Analysis of the Behavior of Electricity Price in Oregon Households

Sam Oliszewski

Oregon State University

Table of Contents

Introduction

This paper provides an analysis of the electricity expenses from the American Community Survey for households in Oregon (ACS). An exploratory analysis was performed to help provide insight on how to solve both the explanatory and prediction problems. The approaches for model determination in the explanatory and predictive settings are described.

Exploratory Analysis

This section discusses the exploratory analysis performed to investigate the ACS dataset.

Assessing the Available Variables

Original Variables. The provided dataset contained fifteen predictor variables. The original variables included in the ACS dataset used for this analysis are: serial number (SERIALNO), type of unit (TYPE), number of people in the household (NP), lot size in acres (ACR), bedrooms in household (BDSP), units in structure (BLD), fuel cost (FULP), gas cost (GASP), house heating fuel type (HFL), number of rooms in household (RMSP), tenure (TEN), property value (VALP), year structure was built (YBL), presence of under age 18 persons (R18), and presence of over age 60 persons (R60). The response variable in this study is the price of electricity per household (ELEP).

Reducing the Number of Predictors. The dataset originally contained fifteen predictor variables and the modified set of predictors used in this analysis contained thirteen. The variables deemed unnecessary for the study were SERIALNO and TYPE. The SERIALNO variable was excluded from the data set because the value of this predictor is an identifier of the observation and has no relationship with the response variable ELEP. The TYPE variable was excluded from the data set because the value is identical for all the observations in the data set and therefore would not contribute to the change in the response variable ELEP. The predictor variables included in the reduced data set used for this analysis are: NP, ACR, bedrooms in household, BDSP, BLD, FULP, GASP, HFL, RMSP, TEN, VALP, YBL, R18, and R60. The response variable in this study is ELEP.

Addressing Missing Values

Explanation of Problem. Since there were several missing values in the data set, the number of observations able to be studied is reduced. This limits the ability to draw meaningful conclusions from the data.

Solution to the Missing Value Problem. To address the issue of missing values, imputation was performed to increase the number of observations able to be studied. There was a mix of integer and factor variables in the dataset and two different methods of imputation were performed as a result— median and mode imputation. For integer variables, median imputation was used because the median was a good measure of the center of the data that preserves the integer value in the result. Since integers are ordered, a median is a reasonable statistic to use for imputation. For the factor values, mode imputation was used. This method was preferred due to many of the level being unordered.

Modifying Variable Structure for Different Goals

BLD in the Explanatory Problem Setting. Since the goal in the explanatory problem is to assess the difference in electricity expenses for people living in houses versus apartments, a variable needed to be selected to determine which household were classified as houses or apartments. Two possible variables were available— TEN and BLD. Ultimately, the BLD variable was determined to be better for this classification. However, the structure of the variable needed to be redone to fit the needs of the question. Therefore, the BLD variable was restructured from the original ten factors into two factor levels: “House” and “Apartment”. The level House was taken from the original levels “One-family house detached” and “One-family house attached”. The level Apartment was taken from the original levels “2 Apartments”, “3-4 Apartments”, “5-9 Apartments”, “10-19 Apartments”, “20-49 Apartments” and “50 or more apartments”. After the variable was restructured, the observations in the dataset containing BLD values “Mobile home or trailer” and “Boat, RV, van, etc.” were removed because they were not relevant to the study.

BLD in the Prediction Problem Setting. In the prediction problem for this study, there is no need to use the restructured BLD variable as discussed previously. It is useful to consider the original factor levels and their associated observations because the question does not specify that there should be a grouping of BLD type. Therefore, the BLD variable in the prediction problem will include the original ten factor levels “Mobile home or trailer”, “One-family house detached”, “One-family house attached”, “2 Apartments”, “3-4 Apartments”, “5-9 Apartments”, “10-19 Apartments”, “20-49 Apartments”, “50 or more apartments”, and “Boat, RV, van, etc.”. All the original observations are included in this problem.

Visualizing Relationships in Data

Figures of Predictors Against Response. The initial exploration of the data included plotting each predictor value against the response to determine any apparent relationships within the dataset. These relationships are depicted after imputation occurred. The resulting plots are shown below:

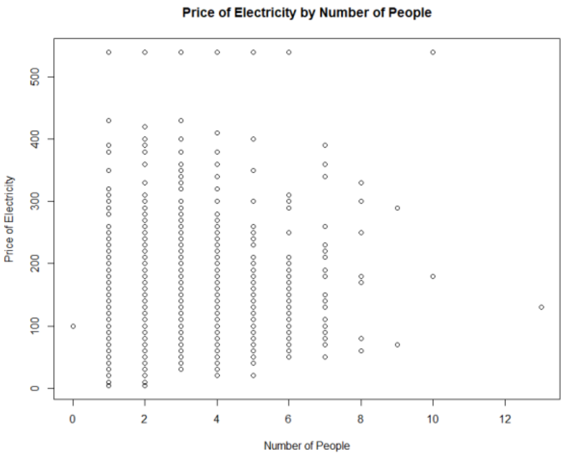

Figure 1. ELEP by NP

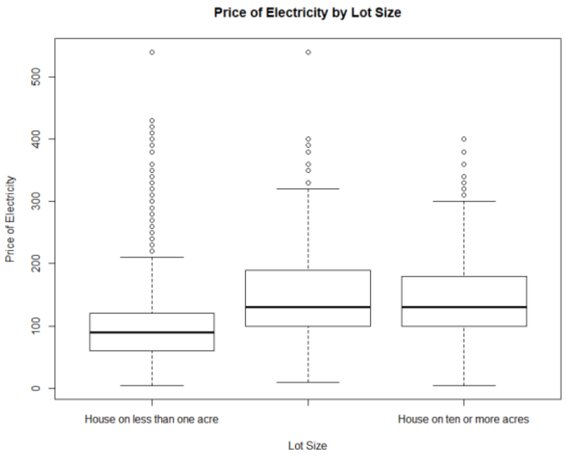

Figure 2. ELEP by ACR

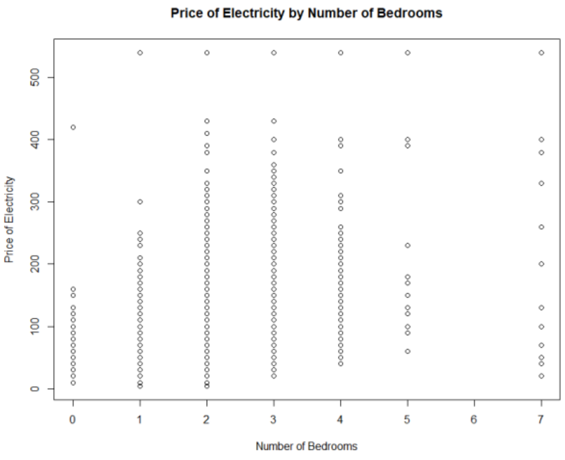

Figure 3. ELEP by BDSP

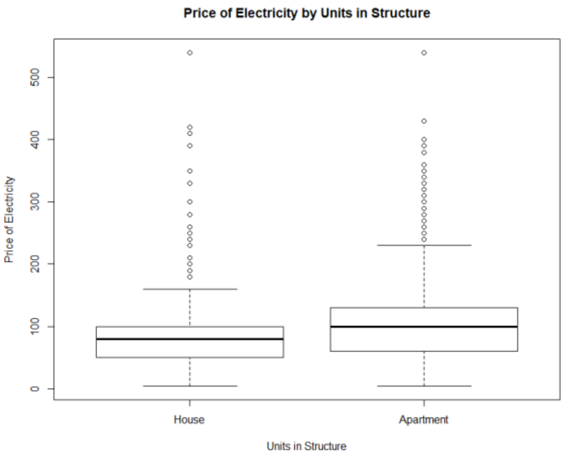

Figure 4. ELEP by BLD

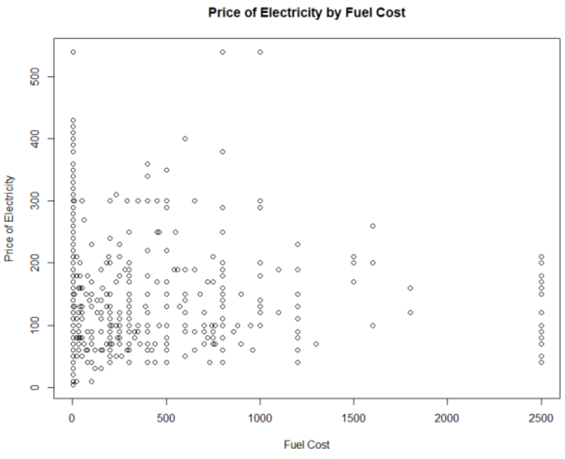

Figure 5. ELEP by FULP

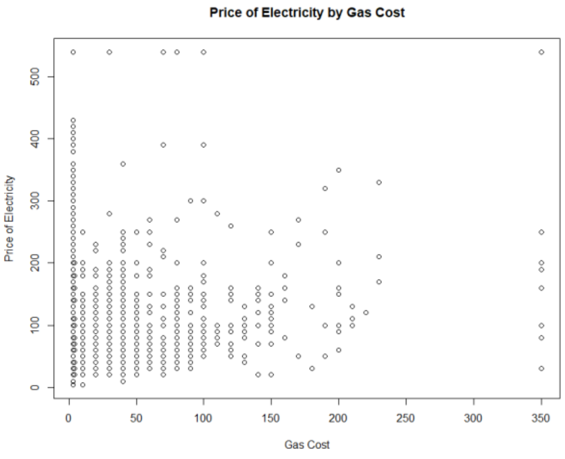

Figure 6. ELEP by GASP

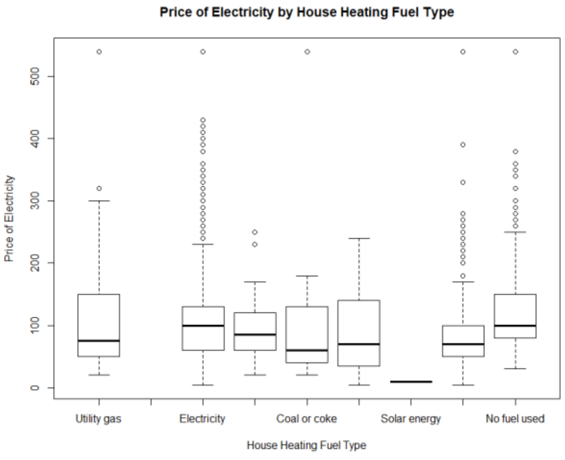

Figure 7. ELEP by HFL

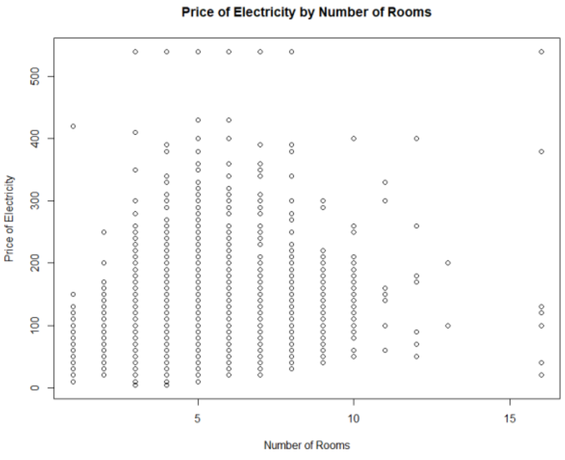

Figure 8. ELEP by RMSP

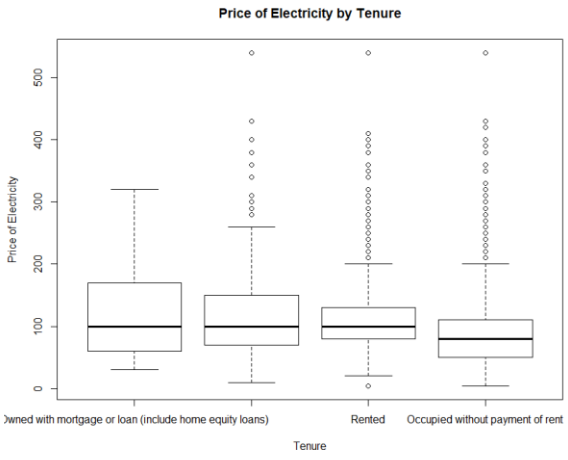

Figure 9. ELEP by TEN

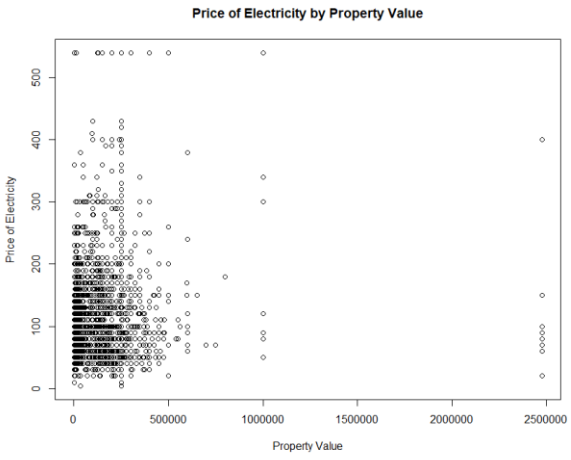

Figure 10. ELEP by VALP

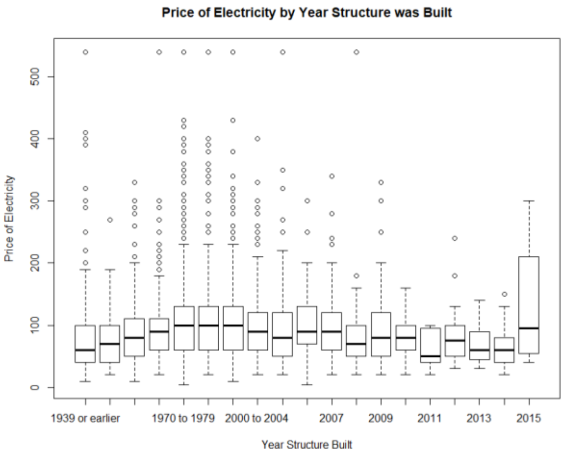

Figure 11. ELEP by YBL

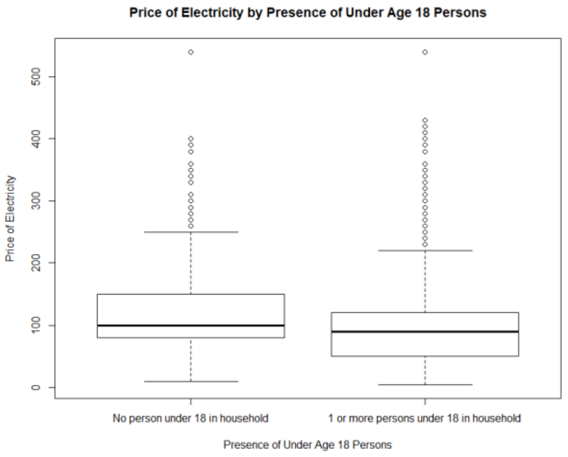

Figure 12. ELEP by R18

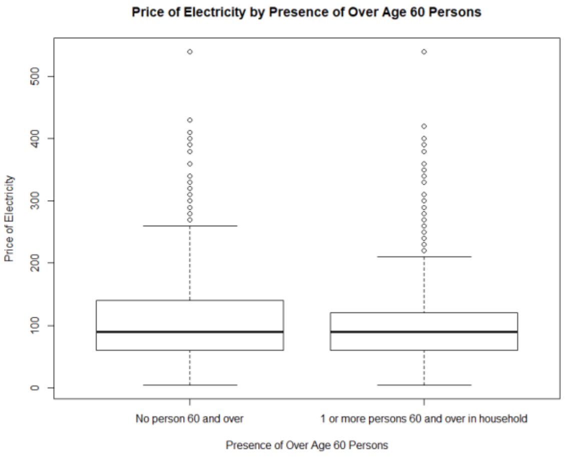

Figure 13. ELEP by R60

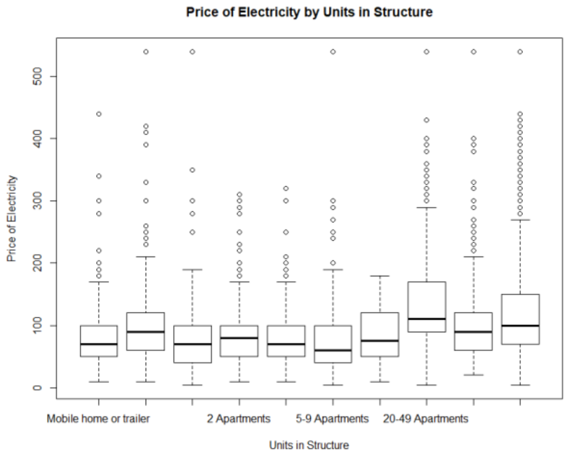

Figure 14. ELEP by Original BLD Factors

Conclusions. After examining these plots, it is seen that there is no strong evidence of impact on the response variable for different values of the predictor for the predictor variables R60, TEN or VALP (most of the observations are clustered). It is seen that there is some evidence of impact on the response variable for different values of the predictor for the predictor variables R18 (slightly higher for no 18 year old persons), YBL (clear outliers for 2008 and earlier years and 2015 is much higher), RMSP (consistency for number of rooms 4 and greater and lower prices for 3 and fewer rooms), HFL (fairly consistent medians across fuel types, but more high values for electricity), GASP (more high values when gas price is almost