & -.235 \\ .068 & -.321 & -.074 & -.319 & -.149 & -.392 & -.168 \\ .166 & -.149 & -.162 & -.143 & -.228 & .030 & .281 \\ .298 & 1.019 & 2.070 & .321 & .112 & -.331 & .072 \\ -1.3 & 0.2 & -.824 & -.313 & .534 & -.209 & -.286 \\ -.337 & -.073 & -1.323 & .965 & .220 & -.009 & -.346 \end{bmatrix} \tag{3}$$

where: $V0bias$ - Biases, X — input value, $W0bias$ — Output Bias, V — weight between input and hidden layer, the active function is $tanh$.

Using the above terms we constructed mathematical formula Mathematically this process could be represented in terms of matrix multiplication and final equation is the following :

$$FYES = [W_{bias(1x1)} + (W_{6x1})^T + \tanh[V_{0bias(7x1)} + (V_{(6x7)})^T * W_{6x1})] * Std_{FYES} + \overline{X}_{FYES}] \tag{4}$$

Model validation is perhaps the most important step in the model building sequence. The validation process include testing whether the model's predictive performance depreciates substantially when applied to data that were not used in model estimation. The data that have not been used to build the model was utilized for validation.

# 5 Conclusion

This work indicates that MNN models have significant potential for use in predicting efficient-home Energy Saving. the model proposed that Energy efficiency is known to be a cost-effective investment when projects cost are higher, but the actual amount of savings differs based on the upgrade in question, also Fuel switching type, and Gas Utility. Although MNN can help to show trends in total energy saving patterns, they cannot explain the various components that contribute to energy saving at the household level. For this type of analysis it is necessary to conduct investigations at the micro-level. Bottom-up methods take a disaggregated approach and estimate energy demand and emissions using high resolution data using a combination of physical, social, behavioural and demographic properties for a household.

# References


[1] Wokje Abrahamse and Linda Steg. How do socio-demographic and psychological factors relate to households' direct and indirect energy use and savings? *Journal of Economic Psychology*, 30(5):711–720, October 2009.

[2] Abdo Abdullah Ahmed Gassar, Geun Young Yun, and Sumin Kim. Data-driven approach to prediction of residential energy consumption at urban scales in London. *Energy*, 187:115973, November 2019.

[3] Fiona Burlig, Christopher Knittel, David Rapson, UC Davis, Mar Reguant, and Catherine Wolfram. Machine Learning from Schools about Energy Efficiency. page 45.

[4] Birgitta Gatersleben, Linda Steg, and Charles Vlek. Measurement and Determinants of Environmentally Significant Consumer Behavior. *Environment and Behavior*, 34(3):335–362, May 2002. Publisher: SAGE Publications Inc.

[5] Andrea L. Hicks and Thomas L. Theis. Residential energy-efficient lighting adoption survey. *Energy Efficiency*, 7(2):323–333, April 2014.

[6] Eunsil Lee, Nam-Kyu Park, and Ju Hyoung Han. Factors Affecting Environmentally Responsible Behaviors in the Use of Energy-efficient Lighting in the Home. *Family and Consumer Sciences Research Journal*, 41(4):413–425, June 2013. Publisher: John Wiley & Sons, Ltd.

[7] Joe O'Doherty, Sean Lyons, and Richard S.J. Tol. Energy-using appliances and energy-saving features: Determinants of ownership in Ireland. *Applied Energy*, 85(7):650–662, July 2008.

[8] Geoffrey K. F. Tso and Kelvin K. W. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761–1768, September 2007.

[9] Sanna Tuomela, Mauricio de Castro Tomé, Netta Iivari, and Rauli Svento. Impacts of home energy management systems on electricity consumption. *Applied Energy*, 299:117310, October 2021.

# Appendices

## A categorical Variable Coding

**Gas Utility**

- 1: NA
- 2: "Central Hudson Gas & Electric"
- 3: "Consolidated Edison"
- 4: "Corning Natural Gas"
- 5: "KeySpan Energy"
- 6: "Long Island Power Authority"
- 7: "Multiple Gas Providers"
- 8: "Municipal"
- 9: "National Fuel Gas Distribution"
- 10: "National Grid"
- 11: "New York State Electric & Gas"
- 12: "No Gas Provider"
- 13: "Orange & Rockland"
- 14: "Rochester Gas & Electric"
- 15: "Saint Lawrence Gas"

