Multiple Linear Regression of Energy Consumption

Nathan S. Lewis

May 10th, 2025

[BTRY 6020 – Final Project]

Introduction

In 2022, the United States consumed approximately 4.07 trillion kilo-Watt-hours (kWh) of electricity. Energy consumption is typically broken down into three categories; residential, industrial, and commercial. The residential sector can be thought of as homes or places where people live and makes up the largest sector of energy consumption in the United States, with most of this energy being used for heating and cooling. The commercial sector can be thought of as businesses, offices, or other business spaces that are not directly manufacturing a good. Energy consumption in commercial spaces is dominated by computers and other electronic office equipment, cooling, and lighting. The industrial sector is defined by the operation of machinery, facilities, and other equipment in order to produce a good, this includes manufacturing plants, construction, and agriculture. Machine usage is the largest consumer of energy in the industrial sector (U.S. Energy Information Administration). The majority of the energy being consumed is from non-renewable sources, commonly referred to as fossil fuels, with 79% of the energy consumed coming from petroleum, natural gas, and coal. The remaining 21% of the energy consumed came from nuclear energy or “renewable” sources like wind, solar, and hydroelectricity (Nonfossil Fuel Energy…).

There are two approaches to cut down on America’s reliance upon fossil fuels: 1) produce more electricity using renewable sources or 2) reduce America’s electricity consumption so that less fossil fuels are needed with current renewable energy production. In an ideal world, both of these objectives would occur simultaneously. For this project, I have chosen to investigate the second option utilizing a multiple linear regression framework on an “Energy Consumption Dataset – Linear Regression

(Sriram, Govindaram)", hereby referred to as the "Energy Consumption Dataset," uploaded to the data science competition platform 'Kaggle (Kaggle)' to explore what covariates drive energy consumption across residential, commercial, and industrial buildings.

The "Energy Consumption Dataset" contains data on seven features; Building Type, Square Footage, Number of Occupants, Appliances Used, Average Temperature, Day of Week, and Energy Consumption. I have chosen to use Energy Consumption, measured in kWh, as the variable to predict. Building type, either residential, commercial, or industrial, square footage, number of occupants, appliances used, average temperature on the outside of the building, and day of week, which is either weekday or weekend, as my predictor variables (covariates).

Methodology

R version 4.4.3 (R Core Team) was used for all analyses within the RStudio interface(RStudio Team). The Energy Consumption Dataset originally contained a split of 1000 datapoints of each variable in a "train" dataset and 100 datapoints of each variable in a "test" dataset, I began by combining these two datasets so that I could use K-fold and leave one out cross validation procedures described later. I then began exploring the dataset by looking at a summary of each of the variables shown in Table 1. I then verified that there were no missing values within the dataset. The Energy Consumption variable seemed to be normally distributed around 4200 kWh (Figure 1). I also plotted box plots of the Energy Consumption variable, as well as the Energy Consumption between the commercial, industrial, and residential as well as the weekday/weekend categories (figures not shown but can be found in the annotated

Table 1. Summary statistics of variables

	Min	First Quartile	Median	Mean	Third Quartile	Max
Square Footage	560	13204	25786	25501	37537	49997
Number of Occupants	1	22	47	48	73	99
Appliances Used	1	13	26	26	38	49
Average Temperature (°C)	10	16	23	23	19	35
Energy Consumption (kWh)	1684	3510	4190	4168	4860	6531

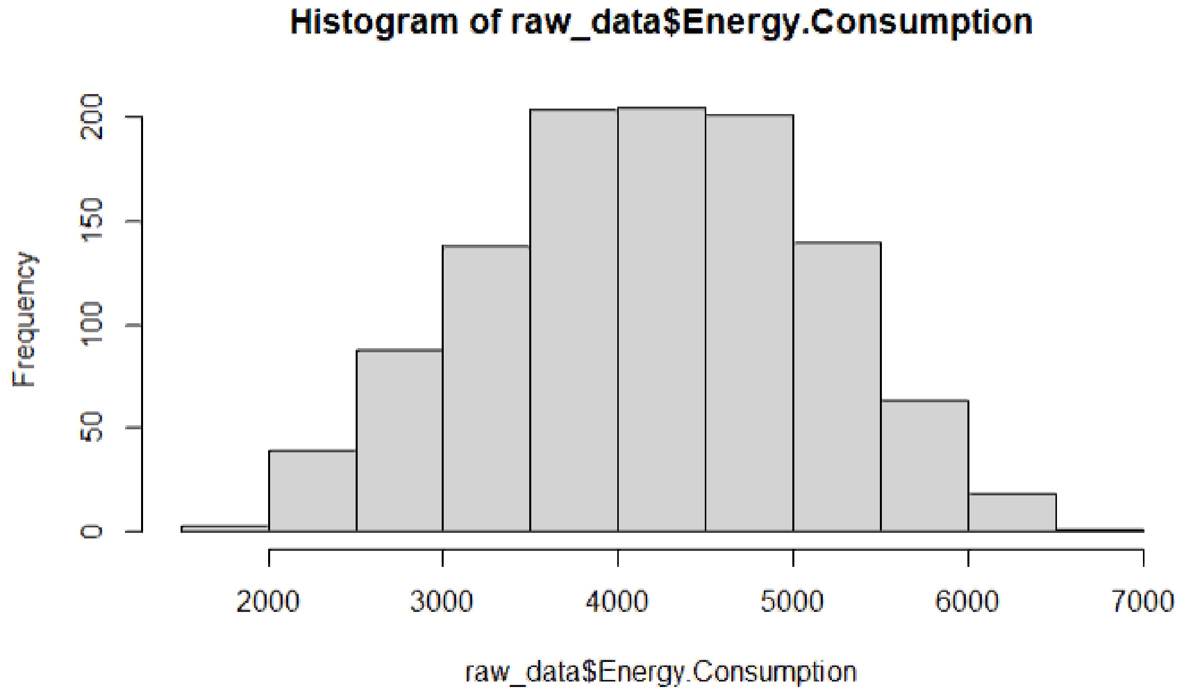

Figure 1. Histogram of Energy Consumption

R code), I was unable to differentiate between the median and spread of the weekday/weekend variable, but could tell a difference among the medians of residential, commercial, and industrial energy consumption, although their spreads seemed close to identical. In the last step of my exploratory data analysis, I plotted a correlation matrix among each of the numerical variables using the R package ‘corrplot (Wei and Simko)’ (Figure 2). To do this, I filtered out the categorical variables from the dataset. As shown in Figure 2, the strongest correlation was between Square Footage and Energy Consumption. Energy Consumption seemed to be mildly correlated with Appliances Used and the Number of Occupants. There did not appear to be any multicollinearity among the predictor variables which was also checked using the “pairs” function in R

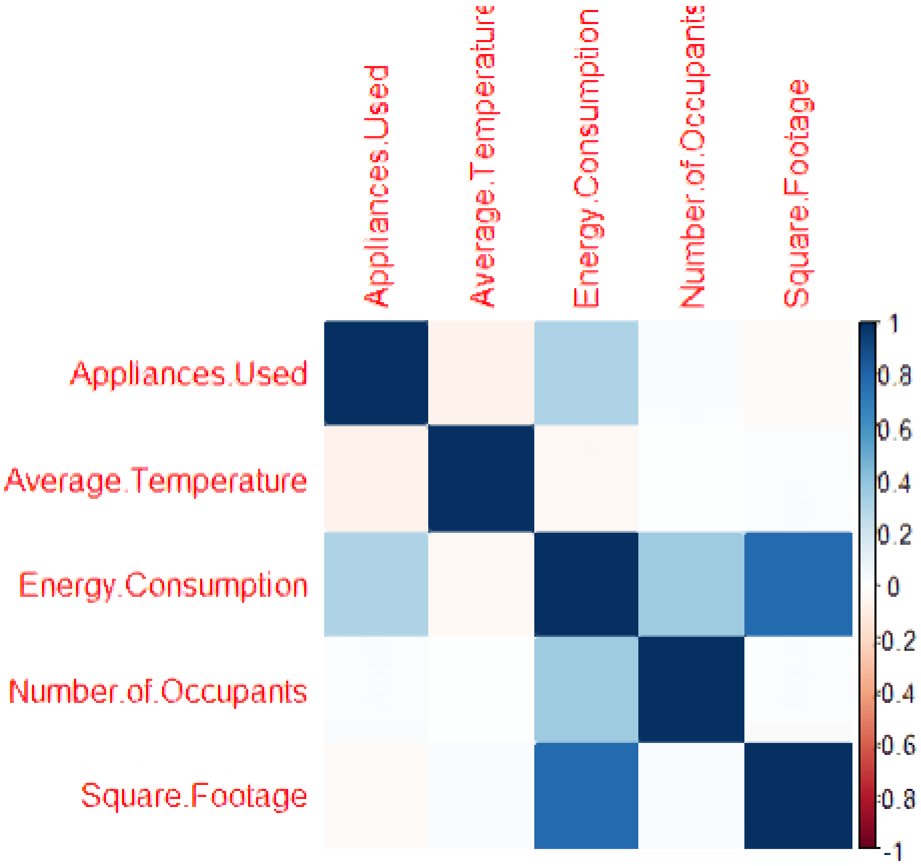

Figure 2. Correlation Matrix of each of the numerical variables

(data not shown). For the linearity assessment, I plotted each of the numerical variables against energy consumption (Figure 3). From these plots, it appeared that all of the predictors had a linear impact on yield. The strongest trend was with square footage, which was the same as Figure 2, appliances used and number of occupants had a slightly increasing effect on energy consumption, and increasing temperature appeared to cause energy consumption to decrease linearly. To check for homoskedasticity, I first plotted the residuals against the fitted values for all six predictor variables. Square footage and building type are shown in Figure 4, while the remaining plots can be found in the detailed R code. For all four of the numerical predictors, the residuals and fitted values appeared random and there was no evidence of funneling. For the two categorical variables, I was unsure of what exactly to be looking for, however, the

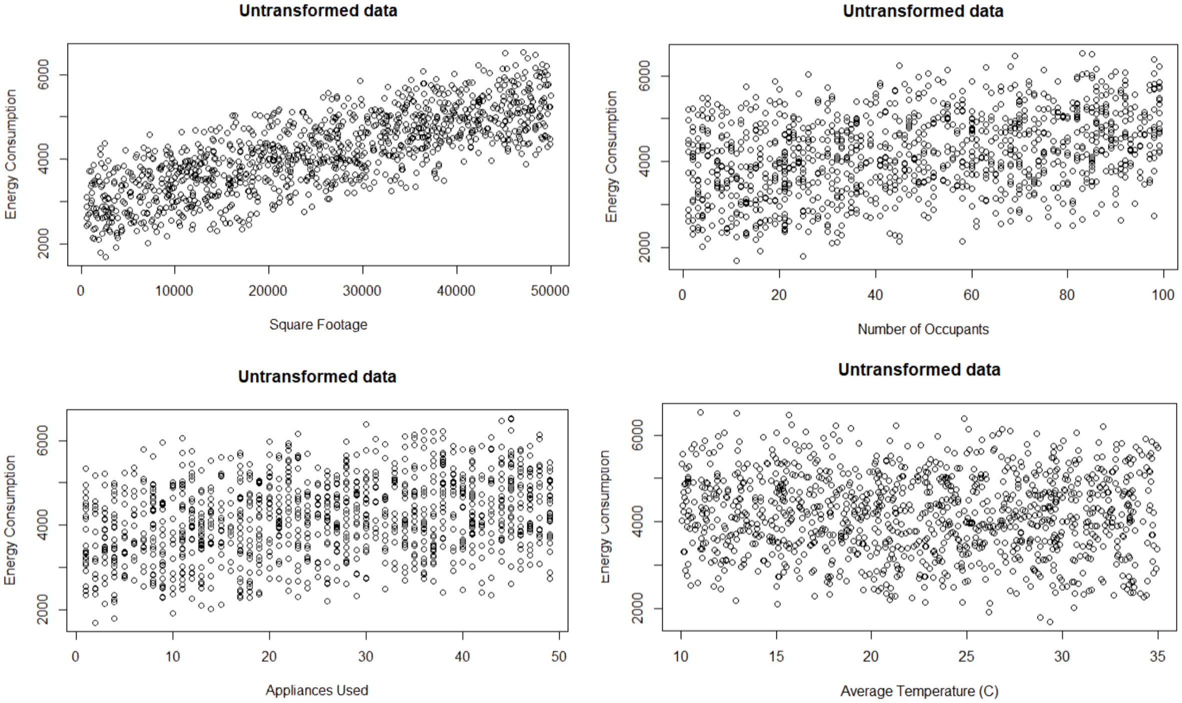

Figure 3. Plots of Energy Consumption against Square Footage (top left), Number of Occupants (top right), Appliances Used (bottom left), and Average Temperature (bottom right)

spread of the residuals appeared to be consistent for each of the building types. The same trend was true for the day of week. To assess for normality of residuals, I plotted Q-Q-plots for each of the predictor variables on energy consumption by fitting a linear model with each of the predictor variables on the Energy Consumption, individually. The plot for square footage is shown in Figure 5 and the same trend was true for number of occupants, building type, appliances used, average temperature, and day of the week. As is illustrated in Figure 5, there was deviance away from the theoretical Q-Q-line for each of the predictor variables, however, since n is large (1100 >> 30), the Central Limit Theorem states that there should be a normal distribution of residuals. I attempted to verify the independence of observations by looking into the meta data on Kaggle, but was unable to find any information on how this dataset was collected, as such, there is no way of verifying that there was an independence of observations and, in theory, the same building could have been collected in multiple months, years, etc. Since the untransformed dataset passed all of the regression assumptions; normality of residuals, linearity, homoskedasticity, and no multicollinearity and that I was I have no idea on the independence of observations, I did not apply any transformations to the dataset, as no

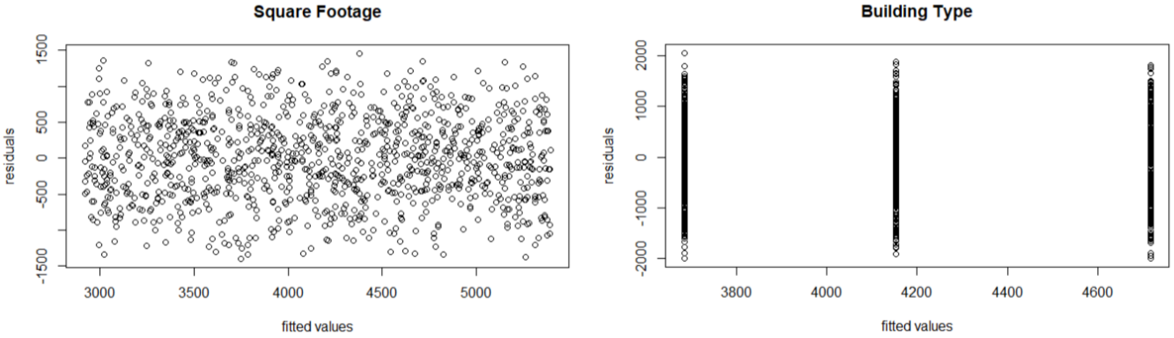

Figure 4. Residual vs Fitted Value plots of Square Footage and Building Type on Energy Consumption

transformations would fix this issue. I could have used mixed models if I knew how the dataset was generated and thus knew how the independence observation was violated, however, this was not the case.

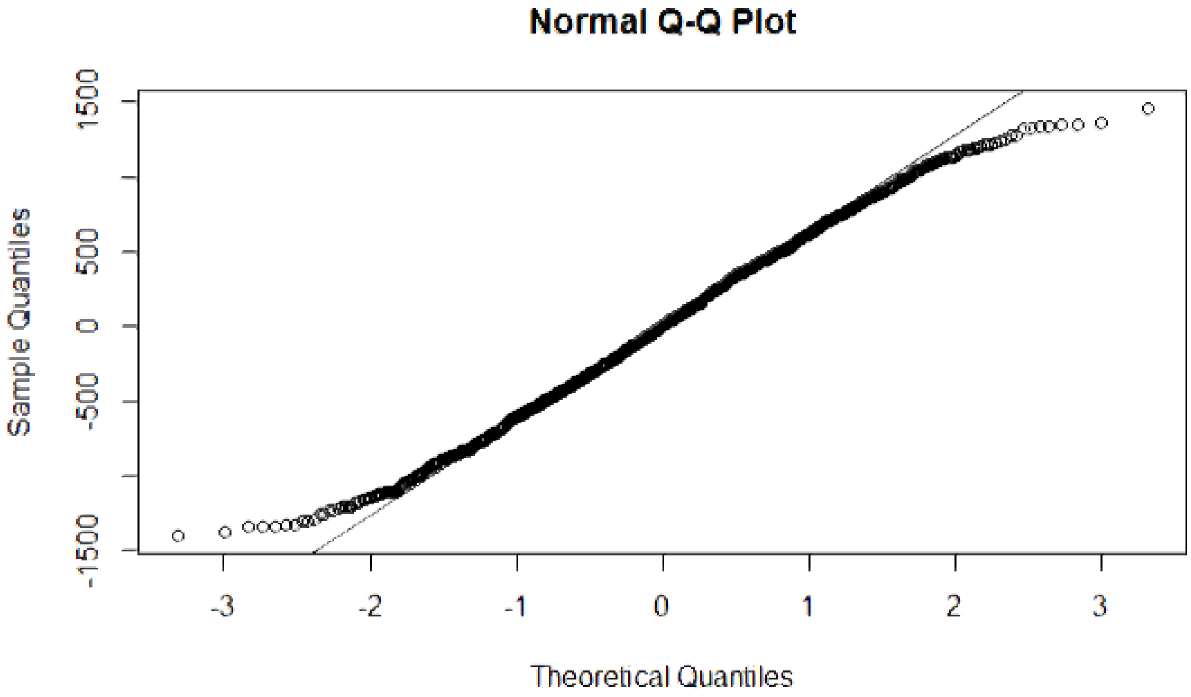

Figure 5. Q-Q Plot of the Square Footage residuals

For variable selection, I initially started with a Forward Selection approach using the ‘lm’ function and assessed the model performance using the Bayesian Information Criterion (BIC). However, when doing the second variable selection procedure using cross validation, I was unable to figure out how to use BIC with this approach and could only figure out how to use the Akaike Information Criterion (AIC). For this reason, I went back to the Forward Selection approach and assessed the model performance using AIC rather than BIC. I would have preferred to use BIC over AIC when assessing model performance since AIC can lead to overfitting and since n is large. For my second approach, I chose to use a cross-validation procedure and added covariates as they improved the model using the ‘glm’ and ‘cv.glm’ functions within the ‘boot’ package (Canty and Ripley). For cross validation, I split the dataset into 11 folds of 100 length and assessed the model performance using both the AIC value as well as the error output from the ‘cv.glm’ function. I continually added covariates to the model until performance did not increase. Not to spoil to much of the results, but I ended up with the same model as the forward selection approach with this procedure, and decided to create a third model for this reason. For the third model, I went back to my code used to generate Figure 4, where I had made a linear model of each covariate on Energy Consumption as a starting point. I saw that the “day of week” and “average temperature” were not significant at alpha = 0.05 when a linear model was fit and fit my third model using only the covariates that were significant at alpha = 0.05 individually. While this is certainly not the most statistically correct approach since I “cherry-picked” predictors that were only significant individually, it was the only way I could think of to create a model different from the forward selection and cross validation approaches. I had attempted to use a backward selection and branch and bound procedure prior to the “cherry-picking” approach, however, these both led to an identical model as the first two generated models. For all three models, after the covariates were selected, I assessed their performance with cross validation using the K-fold method with 11 folds. For the third model, I also assessed the performance using a Leave One Out Cross Validation approach, mainly for my own amusement/curiosity, since my personal computer was able to do this relatively quickly (less than five seconds). This LOOCV result was compared to the K-fold method for third model, but was not compared to any of the other models since it was not used on those models.

Results

Model 1. Forward Selection

For the forward selection model using AIC, the final model fit used all six covariates and the outputs of the summary of the model are shown in Table 2. The AIC value started at 17159 for the initial model with just the square footage, decreased to 10217 in the model with all predictors except Day of Week, and dropped below zero to -6300 when all covariates were included. When the K-fold cross validation procedure was run using 11 folds, the error output from the 'cv.glm' was 0.00019.

Table 2. Summary table of the final Forward Selection model, * indicates significance at alpha = 0

	Estimate	Std. Error	t value	Pr (>|t|)	
(Intercept)	2.05e+3	2.01e-3	1020720	<2e-16	*
Square.Footage	5.00e-2	2.92e-8	1714321	<2e-16	*
Building.TypeIndustrial	5.00e+2	1.03e-3	484955	<2e-16	*
Building.TypeResidential	-5.00e+2	1.01e-3	-486750	<2e-16	*
Number.of.Occupants	1.00e+1	1.43e-5	701815	<2e-16	*
Appliances.Used	2.00e+1	2.95e-5	678706	<2e-16	*
Average.Temperature	-5.00e+0	5.84e-5	-85556	<2e-16	*
Day.of.WeekWeekend	-5.00e+1	8.30e-4	-60266	<2e-16	*

Model 2. Cross Validation approach to Variable Selection, similar to the approach taken in Lab 9 under “Penalized Scores”

For the cross validation approach to variable selection, I mirrored the approach taken in Lab 9 under Penalized scores where covariates were added until model performance decreased. In this case, model performance never decreased and the best model was fit when all covariates were added. The AIC values followed a similar trend as described under Results Model 1, however, I was able to assess the errors at each step using this approach. The error from the ‘cv.glm’ using K = 11 with the K-fold approach began at 347179 when just Square Footage was included and dropped to 0.00019 (identical to Model 1 final model) when all covariates were included. Since the same model was fit, the summary table was identical to that of Model 1 (Table 2) and as a result that table is omitted here.

Model 3. Manual selection of Covariates based on their individual predictions

For the model that was generated using the four covariates that were significant at alpha = 0.05 on their own the final model was fit using the intercept, industrial and residential building types (commercial was in the main model), square footage, number of occupants, and appliances used. I also assessed the model performance using the K-fold cross validation procedure with 11 folds and had an error term of 1865 using this approach. Additionally, I used the LOOCV approach out of curiosity and found a slightly lower error term of 1859.39. The summary output table of this model is shown in Table 3.

Table 3. Summary table of the manually selected model, * indicates significance at alpha = 0

	Estimate	Std. Error	t value	Pr (>|t|)	
(Intercept)	1.91e+3	4.51e+0	424	<2e-16	*
Square.Footage	5.00e-2	9.12e-5	154	<2e-16	*
Building.TypeIndustrial	4.97e+2	3.22e+0	-160	<2e-16	*
Building.TypeResidential	-5.02e+2	3.15e+0	548	<2e-16	*
Number.of.Occupants	9.98e+0	4.46e-2	224	<2e-16	*
Appliances.Used	2.02e+1	9.20e-2	220	<2e-16	*

Discussion

As I am sure you have already realized and what I first became suspicious of after running the forward selection model, I believe that the dataset that I chose for this project was simulated using a random number generation approach for the Square Footage, Number of Occupants, Appliances Used, and Average temperature and then arbitrarily listed the building type and day of week, and then used all of information to actually calculate the energy consumption variable. Without this being the case, I cannot understand how I would have created multiple “perfect” models with no error. What initially tipped me off was when I saw in Table 2 that an Industrial Building Type increased the energy consumption as compared to a Commercial Building Type by exactly 500 and a Residential Building Type decreased the energy consumption by exactly 500 compared to a Commercial Building Type. All of the estimated effects of the covariates in Model 1 and Model 2 went out ended with 00 to the thousandths place, which was another tip off. Lastly, the t values were beyond anything I could have imagined. Summarizing the estimated effects briefly for these two models, an increase in one square foot lead to an increase of 0.05 kWh of energy consumption, adding one occupant lead to an increase of 10 kWh of energy consumption, adding an appliance lead to an increase of 20 kWh of energy consumption, an increase of one °C to the average temperature lead to a decrease of 5 kWh of energy consumption, and the day being a weekend lead to a decrease of 50 kWH as compared to if it were a weekday. The results of Model 3 for adding an additional square foot, occupant, appliance, or the building type were nearly identical for Model 3 and are shown fully in Table 3. Hypothesis tests were computed using the 'summary' function within R and all tests of no correlation (estimate = 0) were rejected at all alpha's tested, since P always equaled 2e-16. Confidence intervals at alpha = 0.05 were created for each covariate in each model and are summarized in Table 4. As is evident, these confidence intervals, especially for Models 1 and 2, are incredibly tiny, which I believe is further evidence that the Energy Consumption variable in this dataset was generated by making a calculation of the other variables and then added to/subtracted from based on the categorical variables.

Table 4. Confidence Intervals for each model and covariate pair.

Model 1.

	2.5%	97.5%
(Intercept)	2049.9977	2050.0056
Square.Footage	0.0499	0.0500
Building.TypeIndustrial	499.9983	500.0023

Building.TypeResidential	-500.0019	-499.9979
Number.Of.Occupants	9.9999	10.0000
Appliances.Used	19.9999	20.0001
Average.Temperature	-5.0001	-4.9999
Day.Of.WeekWeekend	-50.0028	-49.9995
Model 2.		
	2.5%	97.5%
(Intercept)	2049.9977	2050.0056
Square.Footage	0.0499	0.0500
Building.TypeIndustrial	499.9983	500.0023
Building.TypeResidential	-500.0019	-499.9979
Number.Of.Occupants	9.9999	10.0000
Appliances.Used	19.9999	20.0001
Average.Temperature	-5.0001	-4.9999
Day.Of.WeekWeekend	-50.0028	-49.9995
Model 3.		
	2.5%	97.5%
(Intercept)	1902.2737	1919.9377
Building.TypeIndustrial	490.2685	502.9064
Building.TypeResidential	-508.0288	-495.6955
Square.Footage	0.0498	0.0501
Number.Of.Occupants	9.8908	10.0656
Appliances.Used	20.0196	20.3801

Conclusion

The fact that the energy consumption variable was likely calculated from all of the other variables was a major limitation within this study and prevents any conclusions from being drawn since this is more than likely an entirely simulated dataset. In summary, within this dataset, an increase in square footage, number of occupants and appliances used each lead to an increase in energy consumption. An increase to the average temperature outside the building lead to a decrease in energy consumption. The building type being industrial lead to an increase of about 500 kWh of energy consumption as compared to a commercial building type, whereas a residential building type lead to a decrease of about 500 kWh of energy consumption compared to a commercial building type. The day of the week being a weekend lead to a decrease of about 20 kWh as compared to a weekday. While the findings of this study seem to match what is likely going on in the real world, there is no way of confirming that this is the case and the creator of the dataset likely used this implicit bias when creating the dataset. A similar approach as in this study could be taken with a real dataset on energy generation to potentially find ways of reducing the electricity consumption of different building types and I would imagine is controlled by many more covariates than this study took into account.

As a note, I had put about 6 or 7 hours into searching different datasets, going through the exploratory data analysis, regression assumptions verification, and assumptions handling, in addition to writing those respective sections of this written report as well as the introduction before I ran the forward selection model and realized that my dataset was likely simulated and “too good to be true” for linear regression.

Since linear regression was in the name of the dataset in Kaggle, it seemed like a good choice. Since I had put so much time into this dataset, had already switched datasets once after noticing extreme levels of multicollinearity in my other dataset (blueberry yield prediction) and noticed that that dataset was ALSO generated from a simulation, and am in the middle of writing a research proposal for my Qualifying Conference that is critical to my completion of my PhD, I decided to stick with dataset rather than attempting to find a third dataset. I hope that I am not penalized too heavily for using this dataset that lead to a perfect MLR model and believe that my interpretations of each of the covariates, approaches to the regression assumptions, exploratory data analysis approaches, and variable selection methods are correct, even though I am unable to have any practical significance of these findings that are generalizable to the real world.

References

Canty, Angelo, and Brian Ripley. boot: Bootstrap Functions (Originally by Angelo Canty for S). R package version 1.3-30, 2022. https://cran.r-project.org/package=boot

“Kaggle.” Wikipedia, Wikimedia Foundation, 16 Apr. 2025, en.wikipedia.org/wiki/Kaggle.

“Nonfossil Fuel Energy Sources Accounted for 21% of U.S. Energy Consumption in 2022 - U.S. Energy Information Administration (EIA).” Nonfossil Fuel Energy Sources Accounted for 21% of U.S. Energy Consumption in 2022 - U.S. Energy Information Administration (EIA), www.eia.gov/todayinenergy/detail.php?id=56980#:~:text=Fossil%20fuels%E2%80%94petroleum%2C%20natural%20gas,high%2013.2%20quads%20in%202022. Accessed 11 May 2025.

R Core Team. R: A Language and Environment for Statistical Computing. Version 4.4.3, R Foundation for Statistical Computing, 2024. https://www.r-project.org

RStudio Team. RStudio: Integrated Development Environment for R. RStudio, PBC, 2024. https://posit.co

Sriram, Govindaram. “Energy Consumption Dataset - Linear Regression.” Kaggle, 6 Jan. 2025, www.kaggle.com/datasets/govindaramsriram/energy-consumption-dataset-linear-regression?resource=download.

“U.S. Energy Information Administration - EIA - Independent Statistics and Analysis.” Use of Electricity - U.S. Energy Information Administration (EIA), www.eia.gov/energyexplained/electricity/use-of-electricity.php#:~:text=Total%20U.S.%20electricity%20consumption%20in,year%20decreases%20occurred%20after%202007. Accessed 11 May 2025.

Wei, Taiyun, and Viliam Simko. corrplot: Visualization of a Correlation Matrix. R package version 0.92, 2021. https://cran.r-project.org/package=corrplot