Relationship Between Energy Costs and the Amount of Rooms, Electronics, and Appliances a Multifamily Household Uses in 2023*

Kei Matsumoto

December 1, 2025

Energy bills are a significant factor when owning or buying a house. Appliances such as dishwashers, washing machines, and refrigerators are one of many items people use almost every day that consume energy. Aside from appliances, devices like televisions, computers, and even lightbulbs in multiple rooms will also contribute to the overall bill. With this in mind, how does the amount of rooms and appliances affect the energy cost per square feet/year for multifamily houses? A dataset by Fannie Mae was constructed from a survey that records the overall energy/water consumption of multifamily households during 2023 with characteristics of each building being the parameters. In this paper, we will be fitting a multiple linear regression to see if there is a relation between the number of room/units, bathrooms, elevators, appliances, and computers with energy fees measured in dollars per square feet per year. Overall, we were not able to conclude that there is a correlation between any of the parameters with energy cost due to an overabundance of missing values and assumption violations.

1 Introduction

When people are looking at renting an apartment or buying a house, many factors will have to be considered. For example, living space, location, number of rooms, and number of bathrooms are a couple of standard factors that people will look into. In addition to these factors, energy and water consumption are also some components that people may consider due to the energy and water bills being key payments which add on to rent. So can we predict ahead of time what our energy cost will be by the amount of rooms and appliances a household will have? According to the International Energy Agency, there have been long-running policies

[*] Project repository available at: https://github.com/keimatsumoto1/261aFinalProject

for appliances and electrical devices that are “now typically consuming 30% less energy than they would have done otherwise.” (IEA, 2021) These policies have been operating the longest in the United States and the European Union which helps appliances reduce their costs, energy consumption, and CO2 emissions. However, even though these appliances have been reducing the amount of energy consumed over the years, electricity bills are still significantly costly. Saveonenergy conducted a survey in 2022 and found that 62% of the results saw an increase in their electricity bills. (Saveonenergy, 2025) Many outside factors such as people over-using devices when not needed, using old appliances, and using more than necessary electronic devices are considerable when looking back on how people used electricity. The over-abundance of electronic devices in this current generation may contribute to a larger usage of energy even if each device has been improving their efficiency.

By being able to see if energy costs have a relation to the number of rooms and/or electronics, we can possibly be able to predict what our future energy cost will be with the house that we plan to buy/rent. It will allow new homeowners to budget how much money they will have to save for energy bills within a year. In addition, it will allow property owners of complexes like apartments to find out how much the rent will need to be through the amount of electronics the building will require.

This paper will be strictly covering multiple linear regression models as well as variable selection with no further methods. This will allow us to see if there is a correlation between energy costs and the number of room types and/or electronics to see if we can predict future multifamily building energy costs. We will conclude that there will not be a significant model that can be produced through either method due to assumption violations of normality, equal variance, and linearity. Within multiple linear regression, there is no possible way to find our prediction with our full model and through variable selection minimizing Mallow’s cp and maximizing the adjusted r-squared. The full model’s residual plot hints that there could possibly be a better model such as a polynomial regression due to the parabola-like trend being visible, with external issues of an overabundance of missing points being prominent. The multiple linear regression and variable selection will be analyzed through the programming language R and packages such as “ggplot2”, “dplyr”, “Metrics”, and “olsrr” to conduct and visualize our regression models. (R Core Team, 2023)

This paper will include the description of the data set, the description of the methods used, the core assumptions our analysis will require for a valid prediction and inference, and the explanation of our results.

2 Data

The data we are looking into for this paper is provided by Fannie Mae which is the primary provider of data to companies such as Energy Star to analyze a building’s energy efficiency. In contrast to commercial buildings, energy and building data is not federally supported and does not have an open data set to the public. To provide information to engineers, Fannie Mae started collecting multifamily building data in 2012 to further find possible ways to decrease the energy and water consumption of multifamily buildings. (Fannie Mae, 2023) The data set by Fannie Mae consists of a survey conducted in 2023 from March 1st to August 31st where Fannie Mae reaches out to many mortgage companies to ask for specific information for multifamily buildings similar to the national survey of utility consumption. In addition, Fannie Mae will reach out to properties and send property data collectors (PDC) to physically note down the key attributes of the properties. These PDCs must have extensive training and follow the "Property Data Collector Independence Requirements" to not have any bias during the data collection. (Fannie Mae, 2023)

The energy cost of a building is calculated by the total cost of energy spent within one year divided by the total area of the building in square feet. The energy spent is divided by the total area to be able to compare buildings on an equal scale. Larger buildings will need to spend more energy on appliances such as HVAC and lighting which will change the scaling of our model drastically if not divided by square feet. Energy spent includes multiple various energy sources such as electricity, oil, and gas. Unfortunately this data set does not control the appliances' and electronics' age as we do not know if the household is up to date with upgraded technology. As stated before in the introduction, newer appliances have been decreasing their energy usage but we are unable to control this for this paper.