zero), FULP (more high values when gas price is almost zero), BLD (in the original factoring the 20-49 apartments level has a median higher than most other structure types Q3 value, and in the restructured BLD the apartment level is higher than house), BDSP (consistency with 2 or more bedrooms and lower electricity price with 0 and 1 bedrooms), ACR (lower electricity price with house on less than one acre, but more outliers), and NP (consistent electricity prices with 7 and fewer people, but lower values with 8 or more people per household).

Explanatory Problem

Overview

The goal in this section is to determine whether people living in apartments pay less on electricity than those living in houses and by how much. To make this determination, first a model must be constructed to reflect the dataset. This process will involve fitting both a model without interactions and a model with interactions. The models being fit will adjust for NP and BDSP as well as the other original predictors. This is done by including all the terms in each model. This adjusts for the terms because the regression line will have its slope altered based on the values of each term because each predictor has an associated coefficient (beta) that represents the estimated amount by which the mean value of the response variable (ELEP) changes for a unit change in the predictor when all other predictors are held fixed. Once two models are fit, the preferred model must be determined. Once the preferred model has been identified, a summary of the model will be reported to determine the difference in electricity expenses between people living in apartments and people living in houses.

Methods

Fitting a Model without Interactions. A model was fit including all thirteen predictors: NP, ACR, BDSP, BLD, FULP, GASP, HFK, RMSP, TEN, VALP, YBL, R18, and R60. To assess the fit of this model, first a residuals versus fitted plot was generated. This plot is shown below:

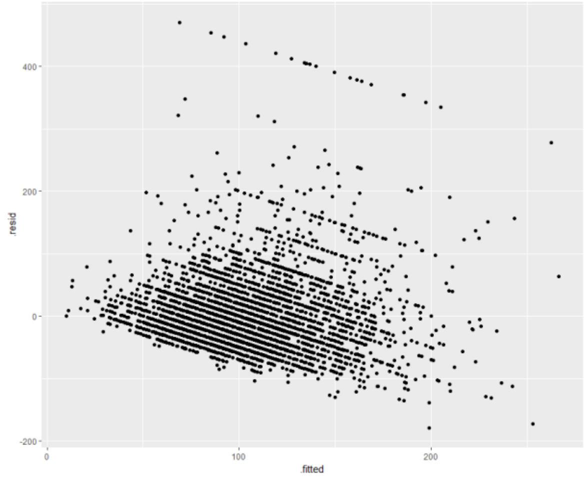

Figure 15. Residual versus Fitted – Model without Interactions

Of note in Figure 15 is the apparent parallel lines created by the plotted residuals. This is likely a reflection of the discrete nature of the response variable ELEP, forcing each value to take on integer values rather than continuous ones. This is also a possible cause of the suspected heteroscedastic behavior causing the residuals to take a shape. With this understanding of the response variable the residual versus fitted plot is reflecting, it is reasonable to assume that the model satisfies the constant variance assumption for multiple linear regression (MLR). Further, the assumption of linearity is also satisfied because the plotted points would generally be random if the response variable were not discrete and thus MLR is still reasonable to pursue. The sample size is sufficiently large to assume that the assumption of normality for each Y around its mean is satisfied and the assumption of independence is also met because the households studied were selected at random and there is no reason to believe that one household's electricity price should be related to another's.

To further assess whether MLR is appropriate to perform, residual versus explanatory variable plots for each of the thirteen explanatory variables were created. These plots are shown below:

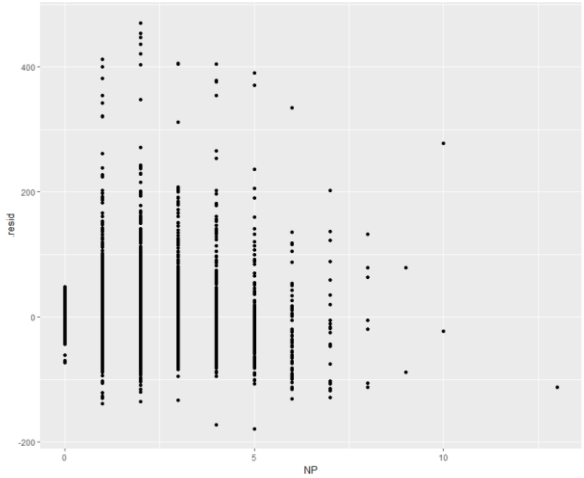

Figure 16. Residuals versus NP

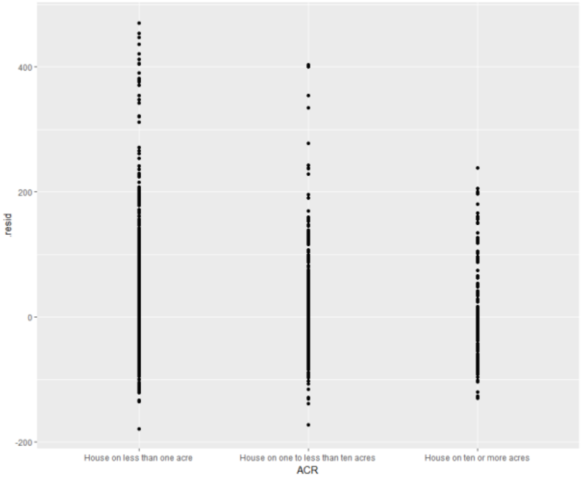

Figure 17. Residuals versus ACR

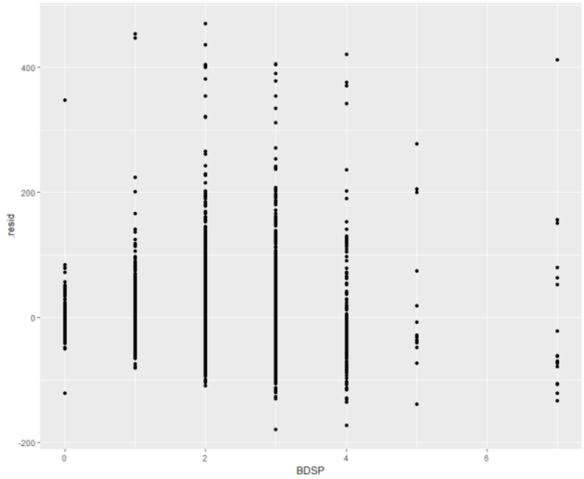

Figure 18. Residuals versus BDSP

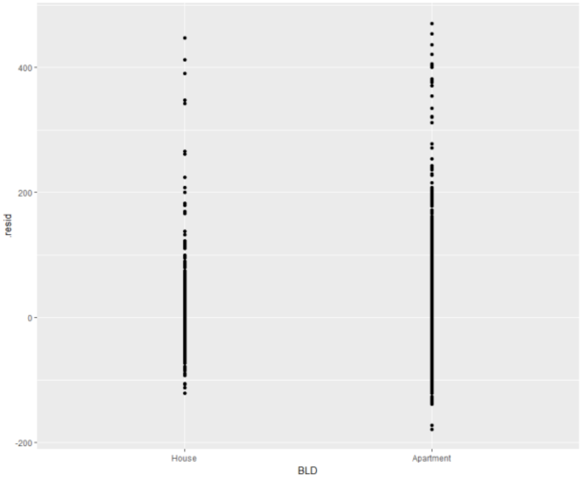

Figure 19. Residuals versus BLD

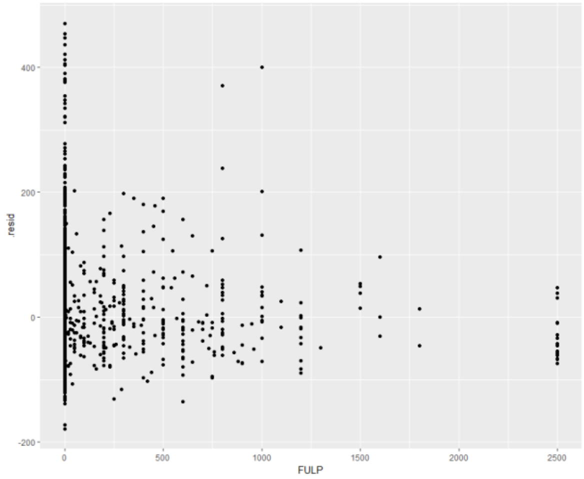

Figure 20. Residuals versus FULP

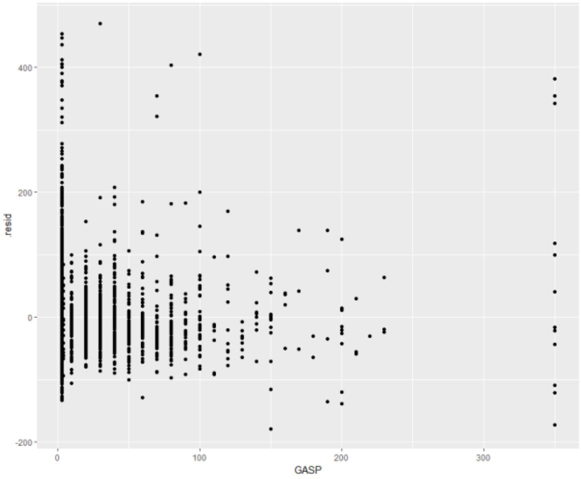

Figure 21. Residuals versus GASP

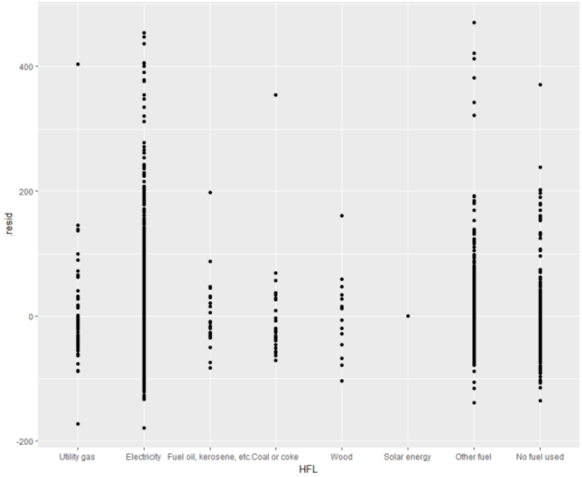

Figure 22. Residuals versus HFL

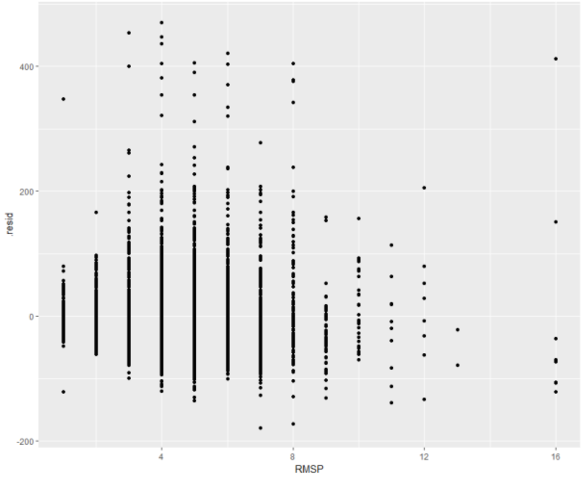

Figure 23. Residuals versus RMSP

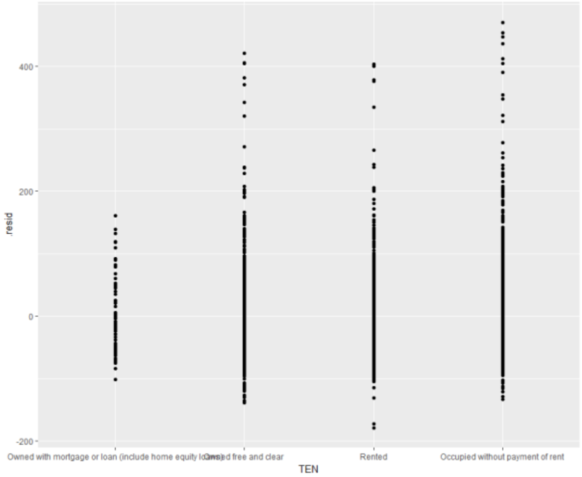

Figure 24. Residuals versus TEN

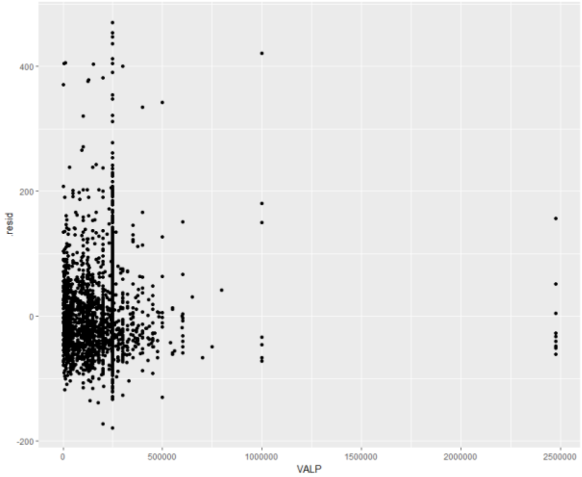

Figure 25. Residuals versus VALP

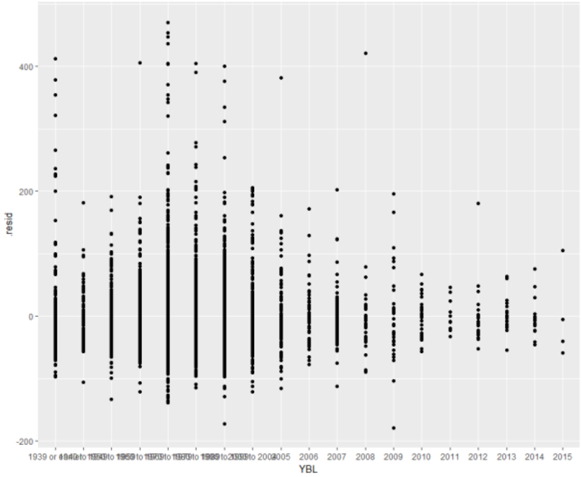

