Ben Tankus

3-16-21

Abstract:

Tankus Industries (TI) was contacted to analyze Oregon housing data from the 2015 American Community Survey. The main task of this analysis is to determine how expensive electricity bills in Oregon are, and what factors contribute to their cost. TI was asked to pay close attention to the relationship between electricity cost of apartments versus houses, and to adjust for the number of bedrooms and occupants in a house. We are then asked to create a model to predict electricity costs for a typical Oregon household.

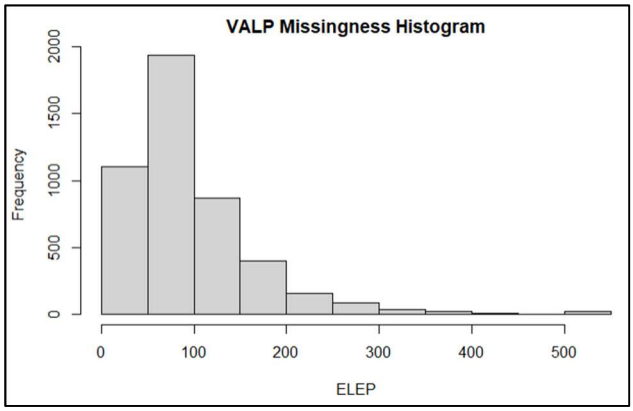
Contents

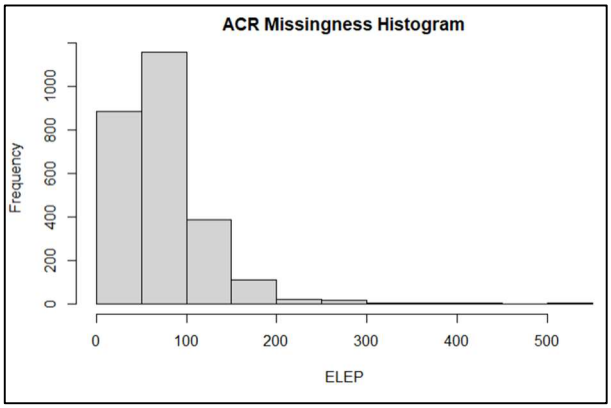
Do People Living in Apartments Pay Less for Electricity than Those Living in Houses?

Does the Data Require Any Cleaning?

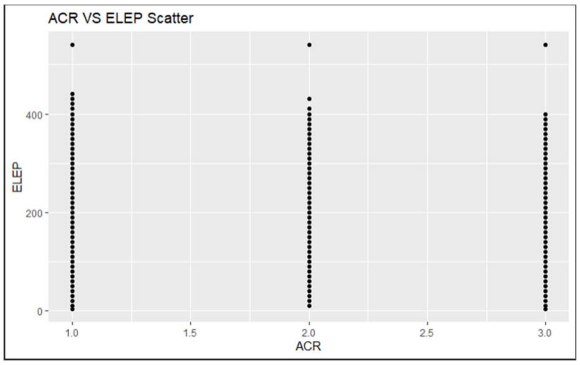
As with most data, this set did require cleaning. Mostly regarding the missing values in the data. ACR and VALP both have many null values, 2586 and 4632 respectively. Looking at the missingness distribution histograms, it seems both VALP and ACR are missing-not-at-random, and clearly have more missing values at the low end of the ELEP histograms.

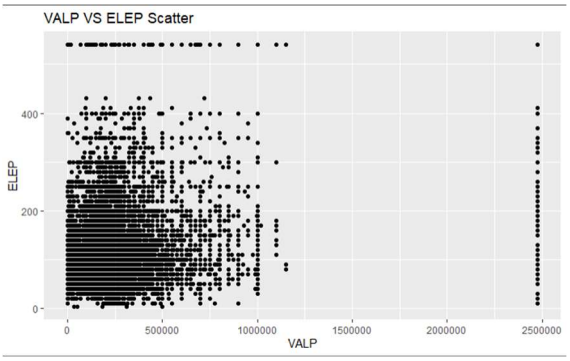
TI believes that these two variables should be removed from the analysis, as property value (VALP) and Lot Size (ACR)





The scatter plots of ACR/VALP VS ELEP also show minimal correlation, meaning it is unlikely that removing ACR and VALP will have a negative impact on the study.