Some factors to consider within this data is that there are many missing values within each parameter which can affect the overall skewness and quality of our analysis. The columns that consisted of the most missing values were computers and bathrooms which can be considered for removal.

In addition, this data is collected from buildings nationwide which can lead to larger variances and skewed results if certain states have more data than others. For energy consumption, location can affect the overall usages of heaters and coolers and different appliances. Within our data set, we found that most of the observations consists of California being 494 out of 2273 entries and then Massachusetts being 180 entries. This highlights that there is a large variance between the amount of data each state has in our study which can lead to our energy costs being skewed which will be further discussed within the Discussion section.

The data consists of the following parameters after being cleaned in R. The data was cleaned to strictly show the parameters of interest which directly include the usage of energy sources within the building:

- Bedrooms: The total number of bedrooms in the building.
- Units: The total amount of units in the building.
- Kitchens: The number of kitchens within the building.
- Dishwashers: The number of dishwashers within the building.
- WasherDryer: The number of washers and dryers within the building. Elevators: The number of elevators within the building.

- Bathrooms: The number of bathrooms within the building.
- Computers: The number of computers within the building.
- EnergyCost: The total cost for the energy usage of the building. Calculated by the total cost divided by the total square feet in one year. ($/sqft/yr)

3 Methods

In this paper, we will be conducting a multiple linear regression to analyze the relationship between the number of rooms, units, elevators, bathrooms, appliances, and computers, with the energy cost of multifamily buildings in 2023.

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \beta_5 X_{5,i} + \beta_6 X_{6,i} + \beta_7 X_{7,i} + \beta_8 X_{8,i} + \varepsilon_i$$

- β_0 will represent the intercept of the model which will be the expected energy cost when all parameters are 0.
- β_1 will represent the coefficient for Bedrooms which will be the expected change in energy cost when the number of bedrooms increases by 1 and all other parameters are held constant.
- β_2 will represent the coefficient for Units which will be the expected change in energy cost when the number of building units increases by 1 and all other parameters are held constant.
- β_3 will represent the coefficient for Kitchens which will be the expected change in energy cost when the number of kitchens increases by 1 and all other parameters are held constant.
- β_4 will represent the coefficient for WasherDryer which will be the expected change in energy cost when the number of washers and dryers increases by 1 and all other parameters are held constant.
- β_5 will represent the coefficient for Dishwashers which will be the expected change in energy cost when the number of dishwashers increases by 1 and all other parameters are held constant.
- β_6 will represent the coefficient for Elevators which will be the expected change in energy cost when the number of elevators increases by 1 and all other parameters are held constant.
- β_7 will represent the coefficient for Bathrooms which will be the expected change in energy cost when the number of bathrooms increases by 1 and all other parameters are held constant.

- β_8 will represent the coefficient for Computers which will be the expected change in energy cost when the number of computers increases by 1 and all other parameters are held constant.
- Y_i will represent the true energy cost of the i-th multifamily household.
- $X_{j,i}$ will represent the given constants of each parameter for the i-th multifamily household where j is 1 - 8.
- ϵ_i will represent the random error with a mean of zero and finite variance of σ^2.

We will also be conducting variable selection with the usage of olsrr's "step_all_possible" function which will calculate a Mallow's CP. We will then choose the combination of parameters that consists of the smallest Mallow's CP and highest adjusted R-squared for prediction which is the following equation. A low Mallow's cp will allow us to have the lowest bias with the best possible fit with our adjusted R-squared.

$$C_p = \frac{SSE_p}{\hat{s^2}} + 2p - n$$

SSE_p represents the sum of squared errors with p parameters. s^2 represents the full model's mean squared error with n representing the number of observations.

