Methodology for Interpretive Model

The methodology used to estimate the interpretive model is a mix of manual and the different model selection techniques i.e. stepwise, forward, backward, and LASSO. Since there are many variables in the data set, as shown earlier, it may take high computational power and time. Therefore, theory and logic are used to subset the variables to what is considered to be relevant in explaining total dollar energy expenditure. The CV press is used as the optimizing criteria to determine the final model for the given model selection techniques. Many iterations of the final subset of variables were considered, table of variables is listed below:

|  |
| --- |
| **Variable Names** |
| Sqft\_100 = TOTSQFT /100 |
| HHAGE |
| ATHOME: 0=No, 1=Yes |
| TEMPHOME: 0=No, 1=Yes |
| TEMPGONE: 0=No, 1=Yes |
| TEMPNITE: 0=No, 1=Yes |
| TOTROOMS |
| YEARMADE |
| AIRCOND: 0=No, 1=Yes |
| EQUIPNOHEAT |
| HDD65 |
| CDD65 |
| HDD30YR |
| CDD30YR |
| AIRCOND\*HDD30YR |
| EQUIPNOHEAT\*CDD30YR |
| AIRCOND\*HDD65 |
| EQUIPNOHEAT\*CDD65 |
| TOTHSQFT |
| TOTCSQFT |
| AIRCOND\*TOTCSQFT |
| EQUIPNOHEAT\*TOTHSQFT |

The logic used is to first determine what appliance utilizes the most energy, and it was determined from experience and anecdotal evidence that heating and cooling uses the most energy. Because of this it is determined that probably this will compose most of the total energy expenditure. Therefore, the indicator for air conditioner (AIRCOND) and heating equipment (EQUIPNOHEAT) are included in the final variable group. Some other factors to consider is of temperature will influence the use of central air and heating. The variables heating (HDD65) and cooling (CDD65) degree days using 65 degrees as the base, and what temp is set when at home, gone, or at night during the winter are good determinants for the behavior of the person when using air and heating. Another factor to consider is the characteristics of the household and building, since these characteristics will influence the efficiency and usage of the air and heating. Therefore, the variables household age (HHAGE), Total Square footage of the building (SQFT\_100=TOTSQFT/100), if the person is a home indicator (ATHOME), Number of rooms (TOTROOMS), when the house is built (YEARMADE), Total heating square footage (TOTHSQFT), Total cooling square footage (TOTCSQFT), and any interactions between the variables.

Once it was determined what model is considered of good fit for explaining the factors that affect total energy expenditures based on Adjusted-RSquare and CV press, further refinements are used to put in or take out variables that maybe relevant to the model. Since we have over 12,000 observations there is a chance that the test statistics could be very sensitive to minor differences, and so practical vs. statistical interpretation is used to determine if the variable is necessary or not. The final estimated model that was chosen is the following:

The interpretation of the variables given in the following:

For every 100 square foot increase of the building leads to a multiplicative change of 1.009758 with a 95% confidence range of 1.00896 to 1.01056 in Median value of total energy expenditure.

For every 1 year increase in the age of the household leads to a multiplicative change of 0.99850911 with a 95% confidence range of 0.998026 to 0.998993 in Median value of total energy expenditure.

When comparing people at home vs. people not at home, people not a home has a multiplicative change of 0.94246687 with a 95% confidence interval of 0.92728 to 0.95790 in Median value of total energy expenditure when compared to people at home.

For every 1 degree increase in temperature when not at home leads to a multiplicative change of 1.0032976 with a 95% confidence range of 1.00271018 to 1.00388491 in Median value of total energy expenditure.

For every 1 room increase in the building leads to a multiplicative change of 1.096288 with a 95% confidence interval range of 1.090605 to 1.1020015 in Median value of total energy expenditure.

For every 1 year increase in the year the house is made leads to a multiplicative change of 0.9975 with a 95% confidence interval range of 0.99717135 to 0.99782736 in Median value of total energy expenditure.

