Final Project Report:

EVD/TVD Consumption Regression

Group 5: Ji Hyun Kang, Marcus Ramirez, Maram Salameh, Dominic Williams

San Jose State University

Table of Contents

1. Background …….....3

2. Problem Statement ……..….3

3. Description Data ….…….3

4. Methods ….……..4

4.1. Part A - Descriptive Statistics & Hypothesis Testing ...…….4

4.2. Part B - Regression Analysis ……....…….5

5. Statistical Analysis Results ...……....…….7

5.1. Part A - Descriptive Statistics & Hypothesis Testing ...…….7

5.2. Part B - Regression Analysis ...……....…….13

6. Conclusions and Recommendations …....……....…….23

1. Background

Wisconsin Power and Light (WPL) studied the effectiveness of two damper devices for improving the efficiency of gas home-heating systems. The first was the electric vent damper, or EVD, which closes of the vent in the chimney, therefore reducing heat loss. The second was the thermally activated vent damper, or TVD, which reduces heat loss through the chimney, by controlling the thermal properties of a set of bimetal fins in the vent.

2. Problem Statement

The task was to analyze the data from WPL's study and run single and multiple regression analyses in order to determine if there is a statistical difference in the effectiveness of the two dampers considering a number of varying characteristics, such as:

1. Type of furnace
2. House characteristics: age and type of house
3. Chimney characteristics: area, shape, height, and type of liner

3. Description of Data

WPL measured and recorded the energy consumption, in BTUs, for 40 EVD-equipped houses and 50 TVD-equipped houses, or 90 total, over a period of several weeks. The independent variables are the 'varying characteristics' listed in the previous section, and the two dependent variables are the energy consumption values measured when the damper is active (BTU IN) and inactive (BTU OUT). Effectiveness can be determined as well to yield a single dependent

variable to be used in the regression. It is important to note that one of the EVD houses was missing the chimney area, chimney shape, and the type of chimney liner; the group removed this house from the population, therefore reducing the EVD-equipped houses to 39, or 89 total for all houses in the study. Also, weather conditions and house-size affected energy consumption and the adjustment-formula is as follows: $\frac{consumption}{weather * house\ area}$

Note: Weather data was not provided.

4. Methods

This section provides the details of the process used in the analysis, which includes the following two subsections: Part A - Descriptive Statistics & Hypothesis Testing and Part B - Regression Analysis.

4.1 Part A - Descriptive Statistics & Hypothesis Testing

The tools used to analyze the distribution of energy consumptions compared to each independent variable were Microsoft Excel, Minitab, and R Studio and the following procedure was followed:

1) The dataset was scrubbed and the EVD house that was missing information was unused in order to run a clean analysis.
2) Scatter plot was used to check the distribution of numerical data like chimney area, chimney height and age. However, it is difficult to identify the relation for Categorical

data. So, histogram was used for categorical data such as type of furnaces, chimney shape, type of chimney liner and type of house via Minitab.

3) R was used to depict the relation between all variables.
4) A pivot table, as well as a bar graph via Excel depicting the respective average of dependent variables (BTU.IN and BTU.OUT) compared to each of the dependent variables, TYPE (type of furnace), CH.AREA (chimney area), CH.SHAPE (chimney shape), CH.HT (chimney height in feet), CH.LINER (type of chimney liner), HOUSE (type of house), AGE (house age in years - 0 to 99+)
5) Next, the group drew conclusions on the effectiveness of the dampers through histograms from Excel were also created and included in the analysis.
6) After measuring effectivity, the group determined the hypothesis test as listed below and calculate the p-value of the hypothesis via Excel:
    a) Null H_0: Effectivity of Damper 1 = Effectivity of Damper 2
    b) Alternative H_1: Effectivity of Damper 1 ≠ Effectivity of Damper 2 (Claim)
7) With the hypothesis and the fact that the data includes two-samples with unequal variances, a two-sided t-test was ran and the p-value was compared to the significant level of $a = 0.05$.

4.2 Part B - Regression Analysis

The group created numerous single and multiple linear regression models in order to determine the factors most affecting the efficiency of the dampers. The tools used to run the

regression analysis were Excel, Minitab, and R Studio and the following procedure was followed:

1) Multiple linear regression was used to predict the value of a variable based on the value of two or more other variables.
    a) Using Minitab and Rstudio, the group developed a linear model by first calculating the difference between BTU IN and BTU OUT and creating a new response variable under Effectivity.
    b) Using lm() function, the group fit a multiple linear model using the different predictor variables. Cor() function produces the correlation between the variables.
    c) Using regsubsets() function the group was able to choose the best fitting regression model to accurately predict Effectivity given the variables excluding BTU IN and BTU OUT.
2) ANOVA (Analysis of Variance):
    a) ANOVA analysis was used to determine whether there is a difference in Effectivity depending on the different types of: HOUSE, DAMPER, CH.AREA, CH.SHAPE, CH.HT, or CH.LINER.
    b) ANOVA analysis was then used to determine difference in energy consumption (BTU.IN and BTU.OUT) depending on the different types of: TYPE(furnaces), HOUSE, CH.AREA, CH.SHAPE, CH.HT, or CH.LINER.

5. Statistical Analysis Results

This section provides the details regarding the results the group determined from the analysis with the following subsections: Part A - Descriptive Statistics & Hypothesis Testing, Part B - Regression Analysis, and Recommendations.