Some drawbacks for variable selection that we must consider are that our coefficients will be biased and overestimated which will invalidate our inferences for variables such as R^2 and confidence interval analysis. We will be conducting variable selection solely as an exploratory comparison with the full model to see if we can determine a model with valid assumptions to predict energy costs. Both model analysis will be conducted through the programming language R utilizing multiple packages such as ggplot2, dplyr, oslrr, Metrics, and readr. (R Core Team, 2023)

4 Assumptions

The assumptions we need to validate for our analysis are the following:

1. Validity where the variables are appropriate for the question we are studying.

2. Representativeness where the sample data represents the population of interest.

3. Linearity where following function must be true: $E(Y_i|X_i) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} ... + \beta_2 X_{8,i}$. Independence of errors where all errors of our model are assumed to be independent.

5. Equal variance of errors where our errors' variances must be equal throughout the model.

6. Normality of errors where our errors must be normally distributed.

7. Multicollinearity where our variables must not have a correlation with each other.

5 Results - Full Model

- $\hat{\beta_1}$ = 0.0017813, p = 0.41144
- $\hat{\beta_2}$ = 0.0097257, p = 0.09984
- $\hat{\beta_3}$ = 0.0002065, p = 0.97038
- $\hat{\beta_4}$ = 0.0003664, p = 0.71969
- $\hat{\beta_5}$ = -0.0046023, p = 0.20642
- $\hat{\beta_6}$ = 0.2681594, p = 0.00284
- $\hat{\beta_7}$ = -0.0052195, p = 0.03726
- $\hat{\beta_8}$ = -0.0101717, p = 0.59898

Within our multiple linear regression, we found that the number of elevators($\hat{\beta_6}$), and bathrooms($\hat{\beta_7}$) had the lowest p-value being all less than 0.05. This highlights that for our full model, the only coefficient that we can conclude that are possibly non-zero is the number of elevators and bathrooms. The other coefficients may truly have a relation to the energy cost of a multifamily, but we are unable to determine this with just multiple linear regression. We will not be considering the intercept of our data as a building with no rooms will not provide support for the research topic.

This means that:

- When all other parameters are held constant, an increase of one elevator will increase the yearly energy cost of the building by 0.2681594 dollars per sqft on average.
- When all other parameters are held constant, an increase of one elevator will decrease the yearly energy cost of the building by 0.0052195 dollars per sqft on average.

Aside from the p-values we found that dishwashers, bathrooms, and computers have a negative slope. These results imply that on average, an increase of each of those electronics or rooms by one with all other parameters held constant would decrease the energy cost. However, dishwashers and computers respectfully have p-values higher than 0.1 which can provide support that the number of dishwashers and computers may not have any influence to energy cost even if the slope parameters are found to be negative. Overall, without considering p-values, we can notice that all of our parameters besides the amount of elevators will not change the overall energy price by a large amount. The amount of elevators have considerably a large change to the energy cost of the overall buildings on average which could lead to elevators being a large consumer of energy.

We found that our adjusted R-squared value resulted to be 0.494 indicating that our model explains only 49.4% of the total variance of our data highlighting that our model may not be suited to provide a good fit for the data.

To check our assumptions we will look into the residual plot between fitted values and the residual values of our full model.

