Modeling Electricity Consumption in the United States

STAT346 Final Project

Sebastian Gomez, Blake Bullwinkel, Mateo Fiorentino

May 23rd, 2020

Abstract

In this study, we examine the relationship between household electricity consuption and several predictor variables reflecting various residential characteristics. We use data from the 2015 U.S. Energy Information Administration Residential Energy Consumption Survey (RECS), which is a nationally representative survey of the energy characteristics, usage patterns, and demographics of 5686 housing units in the United States. Respondents were selected at random using a complex multistage area-probability sample design and answered questionnaires that were completed in one of three ways: in-person computer-assisted personal interviews (CAPI), paper questionnaires sent through the mail, and web questionnaires accessed by a URL.

Through a multiple linear regression model with normal errors, we study electricity consumption in this sample of households and extrapolate our conclusions to the overall United States population. With continous, categorical, and interacting predictor variables, we aim to account for and explain the vairance in the consumption of electricity. In particular, some of our research questions include: Can socioeconomic or racial characteristics help predict annual electricity consumption? How does annual electricity consumption depend on region? What policy might the government implement to decrease annual electricity consumption for the average household? Our findings suggest that race and region are significant, with white, Southern, and urban homeowners consuming more electricity than other demographic groups. Further, the number of household occupants and several measures of household fixtures, including refrigerators and televisions, are all significant predictors of total electricity consumption.

I. Introduction

Economic development and modern routines are highly related to electricity consumption. Over the years, electricity has become a key component of modern daily living and powers a multitude of basic activities such as lighting, cooking, heating, cooling, recreational appliances, and the internet, to name a few. Since the 1950s, electricity consumption in the United States has increased by a factor of 15 and in 2018 exceeded 3.9 trillion kilowatthours. Unsurprisingly, these trends are also closely associated with climate change and higher global temperatures. In the commercial sector, 14% (the largest share) of electricity consumption is directly related to cooling an refrigeration. Likewise, cooling accounts for the largest share of residential sector electricity consumption and increases every year. According to the U.S. Energy Information Administration, average US household electricity consumptions sits at around 11,000 kWh per year, with households in the Northeast consuming the least amount of electricity and Southern households consuming the most.

Our paper seeks to explain the variation in electricity consumption across the country via a multiple linear regression model with normal errors. Because energy consumption has a wide range of important socioeconomic implications raning from climate policy to national security, we believe that our analysis could lead to significant findings with real-world applications. Our analysis is based on data from the 2015 U.S. Energy Information Administration, Residential Energy Consumption Survey (RECS), which collects a nationally representative survey of the energy characteristics, usage patterns, and demographics of housing units in the United States. Using a complex multistage area-probability sample design, we use this nationally representative sample of 5686 housing units in the United States. Respondents answered questionnaires that were completed in one of three ways: in-person computer-assisted personal interviews (CAPI), paper questionnaires sent through the mail, and web questionnaires accessed by a URL. Among all of the variables available, we decided to analyze the following 10 predictors: Census region (Northeast, Midwest, South, West), householder race, annual gross household income, whether the house is in an urban area, number of household members, total number of full bathrooms, number of windows, whether the house has a heated swimming pool, number of refrigerators, number of televisions used, number of lightbulbs installed, and whether the house uses solar energy. The response variable is total site electricity usage in kilowatthours (kWh).

To evaluate the linear associations between these predictors and the response variable, Section I. provides a brief introduction to the research questions, Section II. describes the potential variables and their associated distributions and statistics, and Section III. introduces a multivariate linear model with normal errors to address several important research questions. To choose a final model and assess the validity of the model, Section IV. displays the diagnostic plots and concludes with a necessary transformation to the response variable in order to achieve a linear relationship and model the sample data using a multivariate linear model with normal errors. Section V. presents the final model and interprets the adjusted effects of each predictor variable given the others. Finally, Section VI. summarizes our main findings and conclusions and addresses our original reasearch questions. Further, we end with a discussion on potential areas of improvement and suggestions for future studies that may focus on this research question or use this data.

