Multivariable Regression Modelling for Residential Energy Consumption of United States

Avani Goyal Nathaniel Dirks

ABSTRACT

Building energy datasets and the associated analysis are becoming increasingly important. Having access to large scale datasets enables analysts to discover trends and make conclusions which otherwise may not have occurred. The Residential Energy Consumption Survey (RECS) administered by the United States Energy Information Administration is one such dataset. This dataset captures hundreds of inputs from thousands of residences throughout the United States and presents valuable insights on key predictors of energy consumption. Statistical techniques such as regression modeling are used to analyze this dataset to enable prediction of energy consumption values for a selected regional division. Three top predictor variables are identified and used as independent variables using correlation values. This is validated using cross-validation technique to obtain a more consistent and accurate predictive model.

1. INTRODUCTION

The most recent RECS dataset contains survey responses from the year 2009 and was released in early 2013. Historically this dataset has been collected every four years or so going back to 1978. In this project, data analysis is performed using the 2009 RECS survey data. Survey results include information about residential buildings such as number of windows, square footage, whether or not a garage is heated, etc. Also, data such as total energy consumed, cost, and more specific breakdowns of those are included. Exploring trends among these variables, insights can be gained into the best predictors of total energy use.[1]

2. OVERVIEW

In this project, RECS dataset is analyzed using statistical techniques like exploratory data analysis and multi-variable regression modeling to enable prediction for total energy consumption. Initial diagnostics is done to understand the shape, data types, and content within the dataset. Data from two census divisions are compared to determine whether there are significant regional differences. If regional differences are present, the further modeling is done using data for a specific region. The best variables for predicting total energy consumption are selected using correlation technique. In choosing the variables for correlation, manual filtering is applied to ensure the correct variables are selected. A multivariable-regression model is implemented using the top predictor variables. The dataset is cross-validated by averaging results of multiple different train and test data splits.

3. MULTIVARIABLE REGRESSION MODEL

3.1 Preliminary Analysis

The dataset has a shape of 12083 rows by 931 columns. The first row is the header and contains the title for each column and is not included in the row count. Each row represents a survey response from an individual residential customer. Each column represents a survey input value for things such as number of refrigerators or insulation material as well as cost and energy use. Next, after getting a sense for the contents of the dataset was to look at the data on a regional level. The third column in the dataset designates the census division for the residence. To validate whether or not there are significant regional differences we chose two divisions, see Figure 1 below, and compared the total energy consumed across major categories. The first chosen is division 2, Middle Atlantic, which includes PA, NY, and NJ, being in the northeast and having four distinct seasons and colder winters. In contrast, we also chose division 7, West South Central, containing AR, LA, OK, and TX. These states have much warmer weather overall and milder winters than the northeast.

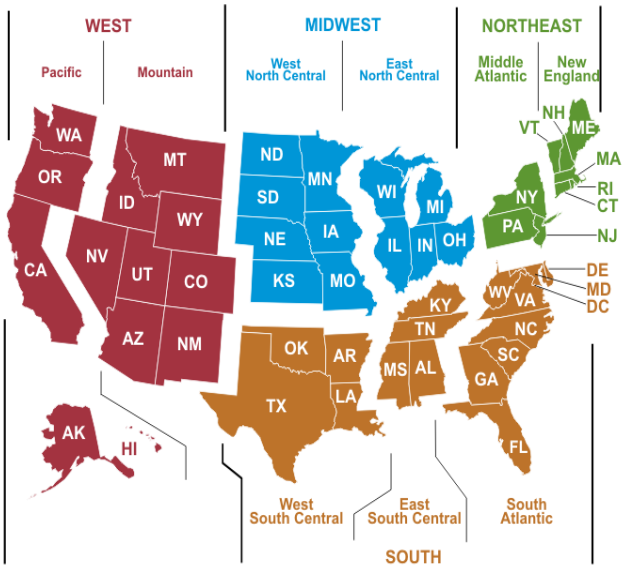

Figure 1. Census Region and Division Map.[2]

For each survey participant, yearly energy consumption in Btu's for categories such as total, space heating, air conditioning, water heating, refrigeration, and other is included. After separating the two regions, the mean of values within each of the six categories was taken. This was done for each region and plotted as seen below in Figure 2. Overall, looking at the average energy use by region, it can be seen that the Mid Atlantic region uses approximately 38% more energy than West South Central. This major increase can be attributed to the fact that Mid Atlantic uses three times as much energy for space heating compared to West South Central. Alternatively, West South Central uses more than six times the energy for air conditioning than Mid Atlantic. These results confirm the regional difference and agree with what would be expected based on where the energy would be used.

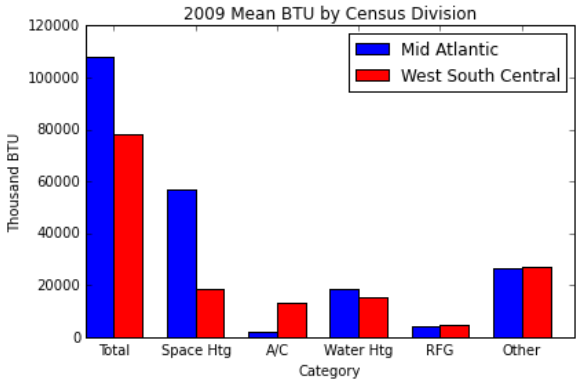

Figure 2. Regional Energy Use by Category.