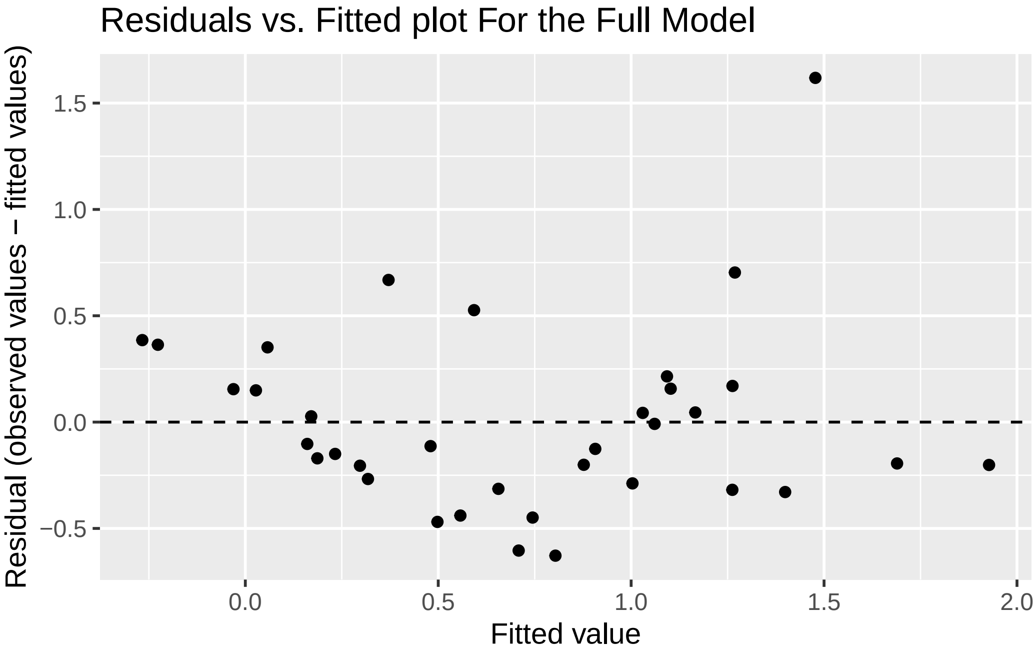

Figure 1: Residual plot of the x axis being the fitted values and the y axis being the residuals for the full model with Energy Cost as the response.

As we can see, the residual plot is not random and we can notice a parabola-shaped trend with a slight right skew. Due to this, the assumptions for linearity and equal variance will be violated suggesting that multiple linear regression is not suitable for this data. The parabola-shaped trend can possibly indicate that a polynomial relationship or other transformations might be suited for this data set. The violation of equal variance may likely cause incorrect inferences within our coefficients leading to coefficients that we may consider correlated being truly uncorrelated. The strength of our accuracies may not be trusted with unequal variances for our model.

Next we will check for the normality assumption by analyzing the QQ-plot of our multiple linear regression.

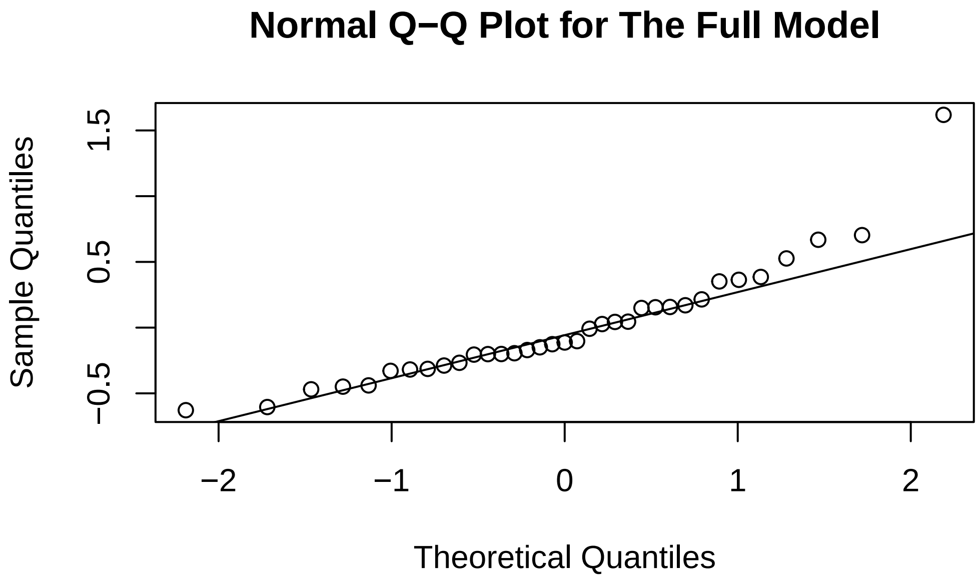

Figure 2: The QQ plot for the full model between the sample (x) and theoretical (y) quantiles with the line representing the theoretical normal distribution

By analyzing the QQ plot of our model, we can see a deviance from the normal distribution on the head of the plot. This violates the assumption that our errors are normally distributed. With errors that are not normally distributed, our p-values for our coefficients may be incorrect leading to incorrect conclusions. We can see that our head of the QQ plot being far from the normal line connects to our previous finding with the residual plot being right skewed and having heteroscedasticity.

An extreme issue within this study is that our data includes too many NA values. The multiple linear regression removed 1556 out of 1591 observations from our data only using 35 total observations. The uncertainty from our data is likely due to the sample size being significantly low violating multiple assumptions such as normality and error independence. Due to the amount of observations being 35, the assumption of normality is critical and influences our p-values for our coefficients making them unreliable.