Figure 26. Residuals versus YBL

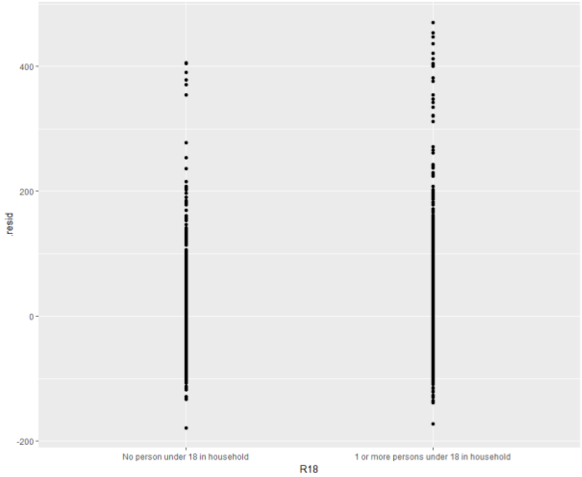

Figure 27. Residuals versus R18

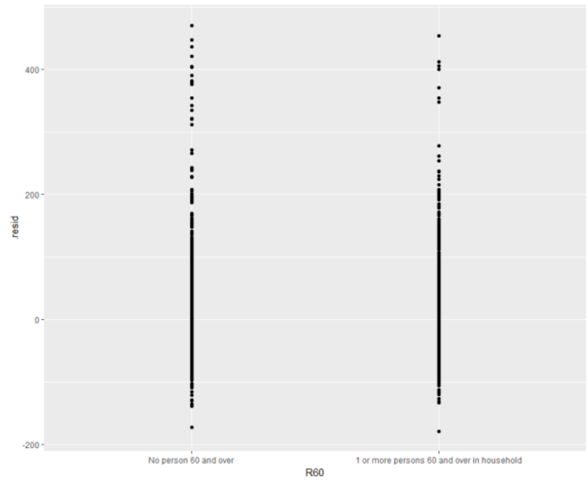

Figure 28. Residuals versus R60

While each of the thirteen residual versus explanatory variable plots showed some values of electricity being higher than the majority of the data points, there are too many points to be considered outliers and do not present any reason to disprove the assumptions for MLR are not met as discussed previously. Further, while the residuals versus predictor plots show non-constant variance, the variance is not due only to single predictor and therefore the residuals versus fitted plot is a more appropriate measure of model fit and does not indicate any MLR assumption violations the would halt inference from proceeding.