**Electric Utility**

- 1: "Central Hudson Gas & Electric"
- 2: "Consolidated Edison"
- 3: "Long Island Power Authority"
- 4: "Municipal (Not Qualified)"
- 5: "National Grid"
- 6: "New York State Electric & Gas"
- 7: "Orange & Rockland"
- 8: "Rochester Gas & Electric"

**Measure Type**

- 1: "Building Shell"
- 2: "Heating and Cooling"
- 3: "Water Heater"

**Home owner received green Job (HRGJ)**

- 2: "Y"
- 1: "N"

**Fuel Type**

- 1: NA
- 2: "Anthracite Coal"
- 3: "Bituminous Coal"
- 4: "Coal"
- 6: "Electricity"
- 7: "Kerosene"
- 9: "Natural Gas"
- 10: "Oil"
- 11: "Propane"
- 12: "Wood"
- 13: "Wood Pellets"

## SPSS Correlations Results

| | | Gasutility | Electricutility | customertype | Totalprojectcost | Totalincentive | Fueltype | yearhomebuilt | sizeofhome | volumeofhome | noofunits | measuretype | FYES |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pearson Correlation | 1 | .647** | -.031** | -.250** | -.098** | -.441** | .118** | -.134** | -.079** | .021** | .003 | -.490** |
| Gasutility | Sig. (2-tailed) | | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .467 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | .647** | 1 | -.042** | -.221** | -.093** | -.149** | .114** | -.150** | -.077** | -.017** | -.001 | -.381** |
| Electricutility | Sig. (2-tailed) | .000 | | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .755 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | -.031** | -.042** | 1 | .006 | -.828** | .026** | .156** | .200** | .128** | -.076** | -.012** | .039** |
| customertype | Sig. (2-tailed) | .000 | .000 | | .190 | .000 | .000 | .000 | .000 | .000 | .000 | .004 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | -.250** | -.221** | .006 | 1 | .339** | .129** | -.119** | .212** | .187** | .034** | .005 | .615** |
| Totalprojectcost | Sig. (2-tailed) | .000 | .000 | .190 | | .000 | .000 | .000 | .000 | .000 | .000 | .267 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | -.098** | -.093** | -.828** | .339** | 1 | .027** | -.205** | -.075** | -.052** | .102** | .013** | .230** |
| Totalincentive | Sig. (2-tailed) | .000 | .000 | .000 | .000 | | .000 | .000 | .000 | .000 | .000 | .002 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | -.441** | -.149** | .026** | .129** | .027** | 1 | -.049** | .070** | .084** | -.022** | -.011* | .239** |
| Fueltype | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | | .000 | .000 | .000 | .000 | .012 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | .118** | .114** | .156** | -.119** | -.205** | -.049** | 1 | -.050** | -.054** | -.054** | -.001 | -.201** |
| yearhomebuilt | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | | .000 | .000 | .000 | .877 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | -.134** | -.150** | .200** | .212** | -.075** | .070** | -.050** | 1 | .853** | .085** | .001 | .231** |
| sizeofhome | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | .000 | | .000 | .000 | .899 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | -.079** | -.077** | .128** | .187** | -.052** | .084** | -.054** | .853** | 1 | .079** | -.003 | .143** |
| volumeofhome | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | | .000 | .496 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | .021** | -.017** | -.076** | .034** | .102** | -.022** | -.054** | .085** | .079** | 1 | -.001 | .021** |
| noofunits | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | | .882 | .000 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | .003 | -.001 | -.012** | .005 | .013** | -.011* | -.001 | .001 | -.003 | -.001 | 1 | .004 |
| measuretype | Sig. (2-tailed) | .467 | .755 | .004 | .267 | .002 | .012 | .877 | .899 | .496 | .882 | | .397 |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |
| | Pearson Correlation | -.490** | -.381** | .039** | .615** | .230** | .239** | -.201** | .231** | .143** | .021** | .004 | 1 |
| FYES | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .397 | |
| | N | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 | 55830 |