Due to our linearity, equal variance, and error normality assumptions being violated and our number of observations being low, multiple linear regression for the full model may not be a suitable model to conduct predictions to see if the number of rooms, electronics, and appliances have an influence on a multifamily building's energy cost.

6 Results - Variable Selection

When conducting variable selection using the olsrr method "ols_step_all_possible," we found that the combination of our parameters with the lowest Mallow's cp was the number of units, number of washers/dryers, the number of elevators within a multifamily building. A low Mallow's cp will provide a model that includes the lowest bias

We will then look into a multiple linear regression model of:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i$$

Where:

- $\beta_1 X_{1,i}$ will represent the coefficient for Units which will be the expected change in energy cost when the number of units increases by 1 and all other parameters are held constant.
- $\beta_2 X_{2,i}$ will represent the coefficient for Elevators which will be the expected change in energy cost when the number of washers/dryers increases by 1 and all other parameters are held constant.
- $\beta_3 X_{3,i}$ will represent the coefficient for Bathrooms which will be the expected change in energy cost when the number of elevators increases by 1 and all other parameters are held constant.

The multiple linear regression using this model resulted with $\hat{\beta_0} = 0.868759$, $\hat{\beta_1}$ being 0.002429 with a p-value of 0.0785, $\hat{\beta_2}$ being 0.058853 with a p-value of 0.0325, and $\hat{\beta_3}$ being -0.001686 with a p-value of 0.0792.

This means that:

- When all parameters are 0, the estimate energy cost on average will be 0.868759 dollars per sqft.
- When all other parameters are held constant, an increase of one unit will increase the yearly energy cost of the building by 0.002429 dollars per sqft on average.
- When all other parameters are held constant, an increase of one elevator will increase the yearly energy cost of the building by 0.058853 dollars per sqft on average.
- When all other parameters are held constant, an increase of one bathroom will increase the yearly energy cost of the building by 0.001686 dollars per sqft on average.

Our p-values for units and bathrooms are greater than 0.05 leading to a possible lack of a correlation with energy cost. On the other hand, the number of elevators includes a low p-value being less than 0.05 similar to the full model highlighting possible relations to energy costs. With the issues with large bias from using variable selection can lead to misinterpretations with these coefficient p-values so we are still unable to say if our three selected parameters are truly related with the energy cost. We see similar trends with the variable selected model and the full model where both p-values for elevators are lower than 0.05 and having larger estimates compared to other parameters. This may lean to correlations between the number of elevators with energy cost that a different model might strongly lean to such as polynomial regression mentioned earlier.

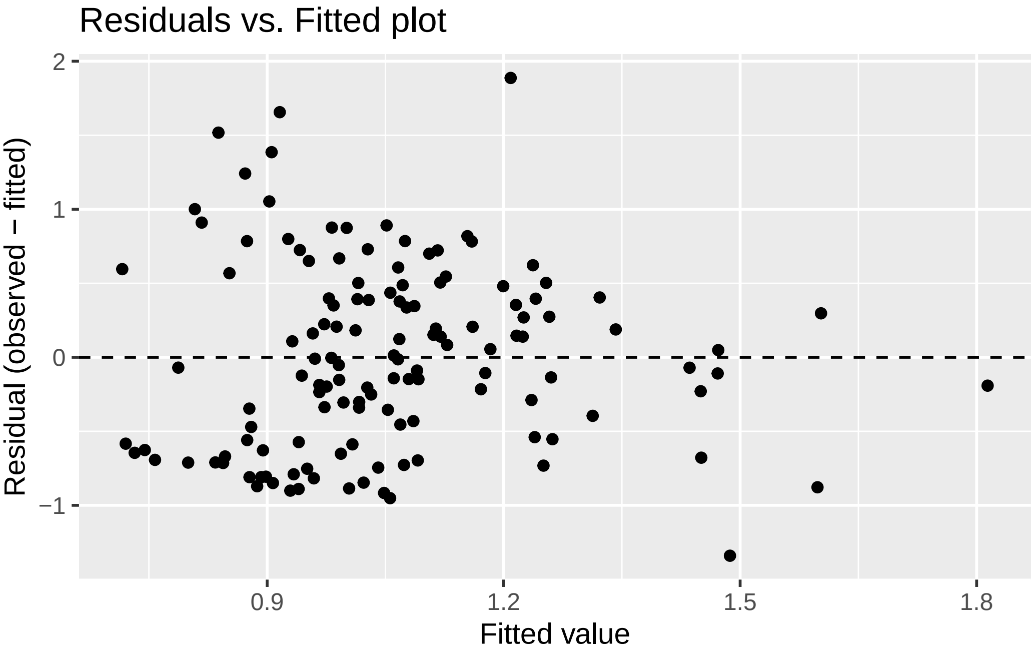

Figure 3: Residual plot of the x axis being the fitted values and the y axis being the residuals for the variable selected model with Enery Cost as the response.