Choosing Practically Relevant Fields

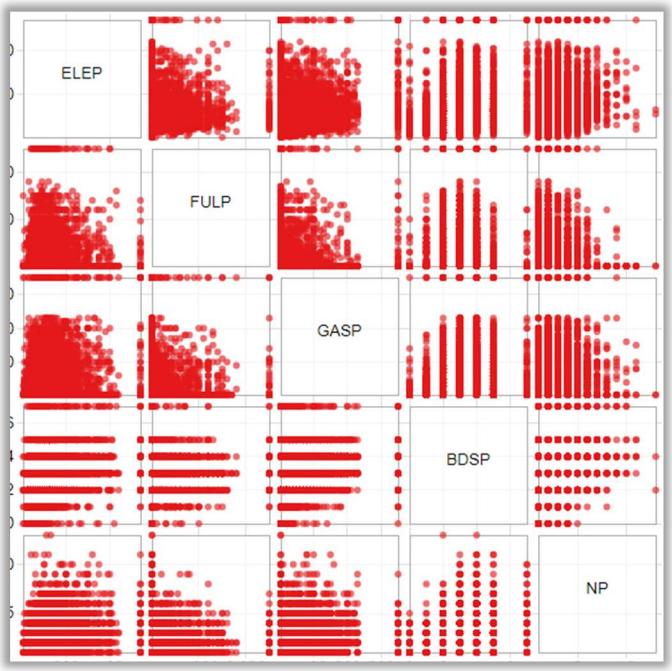
Once the missingness of the data was analyzed, TI looked at a summary of the data to look at the existing fields. From the preliminary examination TI recommends dropping SerialNO, and Type from the analysis. SerialNO is important to organize the data, but this column will not impact electricity price. Along with this, Type is always 1, and should be dropped.

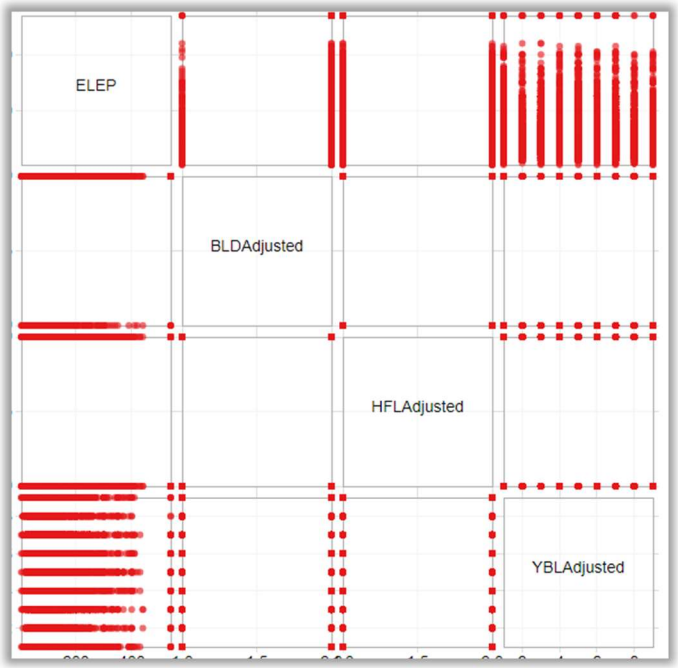
The data was mixed, and is summarized in the table below:

```
    SERIALNO             NP              TYPE          ACR             BDSP            BLD              ELEP
Min.   :     70   Min.   : 1.000   Min.   :1   Length:15166       Min.   :0.000   Length:15166       Min.   :  4.0
1st Qu.: 368628   1st Qu.: 1.000   1st Qu.:1   Class :character   1st Qu.:2.000   Class :character   1st Qu.: 70.0
Median : 748326   Median : 2.000   Median :1   Mode  :character   Median :3.000   Mode  :character   Median :100.0
Mean   : 749620   Mean   : 2.403   Mean   :1                      Mean   :2.789                      Mean   :116.2
3rd Qu.:1126788   3rd Qu.: 3.000   3rd Qu.:1                      3rd Qu.:3.000                      3rd Qu.:150.0
Max.   :1513284   Max.   :13.000   Max.   :1                      Max.   :7.000                      Max.   :540.0

      FULP              GASP             HFL                RMSP             TEN                VALP
Min.   :   1.00   Min.   :  3.00   Length:15166       Min.   : 1.00   Length:15166       Min.   :   1000
1st Qu.:   2.00   1st Qu.:  3.00   Class :character   1st Qu.: 4.00   Class :character   1st Qu.: 160000
Median :   2.00   Median :  3.00   Mode  :character   Median : 6.00   Mode  :character   Median : 250000
Mean   :  85.25   Mean   : 35.68                      Mean   : 6.01                      Mean   : 301966
3rd Qu.:   2.00   3rd Qu.: 50.00                      3rd Qu.: 7.00                      3rd Qu.: 360000
Max.   :2500.00   Max.   :350.00                      Max.   :16.00                      Max.   :2476000
                                                                                         NA's   :4632

     YBL                R18                R60
Length:15166       Length:15166       Length:15166
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character
```