**. Correlation is significant at the 0.01 level (2-tailed).

**Descriptive Statistics**

| | N | Mean | | Variance | Skewness | | Kurtosis | |
|---|---|---|---|---|---|---|---|---|
| | Statistic | Statistic | Std. Error | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Gasutility | 55830 | 8.16 | .022 | 26.602 | -.399 | .010 | -1.433 | .021 |
| Elecutility | 55830 | 5.43 | .008 | 3.915 | -.273 | .010 | -.732 | .021 |
| customertype | 55830 | 1.61 | .002 | .238 | -.453 | .010 | -1.795 | .021 |
| totalprojectcost | 55830 | 8572.80 | 23.628 | 31169565.735 | 2.063 | .010 | 8.209 | .021 |
| totalincentive | 55830 | 1687.31 | 7.219 | 2909413.163 | .961 | .010 | .262 | .021 |
| fueltype | 55830 | 9.04 | .005 | 1.393 | -2.159 | .010 | 14.648 | .021 |
| yearhomebuilt | 55830 | 1945.78 | .160 | 1436.601 | -.932 | .010 | 1.004 | .021 |
| sizeofhome | 55830 | 1909.51 | 2.889 | 466074.787 | .772 | .010 | .120 | .021 |
| masuretype | 55830 | 1.01 | .000 | .010 | 10.382 | .010 | 110.132 | .021 |
| FYES | 55830 | 596.21 | 2.893 | 467243.075 | 2.395 | .010 | 7.425 | .021 |
| HRGJ | 55830 | 1.85 | .002 | .127 | -1.962 | .010 | 1.850 | .021 |
| Valid N (listwise) | 55830 | | | | | | | |

Linear regression

**Variables Entered/Removed**[a]

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | Gasutility, totalincentive, sizeofhome, fueltype, totalprojectcost, Elecutility[b] | . | Enter |

a. Dependent Variable: FYES

b. All requested variables entered.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .712[a] | .507 | .507 | 480.149 |

a. Predictors: (Constant), Gasutility, totalincentive, sizeofhome, fueltype, totalprojectcost, Elecutility

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 13216120488.804 | 6 | 2202686748.134 | 9554.349 | .000[b] |
| | Residual | 12869593153.541 | 55823 | 230542.844 | | |
| | Total | 26085713642.345 | 55829 | | | |

a. Dependent Variable: FYES

b. Predictors: (Constant), Gasutility, totalincentive, sizeofhome, fueltype, totalprojectcost, Elecutility

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 212.790 | 20.551 | | 10.354 | .000 |
| | Elecutility | -19.250 | 1.383 | -.056 | -13.924 | .000 |
| | totalprojectcost | .061 | .000 | .496 | 148.515 | .000 |
| | totalincentive | .013 | .001 | .032 | 10.064 | .000 |
| | fueltype | 14.958 | 1.960 | .026 | 7.632 | .000 |
| | sizeofhome | .074 | .003 | .074 | 23.785 | .000 |
| | Gasutility | -40.563 | .585 | -.306 | -69.381 | .000 |

a. Dependent Variable: FYES

# Multi Neural Network SPSS results

**Case Processing Summary**

| | | N | Percent |
|---|---|---|---|
| Sample | Training | 40040 | 71.7% |
| | Testing | 9810 | 17.6% |
| | Holdout | 5980 | 10.7% |
| Valid | | 55830 | 100.0% |
| Excluded | | 1 | |
| Total | | 55831 | |

