Analysis of Building Energy Consumption Diversity: A Linear Regression Approach

Haoyu, Zhao

• Abstract

This research employs multivariate linear regression model to investigate the relationship between spatial dimensions of commercial buildings and the diversity of energy sources utilized in various activities. Utilizing the 2018 Commercial Building Energy Consumption Survey (CBECS) dataset, we implement a series of statistical methods, considering the complex survey design inherent in the data. Our analysis includes data transformation, multicollinearity checks, variable selection, and interaction effect analysis, providing comprehensive insights into the factors influencing energy consumption diversity in commercial buildings. Additionally, we incorporate control variables to enhance the robustness of our findings. The results contribute valuable information to the understanding of energy consumption patterns in the commercial building sector.

• Introduction

The rising importance of sustainable energy practices underscores the need for a comprehensive understanding of factors influencing building energy consumption diversity. This study aims to explore the intricate relationship between spatial dimensions of buildings and the diversity of energy sources employed, focusing on various building activities.

• Data Features and Variables

This study relies on the 2018 Commercial Building Energy Consumption Survey (CBECS) dataset, encompassing 6,436 observations and 1,249 variables. Our approach involves identifying key features guided by their potential relevance to the research problem and the completeness of data.

Predictors (Spatial Dimensions):
Total square footage in the building, Building shape, Number of floors, Floor-to-ceiling height.

Controls:
Percentage of the exterior wall surface covered with glass, Year of construction.

Activity Category:
Principal building activity

• Data Transformation

In our preliminary model fitting, which involved all the variables, we observed a non-normality issue in the residuals of the initial linear regression model. Simultaneously, we recognized a substantial difference in scale between the range of square footage (from 1,001 to 210,0000) and the smaller scale of our response variable (from 1 to 7). The decision to apply log transformation was prompted by both the non-normality and scale issues.

Despite the transformation, the non-normality persisted, revealing an even stronger linearity between residuals and fitted values (see Figure 1). This pattern suggested the presence of unaccounted variables related to our response. The notably low R-squared value of 0.1443 in the initial model aligns with our findings of unaccounted variables contributing to the observed residual linearity. However, we refrain from further exploration to identify these variables because our primary focus is on qualitative insights rather than quantitative predictive modeling.

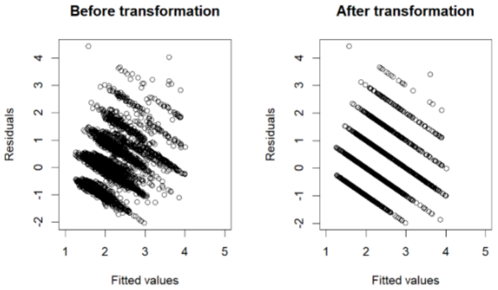

Figure 1 Residuals versus Fitted values.

• Check for Multicollinearity

Given the potential correlation among certain spatial dimensions, particularly total square footage and the number of floors, we conducted checks for multicollinearity using the Generalized Variance Inflation Factor (GVIF). The analysis revealed that square footage and the number of floors had the largest two GVIF values, aligning with our expectation that these variables exhibit the most significant collinearity (See Figure 3 in appendix). However, it's important to note that overall multicollinearity is not severe, with GVIF values close to 1 for all variables. Given this, we consider multicollinearity not to be a significant concern in this context.

• Variable Selection

Variable selection involved a process to identify significant predictors. For continuous variables, we assessed significance directly through p-values. However, for categorical variables, a more nuanced approach was employed. We conducted a series of linear regressions for each categorical variable (denoted by D_i). In each regression, we included one dummy variable representing a specific level of the categorical variable of interest, while the choice of other variables remained the same. This process was repeated for each level of the categorical variable.

$$\text{Response} \sim I(D_i = 1) + \text{Other variables}$$
$$\text{Response} \sim I(D_i = 2) + \text{Other variables}$$
$$\cdots$$

After fitting linear models for all levels of all categorical variables, we consolidated the results. Insignificantly contributing levels were amalgamated, resulting in a refined model. The final model only includes log square footage, number of floors, floor-to-ceiling height, year of construction, and a new categorical variable (denoted by new PBA) obtained by amalgamating insignificant levels of Principal building activity.

• Interaction Effects

Exploring interaction effects, we identified a significant interaction between log square footage and some levels of new PBA. An interaction plot (see Figure 2) highlights the influential levels of new PBA.

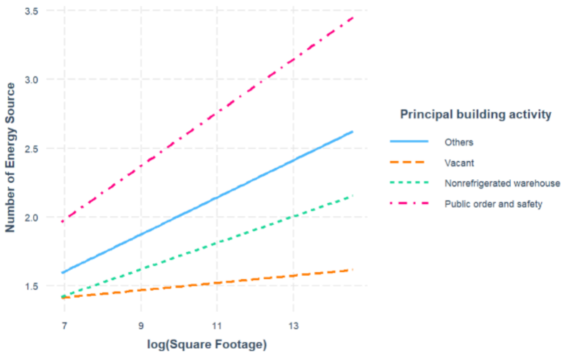

Figure 2 Interaction Plot of new PBA and log square footage

• Results

Our analysis reveals significant relationships between several spatial dimensions, control variables, the principal building activity, and the number of energy sources. The predictors found to be related to the response variable consist of:

Positive Associations: Total square footage in the building, Number of Floors, Floor-to-ceiling height.
Negative Association: Year of construction.
No Association: Building shape, Percentage of the exterior wall surface covered with glass.

Additionally, the categorical variable Principal Building Activity (PBA) is also associated with the diversity of energy sources. Specific activities tend to exhibit distinctive patterns:

Less Energy Use Diversity: Vacant, Office, Nonrefrigerated warehouse, Retail other than mall.
More Energy Use Diversity: Public order and safety, Food service, Inpatient health care, Nursing, Strip shopping center.

• Conclusion

In conclusion, our analysis provides crucial insights into the diversity of energy sources in commercial buildings. Positive associations with spatial dimensions like total square footage, number of floors, and floor-to-ceiling height underscore their impact. Conversely, the year of construction exhibits a negative association, reflecting evolving energy efficiency trends. The categorical variable Principal Building Activity (PBA) reveals distinct patterns among activities. Additionally, our exploration of the significant interaction between log square footage and PBA further refines our understanding. These findings offer practical insights for optimizing energy efficiency in commercial buildings, benefiting architects, policymakers, and stakeholders.

- **Figures and Appendices**

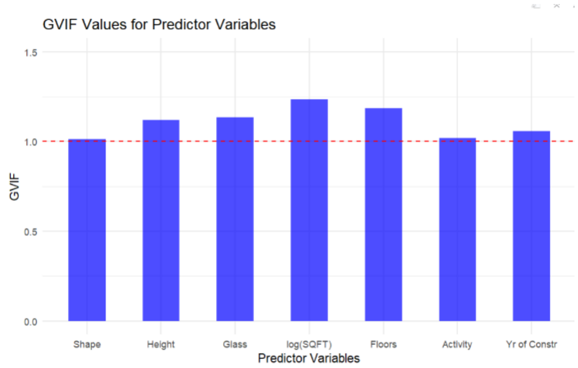

Figure 3 Generalized Variance Inflation Factor of Predictor Variables

R and SAS code for this research can be found at https://github.com/hzhaoar/Stats_506_Project

- **References**

Data source: 2018 Commercial Building Energy Consumption Survey (CBECS) dataset https://www.eia.gov/consumption/commercial/data/2018/index.php?view=microdata