II. Exploratory Data Analysis

Table 1: Numerical summary of the continuous variables

	mean	sd	median	IQR
Total site electricity usage (kWh)	11028.93	7049.73	9549.35	8631.08
Number of household members	2.58	1.43	2.00	1.00
Number of full bathrooms	1.75	0.75	2.00	1.00
Number of windows	12.51	7.13	13.00	10.00
Number of refrigerators used	1.40	0.68	1.00	1.00
Number of televisions used	2.36	1.29	2.00	2.00

First, we describe the summary statistics and distributions of the continous variables. In Figure 1, we see that all of the continuous variables are right-skewed. This is unsurprising because we would expect that most households will have a relatively small number of features like bedrooms, refrigerators and televisions, while only a few households are likely to have higher numbers of these. Note that during data cleaning, we converted some variables to binary variables when they included extraneous information or the majority of observations took only one value (e.g. race). We also split some variables into binary indicator variables when it did not make sense to interpret them as ordered categorical variables (e.g. region).

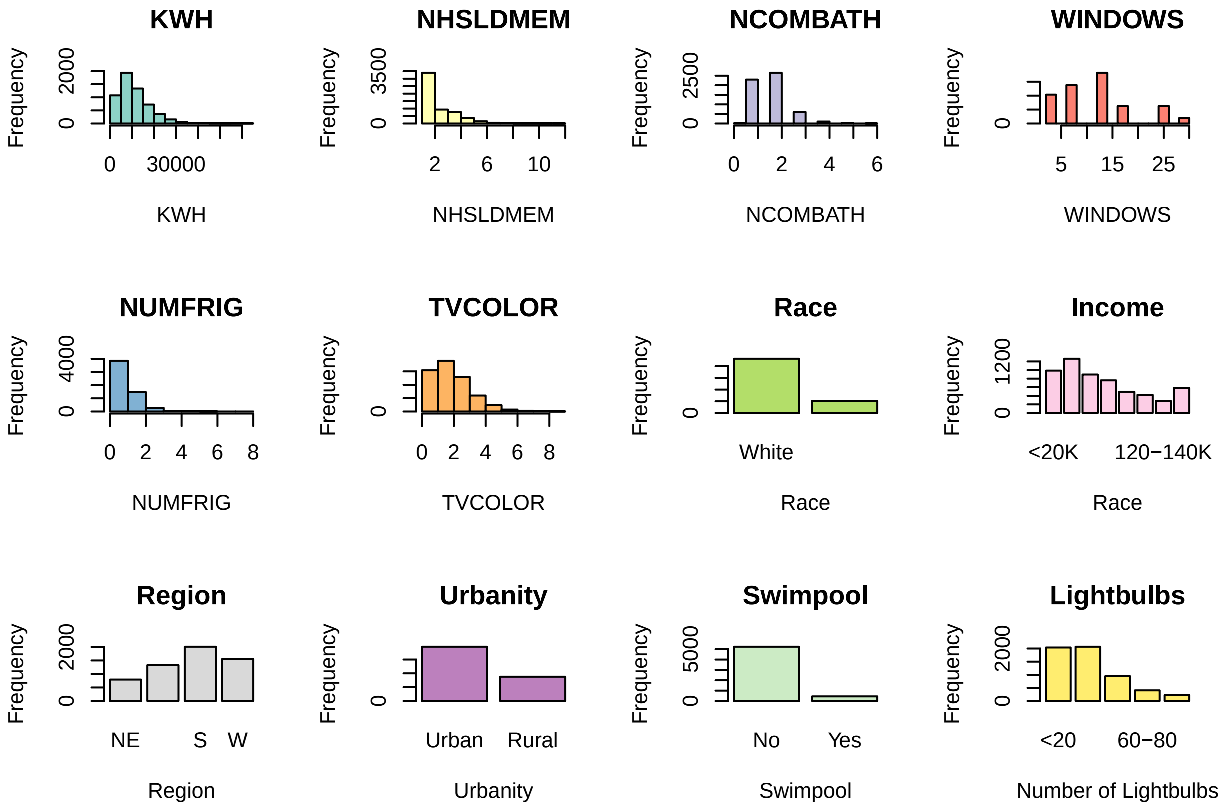

Figure 1: Univariate EDA showing histograms and barplots.