After seeing the spread of the data TI took the steps to rank relevant fields, and potentially remove additional practically irrelevant fields from our data. Once identified, these fields will have a correlation check with ELEP before complete removal. Some categorical fields (BLD, HFL, TEN, YBL, R18, R60) needed converting to numeric for correlation consideration. The full scatter matrix is available after the initial table, and correlation metrics are listed below. None of the practically relevant factors have noteworthy correlation, so TI recommends removing them from the model. RMSP and BDSP are also highly correlated (0.72), so TI recommends removing RMSP to prevent redundant information. FULP, GASP, and YBL may still be relevant after further inspection, so TI recommends leaving them in along with BLD, HFL, BDSP, NP, and ELEP.

Field	Description	Practically Relevant to ELEP?	Correlation to ELEP
SERIALNO	Serial Number	No	N/A
TYPE	Type of Unit	No (always 1)	N/A
R18	Presence of persons under 18	No	0.16
R60	Presence of persons over 59	No	0.013
TEN	Tenure	No	0.077
FULP	Yearly Fuel Cost (other than gas & electricity)	Maybe	0.06
GASP	Gas (monthly cost)	Maybe	0.029
YBL	When structure first built	Maybe	0.01
NP	Number of Persons in the house	Yes (not in current form)	0.28
BLD	Units in Structure	Yes (not in current form)	0.18
HFL	House Heating Fuel	Yes (not in current form)	0.14
RMSP	Number of Rooms	Yes (Redundant)	0.23
BDSP	Number of Bedrooms	Yes	0.26
ELEP	Electricity (monthly Cost)	Yes	1





Cleaning Data and Transitioning Relevant Data to a Computer Digestible Format

Now that the practically relevant fields have been identified, TI must convert them to more useable data, based on the research question. For example, there are many housing types listed under BLD, but the research question only asks for differences between apartments and houses. TI recommends analyzing these data as “Apartment” and “House” values, as shown in the table below. HFL also has many types of fuel and TI similarly recommends using “Electricity” and “Not Electricity” values for this field. TI also recommends grouping houses newer than 2005 into one “2005 to 2015” group to reduce model complexity.

Field	Previous Value	Analyzed Value
BLD	Mobile home or trailer	Removed
BLD	Boat, RV, van, etc.	Removed
BLD	One-family house detached	House
BLD	One-family house attached	House
BLD	2 Apartments	Apartment
BLD	All other Apartment Fields	Apartment
HFL	Electricity	Electricity
HFL	All other non-electricity fields	Not Electricity
YBL	Year-by-Year for 2005+	2005 to 2015

Do People Living in Apartments Pay Less on Electricity than those Living in Houses?