Fitting a Model with Interactions. A model was fit including all thirteen of the following predictors and their interactions: NP, ACR, BDSP, BLD, FULP, GASP, HFK, RMSP, TEN, VALP, YBL, R18, and R60. To assess the fit of this model, a residuals versus fitted plot was generated. This plot is shown below:

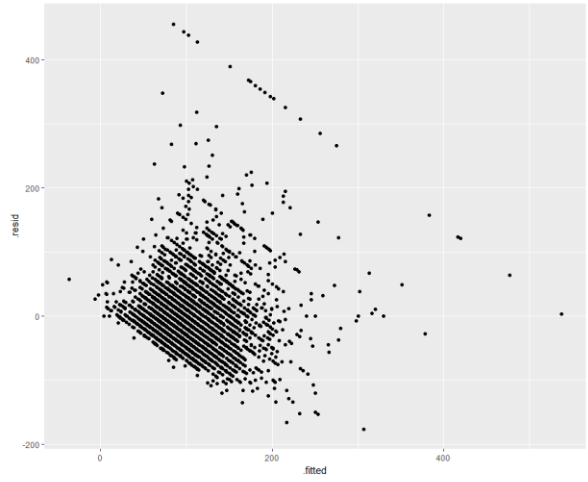

Figure 29. Residuals versus Fitted – Model with Interactions

This plot is very similar to the residuals versus fitted plot for the model without interactions and therefore does not present any evidence that the assumptions for MLR are not met, as were discussed previously.

Choosing the Best Model. After the two models were fit and the assumptions for MLR were proven to be satisfied, a decision needed to be made as to which of the models best fit the data. To make this determination, an ANOVA test was performed comparing the two models. The results of this test found that the model with interactions was preferred (Extra Sum of Squares comparing model with interaction to model without interaction, p-value = 2.2e-16).

Summarizing the Preferred Model. With the best model for the data selected, a summary was generated to determine whether there is a difference in the electricity expenses between Oregonians living in apartments and houses and how much that difference is. This showed that the estimated difference in the BLD term for people living in apartments versus houses is 8.42e01 (84.23), with a standard error of 5.17e01 (51.66), a t-statistic of 1.63, and a p-value of 0.10.

Conclusion

With 95% confidence, there is no evidence that the mean price of electricity for people living in apartments is less than the mean price of electricity for people living in houses. It is estimated that the mean price of electricity for people living in houses is between 1.70e01 (17.05) dollars below and 1.86e02 (185.51) dollars above the mean price of electricity for people living in houses, with a point estimate of 8.42e01 (84.23) (t-test, d f= 4008, t =1.63, p-value = 0.10).

Limitations. Of note on the conclusions from this study is that the values calculated may be somewhat inaccurate because the data was imputed instead of resulting solely from actual observations. This may underestimate the standard error associated with the generated values. Also noteworthy is that the missing data in the dataset requires investigation to see if it was missing completely at random as this is the only type of missing data that would not affect the inferential conclusions.

Prediction Problem

Overview

The goal in this section is to create a model that could be used to predict electricity costs for a household in Oregon. To do this, best subset selection was performed, and both the validation set approach and k-fold cross validation were used to select the best model. Then, a summary of the best model will be reported.

Methods

Validation Set Approach. To determine the best model, thirteen models were fitted that included from one to thirteen predictors. Best subset selection was performed to determine which predictors to include in the model at each model size. Then, to test each model, a specific number of observations aside as a test set and using the remaining observations as a training set. The mean squared error (MSE) was calculated using the training set for each model and the lowest MSE among each calculated for the different model sizes is selected as the preferred model. This indicated the optimal number of predictors to include in the model. The validation set approach resulted in a model that included all thirteen of the original predictors.

K-Fold Cross Validation Approach. To determine the best model, thirteen models were fitted that included from one to thirteen predictors. Best subset selection was performed to determine which predictors to include in the model at each model size. Then, to test each model, the data was divided into k number of groups of even size. In this study, the value of k was equal to 10. Then, 10 experiments were performed where one of the 10 groups is used as a test set, and the remaining k-1 (10 - 1 = 9) groups acted as the training set. The MSE was computed for each model and the lowest test MSE among each calculated for the different model sizes was selected as the preferred model. This indicated the optimal number of predictors to include in the model. The k-fold cross validation approach resulted in a model that included all thirteen of the original predictors.

Conclusion

The best model as chosen by both the validation set approach (MSE = 4047.35) and k-fold cross validation (MSE = 3846.57) includes the following terms:

Term	**Coefficient**	**Standard Error**	**t-statistic**	**p-value**
Intercept	-5.46e00	7.04e00	-0.78	0.44
NP	1.36e01	5.23e-01	26.07	**2e-16**
BLDOne-family house detached	2.16e01	4.40e00	4.92	**8.96e-07**
BLDOne-family house attached	2.67e00	4.53e00	0.59	0.56
BLD2 Apartments	4.35e00	4.02e00	1.08	0.28
BLD3-4 Apartments	-1.84e00	4.04e00	-0.46	0.65
BLD5-9 Apartments	-9.21e-01	4.29e00	-0.22	0.83
BLD10-19 Apartments	2.96e01	1.29e01	2.29	**0.02**
BLD20-40 Apartments	3.95e01	3.67e00	10.76	**2e-16**
BLD 50 or more apartments	2.32e01	3.98e00	5.84	**5.31e-09**
BLDBoat, Rv, van, etc.	3.15e01	3.39e00	9.28	**2e-16**
ACRHouse on one to less than ten acres	2.75e01	1.64e00	16.83	**2e-16**
ACRHouse on ten or more acres	2.62e01	2.44e00	10.73	**2e-16**
BDSP	6.39e00	7.79e-01	8.20	**2.58e-16**
FULP	8.58e-03	2.12e-03	4.04	**5.35e-05**
GASP	1.48e-01	1.24e-02	11.87	**2e-16**
HFLBottled, tank, or LP gas	-2.03e01	4.50e01	-0.45	0.65
HFLElectricity	4.08e01	3.78e00	10.79	**2e-16**
HFLFuel oil, kerosene, etc.	2.00e00	5.07e00	0.39	0.69
HFLCoal or coke	2.56e01	1.01e01	2.54	**0.01**
HFLWood	2.85e01	7.15e00	3.98	**6.81e-05**
HFLSolar energy	-1.47e01	2.27e01	-0.65	0.52
HFLOther fuel	-9.19e00	3.69e00	-2.49	**0.01**
HFLNo fuel used	9.46e00	4.08e00	2.32	**0.02**
RMSP	1.69e00	3.34e-01	5.06	**4.30e-07**
TENOwned free and clear	-8.97e00	4.30e00	-2.08	**0.04**
TENRented	-2.39e00	4.24e00	-0.56	0.57
TENOccupied without payment of rent	-3.54e00	4.34e00	-0.81	0.42
VALP	2.36e-05	2.42e-06	9.76	**2e-16**
YBL1940 to 1949	8.22e00	2.54e00	3.24	**0.00**

YBL1950 to 1959	5.42e00	2.20e00	2.47	**0.01**
YBL1960 to 1969	3.34e00	2.16e00	1.55	0.12
YBL1970 to 1979	8.73e00	1.86e00	4.70	**2.65e-06**
YBL1980 to 1989	4.73e00	2.14e00	2.21	**0.03**
YBL1990 to 1999	-7.32e-01	1.93e00	-0.38	0.70
YBL2000 to 2004	-2.64e00	2.35e00	-1.12	0.26
YBL2005	-4.44e00	3.92e00	-1.13	0.26
YBL2006	-6.25e00	4.37e00	-1.43	0.15
YBL2007	-4.92e00	4.28e00	-1.15	0.25
YBL2008	-7.02e00	5.74e00	-1.22	0.22
YBL2009	2.56e00	6.10e00	0.42	0.67
YBL2010	-7.63e00	6.75e00	-1.13	0.26
YBL2011	9.72e-01	8.69e00	0.11	0.91
YBL2012	-3.89e00	7.84e00	-0.50	0.62
YBL2013	-1.20e01	7.13e00	-1.68	0.09
YBL2014	-1.15e01	7.83e00	-1.46	0.14
YBL2015	-2.27e01	1.42e01	-1.60	0.11
R181 or more persons under 18 in household	4.47e00	1.75e00	2.55	**0.01**
R601 or more persons 60 and over in household	-3.70e00	1.20e00	-3.09	**0.00**

Limitations. While this model was selected as the best fit for the data, there are some limitations to consider. One such limitation to the model is that imputation was performed on the data meaning that several observations are affected by the imputed values, rather than actual observations. As a result, the standard error is likely underestimated. Also of note is that the missing data in the dataset requires investigation to see if it was missing completely at random as this is the only type of missing data that would not affect the inferential conclusions. Another limitation to this predictive model is that these households are only representative of Oregon households and should not be considered representative of households in other states. Also affecting the performance of this predictive model is how the sample size is large and we may run into issues such as small effects being considered statistically significant, raising the question of practical versus statistical significance.

Compare and Contrast

Differences Between Approaches. Two different procedures were followed for the explanatory and prediction problems discussed previously. This is because the nature of the questions was quite different. In the explanatory problem, the question of interest involved comparing people living in houses versus apartments, whereas in the prediction problem, the question of interest involved all Oregon households. This meant that the dataset needed to be restructured and refined to answer the explanatory question, whereas in the prediction problem, the data is in its original structure.

Similarities Between Approaches. Both approaches discussed previously involved the same dataset. This dataset had several missing values and required imputation to expand the observation set and help draw more meaningful conclusions. Both methods also involved determining which predictors were the most important to include in the model.

Challenges Faced. This analysis included several challenging aspects. The first challenge involved imputation. Since there is no simple method for determining the mode of a dataset, a function had to be written to determine this statistic. This challenge applied to both approaches. Determining how to group “people living in houses” versus “people living in apartments” was another challenge, as there were two potential variables that could have been used for the classification— TEN and BLD. Ultimately, BLD seemed more straightforward and was chosen as the deciding factor. This decision was challenging and relevant to both approaches. It also was challenging to interpret the residual plots because so many of the variables were discrete or factor variables and there were interactions in the dataset. It took a bit more effort to determine whether the assumptions for MLR were met based on these plots as a result, as proved to be yet another challenge with this study. Another challenge involved restructuring the dataset for the explanatory problem. The BLD variable needed to have the number of factors reduced from ten to two, in order to draw meaningful inference and determining how to do this was difficult. The toughest part of the prediction problem was determining what data to include when generating the prediction model. The data was reduced to perform the explanatory analysis and ultimately it did not make sense to use the reduced dataset for the prediction model, based on the wording of the question. In the end, there were many decisions that had to be made during this study and each resulted in its own set of challenges to find solutions to.