Looking more closely at Figure 1, we notice that a couple of households have very high electricity consumption while the vast majority have similar amounts on the lower end of the distribution. The majority of respondents self-identified as white (81%). However, there is considerable variation in household income, with 17% of respondents reporting less than 20K, 22% between 20-40K, 15% between 40-60K, 13% between 60-80K, 8% between 80-100K, 7% between 100-120K, 5% between 120-140K, and 10% over 140K. 14% of households live in the Northeast, 23% in the Midwest, 35% in the South, and 27% in the West. 69% of households live in an urban area whereas only 31% live in rural or non-urban areas. In addition, 92% reported not having a heated swimming pool, and 99% reported that they do not use solar energy. Finally, we see large variation in the number of lightbulbs per household with 35% having fewer than 20, 36% between 20 and 40, 16% between 40 and 60, 7% between 60 and 80, and 4% over 80.

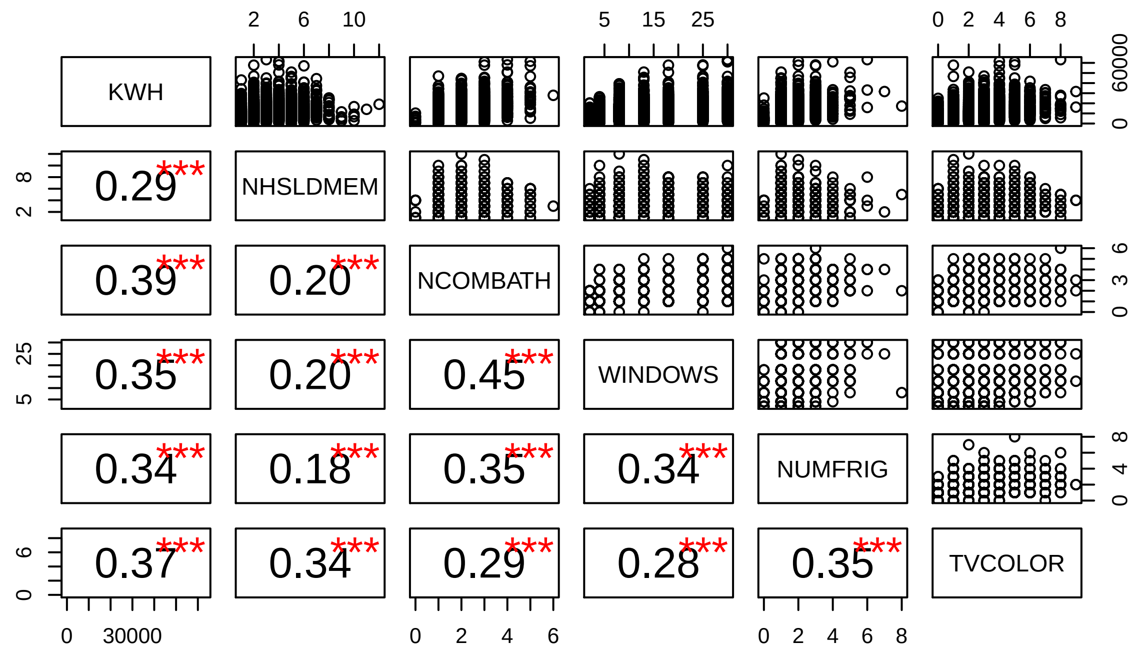

Figure 2: Pairs plot showing bivariate EDA for the continuous variables.