This is the main question TI was asked, and we are now able to answer it. TI has analyzed the data by iteratively adding factors and comparing Residual Sum of Square (RSS) values to judge whether it was a beneficial add. It is unlikely that there is any interaction between terms due to low multicollinearity, and TI did not see any relevant implementation of random effects. Therefore, this will be a multi factor linear regression model. This full analysis is available in the R script included with the submission for further analysis if the reader is interested in that level of detail.

From the base model of BLD, NP, and BDSP, TI also recommends incorporating YBL, and FULP. GASP did lower the RSS value of the model but by a relatively insignificant margin. According to this final model, **electricity is more expensive in a house than in an apartment** with **houses costing roughly double every month**. This model includes number of persons in the house, number of bedrooms, method of house heating, year when the structure was built, and how much the alternative fuel cost per year. For those interested in viewing the exact estimates and error values, please examine the table below. Note, some of the categorical values are not significant, but TI still recommends including them in the model for consistency with the remaining time ranges. This model has Residual standard error of 66.76, Adjusted R-squared of 0.76, and p-value of 2.2e-16, which puts it in a "pretty good fit" range.

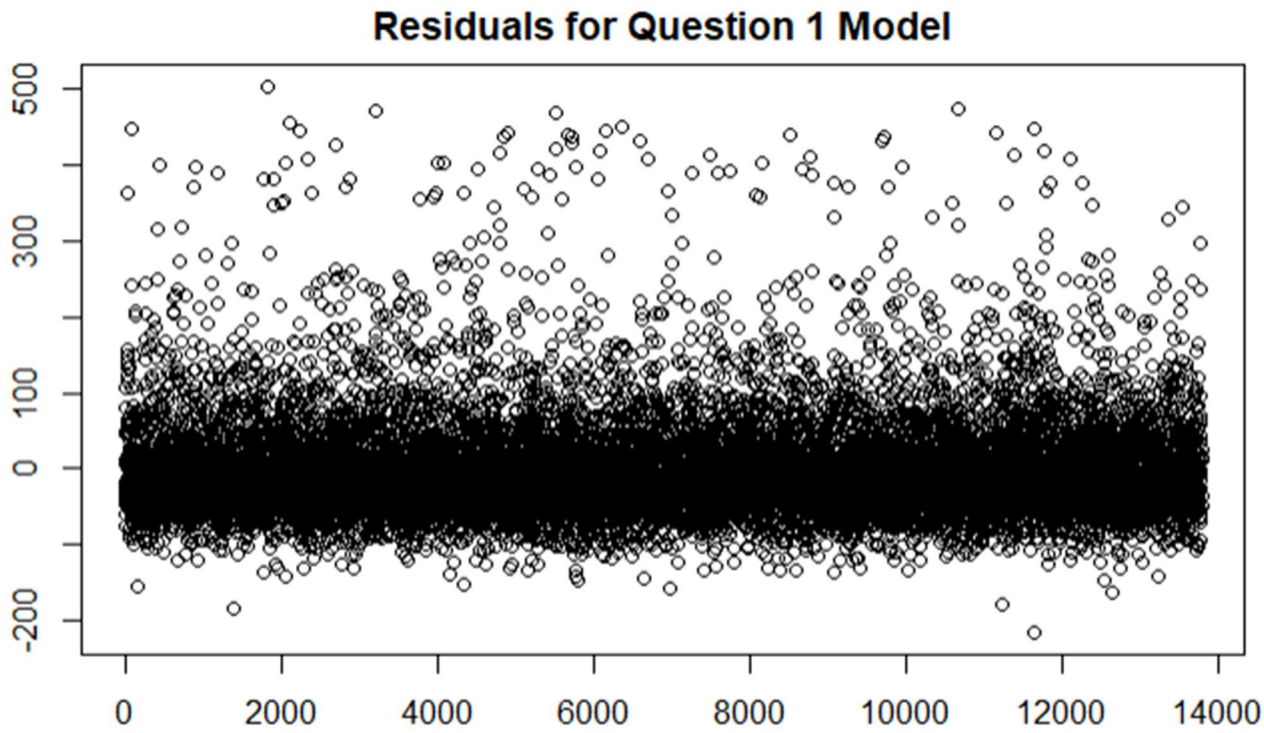
	Estimate	Std. Error	t value	Pr(>|t|)
BLDAdjustedApartment	39.80	2.35	16.90	0.00
BLDAdjustedHouse	76.06	2.68	28.41	0.00
NP	11.89	0.45	26.43	0.00
BDSP	12.91	0.70	18.40	0.00
HFLAdjustedNot Electricity	-41.96	1.29	-32.65	0.00
factor(YBLAdjusted)1940 to 1949	6.91	2.77	2.50	0.01
factor(YBLAdjusted)1950 to 1959	3.63	2.41	1.51	0.13
factor(YBLAdjusted)1960 to 1969	2.47	2.39	1.03	0.30
factor(YBLAdjusted)1970 to 1979	7.21	2.07	3.49	0.00
factor(YBLAdjusted)1980 to 1989	6.15	2.42	2.53	0.01
factor(YBLAdjusted)1990 to 1999	-0.59	2.14	-0.27	0.78
factor(YBLAdjusted)2000 to 2004	-5.07	2.66	-1.91	0.06
factor(YBLAdjusted)2005 to 2015	-7.02	2.43	-2.89	0.00
FULP	0.02	0.00	8.88	0.00