Furthermore, it would be helpful to understand the distribution of total energy consumption for Mid Atlantic division. By plotting a histogram of these values, see Figure 3 below, it can be seen the values approximate a normal distribution. Also, a long right tail shows there are quite a few outliers above 300,000.

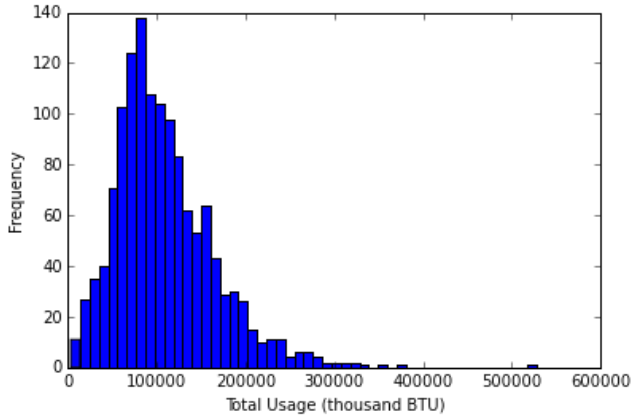

Figure 3. Histogram of Mid-Atlantic Total Energy Usage.

To determine whether or not a piecewise linear model is necessary, the total energy usage values for Mid-Atlantic region are plotted and fit with a linear least square fit line, see Figure 4 below. The results show a strong linear trend suggesting no piecewise linear model is required. Also, the average energy usage for Mid-Atlantic is nearly 105,000.

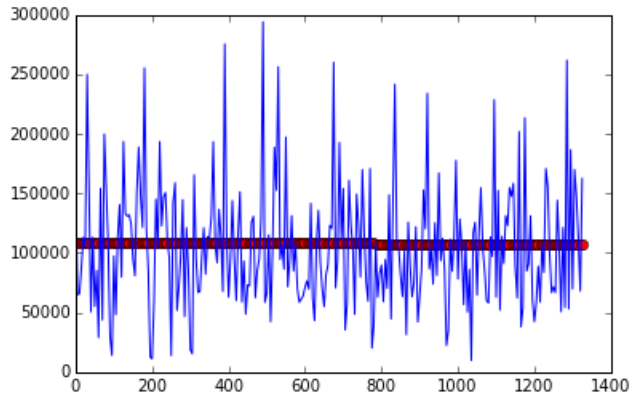

Figure 4. Linear Least Square Fit.

Next, the process of determining which values correlate best with total usage in order to start building the multi-variable regression model. Initially this was done manually, by opening the public layout file and looking through the descriptions of each variable. Multiple variables were selected and tested for correlation against total usage using the numpy.corrcoef function. The top three variables for correlation from the manual selection ended up being total rooms (0.49), total square footage (0.54), and number of windows (0.55). To validate these results and check for other variables we may have missed, we developed a function to iterate through all columns and return the correlation along with index value. We filtered the results for only those with correlation above 0.47. These results ended up showing certain values with high correlation but those are from cost, power, and energy variables and that would be expected. The only other values that made the cut off ended up being the ones we manually selected thus validating the original variable choices, see Figure 5 below for top three variables plotted against total usage.

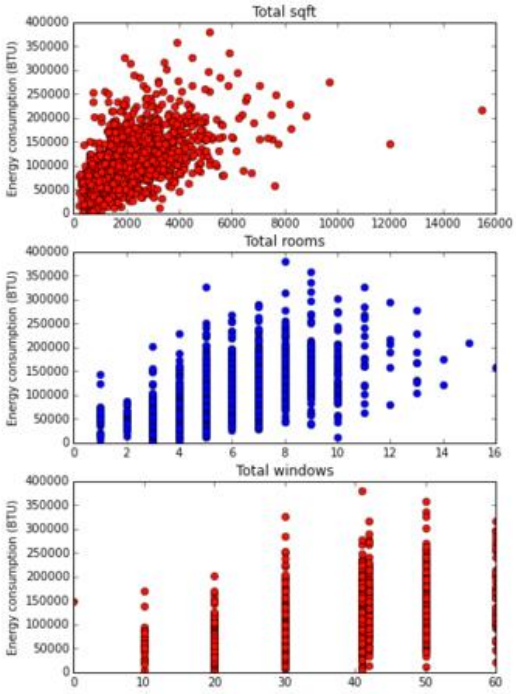

Figure 5. Best Correlation Variables.