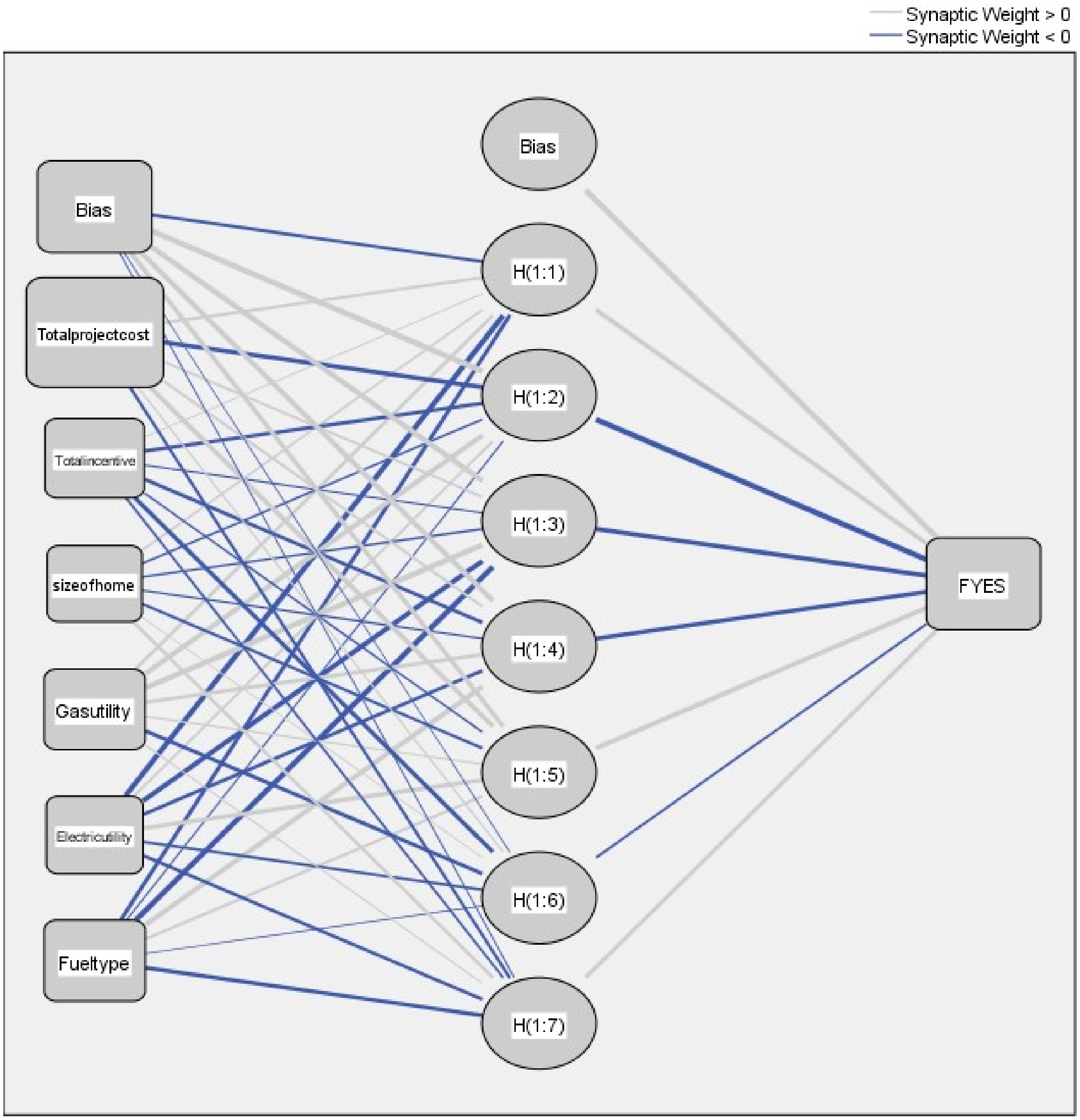
the software modiy the selection for the number of cases of training testing and holdout, better selection

**Network Information**

| | | | |
|---|---|---|---|
| Input Layer | Covariates | 1 | Totalprojectcost |
| | | 2 | Totalincentive |
| | | 3 | sizeofhome |
| | | 4 | Gasutility |
| | | 5 | Electricutility |
| | | 6 | Fueltype |
| | Number of Units[a] | | 6 |
| | Rescaling Method for Covariates | | Standardized |
| Hidden Layer(s) | Number of Hidden Layers | | 1 |
| | Number of Units in Hidden Layer 1[a] | | 7 |
| | Activation Function | | Hyperbolic tangent |
| Output Layer | Dependent Variables | 1 | FYES |
| | Number of Units | | 1 |
| | Rescaling Method for Scale Dependents | | Standardized |
| | Activation Function | | Identity |
| | Error Function | | Sum of Squares |

a. Excluding the bias unit

NI imput layer : vector input X* = (6,1)




Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Identity

**Model Summary**

| | | |
|---|---|---|
| Training | Sum of Squares Error | 9883.648 |
| | Relative Error | .494 |
| | Stopping Rule Used | 1 consecutive step(s) with no decrease in error[a] |
| | Training Time | 0:00:00.29 |
| Testing | Sum of Squares Error | 2396.854 |
| | Relative Error | .479 |
| Holdout | Relative Error | .495 |