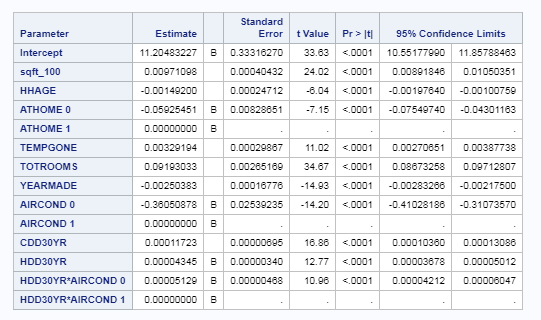
When comparing air conditioner used vs NO air conditioner used, people not using an air conditioner has a multiplicative change of 0.69732145 with a 95% confidence interval of 0.66346322 to 0.73290756 in Median value of total energy expenditure when compared to people using an air conditioner.

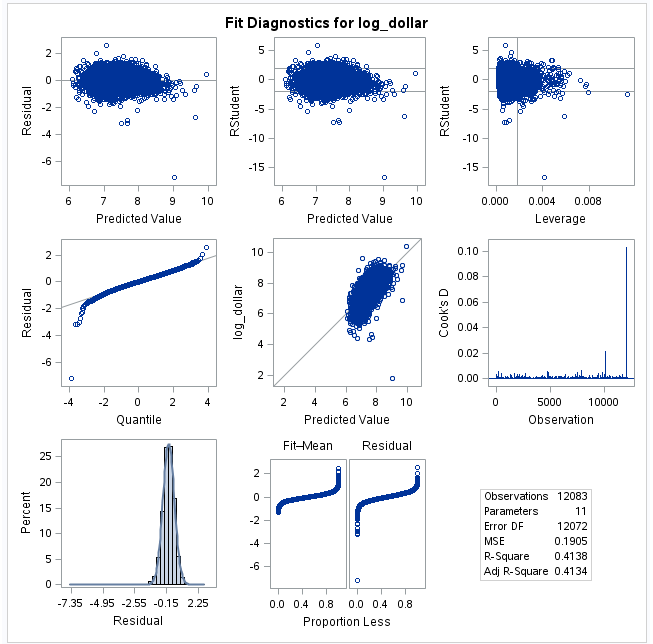
For every 1 day increase in the average 30 year cooling degree days leads to a multiplicative change of 1.00011724 with a 95% confidence interval range of 1.00010361 to 1.00013087 in Median value of total energy expenditure.

For every 1 day increase in the average 30 year heating degree days leads to a multiplicative change of 1.00004345 with a 95% confidence interval range of 1.00003678 to 1.00005012 in Median value of total energy expenditure.

In the interaction of average 30 year heating degree days and air conditioner used, it shows that the effect of average 30 year heating degree days when air conditioner is not used leads to an increase multiplicative change of 1.00005129 with a 95% confidence interval of 1.00004212 to 1.00006047 in the Median value of energy expenditure when compared to average 30 year heating degree days when air conditioner is used.

When looking at the fit statistics of the model, we can see that all variables are significant at the 99% confidence level. We can also see that the residuals are randomly scattered indicating no issue of correlation of residuals. When looking at the leverage and Cook’s D, Leverage does show some concern at the near 0.004, however, when looking at the Cook’s D there does not show of any values indicating of concern where the largest Cook’s D value is 0.10. When looking at the histogram and QQ plots of the residuals, we can see that both show very good indication that the residuals are normally distributed where the histogram shows a normal distribution and the QQ plots show majority of point falling on the 45 degree line.





Some conclusions that can be taken from the interpretive model are that we can see the biggest effect on total dollar energy expenditure is the use of air conditioning in the building. From the model we can see that individuals without a central air unit pay much less compared to individual with central air by a multiplicative change of 0.69 in the median value of log total dollar energy expenditure. The next biggest effect, with a multiplicative change factor of 1.096, is the number of rooms in the building or house since it follows logic that the more rooms the more people that would be living there increase the probability of use of the air and heating unit.

Things to note in this model, is the factors used in this regression do not have large effects when compared to AIRCOND indicating that it may not be really different from zero in a practical sense. Therefore, for future analysis, other variables such as what temperature was set at home, gone, and night during the summer time since we can see that AIRCOND has a significant effect should be included in the model. Other things to consider is maybe finding a better heating unit variable than the one chosen for the subset since the EQUIPNOHEAT has many categories which makes it hard to estimate and have little data in some categories. Another thing to consider is to add region and its interaction with the variables to better parse out the effect of AIRCOND usage since each region has its own climate.