Zuriahn Yun, Jordan Van

Yunz@wwu.edu

Winter 2025, Math 342

March 21st 2025

Quantifying the Effects of Building Features on Energy Usage

This project focuses on multilinear regression analysis to predict energy consumption(kilowatt-hours) based on various building features and environmental factors. Our goal is to develop an accurate model that helps in understanding how different variables contribute to the energy needs of commercial and residential buildings. By analyzing factors such as building type, square foot , number of occupants , appliances used , average temperature and day of the week, we can identify trends and patterns that influence energy usage. This information could be valuable to measure a building's carbon footprint, optimizing energy efficiency or estimating infrastructure costs. By leveraging multilinear regression , we aim to analyze a model that will effectively quantify the impact of our features on energy consumption.

Regression Model:

ENERGYCONSUMPTION = 2500.01 * SQUAREFOOTAGE + 10 * NUMBEROFOCCUPANTS + 19.9999 * APPLICANCESUSED - 4.99975 * AVERAGETEMPERATURE + 49.9939 * DAY - 1000 * TYPEONE - 499.997 TYPETWO

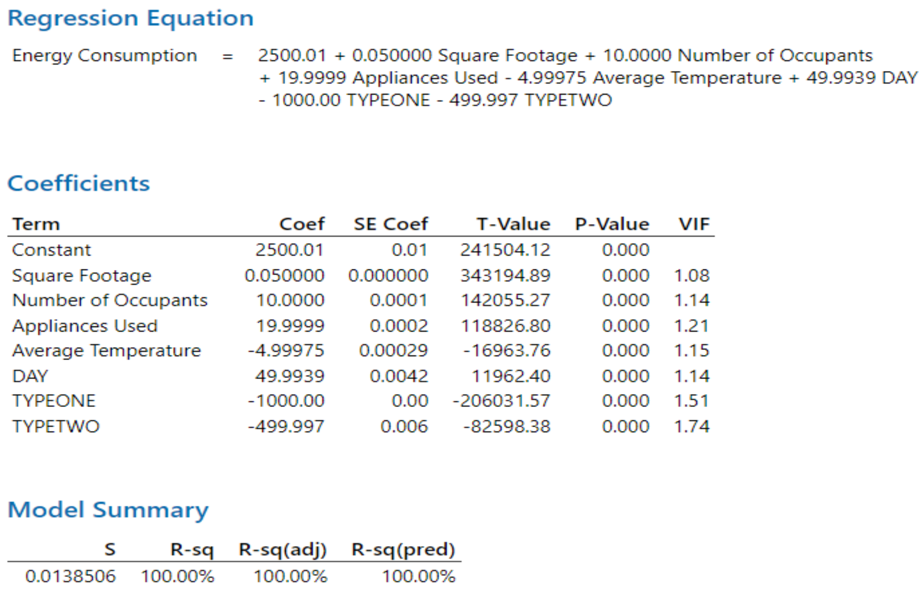
Response

- ENERGYCONSUMPTION(a numerical variable representing the amount of energy consumption measured in Kilowatt-hours)

Predictors

- SQUAREFOOTAGE(a numerical variable representing the square footage of the building measured in square feet)
- NUMBEROFOCCUPANTS(a numerical variable representing the number of people who reside in the building)
- APPLIANCESUSED(a numerical variable representing the number of appliances used in the building)
- AVERAGETEMPERATURE(a numerical variable representing the temperature in or around the building measured in celsius)
- DAY (a dummy variable that is 1 for weekday and 0 weekend)
- BUILDINGTYPE(if both TYPEONE and TYPETWO are 0 the BUILIDINGTYPE is industrial)
    - TYPEONE(a dummy variable that is 1 for residential and 0 if not)
    - TYPETWO(a dummy variable that is 1 for commercial and 0 if not)

Anova Table



Regression Equation

Energy Consumption = 2500.01 + 0.050000 Square Footage + 10.0000 Number of Occupants
+ 19.9999 Appliances Used - 4.99975 Average Temperature + 49.9939 DAY
- 1000.00 TYPEONE - 499.997 TYPETWO

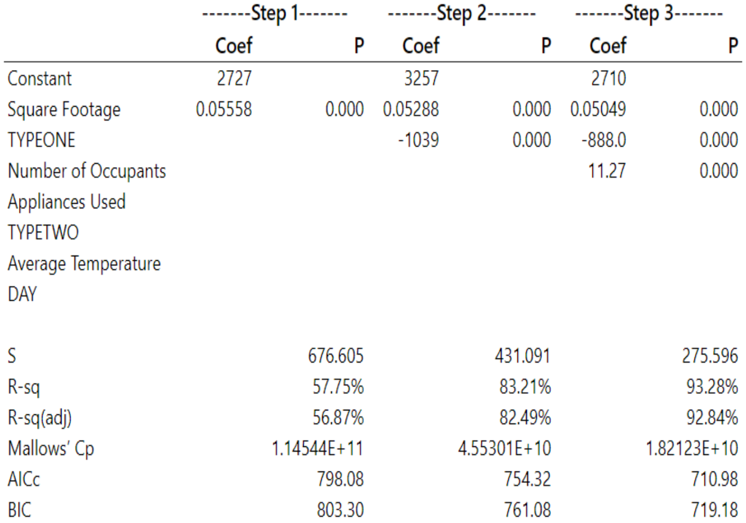
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2500.01	0.01	241504.12	0.000	
Square Footage	0.050000	0.000000	343194.89	0.000	1.08
Number of Occupants	10.0000	0.0001	142055.27	0.000	1.14
Appliances Used	19.9999	0.0002	118826.80	0.000	1.21
Average Temperature	-4.99975	0.00029	-16963.76	0.000	1.15
DAY	49.9939	0.0042	11962.40	0.000	1.14
TYPEONE	-1000.00	0.00	-206031.57	0.000	1.51
TYPETWO	-499.997	0.006	-82598.38	0.000	1.74

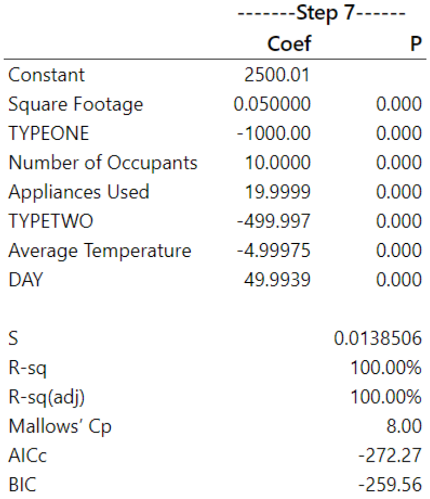
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0138506	100.00%	100.00%	100.00%

Stepwise Regression (Steps 1-3 and Final Step)



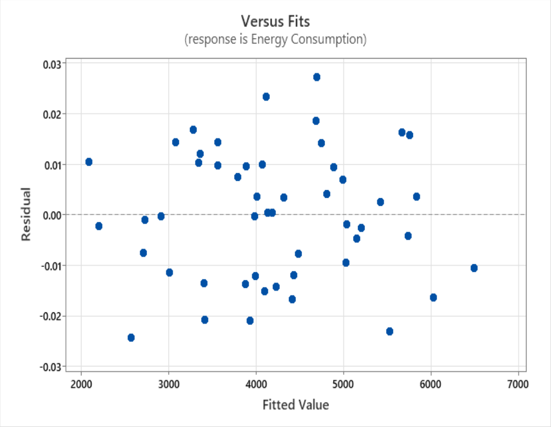
	-------Step 1-------		-------Step 2-------		------Step 3-------	
	Coef	P	Coef	P	Coef	P
Constant	2727		3257		2710	
Square Footage	0.05558	0.000	0.05288	0.000	0.05049	0.000
TYPEONE			-1039	0.000	-888.0	0.000
Number of Occupants					11.27	0.000
Appliances Used						
TYPETWO						
Average Temperature						
DAY						
S		676.605		431.091		275.596
R-sq		57.75%		83.21%		93.28%
R-sq(adj)		56.87%		82.49%		92.84%
Mallows' Cp		1.14544E+11		4.55301E+10		1.82123E+10
AICc		798.08		754.32		710.98
BIC		803.30		761.08		719.18



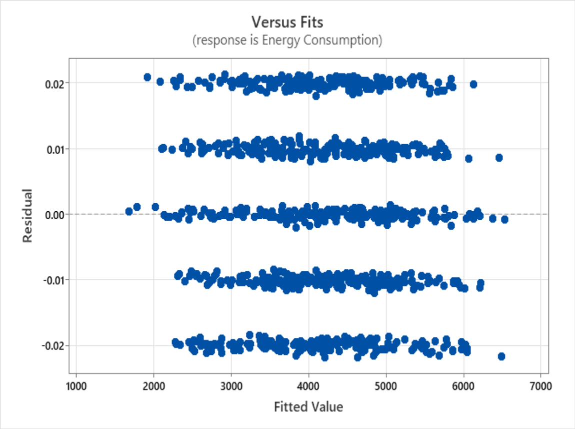
	-------Step 7------	
	Coef	P
Constant	2500.01	
Square Footage	0.050000	0.000
TYPEONE	-1000.00	0.000
Number of Occupants	10.0000	0.000
Appliances Used	19.9999	0.000
TYPETWO	-499.997	0.000
Average Temperature	-4.99975	0.000
DAY	49.9939	0.000
S		0.0138506
R-sq		100.00%
R-sq(adj)		100.00%
Mallows' Cp		8.00
AICc		-272.27
BIC		-259.56

Initially, we conducted a stepwise regression to determine which variables would be most useful in our model. This approach included all available variables. Afterward, the model showed an R^2 of 100.0, which raised some concerns, especially given our large dataset of 1,000 values. We also saw that each variable within our regression model had a p-value of 0.00 indicating they were all significant predictors. We recognized that a perfect fit might not necessarily indicate the best model, as it could point to overfitting. To address this, we checked for multicollinearity, and none of the VIF values exceeded 10, with the highest being 1.74, indicating no problematic multicollinearity. As a result, we proceeded with a final model that utilized a smaller, more relevant subset of the data that still ensured accuracy.

Final Model:



Original Model:



Model Diagnostics

Our original model, based on all 1,000 data points, produced a residual plot with a pattern of horizontal lines(Above Right), which indicated a problem with the fit. To address this, we initially took the natural log of all our variables, but this did not resolve the issue. Next, we tried using a smaller subset of 500 data points, which once again resulted in horizontal lines. The natural log of the model with 500 points resulted in a bowl shaped versus fit graph, which is still not ideal. Finally, we created a model using just 50 data points, and this produced perfectly randomized residuals, indicating a well-fitting model. As a result, we can confidently proceed with this model, as the residuals now appear appropriately distributed. This model keeps our R^2 and also has very similarly coefficient values with next to no deviation from the original model.

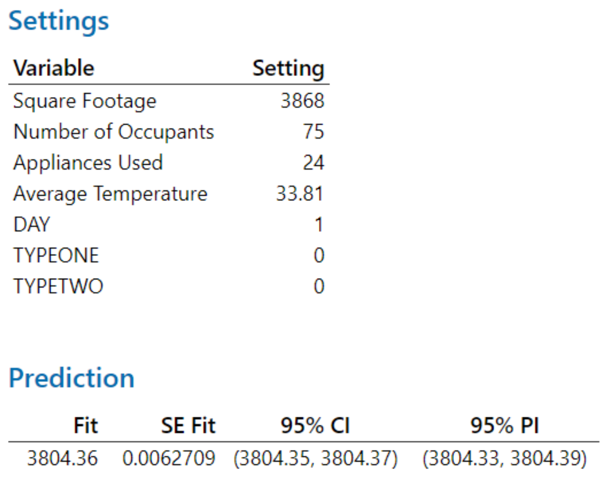
Influential Points

Obs:	SF	NO	AU	AT	DAY	TYPE ONE	TYPE TWO	SRES	HI	COOK
32	46749	22	36	15.41	0	1	0	2.158	0.172	0.121
24	23898	55	16	26.95	0	0	1	-1.696	0.212	0.0965

These two points in our model have the highest Cook's D values. We believe these are the most influential points because they also have the highest residuals, with values of 2.158 and -1.696. While there may be other influential data points in the larger dataset, within the arbitrary subset of 50 points we selected for our model, these two stand out. Neither has an exceptionally high HI value or Cook's D, but they are the most influential among the points in this subset. The SRES of Observation 32 indicates that it is an outlier due to its SRES value greater than 2.

Using the Model

Prediction 1



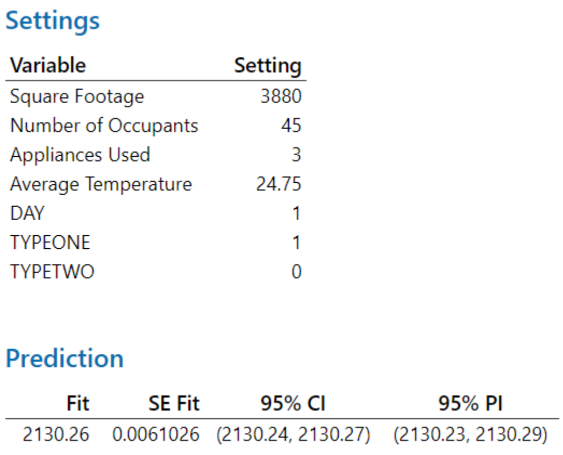
Settings

Variable	Setting
Square Footage	3868
Number of Occupants	75
Appliances Used	24
Average Temperature	33.81
DAY	1
TYPEONE	0
TYPETWO	0

Prediction

Fit	SE Fit	95% CI	95% PI
3804.36	0.0062709	(3804.35, 3804.37)	(3804.33, 3804.39)

Prediction 2



Settings

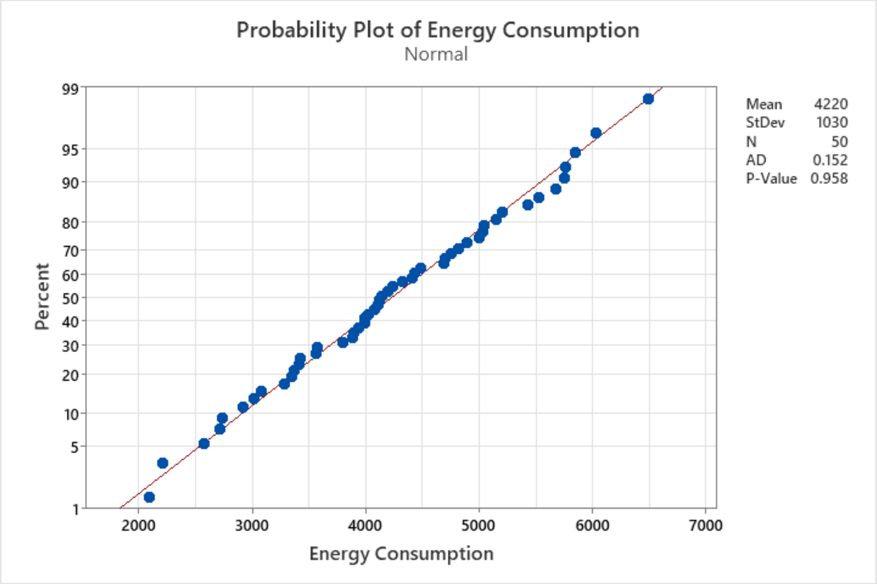
Variable	Setting
Square Footage	3880
Number of Occupants	45
Appliances Used	3
Average Temperature	24.75
DAY	1
TYPEONE	1
TYPETWO	0

Prediction

Fit	SE Fit	95% CI	95% PI
2130.26	0.0061026	(2130.24, 2130.27)	(2130.23, 2130.29)

Above is our first prediction, which has a true value of 3804.34, and the second prediction has a true value of 2130.26. The true value for Prediction One is not within our confidence interval, but it is only 0.01 outside of it. We decided that our true value being only 0.01 away from our prediction interval is not a difference significant enough to render our model useless. It shows that while our model is not perfect it is still very accurate and it also keeps its R^2 value of 100.00. Our second prediction is within our confidence interval.

Anderson Darling



The Anderson Darling test above of our final model has a large P-value indicating that our residuals are normally distributed. This helps indicate that the model's errors are not influenced by any patterns, enhancing the reliability of our model. This helped us to further ensure that our model produces accurate and unbiased estimates.

Conclusion

In conclusion, this model has provided valuable insights into effective ways to estimate the amount of kilowatt-hours a building will require to operate efficiently. Overall, we believe that our dataset contained all the necessary information to create an accurate prediction model. Unsurprisingly, the most influential predictor, and the first one to be selected during stepwise regression, was square footage. This makes sense, as the size of a building is likely the most important factor influencing its energy usage. The most surprising finding, however, was the significance of building type. The model was able to adjust based on the type of building, clearly differentiating between industrial, commercial, and residential structures. This suggests that building type plays a critical role in energy consumption, and it may be worth conducting further analysis to explore this variable in more depth, especially to better understand how each building type impacts energy efficiency. Ultimately, this model provides a solid foundation for understanding the key factors that influence energy consumption.

References

[1] Energy Consumption Dataset

https://www.kaggle.com/datasets/govindaramsriram/energy-consumption-dataset-linear-regression