By observing our residual plot of the model using variable selection, we can notice that there are no significant trends to violated linearity where all the points are randomly scattered. However, the distribution of the points are still skewed to the right and are forming a funnel shape violating the equal variance of errors assumption. This highlights that even when using variable selection to find the best possible combination of parameters, the usage of multiple linear regression may not be suited for this data. Having an unequal variance within our regression can cause misinterpretations of our coefficients. We may be concluding that our coefficients may be insignificant when they may truly be significant and have a correlation with energy costs.

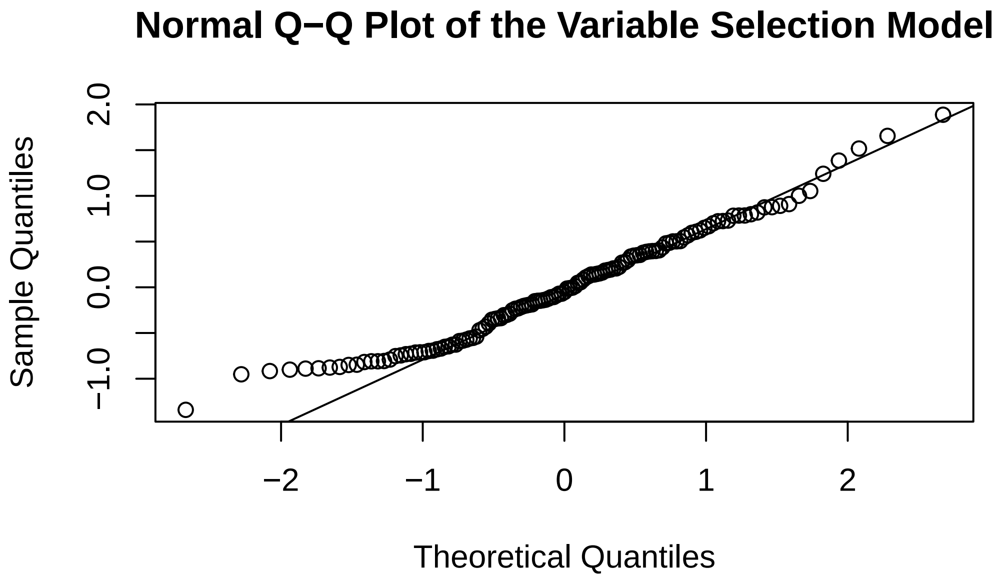

Figure 4: The QQ plot between the sample (x) and theoretical (y) quantiles with the line representing the theoretical normal distribution for the Variable selection regression model with

The QQ plot for the variable selected model is also similar scenario with the original full model where the tail and head of the data are both spread apart from the normal distribution line implying that the errors are not normal for this model as well. In contrast to the full model, there includes a larger amount of observations within the tail that is far from the normal line with most other observations staying fairly normal. By combining the two assumption violations, our data may not be suitable for predictions for energy cost with any combination of multiple linear regression. This suggests that other methods such as non-linear regression models may be preferred when predicting a multifamily building's energy cost from the amount of rooms, electronics, and appliances it includes.

7 Results - Predictions

When calculating predictions, we found that our RMSE for our full and our variable selection models ended up being 1.086643 and 0.8528844 respectively. Although these numbers may seem small at glance, the range of energy cost within our test set will vary from 0.1 to 3 making 1 and 0.85 being large errors within our scale. When comparing the overall accuracy between the full and variable selected model, we can see that our variable selected model predicts the energy costs slightly more accurate than our full model. However, due to both RMSE values being high, neither model will be preferred for predicting the overall future energy costs.

An extreme issue that we found within both the test and training set is the issue with missing values. Our predictions can only look into 12 out of 682 of our observations for our test set and 35 out of 1591 observations in our training set. When looking into the training set's predictive power, we do see similar results with both RMSEs being 1.2529 and .8112727 with poor accuracies overall. Further, the scaffolding issues with our assumption violations, missing values, and poor predictive accuracies will not allow us to see if there is a true relationship between energy costs and room, units, electronic, and appliance numbers.