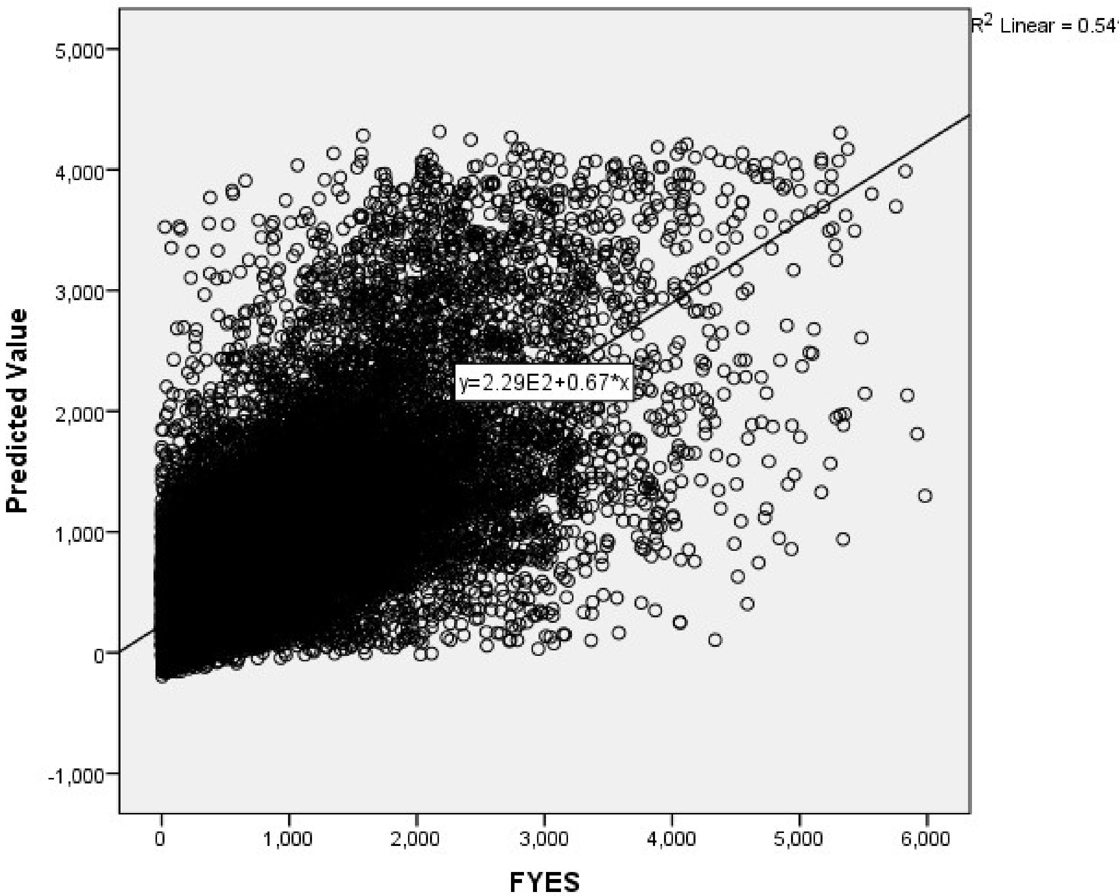
Dependent Variable: FYES

a. Error computations are based on the testing sample.

**Parameter Estimates**

| Predictor | | Predicted | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Hidden Layer 1 | | | | | | | Output Layer |
| | | H(1:1) | H(1:2) | H(1:3) | H(1:4) | H(1:5) | H(1:6) | H(1:7) | FYES |
| Input Layer | (Bias) | -.260 | 3.200 | 1.799 | 1.049 | .566 | -.005 | -.049 | |
| | Totalprojectcost | .254 | -.542 | .193 | .208 | .714 | .109 | -.235 | |
| | Totalincentive | .068 | -.321 | -.074 | -.319 | -.149 | -.392 | -.168 | |
| | sizeofhome | .166 | -.149 | -.162 | -.143 | -.228 | .030 | .281 | |
| | Gasutility | .298 | 1.019 | 2.070 | .321 | .112 | -.331 | .072 | |
| | Electricutility | -1.301 | .199 | -.824 | -.313 | .534 | -.209 | -.286 | |
| | Fueltype | -.337 | -.073 | -1.323 | .965 | .220 | -.009 | -.346 | |
| Hidden Layer 1 | (Bias) | | | | | | | | 2.239 |
| | H(1:1) | | | | | | | | .643 |
| | H(1:2) | | | | | | | | -2.000 |
| | H(1:3) | | | | | | | | -.607 |
| | H(1:4) | | | | | | | | -.405 |
| | H(1:5) | | | | | | | | .821 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | H(1:6) | | | | | | | | -.189 |
| | H(1:7) | | | | | | | | .347 |




**Independent Variable Importance**

| | Importance | Normalized Importance |
|---|---|---|
| Total Proje Cost | .561 | 100.0% |
| Total Incentive | .095 | 17.0% |
| Size of Home | .044 | 7.8% |
| Gas Utility | .116 | 20.7% |
| Electric Utility | .065 | 11.7% |

| | | |
|---|---|---|
| Fuel Type | .119 | 21.1% |