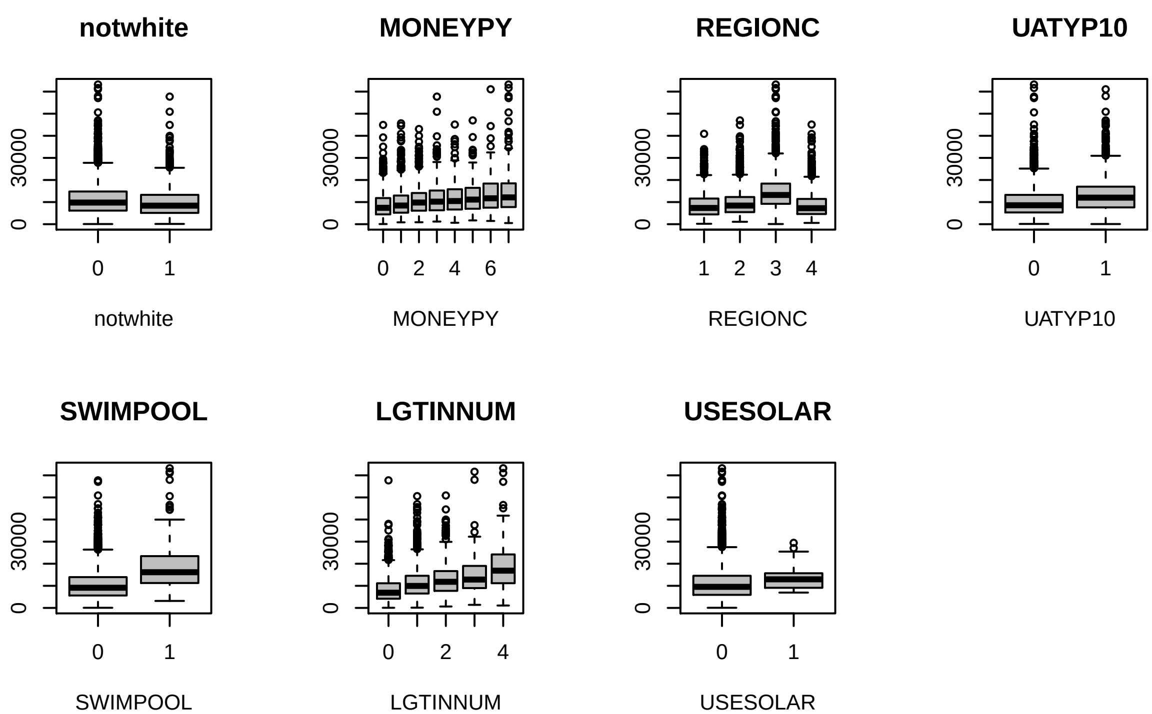

Figure 3: Boxplots showing bivariate EDA for the categorical variables.

Figures 2 and 3 show bivariate EDA via a pairs plot for the continous variables and boxplots for the categorical variables, respectively. None of our continous variables appear to be very closely related to each other, which is a good sign that we do not have a high degree of multicollinearity. We also observe that electricity usage appears to increase as the numbers of bathrooms, windows, refrigerators, and televisions increase. Given that the patterns are not entirely obvious, however, the diagnostic plots and model selection sections will help us decide which predictors are most useful in capturing the variation in electricity usage. The boxplots for the categorical variables in Figure 3 suggest that we can consider both income and number of lightbulbs as ordered categorical variables because we see fairly regular increases in electricity consumption for each category increase. It is not surprising that households located in urban areas or that have heated pools show higher median electricity consumption. Finally, it is noteworthy that electricity consumption does not appear to vary with the race of the respondent and that the Southern region reflects higher median electricity usage.