Parameter Coefficients

	2.5 %	97.5 %
BLDAdjustedApartment	35.18	44.41
BLDAdjustedHouse	70.81	81.31
NP	11.01	12.77
BDSP	11.53	14.28
HFLAdjustedNot Electricity	-44.48	-39.44
factor(YBLAdjusted)1940 to 1949	1.49	12.34
factor(YBLAdjusted)1950 to 1959	-1.08	8.35
factor(YBLAdjusted)1960 to 1969	-2.22	7.15
factor(YBLAdjusted)1970 to 1979	3.16	11.26
factor(YBLAdjusted)1980 to 1989	1.39	10.90
factor(YBLAdjusted)1990 to 1999	-4.79	3.62
factor(YBLAdjusted)2000 to 2004	-10.28	0.13
factor(YBLAdjusted)2005 to 2015	-11.78	-2.25
FULP	0.01	0.02

Parameter Confidence Intervals

The model residuals are more than likely nearly normally distributed as seen on the residual plot below. There does seem to be a small skew, but an insignificant number compared to most of the population.


Residuals for Question 1 Model
500
300
100
0
-200
0
2000
4000
6000
8000
10000
12000
14000

Creating an Electricity Model to Predict Electricity Costs in Oregon

Introduction

Creating a model to *predict* electricity costs is more difficult than just fitting the current data. To fit the data TI plans to fit models using the forwards and exhaustive validation set approaches and manually identify optimal model trends using Mean Squared Error, Adjusted R-Squared, BIC, and CP. These four metrics will be analyzed for each model and compared. TI will then recommend either the exhaustive or forwards validation methods based off these four-comparison metrics. To note, TI removed all apartment data for this part of the analysis as we are only interested in creating a model to predict housing costs for houses. The sample size is still quite large and

Part of the requirement for model prediction is correlation, which is shown in the first row of the scatter matrix plots in the first section. The leftmost plots are the continuous metrics, FULP, GASP, BDSP, and NP, and the rightmost plots are categorical metrics BLD, HFL, and YBL. TI will reiterate correlation trends discovered in the table below.

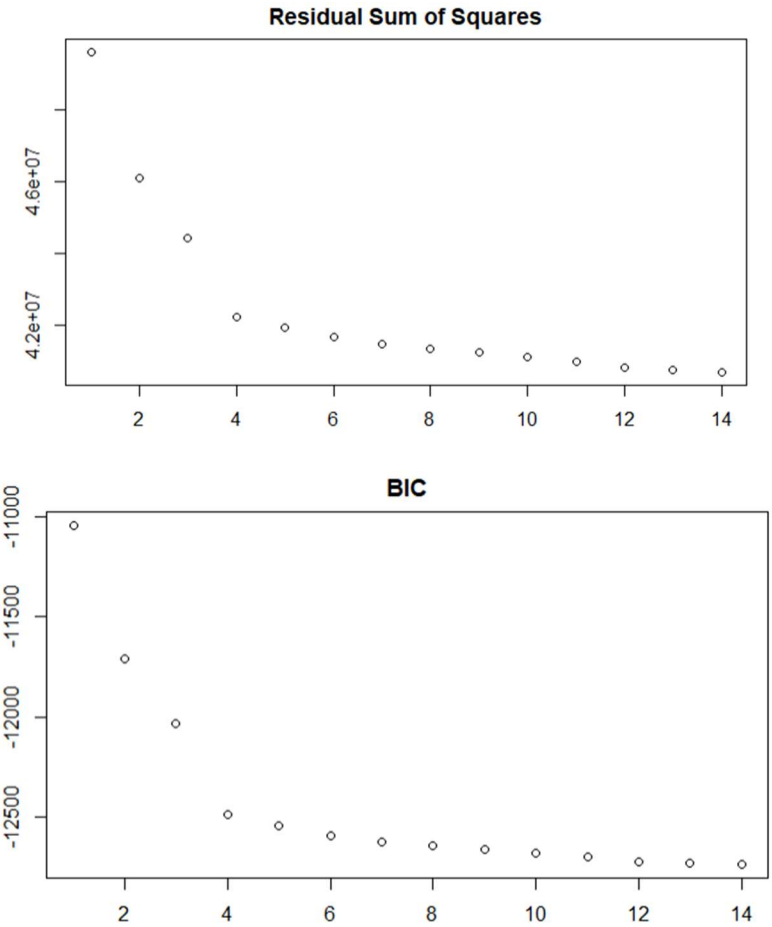
Factor	**Correlation Trend with ELEP**
FULP	Negative
GASP	Negative
BDSP	Positive
NP	Negative
BLD	Positive (House is larger ELEP than Apartment)
HFL	Negative (Electricity is larger than non-electricity)
YBL	Negative (Newer houses cost less to heat)