8 Discussion

When observing the overall data set after conducting variable selection, a major issue within the data set is that each observation includes missing values in at least one column leading to many observations being disregarded within our regressions. A possible solution to this issue would be to remove specific variables from our multiple linear regression. However, due to many combinations of parameters also having similar issues with missing values, the removal of columns would lead to individual simple linear regressions. Although we can use simple linear regression, our conclusions can include a large bias leading to invalid connections between the predictor and the response variables directing us into incorrect results because we are not considering different parameters at the same time.

Another limitation that we have within our data is the variance in locations. The data set consists of observations primarily from California and different amounts of observations in other States. Location can be a large factor when considering our conclusions as a possible reason why our parameters do not correlate to energy cost could be due to different locations using different amounts of appliances. Weather conditions and temperature variances between States is one of many factors that can cause energy consumption variances. A study in 2023 from the U.S. The Energy Information Administration discovered that in 2020, "air conditioning accounted for 28% of total site energy usage in Florida but just 2% in Maine." (EIA, 2023) Due to different states consuming energy with different intensities, it would be difficult to conclude that one specific appliance or electronic device can lead to higher energy costs.

In addition, outside variables that are not in our data set such as car culture can affect the costs of energy. California had the most observations within our data set and according to The U.S. Department of Energy, California consisted of 35% of electric vehicle registrations within the United States in 2023. (U.S. Department of Energy, 2023) Owning an electric vehicle can increase the usage of electricity within a household from overnight and daily charging. Outside variables that are not able to be calculated within this data set can also affect why our regression concluded with a lack of correlations within our full model and any combination of variables.

Due to many outside factors, missing variables, and variance within locations, we cannot conclude whether or not the number of rooms, elevators, bathrooms, appliances, and electronics, have a relationship with the overall energy cost of a multifamily building. There can be many other methods to test to see if the amount of rooms, electronics, and appliances truly do relate to energy costs. Possible methods include tightening the study into separate States or locations to see if electronics in specific states have a relation to energy costs. This can provide less possible missing observations as well as fix the variance within locations. Another possible method can be to use a different model such as a polynomial regression. We noticed a parabola-shaped pattern within our residual plot for our full model highlighting how there can be a possible better fit for a polynomial. This can fix the heteroskedasticity we noticed in our multiple linear regression model validating the assumption of equal variance in errors allowing us to possibly conclude connections between the relationship of energy costs and our parameters.

9 References

Appliance standards and labelling is highly effective at reducing energy use, new joint study finds - News. (n.d.). IEA. https://www.iea.org/news/appliance-standards-and-labelling-is-highly-effective-at-reducing-energy-use-new-joint-study-finds

Electricity Bills By State Monthly Report | SaveOnEnergy®. (n.d.). Www.saveonenergy.com. https://www.saveonenergy.com/resources/electricity-bills-by-state/

Hamner B, Frasco M (2018). *Metrics: Evaluation Metrics for Machine Learning.* doi:10.32614/CRAN.package.Metrics https://doi.org/10.32614/CRAN.package.Metrics, R package version 0.1.4, https://CRAN.R-project.org/package=Metrics.

Hebbali A (2024). *olsrr: Tools for Building OLS Regression Models.* R package version 0.6.1.9000, https://github.com/rsquaredacademy/olsrr, https://olsrr.rsquaredacademy.com/.

How Much Electricity Do My Home Appliances Use? (2019). PublicWebsiteSitefinity. https://www.igs.com/energy-resource-center/energy-101/how-much-electricity-do-my-home-appliances-use

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

“2023 Multifamily Energy & Water Survey | Fannie Mae.” *Fanniemae.com*, 2023, multifamily.fanniemae.com/financing-options/green-financing/2023-multifamily-energy-water-survey. Accessed 10 Dec. 2025.

U.S. Department of Energy. (2022, June). *Alternative Fuels Data Center: Maps and Data - Electric Vehicle Registrations by State.* Afdc.energy.gov; U.S. Department of Energy. https://afdc.energy.gov/data/10962

U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. (n.d.). Www.eia.gov. https://www.eia.gov/pressroom/releases/press535.php

Wickham H, Hester J, Bryan J (2023). _readr: Read Rectangular Text Data_. R package version 2.1.4, https://CRAN.R-project.org/package=readr.