III. Initial Modeling

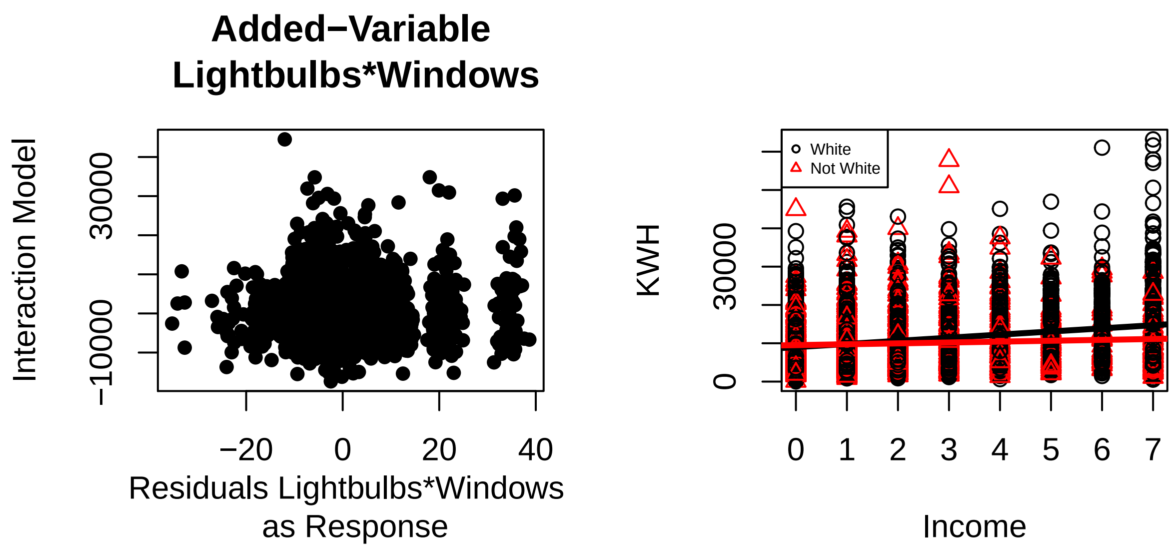

Figure 4: Plots for possible interaction terms

In our initial model, we use total site electricity usage in kWh as our response variable and include all other variables as predictors. Figure 4 shows the added variable plot and slope plot for two possible interaction terms: one between the number of lightbulbs and the number of windows in a household, and another between income and race. First, the added variable plot (which we use because both are continous variables) on the left suggests that an interaction term between the number of lightbulbs and windows will not help our model because the residuals do not show any trend, implying that adding this interaction would not help to explain any additional unexplained variance. Second, given that the lines for the two categories of race intersect in the plot on the right, we can infer that including an interaction term between race and income might help explain some variation in electricity usage. We will further evaluate whether or not to include these interactions in the diagnostics section and through hypothesis testing.

IV. Diagnostics and Model Selection

Figure 5 shows the residual plots for diagnostic purposes. It is evident from the first plot of model residuals and fitted values that the errors are not normal and do not have constant variance. This is corroborated by our boxcox transformation plot, which indicates that we should apply a transformation to the electricity consumption variable of $(KWH)^{1/3}$ in order to fix this issue and satisfy model assumptions. Figure 6 shows the same diagnostic plots after applying this transformation. As we can see, the variance is now much more constant, a significant improvement to our original, untransformed model. While the normal probability plot shows that the residuals are now aligned with the normality line, the tails are still heavy, which might indicate the presence of extreme values. The individual predictor plots suggest that the residuals are randomly scattered around 0 and the variance is constant.