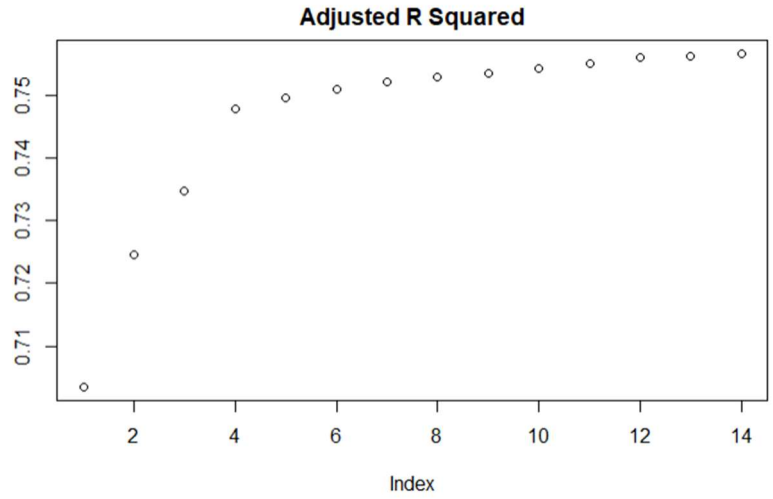
Model Selection

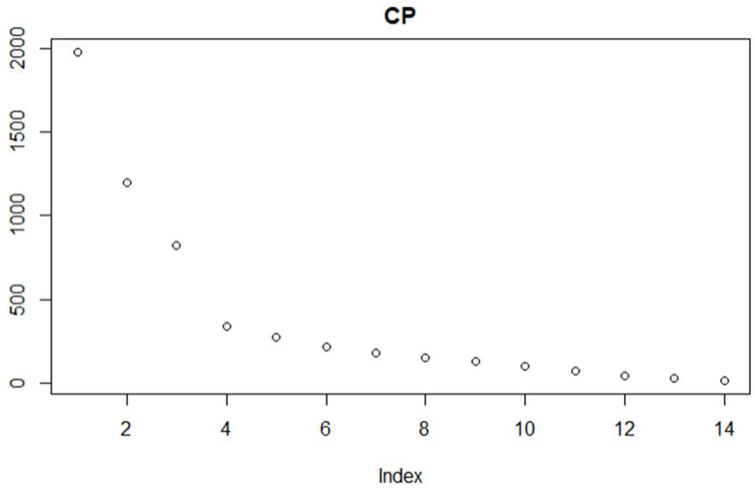
TI discovered during the analysis of the two models that both Forwards and Exhaustive stepwise regression methods produce identical coefficients. This is likely due to the large sample size of the data, allowing for values close to the true mean (Law of Large Numbers). This also eliminates the need for cross-validation. For this reason, the following analysis will be done using Forward stepwise selection to save on computation time. According to the following plots, TI recommends using 13 factors in the analysis. The list below is TI's recommended factors for prediction, along with their relative strength.

FULP	0.03
GASP	0.09
BDSP	24.82
NP	14.71
HFLAdjustedNot Electricity	-39.01
YBLAdjusted1940 to 1949	37.06
YBLAdjusted1950 to 1959	31.17
YBLAdjusted1960 to 1969	27.33
YBLAdjusted1970 to 1979	35.26
YBLAdjusted1980 to 1989	33.64
YBLAdjusted1990 to 1999	23.54
YBLAdjusted2000 to 2004	17.59
YBLAdjusted2005 to 2015	15.62

Coefficients of the Prediction Model







This analysis allows TI to recommend the following model to predict electricity costs for a house in Oregon:

Cost = 0.03 * FULP + 0.09 * GASP + 24.82 * BDSP + 14.71 * NP - 39.01 * HFLAdjusted + 37.06 * YBL 1940 to 1949 + 31.17 * YBL 1950 to 1959 + 27.33 * YBL 1960 to 1969 + 35.26 * YBL 1870 to 1979 + 33.64 * YBL 1980 to 1989 + 23.54 * YBL 1990 to 1999 + 17.59 * YBL 2000 to 2004 + 15.62 * YBL 2005 to 2015

This model has RSS: 58,182,444.45, BIC: -15,382.14, Adjusted R Squared: 0.75, and CP: 44.12.

Compare and Contrast Differences in Approaches

The differences between the previous two questions boil down to intent. In the first section TI is fitting the current data and using these data to determine if there is a difference between electricity costs in building types. This could all be done using the existing data and TI did not have to reach outside the bounds of data. In the second section, the prediction intent makes things considerably more complex. The other large difference is the removal of the apartment data in the second question. We are only interested in creating a model for houses in Oregon, so the apartment data becomes noise and can be removed. This will have an impact on the model coefficients and will change them by a few points in either direction.

Because we may have to predict values outside our current range of influence (for example, how houses built in 2025 could affect electricity cost), prediction methodologies require correlation between the explanatory variables and the dependent variable (ELEP). Due to this added complexity, there was additional issue with the analysis. TI had difficulties with implementing cross-validation on the second question, however we think with the large sample size of the data it may be unnecessary. The two section models were similar and only differed by a few factors, namely GASP. Although GASP may not have been practically significant in the first section, TI believes the positive correlation between ELEP and GASP is a necessary addition to the prediction model to assist in the power of the analysis.

One surprise was how significant the fabrication dates of the houses. It is logical looking back, as house insulation and building standards have significantly increased in the modern day allowing for less heat loss and therefore, less cost, but the initial inspection of the data was difficult to identify this trend. TI does not think we would have been able to identify this trend without conducting stepwise selection.