3.2 Regression Model Method

The model for a generic multivariable regression modeling follows equation (1) below.

$$y_i = \sum_{i=1}^{n} \beta_i x_i \quad (1)$$

With this fairly simple model, the design matrix (X) is sized NxP and contains the selected variables and the beta coefficient vector is the size of the number of variables selected, in this case three. Next the dataset must be split into two parts, called train and test. The train portion is used to create the predicted values and the test data is used to validate how well the prediction models the actual values. How well the model performs is calculated from the coefficient of determination, otherwise known as R^2 as seen in equation 2 below.

$$R^2 = 1 - \frac{(Y - \mathbf{X}\hat{\beta})^T (Y - \mathbf{X}\hat{\beta})}{(Y - \bar{y})^T (Y - \bar{y})} \quad (2)$$

As mentioned previously from the histogram, the distribution has outliers. These outliers can affect the regression accuracy. A maximization function was developed to perform the R^2 calculation on the specific variables in the dataset but limit the total usage value and thus removing outliers above that level. For each instance, R^2 was calculated and plotted with the associated cutoff value. The best R^2 value was found to be 0.38 with a cut off of 310,000.

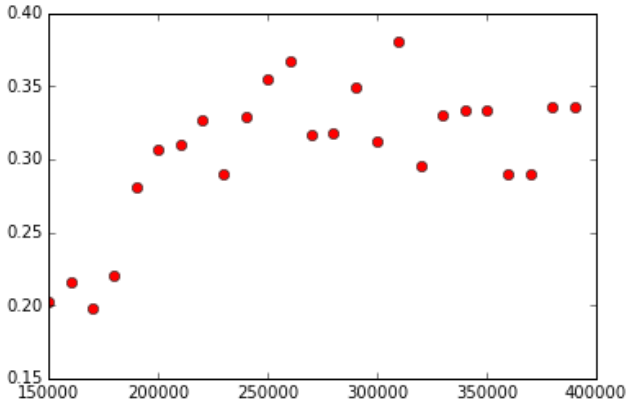

Figure 6. Distribution of R^2 (Coefficient of Determination)

3.3 Model Validation

The model is built for the dataset containing information for Mid-Atlantic region after removing the outliers. This dataset is split into train and test datasets using different algorithms for validation purposes. The algorithm used here sorts the data on 'Total Rooms' and takes every other value to ensure the best representation of the data. A 'validation' function is created that takes the train and test datasets as input and calculates R^2 value. R^2 value is used to represent the goodness-of-fit for the regression model. In addition, comparison plots are created for each model for actual v/s predicted values.

Cross Validation technique is used to further improve the accuracy of the model. The dataset is first split into two-third and one-third sizes in three different manner. Corresponding beta matrices and R^2 values are calculated to check the consistency of values.

3.4 Results and Inferences

The final beta matrix and R^2 value is evaluated as the mean value obtained from each calculation.

The resultant mean value for the datasets used for cross-validation is 0.36, which is a moderate value for fit of the model.
R^2_mean = mean(R^2_1, R^2_2, R^2_3,R^2_4) = 0.36

Three of the four validations provided values greater than average. The average value of 0.36 suggests the model is a moderate fit to predict the total energy consumption.

The resultant matrix of average beta values is as below.
βaverage = np.mean (β1, β2, β3, β4)
βaverage = [7.7 7123.2 1296.6]

Actual and predicted values for the test dataset are compared for each model with the best fit amongst them shown in Figure 7 below.

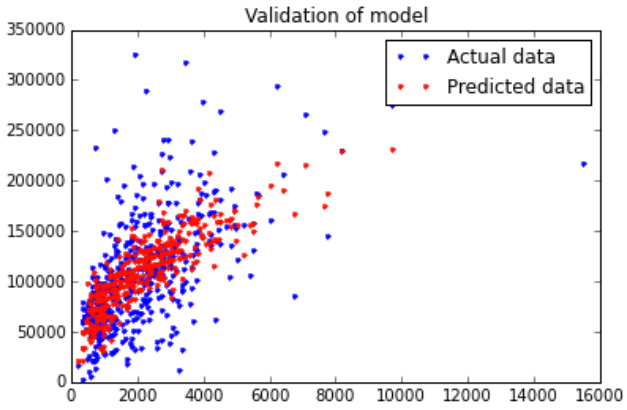

Figure 7. Best Fit Graph

Uncertainty in the result is calculated using 95% confidence interval.

Base value: [7.7 7123.2 1296.6]

Maximum value: [10.7 9344.1 1632.4]

Minimum Value: [4.7 4902.3 960.8]

4. CONCLUSION

This project has created a structure for multivariable regression modelling. This will allow the user to predict any variable of interest based on more than one independent variable. For the scope of this project, the model has been applied for the Mid-Atlantic census division of the RECS 2009 dataset. A specific division was selected because of strong regional variation in data. Out of all independent variables in the microdata, the best correlated variables fell into three specific categories such as number of rooms, area of floor space, and number of windows. The current RECS dataset does not provide the data in a suitable format to build an effective regression model. The resultant cross-validated R^2 value of 0.36 is obtained based on the present format of RECS dataset. This has scope for improvement if the direct variables in the data can be interpreted in a more suitable manner or more suggestive numerical variables are incorporated.

5. REFERENCES

[1] "Residential Energy Consumption Survey (RECS) - U.S. ..." 2015. 14 Dec. 2015 https://www.eia.gov/consumption/residential/about.cfm.

[2] "Census Regions and Divisions of the United States." 2015. 14 Dec. 2015 http://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf