



# ANALYSIS OF THE RELATIONSHIP BETWEEN COMMERCIAL BUILDING FEATURES AND ENERGY CONSUMPTION IN UNITED STATES

STAT8101: Sampling Design and Analysis - Group Project  
Macquarie University, Nott Ryde, NSW 2019

## Abstract

This report critically examines the pattern and determinants of energy consumption in commercial buildings using the data from the 2018 Commercial Buildings Energy Consumption Survey (CBECS). The goal is to identify key factors influencing energy use to help stakeholders optimize energy consumption and promote sustainability. Despite advances in technology, many buildings remain inefficient. We have used various statistical methods to examine the relationships between factors affecting energy consumption. Challenges that we faced include data skewness and confounding variables, we used GLM with Gamma distribution for accurate modeling. The result illustrates climate zone (PUBCLIM), total square foot (SQFT) and principal building activity (PBA) have potential relationship with energy consumption (MFBTU). Future surveys should include more detailed equipment data for further analysis.

Keywords: Commercial Building Energy Consumption Survey, Energy Consumption, Survey Methodology, General Linear Model, Regression

Keith Kwan Ho Ching (47748249)  
Rodulfo II Dela Paz Alfafara (47747420)  
Sujit K C (47521686)  
Wai Kin Wong (46553924)  
Lubaba Reza (47489758)


## Declaration


We, the undersigned, declare that this project is the result of our own work. Each member of our team contributed significantly to the completion of this project, as outlined below:


- Abstract: Sujit K C
- Introduction: Sujit K C
- Survey Methodology: Lubaba Reza
- Data Description: Keith Kwan Ho Ching, Rodolfo II Dela Paz Alfafara, Wai Kin Wong
- Data Analysis: Keith Kwan Ho Ching, Rodolfo II Dela Paz Alfafara, Wai Kin
- Discussion and Conclusion: Keith Kwan Ho Ching, Rodolfo II Dela Paz Alfafara, Wai Kin
- Writing – Original Draft Preparation: Keith Kwan Ho Ching, Rodolfo II Dela Paz Alfafara, Sujit K C, Wai Kin Wong, Lubaba Reza
- Writing – Review & Editing: Original Draft Preparation: Keith Kwan Ho Ching, Rodolfo II Dela Paz Alfafara, Sujit K C, Wai Kin Wong, Lubaba Reza

Each member has reviewed and approved the final version of the project.

Signed,

Keith Kwan Ho Ching (47748249) 

Rodolfo II Dela Paz Alfafara (47747420) 

Sujit K C (47521686) 

Wai Kin Wong (46553924) 

Lubaba Reza (47489758) 

Date: 24-May-2024

## Contents

1	INTRODUCTION .....	1
1.1	Project Aims .....	1
1.2	Project Background.....	1
2	Survey methodology .....	2
2.1	Commercial Buildings Energy Consumption Survey Methodology .....	2
2.2	Energy Supplier Survey Methodology .....	4
2.3	Data processing and quality checks.....	4
3	Data Description .....	6
3.1	Data Analysis.....	6
3.1.1	<b>Data Analysis - Continuous Variables .....</b>	<b>7</b>
3.1.1.1	<b>Data Analysis - Continuous Variables – Relationship Analysis.....</b>	<b>9</b>
3.1.2	<b>Data Analysis – Categorical Variables.....</b>	<b>10</b>
3.1.2.1	<b>Data Analysis – Categorical Variables – Relationship Analysis .....</b>	<b>11</b>
4	Data Analysis – Modelling .....	14
4.1	Data Analysis – Modelling Result.....	15
5	Interpretation of Result and Discussion.....	18
5.1	Interpretation of Result.....	18
5.2	Discussion and Conclusion.....	19
6	References .....	20

Appendix – R Codes for Modelling

# **1 INTRODUCTION**

## **1.1 Project Aims**

The project aims to critically analyze the pattern and determinants of energy consumption in commercial buildings. We aim to identify and highlight key factors that are associated with energy consumption, which can be used by other researchers for studies on related topics. Also, the report seeks to provide actionable insights that can help stakeholders, including building managers, lawmakers, and energy service companies to understand energy usage, thereby helping them in formulating relevant strategies and policies to optimize energy efficiency, reduce costs, and promote sustainable practices in the commercial building sector.

## **1.2 Project Background**

Energy is a fundamental and crucial component in the modern global economy. Industries, commercial facilities, health, education, and people use energy for day-to-day activities. Consequently, energy supply and consumption play crucial roles in the world and maintaining quality of life. Commercial buildings are major consumers of energy in urban areas. The demand for energy in these buildings is caused by various factors such as building size, floor space, geographical location, and types of activities conducted within. The energy consumption on these commercial buildings is increasing every year, and there has been growing concern on improving energy efficiency in commercial buildings, to overcome problems such as huge energy bills, adverse effect on environment, ethical and regulatory pressures. Although there has been major advancement in building technologies and energy management systems, many commercial buildings still have inefficient energy usage, which results in excessive energy consumption. Identifying these factors which cause high energy consumption and formulating effective solutions are necessary for achieving sustainability goals and protecting the earth. With this in mind, this report aims to perform a critical statistical analysis on a survey dataset of energy consumption by commercial buildings. This study will focus on the relationships among the factors that affect energy consumption with the hopes of being able to provide valuable and actionable insights that can be used to support global energy sustainable practices.

## **2 Survey methodology**

This paper used data from the most recent Commercial Buildings Energy Consumption Survey, CBECS 2018, conducted by the U.S. Energy Information Administration (EIA). The survey was conducted in two stages: the Buildings Survey and the Energy Supplier Survey. (EIA, n.d.) [6] After constructing a sampling frame and selecting a sample of buildings, the Buildings Survey was conducted both in person and online. When conducting the Buildings Survey, the respondents were asked to provide the names of their energy suppliers and their account numbers to conduct the Energy Suppliers Survey. After information was collected from both surveys, the data was used to model energy end-use consumption. (EIA, n.d.) [6]

### **2.1 Commercial Buildings Energy Consumption Survey Methodology**

Since there are no comprehensive databases of commercial buildings in the U.S., a sampling frame had to be created for this survey. The sampling frame consisted of two parts: the area frame and the list frames. Five different lists of large buildings were used to create the list frames portion, while a multi-stage area probability sampling of selected geographic areas was used to construct the area frame. The whole country was divided into 687 counties or groups of counties which were the Primary Sampling Units (PSUs). (EIA, 2023) [9] The PSUs were selected with probabilities proportional to their commercial activity. 151 PSUs were selected containing 8,559 census tracts or groups of census tracts, which were the Secondary Sampling Units (SSUs). 764 SSUs were selected with selection probabilities proportional to an estimate of the number of commercial buildings within the SSU. Finally, field visits were used to create an area frame of all the commercial buildings in these 764 SSUs, supplemented by virtual listings such as satellite images, GIS, and other databases. Selecting samples from only the area frame would not provide an adequate quantity of large buildings, thereby separate lists of buildings larger than 200,000 square feet were used to expand the sampling frame. These lists combined with the area frame were used to create the sampling frame for this survey. 16000 buildings were selected from the sampling frame after dividing it into subgroups of buildings based on the sampling frame, PSU, SSU, building size, and building type. Subgroups with a higher variation in total energy consumption were selected more frequently than subgroups with a lower variation in energy consumption to ensure a high level of variation in the sample selected. (EIA, 2023) [9] This is a great technique to ensure that the sample selected provides a good representation of the energy consumption in commercial buildings in the entire nation.

This survey has a well-defined eligibility criteria for buildings. The eligibility of buildings was decided using three criteria: building definition, building use, and building size. To be eligible, a building could not be under construction, dilapidated, or condemned, more than 50% of the building's floor space had to be used for commercial activities and the building had to be greater than 1,000 square feet. (EIA, 2023) [10] The sampled buildings were assigned base weights which were later adjusted to reflect buildings that did not respond and accommodate for buildings that were later proven to be ineligible. The sum of these adjusted weights was used to estimate the total number of commercial buildings.

Data was collected using voluntary interviews with residents, managers, or building owners in person, via a phone call or online. Having multiple modes of data collection is a great way to minimize nonresponse. Before conducting the survey, the interviewers were trained, and the questionnaire and survey program were pretested. (EIA, 2023) [7] 175 interviewers supervised by field supervisors, regional managers, and a field director collected data from April 2019 to January 2020. Preceding the survey, interviewers visited the sampled buildings to identify the exact location and ensure that the buildings satisfied the eligibility requirements for the survey. A broad range of topics were covered by the questionnaire: physical characteristics and building use patterns, types of energy used and usage activities, types of equipment used, energy management practices, amount of electricity, natural gas, fuel oil, and district heat used and associated expenditures. (EIA, 2023) [7]

Of the initial 16000 sampled buildings, 6436 were included in the final data set with completed responses. There were 5,915 nonresponses and 20% of the sampled buildings were later identified as ineligible. (EIA, 2023) [8] Nonresponses mainly occurred because either a respondent could not be identified, or they refused to participate. This survey handled nonresponses very meticulously, on a case-by-case basis with proper guidance from the supervisors through weekly phone calls. Refusals were the main cause of nonresponse. EIA and the survey team either made phone calls or sent emails or letters to persuade the respondents to reconsider. In-person rejection conversion attempts were also undertaken by interviewers several times. By having a proper strategy for handling nonresponse cases, this survey reduced its nonresponse bias to a negligible amount. (EIA, 2023) [8]

The sampling frame of the Buildings Survey ensures that a comprehensive list of commercial buildings in the areas in the U.S. with significant commercial activity are included to avoid underestimating energy consumption. However, this is a complex and time-consuming procedure which could be expensive and require many resources.

## **2.2 Energy Supplier Survey Methodology**

The Energy Supplier Survey was used to collect monthly energy usage and cost data of the buildings that participated in the Commercial Buildings Energy Consumption Survey from their electricity, natural gas, fuel oil and district heat (steam or hot water) suppliers. This survey was conducted to reduce the inconvenience for the buildings survey respondents, requiring them to provide only annual data, while monthly data was collected from the energy suppliers. The data was collected from April 2020 for 16 billing periods from November 2017 to February 2019. (EIA, 2023) [11] Data was collected through either an interactive data collection web page, an Excel spreadsheet downloaded and completed by the supplier, the supplier's own electronic file, or data collection forms filled out by the supplier and mailed to the survey team. For this survey, a total of 610 energy suppliers were contacted which consisted of 317 electric utilities, 174 natural gas utilities, 101 heating oil distributors, and 18 district heat suppliers. (EIA, 2023) [11]

One of the strengths of the ESS was its high response rate at the supplier level since participation was mandatory. Electricity and natural gas both had greater than 90% response rates at the supplier level. (EIA, 2023) [11] However, nonresponses still occurred when suppliers failed to find the customer's records in their database, or the supplier of a building could not be identified. In addition, the reliance on supplier records in data collection may introduce bias due to potential inaccuracies or inconsistencies in those records.

## **2.3 Data processing and quality checks**

The data for both the CBECS Buildings Survey and the Energy Supplier Survey (ESS) was reviewed and processed during the survey and after completion to ensure data quality. (EIA, 2023) [12] The computerized survey instrument used for data collection was programmed to minimize errors by only accepting appropriate responses and providing acceptable ranges for data where applicable. For self-administered online responses the data was processed after collection. This process involved conducting data consistency checks, such as verifying

whether the building size reported by the respondent aligned with the sampling size category and confirming that the correct building was interviewed. The responses were further investigated to confirm the eligibility of the buildings and to ensure that no critical questions were left unanswered. (EIA, 2023) [12]

For the Buildings Survey, hot-deck imputation is used to handle nonresponse for questions. Using hot-decking, when a building has a missing value for a particular question, another building with similar characteristics was chosen at random to impute that missing value. The principal building activity, square footage category, year constructed category, and census region were the most used attributes. (EIA, 2023) [12] For the ESS data a sequence of assessments was performed, including assessments for inadequate data, missing units, unusually low costs, and exceptionally high or low prices. These errors were rectified by contacting the suppliers and through manual review by the analysts.

The Buildings Survey collected annual energy consumption data while the Energy Supplier Survey collected monthly energy consumption data. To reconcile the consumption data, the data from both surveys were annualized and disaggregated when a building's reported consumption figure included the consumption of additional buildings. Afterward, the data was examined to decide which survey to use for the final consumption value for each case. For each building, the intensity (consumption per square footage) was compared with the same building's intensity the 2012 CBECS, and the building's reported energy usage was compared with expected usage estimated using engineering-based models and the building's characteristics. (EIA, 2023) [12] For each case, the survey with consumption data closest to the expected usage was chosen as the data source for that case.

This survey obtains high data quality by implementing data consistency checks and reviews every case for errors and inconsistencies. The survey methodology also employs the hot-deck imputation for estimating missing values. It preserves the integrity and completeness of the dataset. However, it may introduce potential bias for using hot-deck imputation as the assumption of buildings with similar characteristics will have similar values for missing data, may not always hold true.



### 3 Data Description

Predicting energy consumption for commercial buildings is our key objective. In the CBECS 2018 dataset, the continuous variable MFTBU which represents the Annual Energy Consumption (thous btu) is our variable of interest.

The full data set consists of 630 survey variables, we are picking four continuous variables and two categorical variables as predictors to identify the relationship between the response variable (MFTBU). Details are highlighted in section 3.1.1 and 3.1.2.

#### 3.1 Data Analysis

Throughout this document the survey design “CBECS\_des” is being used to generate various diagrams and estimates. The design is highlighted below, more details can be found in the Appendix section.

```
#read the dataset
CBECS <- read.csv("cbeecs2018_final_public.csv", header = TRUE)

#applied the weights

CBECS_des <- CBECS %>%
  as_survey_rep(weights = FINALWT,
    repweights = FINALWT1:FINALWT151,
    type = "JK2",
    mse = TRUE)
```

In the CBECS survey, jackknife method is used for estimating standards errors ,hence, our design type is set to “JK2”. (EIA,n.d.)[3] In addition, we also use some reference codes from Zimmer, Powell & Velásquez (2024).

### 3.1.1 Data Analysis - Continuous Variables

Four continuous variables are selected for our analysis, a summary is highlighted below:

- **MFBTU** is the target response variable, which represents the Annual Energy Consumption (thous btu). According to the Consumption and expenditures report [1], it has relationships with heating, cooling, and building size, as such, we choose to observe the following continuous variables:
- **SQFT** is one of the possible predictors, which represents the total square foot of a building.
- **HDD65** is one of the possible predictors, which represents heating degree days (base 65). This field indicates how cold the weather is over a day or a series of days.
- **CDD65** is one of the possible predictors, which represents cooling degree days (base 65). This field indicates how hot the weather is over a day or a series of days.

Figure 1 highlights the distribution of the chosen continuous variables. The large standard deviation for both MFBTU and SQFT indicates significant variability in the data. The histograms in Figure 2 also indicate that the data are right-skewed.

	MFBTU_mean	MFBTU_mean_se	MFBTU_median	MFBTU_median_se	MFBTU_SD
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1209619.	41565.	276359	13405.	5814350.
	SQFT_mean	SQFT_mean_se	SQFT_median	SQFT_median_se	SQFT_SD
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	16310.	421.	5400	253.	52149.
	CDD65_mean	CDD65_mean_se	CDD65_median	CDD65_median_se	CDD65_SD
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1673.	62.1	1490	68.6	1004.
	HDD65_mean	HDD65_mean_se	HDD65_median	HDD65_median_se	HDD65_SD
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	4529.	135.	4568	194.	2250.

Figure 1. Summary statistics of continuous variables.

Figure 2 illustrates significant skewness of distribution for the four continuous variables. Both MFBTU and SQFT are extremely right skewed.

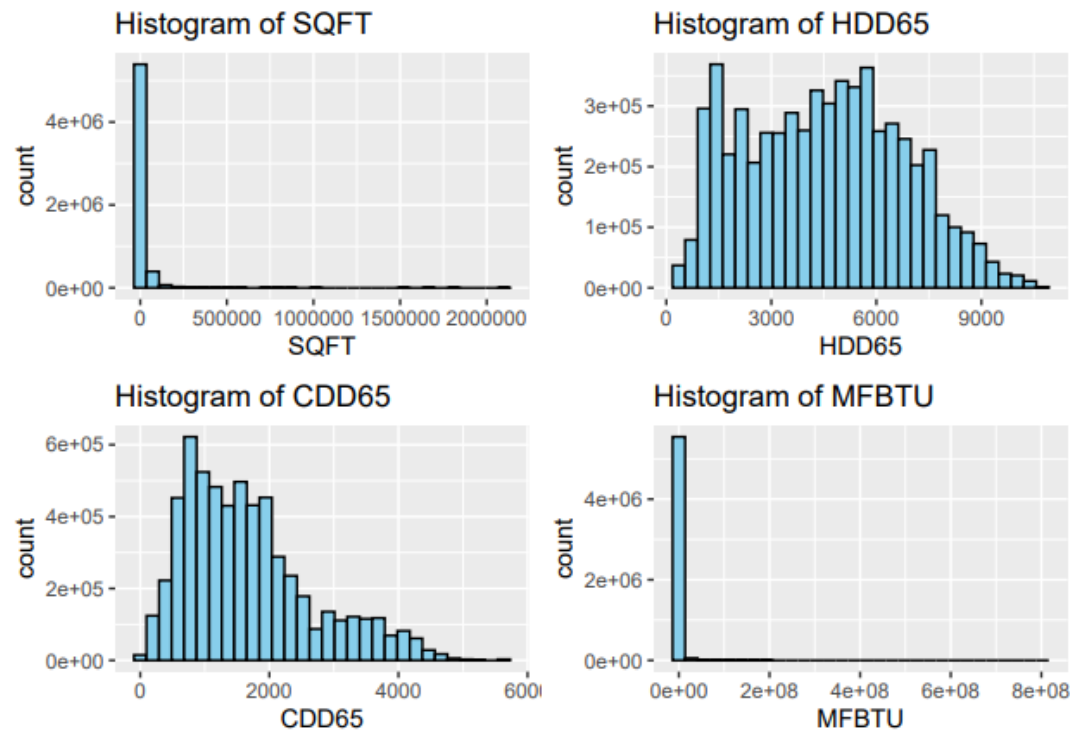


Figure 2 .Histograms of continuous variables.

Table 1 illustrates that only 79 out of 6436 rows have missing MFBTU. Since the missing data only represents a small fraction of the total data in the CBECS dataset, those missing rows will be ignored in the modelling phase.

Column	Missing_values
MFBTU	79
SQFT	0
CDD65	0
HDD65	0

Table 1 . Missing data of continuous variables

Due to skewness of the four continuous variables, log transformations are performed on those variables before conducting correlation analysis.

### 3.1.1.1 Data Analysis - Continuous Variables – Relationship Analysis

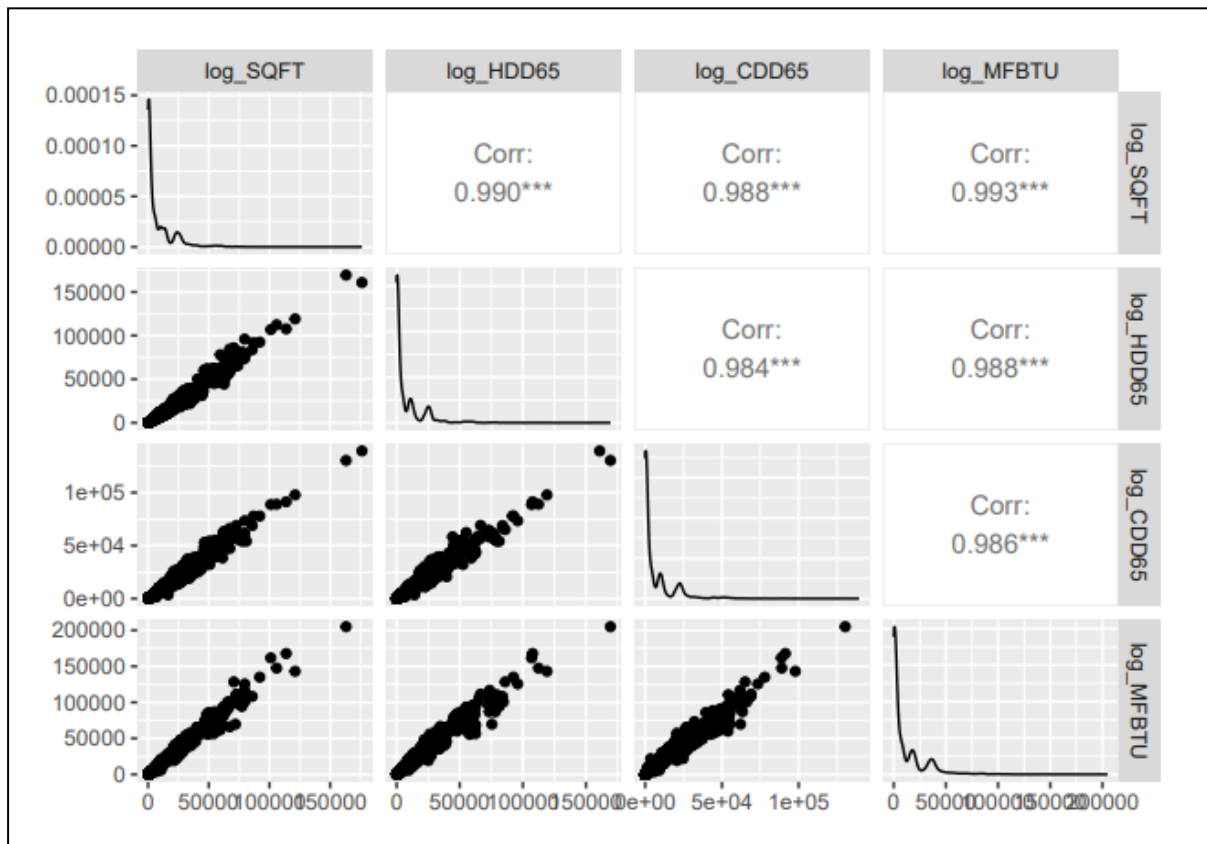


Figure 3 .Correlation Matrix

Figure 3 depicts the relationship between the selected predictors and the response variable. Apparently, log\_MFBTU has a strong positive relationship with every other continuous predictor. However, it is important to note that all continuous predictors have strong relationships with one another, which indicates that they are confounding variables.

In addition, log\_MFBTU shows a symmetric pattern around a horizontal line at zero for all variable relationships. Moreover, the residuals are evenly spread across, which may indicate a semblance of normality for all predictors.

Based on the correlation matrix, log\_SQFT has the best linear relationship with log\_MFBTU. It can be treated as the best predictor for the response variable. Since log\_HDD65 and log\_CDD65 appear to be confounding variables of log\_SQFT, these two variables will be removed for future modelling.

In subsequent sections, other categorical predictors will be further analysed to determine whether they can be used together with log\_SQFT for modelling.

### 3.1.2 Data Analysis – Categorical Variables

Two categorical variables are selected for our analysis, a summary is highlighted below:

- **PUBCLIM** is one of the possible predictors, which represents the climate zone where the building is located. Categories here are representing the climate zone, which is usually very cold, very hot, or in between.
- **PBA** is one of the possible predictors, which represents the principal building activity being performed in the building.

PUBCLIM contains six categories that identify the climate zone of each sample. The categories are as follows: 1 represents "Cold or Very Cold" climate zone, 2 represents "Cool" climate zone, 3 represents "Mixed Mild" climate zone, 4 represents "Warm" climate zone, 5 represents "Hot or Very Hot" climate zone, and 7 represents "Withheld for Confidentiality". There are no missing responses for the PUBCLIM variable.

PUBCLIM	1	2	3	4	5	7
n	422	1428	1479	1483	772	852

Table 2 . PUBCLIM Distribution

Table 2 illustrates the distribution of PUBCLIM category. Most of the samples are in the Cool, Mixed Mild, and Warm climate zones. However, a significant proportion of the samples have not disclosed their climate zone and are categorized as "7. Withheld for Confidentiality." This category has a notably higher median and greater variance (refer to Figure 7)., and for this reason, further analysis will be carried out to identify whether category “7” is suitable for modelling in section 4.

Table 3 illustrates the distribution of PBA which contains 20 categories which identify the principal building activity of each sample. There is no missing data in this field and the average count for this category is 321.8.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]
PBA	1	2	4	5	6	7	8	11	12	13	14	15	16	17	18	23	24	25	26	91
n	117	1329	69	753	91	105	217	23	271	481	936	218	276	114	418	208	35	339	350	86

Table 3. PBA Distribution

### 3.1.2.1 Data Analysis – Categorical Variables – Relationship Analysis

To facilitate further analysis, PBA is treated as a factor and different levels are combined to achieve a balanced level (near average level). This is possible as PBA is nominal data. Finally, categories 4, 6 and 7 are regrouped into category 1. Categories 11, 17, 24 are regrouped into category 8. Categories 23 and 29 are regrouped into category 15 and category 12 is renamed as category 16.

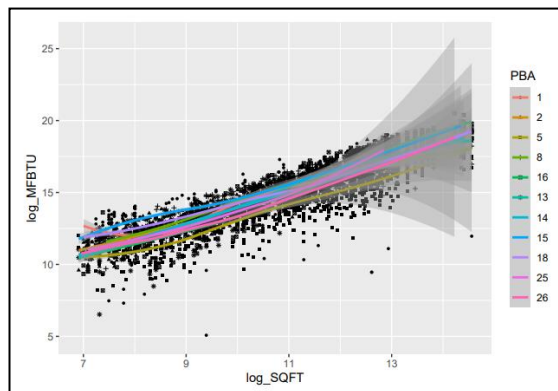


Figure 4. log\_MFBTU vs log\_SQFT (PBA)

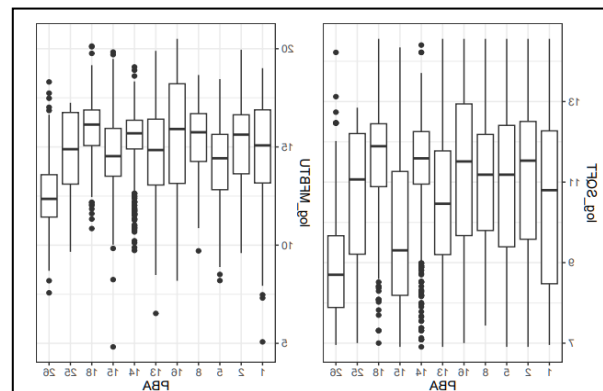


Figure 5. Boxplot - PBA vs log\_SQFT & log\_MFBTU

Based on Figure 4, it is evident that the change in response to the changing log\_SQFT becomes different as the PBA factor levels change. For this reason, it appears that log\_SQFT and PBA can be promising predictors. From the boxplot plot in Figure 5, it confirms that as PBA levels

change, there will be visible change in the range of values for both log\_SQFT and log\_MFBTU. This indicates that log\_SQFT and PBA can be used together as predictors.

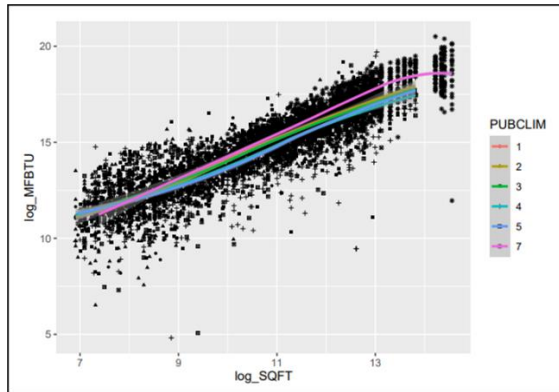


Figure 6 .log\_MFBTU vs log\_SQFT (PUBCLIM)

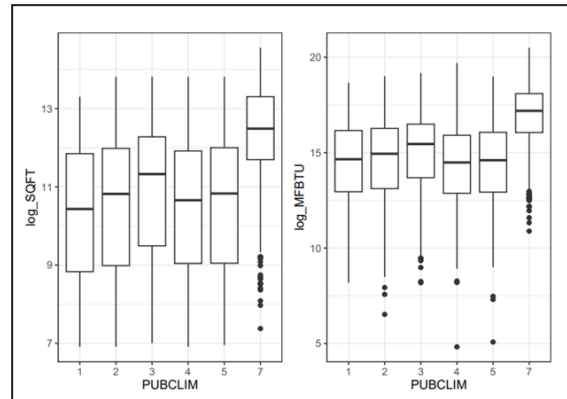


Figure 7. Boxplot - PUBCLIM vs log\_SQFT & log\_MFBTU

From Figure 6 and 7, it appears that change in response to the changing log\_SQFT becomes different as the factor levels change for all categorical variables. For this reason, it appears that log\_SQFT along with the two categorical variables can be promising predictors of log\_MFBTU.

Finally, an analysis was conducted to explore the relationship between the categorical variables PUBCLIM and PBA. Due to the sparsity caused by category 7 of PUBCLIM, a proper chi-square test could not be performed. Therefore, category 7 was temporarily removed from the analysis as shown in Figure 8.

PBA		1	2	4	5	6	7	8	11	12	13	14	15	17	18	23	25	26	91
PUBCLIM	1	3	84	1	51	11	4	16	1	15	31	80	21	19	5	9	26	39	6
	2	31	282	18	169	26	15	52	5	60	139	226	56	36	31	53	85	119	25
	3	28	410	23	187	18	15	69	5	63	109	232	52	32	43	34	71	67	21
	4	34	326	23	241	22	17	58	8	95	120	138	53	20	49	73	92	86	28
	5	20	184	4	93	14	7	22	4	38	73	106	36	7	17	39	65	38	5

Figure 8. Frequency table of PUBCLIM and PBA

Design-based Wald test of association

```
data: svychisq(formula = ~PUBCLIM + PBA, design = CBECS_des, statistic = "Wald", na.rm = TRUE)
F = 1.1683, ndf = 68, ddf = 5583, p-value = 0.1635
```

*Figure 9 .Chi-square analysis- PUBCLIM vs PBA*

A Chi-square analysis was conducted between PUBCLIM and PBA, as shown in Figure 9. The analysis indicates a p-value of 0.1635, which is greater than the significance level of 0.05. Therefore, there is no statistically significant association between the two categorical variables. Despite the lack of significance, PUBCLIM remains in the model due to its potential relationship with log\_SQFT in explaining energy consumption. Consequently, PUBCLIM and PBA can be utilized as independent predictors along with log\_SQFT in our final model.

For now, category 7 will remain in the dataset to preserve the dataset integrity for further analysis. Removing this category at this stage could impact the overall dataset and introduce bias into the analysis. Hence, the removal of category 7 will be addressed at a later stage. The final modelling will be conducted in Section 4.



#### 4 Data Analysis – Modelling

There are many methods to predict  $\log\_MFBTU$ . Initially, Linear Regression was performed using  $\log\_MFBTU$  and  $\log\_SQFT$ . However, the homoscedasticity assumption cannot be satisfied. As illustrated in Figure 10, the residuals do not have a constant standard deviation.

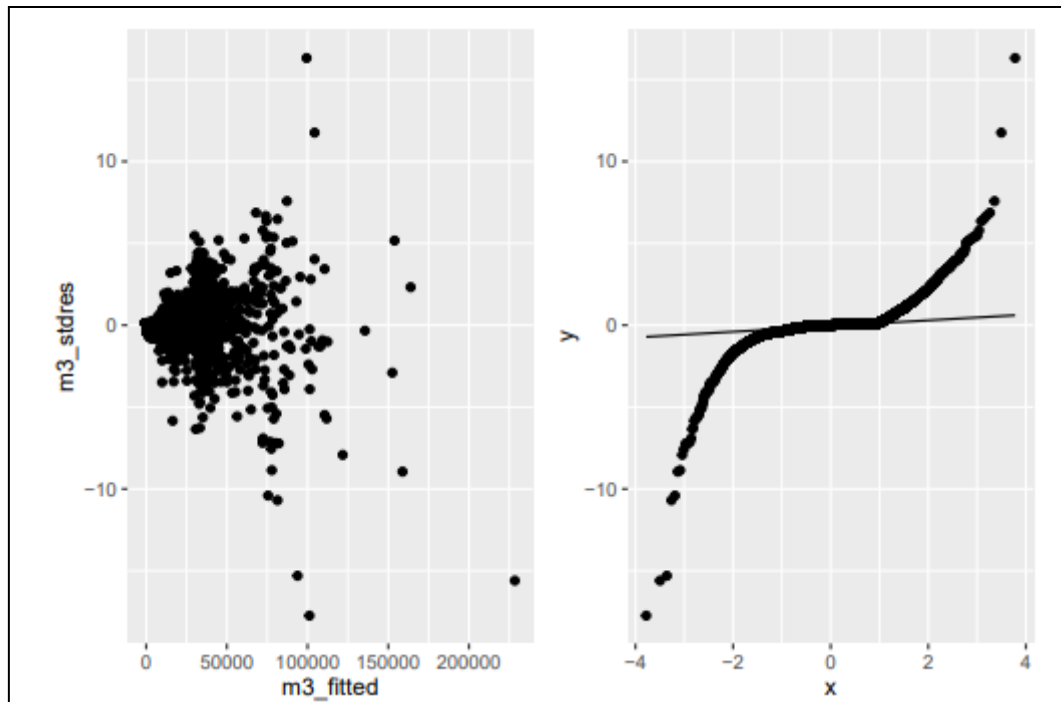


Figure 10 .Residual plot -  $\log(MFBTU)$  vs  $\log(SQFT)$

To overcome this issue, Generalized Linear Model (GLM), using Gamma distribution was used in the modelling process. To investigate if the fitted model meets the Gamma distribution, the QQ plots of residuals, Residuals vs Predicted and Histogram using the DHARMA package were analysed.

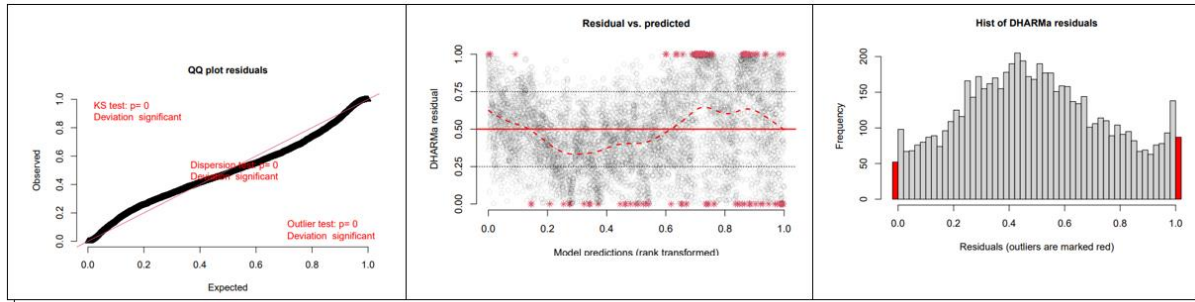


Figure 11 .Residual Analysis for GLM – Gamma Distribution

Based on the QQ plot in Figure 11, we can observe that the KS test is significant, which means the model does not fit well with the uniform distribution. For the Residuals vs. Predicted plot, it appears that the residuals are found to be around the horizontal line at 0.5 level, which hints that although the model does not perfectly fit the uniform distribution, it does exhibit some conformity to this distribution. This is further supported by the histogram where despite having bars of varying lengths and a lack of perfect uniform distribution, it is indicative that the model somewhat follows a uniform distribution. Alternative distributions, such as logarithmic, inverse, and Gaussian, were also tested, but they performed significantly worse than the current choice.

#### 4.1 Data Analysis – Modelling Result

Different GLM- Gamma models were analysed which included the Main Effects and the interactions among the covariates. Based on the AIC value, model 6 with AIC value 84323 (lowest AIC) is the best model. The AIC results are summarised in Figure 3. This model is corresponding to:

$$\log S QFT + PBA + PUBCLIM + \log S QFT * PBA + \log S QFT * PUBCLIM$$

##	Model_Number	Predictors
## 1	1	log_SQFT
## 2	2	PBA
## 3	3	log_SQFT + PBA + log_SQFT*PBA
## 4	4	PUBCLIM
## 5	5	log SQFT + PUBCLIM + log SQFT*PUBCLIM
## 6	6	log_SQFT + PBA + PUBCLIM + log_SQFT*PBA + log_SQFT*PUBCLIM
## 7	7	log_SQFT + PBA
## 8	8	log_SQFT + PUBCLIM
## 9	9	log_SQFT + PBA + PUBCLIM
##	AIC	
## 1	86041	
## 2	126937	
## 3	84377	
## 4	126788	
## 5	85995	
## 6	84323	
## 7	85767	
## 8	86030	
## 9	85753	

Figure 12 .AIC results

The impact of each level from the categorical variables PBA and PUBCLIM were further analysed to fine tune the model. Figure 4 highlights the result of the model. From the summary result, the **residual deviance to the degrees of freedom ratio** was calculated, which amounts to 0.004960316. This very small number indicates that the model is a good fit.

```
## (Dispersion parameter for Gamma family taken to be 0.004801233)
##
## Null deviance: 18472.406 on 6356 degrees of freedom
## Residual deviance: 31.374 on 6325 degrees of freedom
## (79 observations deleted due to missingness)
## AIC: 84323
##
## Number of Fisher Scoring iterations: 6

From the summary above, we are able to identify the following:

residual_deviance <- 31.374
degrees_of_freedom <- 6325

dev_df_ratio <- residual_deviance/degrees_of_freedom
dev_df_ratio

## [1] 0.004960316
```

Figure 13 .Summary of the initial model

The result (full result is listed in the Appendix section) also indicated that:

- The very low value in **residual deviance to the degrees of freedom** might also indicate that the model is overfitting.
- SQFT Main effect is significant with p-value of  $< 2e-16$ .
- PUBCLIM Main effect level 3 is significant with p-value of 0.024770. However, PUBCLIM levels 1-3 have significant interactions with SQFT. Therefore, only levels 5 and 7 will be removed.
- PBA Main effect levels 1,14, 16 and 18 are significant. However, the only levels that do not have significant interactions with SQFT are 18 and 25. Therefore the only level we will exclude is level 25.

To further fine tune the model, levels 5 and 7 from PUBCLIM were excluded from the model and level 25 from PBA was also excluded from the model. After refitting the model, we obtain a slightly higher **residual deviance to the degrees of freedom** ratio which is still relatively small i.e. 0.05295905. However, it is important to note that the AIC value of the fine-tuned model is much smaller, which is 61791, which means that the model has significantly improved after the fine-tuning. The refined result is highlighted in Figure 14.

```
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS_filtered$PUBCLIM,
##   ref = "4") + relevel(CBECS_filtered$PBA, ref = "2") + weighted_SQFT *
##   relevel(CBECS_filtered$PUBCLIM, ref = "4") + weighted_SQFT *
##   relevel(CBECS_filtered$PBA, ref = "2"), family = Gamma(link = "identity"))
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.005148707)
##
## Null deviance: 11764.247  on 4469  degrees of freedom
## Residual deviance: 23.535  on 4444  degrees of freedom
## (68 observations deleted due to missingness)
## AIC: 61791
##
## Number of Fisher Scoring iterations: 5
```

Figure 14 .Result of fine-tuned model

## 5 Interpretation of Result and Discussion

### 5.1 Interpretation of Result

Based on our best performing model, the effect of SQFT on MFBTU varies by levels of Public and PBA. We have our equation as:  $Y = \beta_0 + \beta_1 X_{i_1} + \beta_2 X_{i_2}$  where Y is the MFBTU response;  $\beta_0$  is the intercept;  $\beta_1$  is the SQFT coefficient;  $X_{i_1}$  is the effect from PUBCLIM levels; and  $X_{i_2}$  is the effect from PBA levels. From this, we have our fitted equation as:  $MFBTU = 0.00009 + 4.039X_{i_1} + 4.039X_{i_2}$

From the summary of the model, we can observe that the effect of SQFT is affected by its interactions with both PUBCLIM & PBA. For certain levels of PUBCLIM and PBA, the effect of SQFT on MFBTU is either slightly increased or decreased. The intercept of -9.366 is the expected value of weighted\_MFBTU when all predictors are at their reference levels and weighted\_SQFT is 0, which makes sense because if SQFT is 0, it means that there is no building in place, and therefore, should have no corresponding energy consumption. Recall that we applied log transformations on the data, hence, the actual value of the intercept is  $e^{-9.366} = 0.00009$ , which is a very small number close to 0. The coefficient of weighted\_SQFT which is  $e^{1.396} = 4.039$  means that for each unit increase in weighted\_SQFT, the weighted\_MFBTU is expected to increase by approximately 4.039, holding other variables constant. Considering SQFT's interactions with PUBCLIM & PBA, this predicted increase of 4.039, will change depending on the PUBCLIM/PBA levels SQFT interacts with. For example, a hypothetical observation with PUBCLIM and PBA level 1, will slightly increase the approximate effects of SQFT by  $e^{0.031} = 1.031$  and  $e^{0.041} = 1.042$  respectively as these are the coefficient estimates of both PUBCLIM & PBA at level 1. Plotting these values in our fitted equation, for one unit increase in SQFT, we will have:

$MFBTU = 0.00009 + 4.039 \cdot 1.031 + 4.039 \cdot 1.042 = 8.373$  The result means that for one unit increase in SQFT that have interactions with PUBCLIM & PBA level 1; there will be a corresponding 8.373 unit increase in MFBTU. All in all, our results show that there are significant differences in weighted\_MFBTU for different levels of PUBCLIM and PBA, indicating that these categorical variables are also important predictors of MFBTU.

## 5.2 Discussion and Conclusion

During our study, we have used different statistical techniques to identify the relationship between total energy consumption with certain possible continuous variables and categorical variables. It is important to note that our study is by no means to be comprehensive, as certain variables such as Equipment Types are meant to be useful predictors, but the dataset lacks comprehensive information on those fields which hindered our thorough analysis on those variables for predicting energy consumption.

The initial statistics analysis such as descriptive statistics and graphical analysis both indicated strong skewness of the data which made the regression analysis more challenging. In addition, two continuous variable “HDD65” and “CDD65” are confounding with “SQFT” and we have decided to only use “SQFT” as the continuous predictor.

For the categorical predictors, our study indicated that level 7 from the climate zone “PUBCLIM” also induced interaction with “PBA”. Further analysis indicated that level 7 may not actually presenting a specific climate zone as it is classified as “Withheld with confidentiality”. Our final model removed the data with this level for analysis and the result is very promising.

The skewness of the data also created an issue in using linear regression as the assumption of homoscedasticity is not satisfied. Finally, we used GLM, following Gamma distribution, to create the model and the model has achieved a very strong degree of accuracy with the input data.

As mentioned earlier, the variability of the chosen continuous data provided challenges in the modelling process. Other possible options that could have been explored further are the use of equipment types to predict total energy use. As highlighted in the Consumption and expenditures reports (EIA, n.d.) [1], Total Energy Consumption composes of many different components such as space heating, space colling, water heating, cooking etc. However, the dataset does not provide data elements that specifically cater to energy consumption of the different equipment types which make modelling with equipment type extremely difficult. Hopefully, the next CBECS survey can provide more information that can make extensive studies on the energy consumption brought by the various equipment types, a possibility. Finally, many researchers are now using machine learning algorithms to provide better prediction results; however, it is out of the scope of this assignment.

## 6 References

[1] Consumption and expenditures report.

<https://www.eia.gov/consumption/commercial/data/2018/pdf/CBECS%202018%20CE%20Release%202%20Flipbook.pdf>

[2] Buildings characteristics report.

[https://www.eia.gov/consumption/commercial/data/2018/pdf/CBECS\\_2018\\_Building\\_Characteristics\\_Flipbook.pdf](https://www.eia.gov/consumption/commercial/data/2018/pdf/CBECS_2018_Building_Characteristics_Flipbook.pdf)

[3] User's Guide to the 2018 CBECS Public Use Microdata File:

<https://www.eia.gov/consumption/commercial/data/2018/pdf/Users%20Guide%20to%20the%202018%20CBECS%20Public%20Use%20Microdata%20File.pdf>

[4] Zimmer, S., Powell, R.J., and Velásquez, S. (2024). Exploring Complex Survey Data Analysis Using R.

<https://tidy-survey-r.github.io/tidy-survey-book/references.html>

[5] microdata/dataset.

<https://www.eia.gov/consumption/commercial/data/2018/index.php?view=microdata>

[6] Methodology.

<https://www.eia.gov/consumption/commercial/data/2018/index.php?view=methodology>

[7] How We Collected Data Using the 2018 CBECS Buildings Survey.

<https://www.eia.gov/consumption/commercial/reports/2018/data-collection-buildings.php>

[8] Response Rates and Nonresponse Bias in the 2018 CBECS Buildings Survey.

<https://www.eia.gov/consumption/commercial/reports/2018/response.php>

[9] How We Chose Buildings for the 2018 CBECS

<https://www.eia.gov/consumption/commercial/reports/2018/methodology/sampling.php>

[10] Determining Building Eligibility

<https://www.eia.gov/consumption/commercial/reports/2018/building-eligibility.php>

[11] How We Collected Data Using the 2018 CBECS Energy Supplier Survey:

<https://www.eia.gov/consumption/commercial/reports/2018/data-collection-energy.php>

[12] How We Reviewed Data to Ensure Quality of the 2018 CBECS

<https://www.eia.gov/consumption/commercial/reports/2018/data-quality.php>

# APPENDIX

## Data Analysis - Modelling

In this portion of our report, we will be employing linear regression to predict Energy Consumption from Commercial Building features in the data; specifically, (1) Total Floorspace; (2) Building Activity & (3) Geographical Location (average temperature). Consequently, results will be plotted and interpreted appropriately.

For Total Floorspace, we will be making use of the data field, SQFT - Square Footage, which is a continuous variable depicting the total square footage utilized by the building.

For Building Activity, we will be making use of the data field, PBA - Principal building activity, which is a categorical variable that represents the type of activity that is being performed in the building.

For Geographical Location (Average Temperature), we will be utilizing PUBCLIM - ASHRAE climate zone, which is a categorical variable that represents the climate zone where the building is located.

For our variable response of interest, we will be observing the MFBTU continuous variable data field which represents the Annual energy consumption in (thous btu) of the commercial buildings.

```
#read the dataset  
CBECS <- read.csv("cbeecs2018_final_public.csv", header = TRUE)
```

```
#applied the weights
```

```
CBECS_des <- CBECS %>%  
as_survey_rep(weights = FINALWT,  
repweights = FINALWT1:FINALWT151,  
type = "JK2",  
mse = TRUE)
```

```
## Warning in svrepdesign.default(variables = variables, repweights = repweights,  
## : with type JK2 scale= and rscales= are not needed and will be ignored
```

```
#assigned the sampling weight to samp_weight  
samp_weight <- CBECS_des$pweights
```

```
#Created new variables log_SQFT, log_HDD65, log_CDD65, log_MFBTU
```

```
#Created new columns in the dataset containing log transformed values of the numerical values
```

```
CBECS$log_SQFT <- log(CBECS$SQFT)  
CBECS$log_HDD65 <- log(CBECS$HDD65)  
CBECS$log_CDD65 <- log(CBECS$CDD65)  
CBECS$log_MFBTU <- log(CBECS$MFBTU)
```

```
#reiterated the survey design object as we now have new columns  
CBECS_des <- CBECS %>%
```



```
as_survey_rep(weights = FINALWT,
repweights = FINALWT1:FINALWT151,
type = "JK2",
mse = TRUE)
```

```
## Warning in svrepdesign.default(variables = variables, repweights = repweights,
## : with type JK2 scale= and rscales= are not needed and will be ignored
```

```
#assigned the sampling weight to samp_weight
samp_weight <- CBECS_des$pweights
```

Given that we have a lot of levels for PBA, we will first investigate if we can group some levels together.

## PBA

```
#Calculated counts for each category
category_counts <- table(CBECS$PBA)
category_counts
```

```
##
##      1      2      4      5      6      7      8     11     12     13     14     15     16     17     18     23
## 117 1329     69    753     91    105    217     23    271    481    936    218    276    114    418    208
##    24     25     26     91
##    35    339    350     86
```

```
#Got the average of the counts
average_count <- mean(category_counts)
average_count
```

```
## [1] 321.8
```

Based on the results above, we can see that there are several levels of PBA that are below the mean frequency. Given that the PBA levels pertain to Principal building activity which is not ordinal, then, we can combine levels that are below the mean value regardless of position.

```
#defined PBA as factors
CBECS$PBA <- as.factor(CBECS$PBA)

#Replaced levels "4", "6", "7", with "1"
levels(CBECS$PBA)[levels(CBECS$PBA) %in% c("4", "6", "7")] <- "1"
```

```
#Calculated counts for each category
category_counts <- table(CBECS$PBA)
category_counts
```

```
##
##      1      2      5      8     11     12     13     14     15     16     17     18     23     24     25     26
## 382 1329    753    217     23    271    481    936    218    276    114    418    208     35    339    350
##    91
##    86
```

```
#Got the average of the counts
average_count <- mean(category_counts)
average_count
```

```
## [1] 378.5882
```

We have combined 3 levels to get a count of 382 which is only slightly higher than the mean. Now our new frequency mean is 378.5882. We shall now combine levels that are still lower than this value until the combined level reaches a value count that is near this new mean value.

```
#Replaced levels "11", "17", "24", with "8"
levels(CBECS$PBA)[levels(CBECS$PBA) %in% c("11", "17", "24")] <- "8"
```

```
#Calculated counts for each category
category_counts <- table(CBECS$PBA)
category_counts
```

```
##
##      1      2      5      8     12     13     14     15     16     18     23     25     26     91
## 382 1329   753   389   271   481   936   218   276   418   208   339   350   86
```

```
#Got the average of the counts
average_count <- mean(category_counts)
average_count
```

```
## [1] 459.7143
```

We will just repeat the process until all count values are relatively close to the mean count.

```
#Replaced levels "23", "91", with "15"
levels(CBECS$PBA)[levels(CBECS$PBA) %in% c("23", "91")] <- "15"
```

```
#Calculated counts for each category
category_counts <- table(CBECS$PBA)
category_counts
```

```
##
##      1      2      5      8     12     13     14     15     16     18     25     26
## 382 1329   753   389   271   481   936   512   276   418   339   350
```

```
#Got the average of the counts
average_count <- mean(category_counts)
average_count
```

```
## [1] 536.3333
```

```
#Replaced levels "12", with "16"
levels(CBECS$PBA)[levels(CBECS$PBA) %in% c("12")] <- "16"
```

```
#Calculated counts for each category
category_counts <- table(CBECS$PBA)
category_counts
```

```
##
##      1      2      5      8     16     13     14     15     18     25     26
## 382 1329  753  389  547  481  936  512  418  339  350
```

```
#Got the average of the counts
average_count <- mean(category_counts)
average_count
```

```
## [1] 585.0909
```

Now we can stop since combining the two levels with the lowest count will already exceed the mean value by a margin larger than the current difference between the mean and the two lowest values individually.

Next, we will also perform the same steps on PUBCLIM as another categorical variable.

```
#PUBCLIM
```

```
#Calculated counts for each category
category_counts <- table(CBECS$PUBCLIM)
category_counts
```

```
##
##      1      2      3      4      5      7
## 422 1428 1479 1483  772  852
```

For PUBCLIM, we will not be grouping levels since disparity is not that large.

Now, we can perform the regression of log\_MFBTU on log\_SQFT which is our promising predictor.

## Linear Regression

### Single with Weight

```
#Applied the weighted variables on the linear regression
weighted_SQFT <- samp_weight * CBECS$log_SQFT
weighted_MFBTU <- samp_weight * CBECS$log_MFBTU
weighted_MFBTU <- weighted_MFBTU + 0.0001
weighted_SQFT <- weighted_SQFT + 0.0001

library(mgcv)
```

```
## Warning: package 'mgcv' was built under R version 4.2.3
```

```
## This is mgcv 1.8-42. For overview type 'help("mgcv-package")'.
```

```
##
## Attaching package: 'mgcv'

## The following object is masked from 'package:mosaic':
##
##      cnorm

# Fit the linear regression model with weighted variables
m1 <- glm(weighted_MFBTU ~ weighted_SQFT, family = Gamma(link = "identity"))
m1_stdres <- rstandard(m1)
m1_fitted <- fitted(m1)
summary(m1)

##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT, family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82188  -0.04389  -0.00599   0.03498   0.38953
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.676101   0.370213  -15.33  <2e-16 ***
## weighted_SQFT  1.400302   0.001681  833.22  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.006488715)
##
##      Null deviance: 18472.406  on 6356  degrees of freedom
## Residual deviance:   41.489  on 6355  degrees of freedom
## (79 observations deleted due to missingness)
## AIC: 86041
##
## Number of Fisher Scoring iterations: 5
```

Assumptions of the model have been plotted and discussed in the Data Analysis-Modeling part of the paper.

With a p-value of 2e-16, the model is significant, which means that our predictor log\_SQFT is a significant predictor of log\_MFBTU as the p-value is a very small value close to 0.

Next, we will see how well the model fares with the categorical variables.

#Regression PBA

First, we will obtain the regression with PBA. Before we perform the regression, we will look at the frequency of each category so we can determine which category to select as the reference category.

```
#Got the table to see number of counts
table(CBECS$PBA)
```

```
##
##      1      2      5      8     16     13     14     15     18     25     26
## 382 1329   753   389   547   481   936   512   418   339   350
```

Referring to the table above, we can see that category 2 is the level with the highest number of counts which makes sense for us to select this as the reference category.

## PBA

```
#fitted regression model for PBA only
m2 <- glm(weighted_MFBTU ~ relevel(CBECS$PBA, ref = "2"), family = Gamma(link = "identity"))
m2_stdres <- rstandard(m2)
m2_fitted<-fitted(m2)
summary(m2)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ relevel(CBECS$PBA, ref = "2"),
##      family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2102  -2.2067  -1.0696   0.2329   7.0555
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9210.2      413.3  22.283 < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")1    4430.7      1262.5   3.510 0.000452 ***
## relevel(CBECS$PBA, ref = "2")5    3993.9       909.0   4.394 1.13e-05 ***
## relevel(CBECS$PBA, ref = "2")8   -3321.5       639.9  -5.191 2.16e-07 ***
## relevel(CBECS$PBA, ref = "2")16    895.4       818.9   1.093 0.274235
## relevel(CBECS$PBA, ref = "2")13   3636.2      1043.6   3.484 0.000497 ***
## relevel(CBECS$PBA, ref = "2")14  -3048.4       528.6  -5.767 8.45e-09 ***
## relevel(CBECS$PBA, ref = "2")15   5554.9      1148.7   4.836 1.36e-06 ***
## relevel(CBECS$PBA, ref = "2")18  -3685.1       605.2  -6.089 1.21e-09 ***
## relevel(CBECS$PBA, ref = "2")25   3674.6      1218.8   3.015 0.002581 **
## relevel(CBECS$PBA, ref = "2")26  20390.0      2625.0   7.768 9.26e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 2.676572)
##
##      Null deviance: 18472  on 6356  degrees of freedom
## Residual deviance: 17282  on 6346  degrees of freedom
##      (79 observations deleted due to missingness)
## AIC: 126937
##
## Number of Fisher Scoring iterations: 3
```

Here, we see that while regressing on PBA, most levels are significant, regressing on PBA alone may not be enough because it has a high AIC value. With this said, we will proceed to regress with the interaction between SQFT & PBA.

## SQFT + PBA

*#fitted regression model including interaction terms for PBA*

```
m3 <- glm(weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PBA, ref = "2") + weighted_SQFT * relevel(CBECS$PBA, ref = "2"),
m3_stdres <- rstandard(m3)
m3_fitted<-fitted(m3)
summary(m3)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PBA,
##       ref = "2") + weighted_SQFT * relevel(CBECS$PBA, ref = "2"),
##       family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88388  -0.03726  -0.00397   0.03216   0.37302
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept)      -7.8740411   0.6671392 -11.803
## weighted_SQFT       1.4020006   0.0033805  414.733
## relevel(CBECS$PBA, ref = "2")1    2.7903597   1.1009337   2.535
## relevel(CBECS$PBA, ref = "2")5    1.2985151   1.4960136   0.868
## relevel(CBECS$PBA, ref = "2")8   -0.2976475   1.5599691  -0.191
## relevel(CBECS$PBA, ref = "2")16  10.3735198   1.2191657   8.509
## relevel(CBECS$PBA, ref = "2")13   2.4548096   1.4968508   1.640
## relevel(CBECS$PBA, ref = "2")14   4.1331914   1.0950030   3.775
## relevel(CBECS$PBA, ref = "2")15  -0.4690380   1.9250594  -0.244
## relevel(CBECS$PBA, ref = "2")18   3.0480973   1.2223815   2.494
## relevel(CBECS$PBA, ref = "2")25   2.0975306   2.4779604   0.846
## relevel(CBECS$PBA, ref = "2")26  -8.0512863   4.3745049  -1.841
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")1    0.0319239   0.0068787   4.641
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")5   -0.1044982   0.0051814 -20.168
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")8    0.0255764   0.0070458   3.630
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")16  -0.0275641   0.0064490  -4.274
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")13   0.0155776   0.0062491   2.493
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")14  -0.0333098   0.0050704  -6.569
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")15   0.1323866   0.0060535  21.870
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")18  -0.0039572   0.0075123  -0.527
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")25   0.0005824   0.0072568   0.080
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")26   0.0198728   0.0064683   3.072
##
##              Pr(>|t|)
## (Intercept)      < 2e-16 ***
## weighted_SQFT      < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")1    0.011283 *
## relevel(CBECS$PBA, ref = "2")5    0.385436
## relevel(CBECS$PBA, ref = "2")8    0.848686
## relevel(CBECS$PBA, ref = "2")16   < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")13   0.101058
## relevel(CBECS$PBA, ref = "2")14   0.000162 ***
## relevel(CBECS$PBA, ref = "2")15   0.807511
## relevel(CBECS$PBA, ref = "2")18   0.012672 *
```

```
## relevel(CBECS$PBA, ref = "2")25          0.397320
## relevel(CBECS$PBA, ref = "2")26          0.065741 .
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")1 3.54e-06 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")5 < 2e-16 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")8 0.000286 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")16 1.95e-05 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")13 0.012700 *
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")14 5.45e-11 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")15 < 2e-16 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")18 0.598378
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")25 0.936034
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")26 0.002133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.004848124)
##
## Null deviance: 18472.406 on 6356 degrees of freedom
## Residual deviance: 31.744 on 6335 degrees of freedom
## (79 observations deleted due to missingness)
## AIC: 84377
##
## Number of Fisher Scoring iterations: 6
```

The main effects of PBA categories other than 1, 16, 14 & 18 are not significant. However, their interaction with SQFT are all significant except for only 18 and 25. Therefore, we will retain them as this may mean that they are only not significantly different from the reference level but may be significantly different from other levels.

Below, we also get the 95% CI for the regression coefficients

## CI

```
#Got the 95% confidence intervals for the regression coefficients
confint(m3)
```

```
## Waiting for profiling to be done...
```

```
##              2.5 %      97.5 %
## (Intercept) -9.089706529 -6.63701480
## weighted_SQFT 1.395535234 1.40848620
## relevel(CBECS$PBA, ref = "2")1 0.785484958 4.82579499
## relevel(CBECS$PBA, ref = "2")5 -1.493195483 4.14034882
## relevel(CBECS$PBA, ref = "2")8 -3.254932117 2.68763718
## relevel(CBECS$PBA, ref = "2")16 8.023904617 12.72695081
## relevel(CBECS$PBA, ref = "2")13 -0.412720394 5.34851282
## relevel(CBECS$PBA, ref = "2")14 2.051455564 6.21839502
## relevel(CBECS$PBA, ref = "2")15 -4.048059147 3.20779282
## relevel(CBECS$PBA, ref = "2")18 0.705300219 5.39073753
## relevel(CBECS$PBA, ref = "2")25 -2.681848059 6.93901085
## relevel(CBECS$PBA, ref = "2")26 -16.333977688 0.60794223
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")1 0.018685146 0.04523472
```

```
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")5    -0.114497888 -0.09448249
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")8      0.011955252  0.03928372
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")16    -0.040105715 -0.01495609
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")13     0.003457577  0.02775746
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")14    -0.043121064 -0.02348131
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")15     0.120660829  0.14416140
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")18    -0.018524197  0.01072250
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")25    -0.013500693  0.01476776
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")26     0.007316892  0.03250126
```

The coefficient of log\_SQFT is within the confidence interval with a narrow interval; not all terms including interaction terms have coefficients within their respective confidence intervals. It is also important to note that some terms have a wide interval where we have to be extra careful of our predictions.

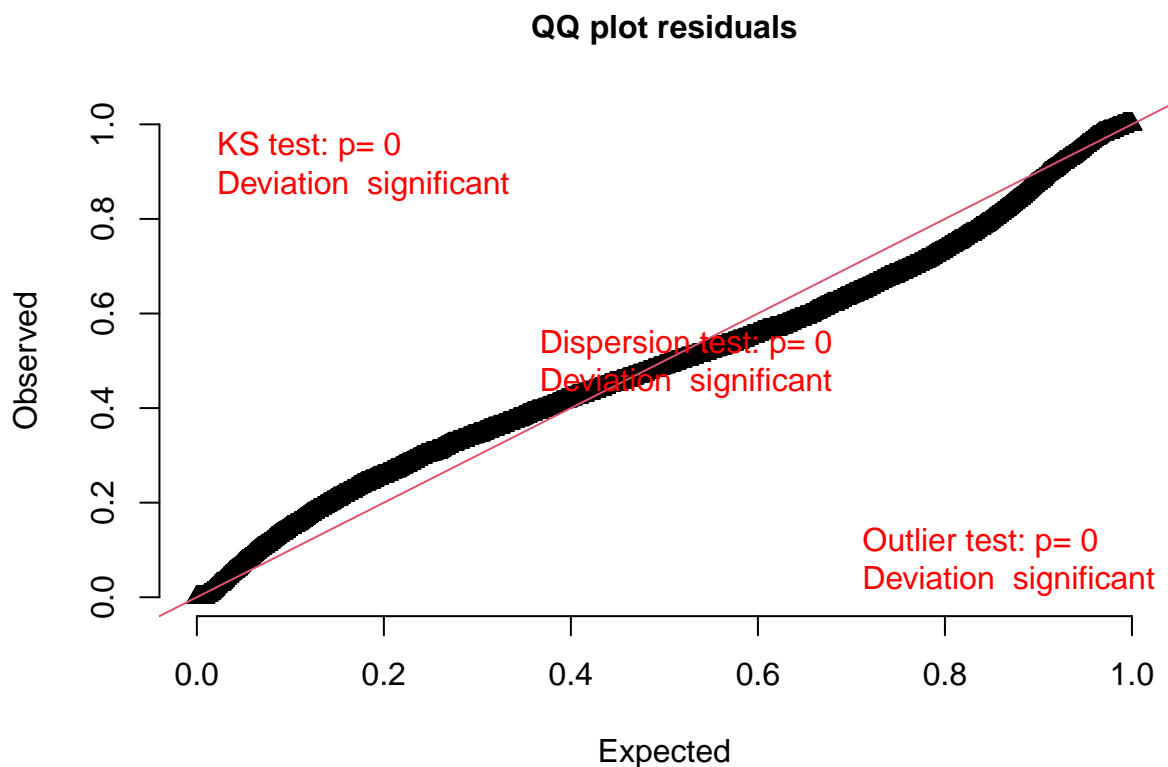
## Assumptions

```
#checked the assumptions
library(DHARMA)
```

```
## Warning: package 'DHARMA' was built under R version 4.2.3
```

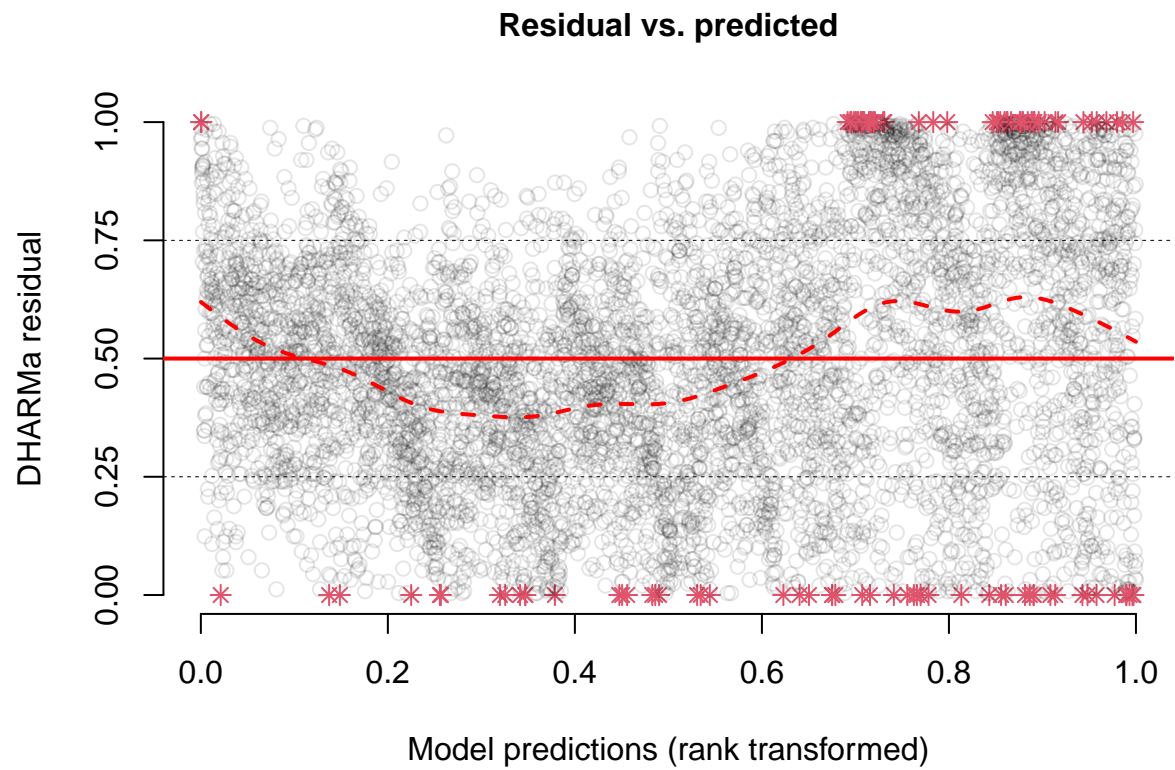
```
## This is DHARMA 0.4.6. For overview type '?DHARMA'. For recent changes, type news(package = 'DHARMA')
```

```
simulationOutput <- simulateResiduals(fittedModel = m3, plt = F)
plotQQunif(simulationOutput)
```

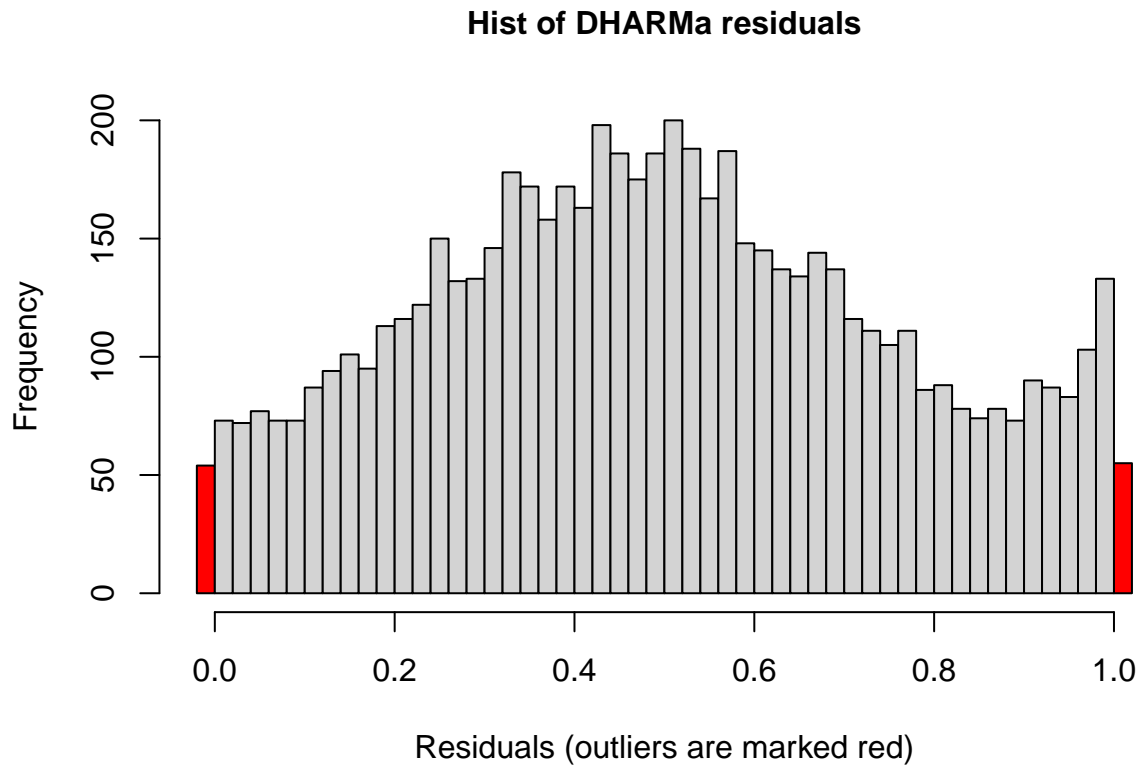




```
plotResiduals(simulationOutput)
```



```
hist(simulationOutput)
```



These plots result to the same conclusions as the interpretation of the plots found in Data Analysis - Modeling part of the paper.

## Regression PUBCLIM

We will now repeat the same steps on PUBCLIM.

### PUBCLIM

```
#fitted regression model for PUBCLIM only
CBECS$PUBCLIM <- as.factor(CBECS$PUBCLIM)
m21 <- glm(weighted_MFBTU ~ relevel(CBECS$PUBCLIM, ref = "4"), family = Gamma(link = "identity"))
m21_stdres <- rstandard(m21)
m21_fitted<-fitted(m21)
summary(m21)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ relevel(CBECS$PUBCLIM, ref = "4"),
##      family = Gamma(link = "identity"))
##
## Deviance Residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -3.187 -1.909 -1.025   0.363   6.261
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          12660.5      553.0  22.893 < 2e-16 ***
## relevel(CBECS$PUBCLIM, ref = "4")1    2128.5      1329.3    1.601   0.109
## relevel(CBECS$PUBCLIM, ref = "4")2    1726.8       844.8    2.044   0.041 *
## relevel(CBECS$PUBCLIM, ref = "4")3   -2838.0       698.6   -4.062 4.91e-05 ***
## relevel(CBECS$PUBCLIM, ref = "4")5    -277.7       929.4   -0.299   0.765
## relevel(CBECS$PUBCLIM, ref = "4")7  -10259.3       569.7  -18.007 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 2.772356)
##
##      Null deviance: 18472  on 6356  degrees of freedom
## Residual deviance: 17001  on 6351  degrees of freedom
##      (79 observations deleted due to missingness)
## AIC: 126788
##
## Number of Fisher Scoring iterations: 3
```

Based on the summary above, only level 1 and level 5 are not significant levels of PUBCLIM. Here we see that the model also has high AIC values which hints that it is not enough to regress only on our categorical variables PBA & PUBCLIM. With this said, we will proceed to regress with the interaction between SQFT & PUBCLIM.

##SQFT + PUBCLIM

```
#Got the table to see number of counts
table(CBECS$PUBCLIM)
```

```
##
##      1      2      3      4      5      7
##  422 1428 1479 1483   772   852
```

Referring to the table above, we can see that category 4 is the level with the highest number of counts which makes sense for us to select this as the reference category.

```
#fitted regression model including interaction terms for PUBCLIM
m6 <- glm(weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,ref="4") + weighted_SQFT * relevel(CBECS$PUBCLIM,ref="4"),
m6_stdres <- rstandard(m6)
m6_fitted<-fitted(m6)
summary(m6)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,
##      ref = "4") + weighted_SQFT * relevel(CBECS$PUBCLIM, ref = "4"),
##      family = Gamma(link = "identity"))
##
## Deviance Residuals:
```

```
##      Min      1Q      Median      3Q      Max
## -0.81813 -0.04350 -0.00576  0.03448  0.39667
##
## Coefficients:
##
##              Estimate Std. Error t value
## (Intercept)      -6.386567   0.856188  -7.459
## weighted_SQFT       1.391591   0.003277 424.638
## relevel(CBECS$PUBCLIM, ref = "4")1    -4.026476   2.137554  -1.884
## relevel(CBECS$PUBCLIM, ref = "4")2    -0.694565   1.255663  -0.553
## relevel(CBECS$PUBCLIM, ref = "4")3     1.199943   1.115389   1.076
## relevel(CBECS$PUBCLIM, ref = "4")5    -1.345043   1.472367  -0.914
## relevel(CBECS$PUBCLIM, ref = "4")7     4.215398   1.255090   3.359
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")1  0.033252   0.007271   4.573
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")2  0.021021   0.004782   4.396
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")3  0.008728   0.004722   1.848
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")5  0.002025   0.005803   0.349
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")7 -0.009528   0.007004  -1.360
##
##              Pr(>|t|)
## (Intercept)      9.86e-14 ***
## weighted_SQFT      < 2e-16 ***
## relevel(CBECS$PUBCLIM, ref = "4")1      0.059653 .
## relevel(CBECS$PUBCLIM, ref = "4")2      0.580183
## relevel(CBECS$PUBCLIM, ref = "4")3      0.282054
## relevel(CBECS$PUBCLIM, ref = "4")5      0.361002
## relevel(CBECS$PUBCLIM, ref = "4")7      0.000788 ***
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")1 4.90e-06 ***
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")2 1.12e-05 ***
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")3 0.064599 .
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")5 0.727174
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")7 0.173763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.00643468)
##
##      Null deviance: 18472.406  on 6356  degrees of freedom
## Residual deviance:   41.061  on 6345  degrees of freedom
## (79 observations deleted due to missingness)
## AIC: 85995
##
## Number of Fisher Scoring iterations: 5
```

The main effects of PUBCLIM categories other than 7 are not significant. However, their interaction with SQFT are significant for levels 1 and 2.

## CI

```
#Got the 95% confidence intervals for the regression coefficients
confint(m6)
```

```
## Waiting for profiling to be done...
```

	2.5 %	97.5 %
## (Intercept)	-7.9326701569	-4.794938623
## weighted_SQFT	1.3852644250	1.397946873
## relevel(CBECS\$PUBCLIM, ref = "4")1	-8.1003631057	0.104381730
## relevel(CBECS\$PUBCLIM, ref = "4")2	-3.0530212779	1.656516105
## relevel(CBECS\$PUBCLIM, ref = "4")3	-0.8642122850	3.250138215
## relevel(CBECS\$PUBCLIM, ref = "4")5	-4.1344702835	1.452322150
## relevel(CBECS\$PUBCLIM, ref = "4")7	1.8295652975	6.584118056
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")1	0.0191250494	0.047481327
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")2	0.0117376980	0.030310049
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")3	-0.0004088095	0.017865951
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")5	-0.0092487218	0.013338733
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")7	-0.0231320581	0.004151269

Similar to the CI for SQFT & PBA, SQFT here is within a narrow interval while there are other levels that have wide intervals.

#Multiple Regression on all Covariates of interest

Since we have now settled that all categorical variables have significant interactions with log\_SQFT; we will be looking at the performance of combining all covariates of interest.

*#fitted regression model including interaction terms for PUBCLIM and PBA*

```
m7 <- glm(weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,ref="4") + relevel(CBECS$PBA,ref="2")
m7_stdres <- rstandard(m7)
m7_fitted<-fitted(m7)
summary(m7)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,
##   ref = "4") + relevel(CBECS$PBA, ref = "2") + weighted_SQFT *
##   relevel(CBECS$PUBCLIM, ref = "4") + weighted_SQFT * relevel(CBECS$PBA,
##   ref = "2"), family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87987  -0.03711  -0.00378   0.03252   0.38138
##
## Coefficients:
##
##              Estimate Std. Error t value
## (Intercept)      -9.131877    1.008502  -9.055
## weighted_SQFT       1.395657    0.004208 331.676
## relevel(CBECS$PUBCLIM, ref = "4")1    -2.331856    1.894827  -1.231
## relevel(CBECS$PUBCLIM, ref = "4")2     0.427203    1.114226   0.383
## relevel(CBECS$PUBCLIM, ref = "4")3     2.257555    1.005367   2.246
## relevel(CBECS$PUBCLIM, ref = "4")5    -0.851861    1.306953  -0.652
## relevel(CBECS$PUBCLIM, ref = "4")7    -1.014600    1.280565  -0.792
## relevel(CBECS$PBA, ref = "2")1         3.499425    1.132513   3.090
## relevel(CBECS$PBA, ref = "2")5         2.369086    1.511425   1.567
## relevel(CBECS$PBA, ref = "2")8         0.317226    1.561938   0.203
## relevel(CBECS$PBA, ref = "2")16        13.119877    1.452518   9.033
## relevel(CBECS$PBA, ref = "2")13         3.342620    1.507411   2.217
## relevel(CBECS$PBA, ref = "2")14         5.642250    1.172779   4.811
```

```

## relevel(CBECS$PBA, ref = "2")15      0.968353  1.949152  0.497
## relevel(CBECS$PBA, ref = "2")18      4.777221  1.348975  3.541
## relevel(CBECS$PBA, ref = "2")25      4.034561  2.492798  1.618
## relevel(CBECS$PBA, ref = "2")26     -7.177404  4.360668 -1.646
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")1  0.029484  0.006303  4.678
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")2  0.016673  0.004152  4.015
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")3  0.009803  0.004100  2.391
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")5 -0.002595  0.005025 -0.516
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")7  0.002665  0.006751  0.395
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")1      0.031089  0.006885  4.515
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")5     -0.106473  0.005185 -20.533
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")8      0.022956  0.007035  3.263
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")16    -0.028189  0.006545 -4.307
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")13     0.013523  0.006243  2.166
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")14    -0.036476  0.005165 -7.062
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")15     0.130320  0.006054 21.528
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")18    -0.001606  0.008227 -0.195
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")25    -0.002229  0.007249 -0.307
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")26     0.015782  0.006486  2.433
##
## (Intercept) < 2e-16 ***
## weighted_SQFT < 2e-16 ***
## relevel(CBECS$PUBCLIM, ref = "4")1  0.218502
## relevel(CBECS$PUBCLIM, ref = "4")2  0.701430
## relevel(CBECS$PUBCLIM, ref = "4")3  0.024770 *
## relevel(CBECS$PUBCLIM, ref = "4")5  0.514559
## relevel(CBECS$PUBCLIM, ref = "4")7  0.428212
## relevel(CBECS$PBA, ref = "2")1      0.002010 **
## relevel(CBECS$PBA, ref = "2")5      0.117059
## relevel(CBECS$PBA, ref = "2")8      0.839065
## relevel(CBECS$PBA, ref = "2")16     < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")13     0.026627 *
## relevel(CBECS$PBA, ref = "2")14     1.54e-06 ***
## relevel(CBECS$PBA, ref = "2")15     0.619342
## relevel(CBECS$PBA, ref = "2")18     0.000401 ***
## relevel(CBECS$PBA, ref = "2")25     0.105608
## relevel(CBECS$PBA, ref = "2")26     0.099826 .
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")1 2.96e-06 ***
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")2 6.01e-05 ***
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")3 0.016841 *
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")5 0.605566
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")7 0.693029
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")1      6.44e-06 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")5     < 2e-16 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")8      0.001109 **
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")16     1.68e-05 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")13     0.030349 *
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")14     1.82e-12 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")15     < 2e-16 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")18     0.845227
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")25     0.758507
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")26     0.014983 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
##
## (Dispersion parameter for Gamma family taken to be 0.004801233)
##
## Null deviance: 18472.406 on 6356 degrees of freedom
## Residual deviance: 31.374 on 6325 degrees of freedom
## (79 observations deleted due to missingness)
## AIC: 84323
##
## Number of Fisher Scoring iterations: 6
```

## CI

```
#Got the 95% confidence intervals for the regression coefficients
confinf(m7)
```

```
## Waiting for profiling to be done...
```

	2.5 %	97.5 %
## (Intercept)	-11.040518343	-7.210390684
## weighted_SQFT	1.387548805	1.403795264
## relevel(CBECS\$PUBCLIM, ref = "4")1	-5.984153976	1.361569449
## relevel(CBECS\$PUBCLIM, ref = "4")2	-1.700027751	2.553824559
## relevel(CBECS\$PUBCLIM, ref = "4")3	0.356158365	4.155766573
## relevel(CBECS\$PUBCLIM, ref = "4")5	-3.364306280	1.669643479
## relevel(CBECS\$PUBCLIM, ref = "4")7	-3.501978106	1.467806534
## relevel(CBECS\$PBA, ref = "2")1	1.402298987	5.627145116
## relevel(CBECS\$PBA, ref = "2")5	-0.468650170	5.255881693
## relevel(CBECS\$PBA, ref = "2")8	-2.656264834	3.319109747
## relevel(CBECS\$PBA, ref = "2")16	10.281783647	15.958868469
## relevel(CBECS\$PBA, ref = "2")13	0.439663641	6.272646293
## relevel(CBECS\$PBA, ref = "2")14	3.375659527	7.913688638
## relevel(CBECS\$PBA, ref = "2")15	-2.671168038	4.701552210
## relevel(CBECS\$PBA, ref = "2")18	2.167439155	7.394094233
## relevel(CBECS\$PBA, ref = "2")25	-0.779780973	8.909043679
## relevel(CBECS\$PBA, ref = "2")26	-15.444308210	1.460623814
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")1	0.017210363	0.041833238
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")2	0.008593188	0.024756302
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")3	0.001848597	0.017757685
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")5	-0.012383069	0.007223219
## weighted_SQFT:relevel(CBECS\$PUBCLIM, ref = "4")7	-0.010481006	0.015856188
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")1	0.017809890	0.044438517
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")5	-0.116507930	-0.096426182
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")8	0.009332657	0.036661874
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")16	-0.040964579	-0.015345948
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")13	0.001390325	0.025712289
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")14	-0.046520734	-0.026416966
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")15	0.118571211	0.142113802
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")18	-0.017559856	0.014450004
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")25	-0.016312219	0.011953426
## weighted_SQFT:relevel(CBECS\$PBA, ref = "2")26	0.003176854	0.028456155

## Multiple Regression Main Effects Only

We have previously identified that we have good performing models that deal with interaction effects. Now, we will also look at how models fare if only the main effects are included.

### PBA Main Effect

*#fitted regression model including main effects of log\_SQFT & PBA*

```
m79 <- glm(weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PBA,ref="2"), family = Gamma(link = "identity"))
m79_stdres <- rstandard(m79)
m79_fitted<-fitted(m79)
summary(m79)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PBA,
##       ref = "2"), family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82278  -0.04167  -0.00644   0.03219   0.38894
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -7.82589    0.62295  -12.563  < 2e-16 ***
## weighted_SQFT       1.40158    0.00166  844.105  < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")1    4.83074    1.06740   4.526 6.13e-06 ***
## relevel(CBECS$PBA, ref = "2")5  -14.02736    1.35502  -10.352  < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")8    3.12563    1.41635   2.207  0.02736 *
## relevel(CBECS$PBA, ref = "2")16   6.99942    1.04468   6.700 2.26e-11 ***
## relevel(CBECS$PBA, ref = "2")13   4.54473    1.41969   3.201  0.00138 **
## relevel(CBECS$PBA, ref = "2")14  -0.13822    0.99270  -0.139  0.88927
## relevel(CBECS$PBA, ref = "2")15  16.20302    2.08298   7.779 8.49e-15 ***
## relevel(CBECS$PBA, ref = "2")18   2.61577    1.00091   2.613  0.00899 **
## relevel(CBECS$PBA, ref = "2")25   2.25993    2.30276   0.981  0.32643
## relevel(CBECS$PBA, ref = "2")26  -3.53256    4.74999  -0.744  0.45709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.006215644)
##
## Null deviance: 18472.406  on 6356  degrees of freedom
## Residual deviance: 39.616  on 6345  degrees of freedom
## (79 observations deleted due to missingness)
## AIC: 85767
##
## Number of Fisher Scoring iterations: 6
```



## PUBCLIM Main Effect

```
#fitted regression model with only main effects of log_SQFT and PUBCLIM
```

```
m71 <- glm(weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,ref="4"), family = Gamma(link = "identity"))
m71_stdres <- rstandard(m71)
m71_fitted<-fitted(m71)
summary(m71)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,
##   ref = "4"), family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82173  -0.04345  -0.00616   0.03427   0.38952
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -7.340164    0.782833  -9.376 < 2e-16 ***
## weighted_SQFT     1.400476    0.001682 832.655 < 2e-16 ***
## relevel(CBECS$PUBCLIM, ref = "4")1  0.719266    1.847977   0.389 0.69713
## relevel(CBECS$PUBCLIM, ref = "4")2  1.847529    1.097058   1.684 0.09222 .
## relevel(CBECS$PUBCLIM, ref = "4")3  2.137695    0.973814   2.195 0.02819 *
## relevel(CBECS$PUBCLIM, ref = "4")5 -1.311536    1.256403  -1.044 0.29658
## relevel(CBECS$PUBCLIM, ref = "4")7  3.022784    0.940822   3.213 0.00132 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.006476161)
##
## Null deviance: 18472.406 on 6356 degrees of freedom
## Residual deviance: 41.352 on 6350 degrees of freedom
## (79 observations deleted due to missingness)
## AIC: 86030
##
## Number of Fisher Scoring iterations: 5
```

## Main Effects of PBA & PUBCLIM

```
#fitted regression model with only main effects of log_SQFT, PUBCLIM and PBA
```

```
m17 <- glm(weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,ref="4") + relevel(CBECS$PBA,ref="2"))
m17_stdres <- rstandard(m17)
m17_fitted<-fitted(m17)
summary(m17)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,
```

```
##      ref = "4") + relevel(CBECS$PBA, ref = "2"), family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -0.82275  -0.04155  -0.00664   0.03198   0.38896
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -10.452926    0.969209  -10.785  < 2e-16 ***
## weighted_SQFT      1.401797    0.001661  843.736  < 2e-16 ***
## relevel(CBECS$PUBCLIM, ref = "4")1    2.247727    1.850839    1.214  0.224626
## relevel(CBECS$PUBCLIM, ref = "4")2    3.307667    1.097449    3.014  0.002589 **
## relevel(CBECS$PUBCLIM, ref = "4")3    3.743753    0.990788    3.779  0.000159 ***
## relevel(CBECS$PUBCLIM, ref = "4")5    0.018119    1.254860    0.014  0.988480
## relevel(CBECS$PUBCLIM, ref = "4")7    1.023454    1.127444    0.908  0.364037
## relevel(CBECS$PBA, ref = "2")1      5.669949    1.095367    5.176  2.33e-07 ***
## relevel(CBECS$PBA, ref = "2")5     -13.192954    1.369615   -9.633  < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")8      3.416614    1.420707    2.405  0.016207 *
## relevel(CBECS$PBA, ref = "2")16     8.569071    1.338359    6.403  1.64e-10 ***
## relevel(CBECS$PBA, ref = "2")13     5.113669    1.430779    3.574  0.000354 ***
## relevel(CBECS$PBA, ref = "2")14     0.564784    1.058360    0.534  0.593609
## relevel(CBECS$PBA, ref = "2")15    17.453771    2.104543    8.293  < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")18     3.938775    1.136646    3.465  0.000533 ***
## relevel(CBECS$PBA, ref = "2")25     3.520309    2.322594    1.516  0.129650
## relevel(CBECS$PBA, ref = "2")26    -3.380488    4.745445   -0.712  0.476265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.006197877)
##
##      Null deviance: 18472.406  on 6356  degrees of freedom
## Residual deviance:   39.469  on 6340  degrees of freedom
##      (79 observations deleted due to missingness)
## AIC: 85753
##
## Number of Fisher Scoring iterations: 7
```

Next, we shall now compare the models' AIC scores to determine the best fitting model.

## Model Selection

```
#Created AIC table with the AIC values of the significant models
AIC_table <- data.frame(Model_Number = c(1, 2, 3, 4, 5, 6, 7, 8, 9), Predictors = c("log_SQFT", "PBA",
AIC_table

##      Model_Number      Predictors
## 1              1      log_SQFT
## 2              2              PBA
## 3              3 log_SQFT + PBA + log_SQFT*PBA
## 4              4      PUBCLIM
## 5              5 log_SQFT + PUBCLIM + log_SQFT*PUBCLIM
```

```
## 6      6 log_SQFT + PBA + PUBCLIM + log_SQFT*PBA + log_SQFT*PUBCLIM
## 7      7                                log_SQFT + PBA
## 8      8                                log_SQFT + PUBCLIM
## 9      9                                log_SQFT + PBA + PUBCLIM
##      AIC
## 1 86041
## 2 126937
## 3 84377
## 4 126788
## 5 85995
## 6 84323
## 7 85767
## 8 86030
## 9 85753
```

Based on the table above, we can now confirm that the best model is model number 6 with all covariates of interest including interaction terms. It has the lowest AIC of 84323.

## Model Evaluation

Now, we will proceed to evaluate our best model.

```
#got summary of the model
summary(m7)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS$PUBCLIM,
##   ref = "4") + relevel(CBECS$PBA, ref = "2") + weighted_SQFT *
##   relevel(CBECS$PUBCLIM, ref = "4") + weighted_SQFT * relevel(CBECS$PBA,
##   ref = "2"), family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87987  -0.03711  -0.00378   0.03252   0.38138
##
## Coefficients:
##                                     Estimate Std. Error t value
## (Intercept)                      -9.131877    1.008502  -9.055
## weighted_SQFT                      1.395657    0.004208 331.676
## relevel(CBECS$PUBCLIM, ref = "4")1  -2.331856    1.894827  -1.231
## relevel(CBECS$PUBCLIM, ref = "4")2    0.427203    1.114226   0.383
## relevel(CBECS$PUBCLIM, ref = "4")3    2.257555    1.005367   2.246
## relevel(CBECS$PUBCLIM, ref = "4")5  -0.851861    1.306953  -0.652
## relevel(CBECS$PUBCLIM, ref = "4")7  -1.014600    1.280565  -0.792
## relevel(CBECS$PBA, ref = "2")1        3.499425    1.132513   3.090
## relevel(CBECS$PBA, ref = "2")5        2.369086    1.511425   1.567
## relevel(CBECS$PBA, ref = "2")8        0.317226    1.561938   0.203
## relevel(CBECS$PBA, ref = "2")16       13.119877    1.452518   9.033
## relevel(CBECS$PBA, ref = "2")13       3.342620    1.507411   2.217
## relevel(CBECS$PBA, ref = "2")14       5.642250    1.172779   4.811
## relevel(CBECS$PBA, ref = "2")15       0.968353    1.949152   0.497
```

```

## relevel(CBECS$PBA, ref = "2")18      4.777221  1.348975  3.541
## relevel(CBECS$PBA, ref = "2")25      4.034561  2.492798  1.618
## relevel(CBECS$PBA, ref = "2")26     -7.177404  4.360668 -1.646
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")1  0.029484  0.006303  4.678
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")2  0.016673  0.004152  4.015
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")3  0.009803  0.004100  2.391
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")5 -0.002595  0.005025 -0.516
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")7  0.002665  0.006751  0.395
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")1      0.031089  0.006885  4.515
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")5     -0.106473  0.005185 -20.533
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")8      0.022956  0.007035  3.263
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")16    -0.028189  0.006545 -4.307
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")13     0.013523  0.006243  2.166
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")14    -0.036476  0.005165 -7.062
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")15     0.130320  0.006054 21.528
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")18    -0.001606  0.008227 -0.195
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")25    -0.002229  0.007249 -0.307
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")26     0.015782  0.006486  2.433
## Pr(>|t|)
## (Intercept) < 2e-16 ***
## weighted_SQFT < 2e-16 ***
## relevel(CBECS$PUBCLIM, ref = "4")1  0.218502
## relevel(CBECS$PUBCLIM, ref = "4")2  0.701430
## relevel(CBECS$PUBCLIM, ref = "4")3  0.024770 *
## relevel(CBECS$PUBCLIM, ref = "4")5  0.514559
## relevel(CBECS$PUBCLIM, ref = "4")7  0.428212
## relevel(CBECS$PBA, ref = "2")1      0.002010 **
## relevel(CBECS$PBA, ref = "2")5      0.117059
## relevel(CBECS$PBA, ref = "2")8      0.839065
## relevel(CBECS$PBA, ref = "2")16     < 2e-16 ***
## relevel(CBECS$PBA, ref = "2")13     0.026627 *
## relevel(CBECS$PBA, ref = "2")14     1.54e-06 ***
## relevel(CBECS$PBA, ref = "2")15     0.619342
## relevel(CBECS$PBA, ref = "2")18     0.000401 ***
## relevel(CBECS$PBA, ref = "2")25     0.105608
## relevel(CBECS$PBA, ref = "2")26     0.099826 .
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")1 2.96e-06 ***
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")2 6.01e-05 ***
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")3 0.016841 *
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")5 0.605566
## weighted_SQFT:relevel(CBECS$PUBCLIM, ref = "4")7 0.693029
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")1      6.44e-06 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")5     < 2e-16 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")8      0.001109 **
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")16     1.68e-05 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")13     0.030349 *
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")14     1.82e-12 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")15     < 2e-16 ***
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")18     0.845227
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")25     0.758507
## weighted_SQFT:relevel(CBECS$PBA, ref = "2")26     0.014983 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```

```
## (Dispersion parameter for Gamma family taken to be 0.004801233)
##
## Null deviance: 18472.406 on 6356 degrees of freedom
## Residual deviance: 31.374 on 6325 degrees of freedom
## (79 observations deleted due to missingness)
## AIC: 84323
##
## Number of Fisher Scoring iterations: 6
```

From the summary above, we are able to identify the following:

```
residual_deviance <- 31.374
degrees_of_freedom <- 6325

dev_df_ratio <- residual_deviance/degrees_of_freedom
dev_df_ratio
```

```
## [1] 0.004960316
```

The model appears to fit the data exceptionally well. However, this very low value might also indicate that the model is overfitting.

SQFT Main effect is significant with p-value of  $< 2e-16$

PUBCLIM Main effect level 3 is significant with p-value of 0.024770. However, PUBCLIM levels 1-3 have significant interactions with SQFT. Therefore, the only levels we will remove are 5 and 7.

PBA Main effect levels 1, 16, 13, 14 and 18 are significant. However, the only levels that do not have significant interactions with SQFT are 18 and 25. Therefore the only level we will exclude is level 25.

#Refined Model

In order to refine our model, we will now exclude the levels mentioned above.

##Filtering

```
library(survey)
library(srvyr)
library(dplyr)

#Filtered dataset
CBECS_filtered <- CBECS %>%
  filter(PUBCLIM != 7) %>%
  filter(PUBCLIM != 5) %>%
  filter(PBA != 25)

#Dropped filtered out levels
CBECS_filtered <- droplevels(CBECS_filtered)

#Updated survey design object with filtered dataset
CBECS_des <- CBECS_filtered %>%
  as_survey_rep(weights = FINALWT,
                 repweights = FINALWT1:FINALWT151,
                 type = "JK2",
                 mse = TRUE)
```

```
## Warning in svrepdesign.default(variables = variables, repweights = repweights,
## : with type JK2 scale= and rscales= are not needed and will be ignored
```

```
#Checked the levels of PUBCLIM to ensure level 5 & 7 is removed
levels(CBECS_des$variables$PUBCLIM)
```

```
## [1] "1" "2" "3" "4"
```

```
#Checked the levels of PBA to ensure level 25 is removed
levels(CBECS_des$variables$PBA)
```

```
## [1] "1" "2" "5" "8" "16" "13" "14" "15" "18" "26"
```

## Refitting

```
#reassigned the sampling weight
samp_weight <- CBECS_des$pweights

weighted_SQFT <- samp_weight * CBECS_filtered$log_SQFT
weighted_MFBTU <- samp_weight * CBECS_filtered$log_MFBTU
weighted_MFBTU <- weighted_MFBTU + 0.0001
weighted_SQFT <- weighted_SQFT + 0.0001

# Fit the linear regression model with the filtered dataset
final_m <- glm(weighted_MFBTU ~ weighted_SQFT + relevel(CBECS_filtered$PUBCLIM,ref="4") + relevel(CBECS_filtered$PBA,ref="2"), family = Gamma(link = "identity"))

final_m_stdres <- rstandard(final_m)
final_m_fitted<-fitted(final_m)
summary(final_m)
```

```
##
## Call:
## glm(formula = weighted_MFBTU ~ weighted_SQFT + relevel(CBECS_filtered$PUBCLIM,
## ref = "4") + relevel(CBECS_filtered$PBA, ref = "2") + weighted_SQFT *
## relevel(CBECS_filtered$PUBCLIM, ref = "4") + weighted_SQFT *
## relevel(CBECS_filtered$PBA, ref = "2"), family = Gamma(link = "identity"))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87792  -0.03829  -0.00484   0.03252   0.38196
##
## Coefficients:
##              Estimate Std. Error
## (Intercept)    -9.366116    1.096378
## weighted_SQFT     1.396020    0.004606
## relevel(CBECS_filtered$PUBCLIM, ref = "4")1    -1.549321    2.009731
## relevel(CBECS_filtered$PUBCLIM, ref = "4")2     1.036720    1.184940
## relevel(CBECS_filtered$PUBCLIM, ref = "4")3     2.534511    1.066830
## relevel(CBECS_filtered$PBA, ref = "2")1         4.567766    1.289528
## relevel(CBECS_filtered$PBA, ref = "2")5         1.881445    1.776048
```

```

## relevel(CBECS_filtered$PBA, ref = "2")8      1.246275  1.779684
## relevel(CBECS_filtered$PBA, ref = "2")16      0.072597  4.910255
## relevel(CBECS_filtered$PBA, ref = "2")13      3.427170  1.708213
## relevel(CBECS_filtered$PBA, ref = "2")14      2.911974  1.471316
## relevel(CBECS_filtered$PBA, ref = "2")15      2.112087  2.309106
## relevel(CBECS_filtered$PBA, ref = "2")18      1.756613  1.870642
## relevel(CBECS_filtered$PBA, ref = "2")26     -8.207110  4.994432
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")1  0.030576  0.006783
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")2  0.016779  0.004450
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")3  0.009749  0.004379
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")1      0.041335  0.008022
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")5     -0.103544  0.005842
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")8      0.032764  0.007870
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")16    -0.039800  0.007821
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")13     0.013665  0.007023
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")14    -0.038803  0.005986
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")15     0.125799  0.006787
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")18     0.012431  0.011584
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")26     0.014818  0.007168
##
## t value Pr(>|t|)
## (Intercept)      -8.543 < 2e-16 ***
## weighted_SQFT    303.079 < 2e-16 ***
## relevel(CBECS_filtered$PUBCLIM, ref = "4")1     -0.771 0.440802
## relevel(CBECS_filtered$PUBCLIM, ref = "4")2       0.875 0.381668
## relevel(CBECS_filtered$PUBCLIM, ref = "4")3       2.376 0.017556 *
## relevel(CBECS_filtered$PBA, ref = "2")1          3.542 0.000401 ***
## relevel(CBECS_filtered$PBA, ref = "2")5          1.059 0.289501
## relevel(CBECS_filtered$PBA, ref = "2")8           0.700 0.483790
## relevel(CBECS_filtered$PBA, ref = "2")16          0.015 0.988205
## relevel(CBECS_filtered$PBA, ref = "2")13          2.006 0.044886 *
## relevel(CBECS_filtered$PBA, ref = "2")14          1.979 0.047859 *
## relevel(CBECS_filtered$PBA, ref = "2")15          0.915 0.360411
## relevel(CBECS_filtered$PBA, ref = "2")18          0.939 0.347760
## relevel(CBECS_filtered$PBA, ref = "2")26         -1.643 0.100402
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")1  4.508 6.72e-06 ***
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")2  3.771 0.000165 ***
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")3  2.227 0.026028 *
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")1      5.153 2.68e-07 ***
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")5    -17.725 < 2e-16 ***
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")8      4.163 3.20e-05 ***
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")16    -5.089 3.75e-07 ***
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")13     1.946 0.051765 .
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")14    -6.482 1.00e-10 ***
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")15    18.534 < 2e-16 ***
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")18     1.073 0.283268
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")26     2.067 0.038758 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.005148707)
##
## Null deviance: 11764.247 on 4469 degrees of freedom
## Residual deviance: 23.535 on 4444 degrees of freedom
## (68 observations deleted due to missingness)

```

```
## AIC: 61791
##
## Number of Fisher Scoring iterations: 5
```

From the result, we can clearly see that the model truly improved as its AIC value now got significantly lower with a value of 61791. Moreover, despite still having some main effects which are insignificant, all interaction terms are now significant except for one level.

From the summary above, we are able to identify the following:

```
residual_deviance <- 23.535
degrees_of_freedom <- 4444

dev_df_ratio <- residual_deviance/degrees_of_freedom
dev_df_ratio
```

```
## [1] 0.005295905
```

The model still has a very low value of dev to df ratio, which means it is exceptionally performing well although it might still be an indication of overfitting.

Next, we will try to explain the effect of SQFT conditioned on PBA and PUBCLIM when it predicts the value of MFBTU.

```
#got summaries of the significant predictors
summary(CBECS_filtered$log_SQFT)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      6.909   9.116  10.915   10.592  12.067   13.816
```

```
summary(CBECS_filtered$PUBCLIM)
```

```
##      1      2      3      4
##    396  1343  1408  1391
```

```
summary(CBECS_filtered$PBA)
```

```
##      1      2      5      8     16     13     14     15     18     26
##    289  1102   648   321   233   399   676   431   128   311
```

```
#got summary of coefficients of the model
summary(final_m)$coefficients
```

```
##
## (Intercept)                                Estimate
## weighted_SQFT                             1.396019959
## relevel(CBECS_filtered$PUBCLIM, ref = "4")1  -1.549320613
## relevel(CBECS_filtered$PUBCLIM, ref = "4")2    1.036720203
## relevel(CBECS_filtered$PUBCLIM, ref = "4")3    2.534510789
## relevel(CBECS_filtered$PBA, ref = "2")1         4.567766140
## relevel(CBECS_filtered$PBA, ref = "2")5         1.881444731
```



```

## relevel(CBECS_filtered$PBA, ref = "2")8 1.246274778
## relevel(CBECS_filtered$PBA, ref = "2")16 0.072597113
## relevel(CBECS_filtered$PBA, ref = "2")13 3.427169698
## relevel(CBECS_filtered$PBA, ref = "2")14 2.911973778
## relevel(CBECS_filtered$PBA, ref = "2")15 2.112086662
## relevel(CBECS_filtered$PBA, ref = "2")18 1.756612889
## relevel(CBECS_filtered$PBA, ref = "2")26 -8.207110031
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")1 0.030576212
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")2 0.016779466
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")3 0.009749358
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")1 0.041334933
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")5 -0.103544484
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")8 0.032763870
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")16 -0.039800473
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")13 0.013664948
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")14 -0.038802545
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")15 0.125799306
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")18 0.012430719
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")26 0.014818284
## Std. Error
## (Intercept) 1.096378325
## weighted_SQFT 0.004606133
## relevel(CBECS_filtered$PUBCLIM, ref = "4")1 2.009730839
## relevel(CBECS_filtered$PUBCLIM, ref = "4")2 1.184939853
## relevel(CBECS_filtered$PUBCLIM, ref = "4")3 1.066830479
## relevel(CBECS_filtered$PBA, ref = "2")1 1.289527844
## relevel(CBECS_filtered$PBA, ref = "2")5 1.776048265
## relevel(CBECS_filtered$PBA, ref = "2")8 1.779683634
## relevel(CBECS_filtered$PBA, ref = "2")16 4.910254742
## relevel(CBECS_filtered$PBA, ref = "2")13 1.708213130
## relevel(CBECS_filtered$PBA, ref = "2")14 1.471316018
## relevel(CBECS_filtered$PBA, ref = "2")15 2.309105749
## relevel(CBECS_filtered$PBA, ref = "2")18 1.870642224
## relevel(CBECS_filtered$PBA, ref = "2")26 4.994431889
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")1 0.006782962
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")2 0.004449648
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")3 0.004378701
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")1 0.008021739
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")5 0.005841611
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")8 0.007870271
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")16 0.007820972
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")13 0.007023477
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")14 0.005986115
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")15 0.006787321
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")18 0.011583532
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")26 0.007167717
## t value
## (Intercept) -8.5427777
## weighted_SQFT 303.0785083
## relevel(CBECS_filtered$PUBCLIM, ref = "4")1 -0.7709095
## relevel(CBECS_filtered$PUBCLIM, ref = "4")2 0.8749138
## relevel(CBECS_filtered$PUBCLIM, ref = "4")3 2.3757390
## relevel(CBECS_filtered$PBA, ref = "2")1 3.5422005
## relevel(CBECS_filtered$PBA, ref = "2")5 1.0593432

```

```

## relevel(CBECS_filtered$PBA, ref = "2")8 0.7002788
## relevel(CBECS_filtered$PBA, ref = "2")16 0.0147848
## relevel(CBECS_filtered$PBA, ref = "2")13 2.0062893
## relevel(CBECS_filtered$PBA, ref = "2")14 1.9791627
## relevel(CBECS_filtered$PBA, ref = "2")15 0.9146773
## relevel(CBECS_filtered$PBA, ref = "2")18 0.9390427
## relevel(CBECS_filtered$PBA, ref = "2")26 -1.6432520
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")1 4.5077963
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")2 3.7709644
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")3 2.2265412
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")1 5.1528645
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")5 -17.7253318
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")8 4.1629915
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")16 -5.0889423
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")13 1.9456102
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")14 -6.4820912
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")15 18.5344576
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")18 1.0731372
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")26 2.0673644
## Pr(>|t|)
## (Intercept) 1.776684e-17
## weighted_SQFT 0.000000e+00
## relevel(CBECS_filtered$PUBCLIM, ref = "4")1 4.408016e-01
## relevel(CBECS_filtered$PUBCLIM, ref = "4")2 3.816681e-01
## relevel(CBECS_filtered$PUBCLIM, ref = "4")3 1.755601e-02
## relevel(CBECS_filtered$PBA, ref = "2")1 4.008782e-04
## relevel(CBECS_filtered$PBA, ref = "2")5 2.895011e-01
## relevel(CBECS_filtered$PBA, ref = "2")8 4.837898e-01
## relevel(CBECS_filtered$PBA, ref = "2")16 9.882045e-01
## relevel(CBECS_filtered$PBA, ref = "2")13 4.488588e-02
## relevel(CBECS_filtered$PBA, ref = "2")14 4.785932e-02
## relevel(CBECS_filtered$PBA, ref = "2")15 3.604107e-01
## relevel(CBECS_filtered$PBA, ref = "2")18 3.477599e-01
## relevel(CBECS_filtered$PBA, ref = "2")26 1.004016e-01
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")1 6.718947e-06
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")2 1.647314e-04
## weighted_SQFT:relevel(CBECS_filtered$PUBCLIM, ref = "4")3 2.602789e-02
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")1 2.676588e-07
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")5 5.566970e-68
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")8 3.200533e-05
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")16 3.749277e-07
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")13 5.176455e-02
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")14 1.003096e-10
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")15 6.248845e-74
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")18 2.832678e-01
## weighted_SQFT:relevel(CBECS_filtered$PBA, ref = "2")26 3.875759e-02

```

For the comprehensive interpretation of the final model, please refer to the Model Interpretation part of the paper prior the Discussion and Conclusion.