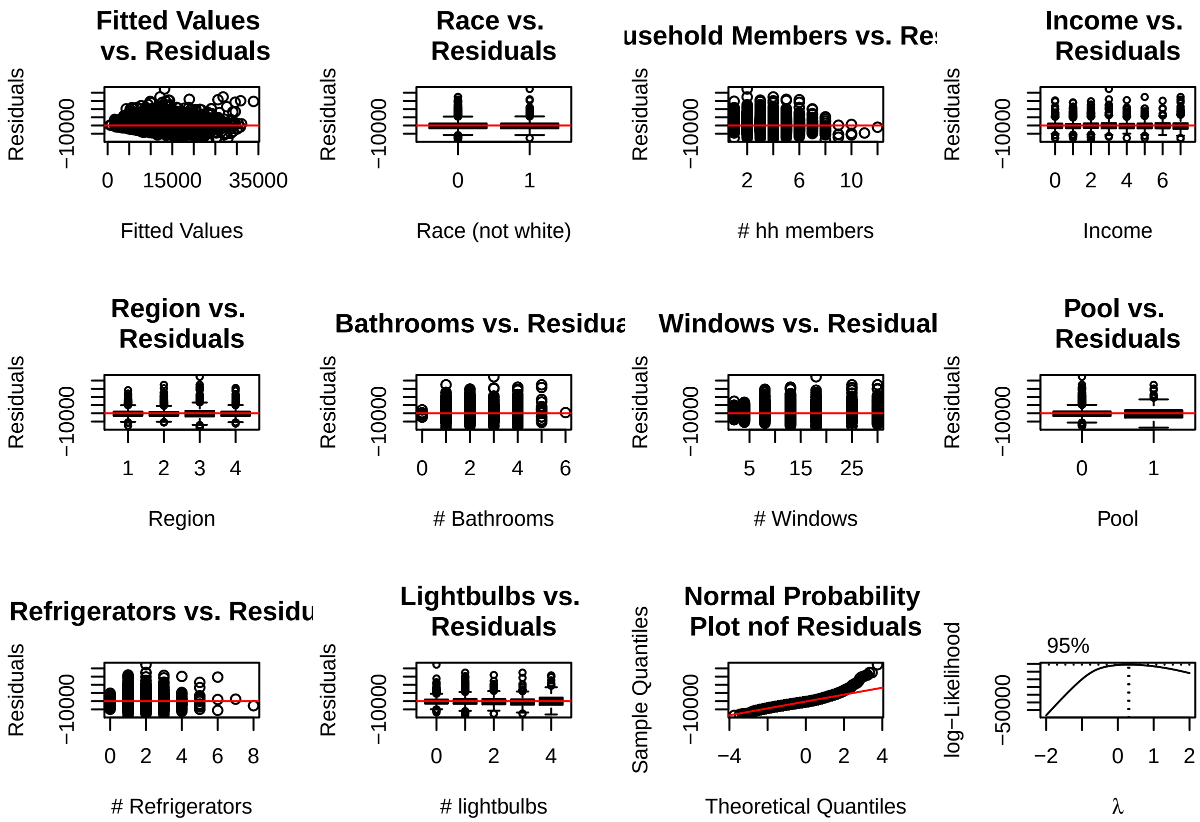

Figure 5: Diagnostic plots before transformation

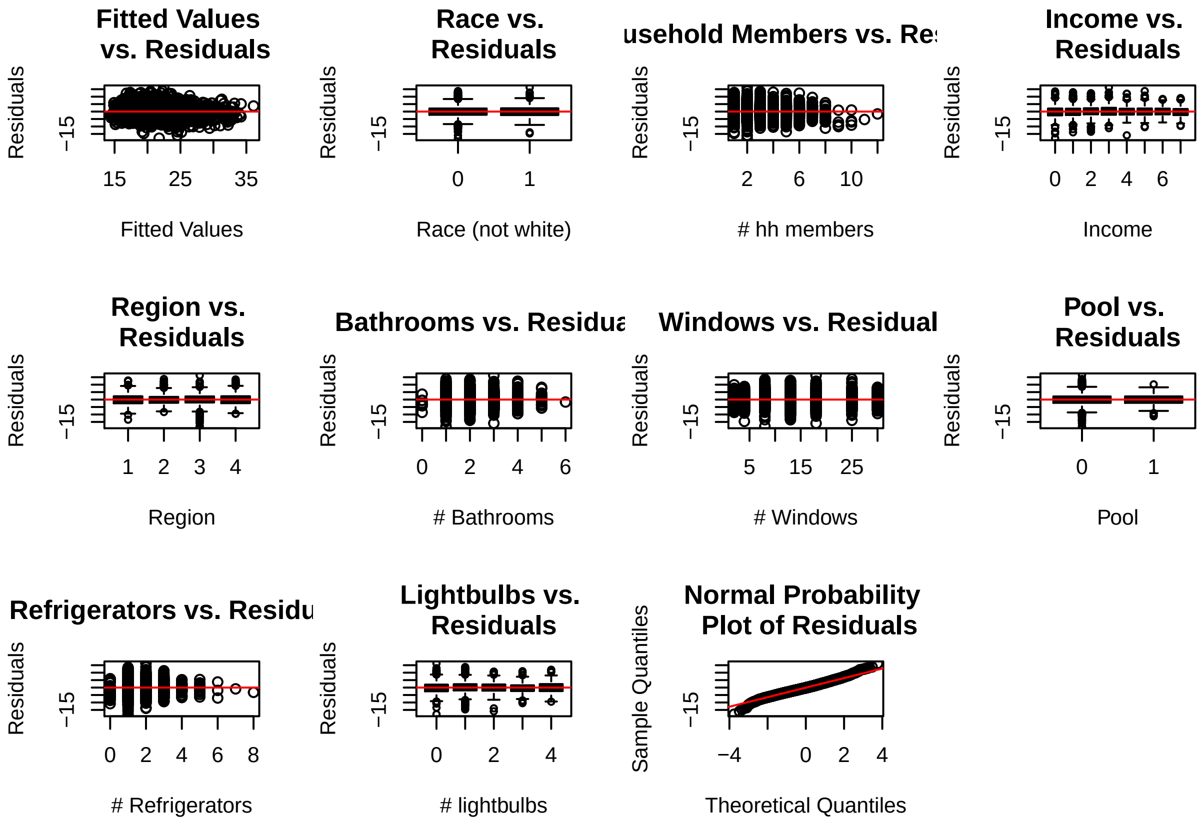

Figure 6: Diagnostic plots after transformation

To finalize our model, we consider adding interaction terms and removing predictor variables using hypothesis testing. We first use a t-test on the interaction between income and race variable. The null hypothesis is that there is no relationship between this interaction and transformed energy consumption. We obtain a p-value of 0.388 and therefore fail to reject the null hypothesis. Hence, we do not include this interaction in our model. Next, we use a partial F-test to see whether we can remove the income and solar usage variables. The null hypothesis is that there is no adjusted relationship between either of these predictors and transformed price. The test has an F-statistic of 0.164 with 2 and 5671 degrees of freedom. The p-value associated with this test is 0.152, which is greater than a significance level of $\alpha = 0.10$. Hence, we fail to reject the null hypothesis and choose to remove both of these variables from the model.

V. Final Model Inference and Results

Estimates for our final model and 95% confidence intervals for each parameter are shown in Table 2. The p-values for all 12 predictors is significant, indicating that each variable has a significant linear relationship with the response $(KWH)^{1/3}$ even after adjusting for the other 11 variables.

Table 2: Estimated coefficients for the final linear regression model

	Estimates	Standard.Error	p.value	CI.lower	CI.upper
(Intercept)	16.764	0.174	0.000E+00	16.423	17.105
notwhite	-0.544	0.120	5.722E-06	-0.778	-0.309
NHSLDMEM	0.540	0.034	4.664E-56	0.474	0.607
NE	-3.760	0.146	4.372E-139	-4.045	-3.474
MW	-2.931	0.123	1.497E-120	-3.171	-2.691
WE	-3.413	0.117	4.542E-174	-3.642	-3.183
UATYP10	1.456	0.100	1.683E-47	1.261	1.651
NCOMBATH	0.608	0.078	7.720E-15	0.455	0.761
WINDOWS	0.061	0.008	2.073E-14	0.046	0.077
SWIMPOOL	2.293	0.174	5.615E-39	1.951	2.634
NUMFRIG	0.855	0.076	9.183E-29	0.705	1.005
TVCOLOR	0.523	0.040	2.056E-38	0.445	0.602
LGTINNUM	0.460	0.057	4.954E-16	0.349	0.571

We now interpret our final model parameters in context:

- The "notwhite" coefficient indicates that households in which the respondent was not white are expected to consume $0.5436^3 = 0.1606$kWh less electricity than households with white respondents, holding all other variables constant.
- The "NHSLDMEM" coefficient indicates that an increase in the number of household members by one is associated with an increase in expected electricity consumption of $0.5402^3 = 0.1576$kWh, holding all other variables constant.
- The "NE," "MW," and "WE" coefficients indicate that households located in the Northeast, Midwest, and Western United States are expected to consume $3.760^3 = 53.15$kWh, $2.931^3 = 25.17$kWh, and $3.413^3 = 39.75$kWh less electricity than Southern households, respectively, holding all other variables constant.
- The "UATYP10" coefficient indicates that households located in urban areas are expected to consume $1.456^3 = 3.086$kWh more electricity than non-urban households, holding all other variables constant.
- The "NCOMBATH" coefficient indicates that an increase in the number of bathrooms in a household by one is associated with an increase in expected electricity consumption of $0.6077^3 = 0.2244$kWh, holding all other variables constant.
- The "WINDOWS" coefficient indicates that an increase in the number of windows in a household by one is associated with an increase in expected electricity consumption of $0.06138^3 = 0.0002312$kWh, holding all other variables constant.
- The "SWIMPOOL" coefficient indicates that households with heated swimming pools are expected to consume $2.293^3 = 12.05$kWh more electricity than households without heated pools, holding all other variables constant.
- The "NUMFRIG" coefficient indicates that an increase in the number of refrigerators in a household by one is associated with an increase in expected electricity consumption of $0.8553^3 = 0.6256$kWh, holding all other variables constant.
- The "TVCOLOR" coefficient indicates that an increase in the number of televisions in a household by one is associated with an increase in expected electricity consumption of $0.5230^3 = 0.1430$kWh, holding all other variables constant.
- The "LGTINNUM" coefficient indicates that an increase in the number of lightbulbs in a household by one level (20 lightbulbs) is associated with an increase in expected electricity consumption of $0.4604^3 = 0.09759$kWh, holding all other variables constant.

VI. Discussion

We sought to sought to understand the determinants of electrical energy consumption among American households.Our twelve-variable model explains 46.27% of the variability in transformed electricity usage ($R^2 = 0.4627$), which suggests that while these variables are significant, there remains a sizable portion left unexplained and a possible area for further inquiry. One of the most significant predictors of higher energy usage was whether a home was located in the south; this makes sense: areas at lower latitudes tend to experience warmer temperatures, which place greater energy demands on cooling systems. Perhaps one way to test this hypothesis would be to regress total electricity consumption against the relative shares attributed to cooling, heating, etc., although there may be data limitations. Another key finding was that the presence of a swimming pool was one of the most significant determinants of annual electrical consumption. Thus, if the U.S. government wanted to reduce energy consumption, it might consider imposing some type of tax on swimming pools, although this would obviously be fraught with controversy. Another way in which our analysis could be extended would be to include square footage as another variable. Many of our predictors (such as number of windows, lightbulbs, etc.) seem like they would be correlated with larger homes, so it would be interesting to see if they remain significant after accounting for this variable. Finally, given that many of the concerns related to energy usage revolve around its implications for the future, performing this type of statistical analysis on time series data could yield valuable information about energy usage trends of concern.

References

Annual Energy Outlook 2019, US Energy Information Administration, January 2019.