5.1 Part A - Descriptive Statistics & Hypothesis Testing

1) Descriptive Statistics for Independent Variables

The independent variables are not normally distributed.

Statistics

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness
TYPE	89	1.21348	0.05865	0.55334	1.00000	1.00000	1.00000	1.00000	3.00000	2.53040
CH AREA	89	62.562	3.448	32.531	28.000	28.000	64.000	80.000	168.000	1.063
CH SHAPE	89	1.76404	0.08152	0.76904	1.00000	1.00000	2.00000	2.00000	3.00000	0.43476
CH HT	89	21.8764	0.6250	5.8965	14.0000	17.0000	20.0000	27.0000	39.0000	0.6383
CH LINER	89	1.01124	0.07909	0.74612	0.00000	0.00000	1.00000	2.00000	2.00000	-0.01823
HOUSE	89	1.8427	0.1071	1.0102	1.0000	1.0000	2.0000	2.0000	5.0000	1.5419
AGE	89	38.326	3.306	31.185	1.000	12.000	30.000	60.000	99.000	0.675

Table 1: Descriptive stats for independent variables

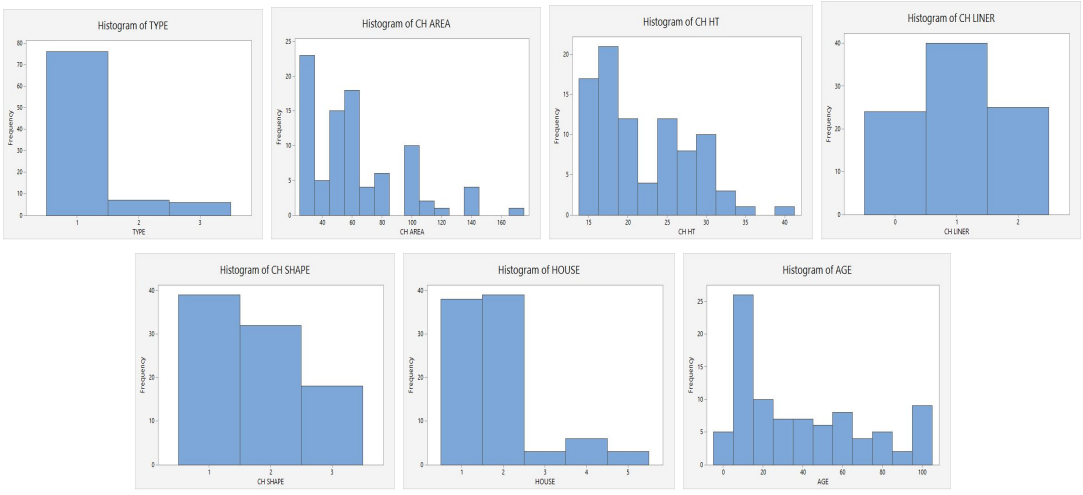


Figure 1

2) Descriptive Statistics for Dependent Variables

The dependent variables are normally distributed with BTU IN & OUT having an outlier. The P-value of normality with BTU IN & OUT is more than 0.05 from Minitab. The effectivity (=BTU IN - BTU OUT) are not normally distributed and has 3 outliers. The P-value of normality with effectivity is less than 0.05. The reason is that the ranges of BTU IN & OUT vary. And the team removed some outliers from effectivity data to reach a meaningful conclusion.

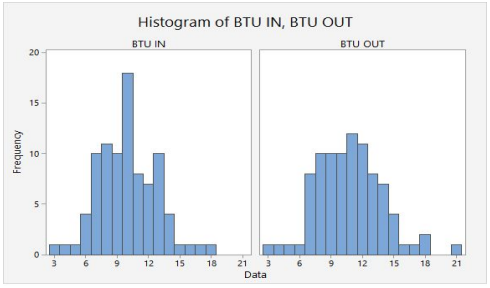


Figure 2

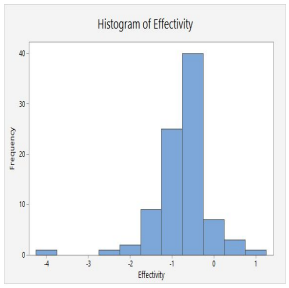
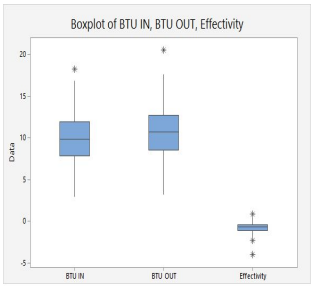


Figure 3



Summary Statistics

Variable	N	Minimum	Q1	Median	Q3	Maximum	95% Median CI
BTU IN	89	2.9700	7.9000	9.8300	11.9300	18.2600	(9.2289, 10.3682)
BTU OUT	89	3.2000	8.5400	10.7200	12.7450	20.5500	(9.6700, 11.4635)
Effectivity	89	-3.98000	-1.06500	-0.70000	-0.41000	0.87000	(-0.80206, -0.60794)

Table 2: Descriptive stats for dependent variables

3) The correlation between variables

There is no strong linear relationship between any of the independent variables only through scatter plot, excep BTU IN and BTU OUT. So, the group used bar graph to

understand the distribution of every consumption and calculated the exact correlation value through software in part B.

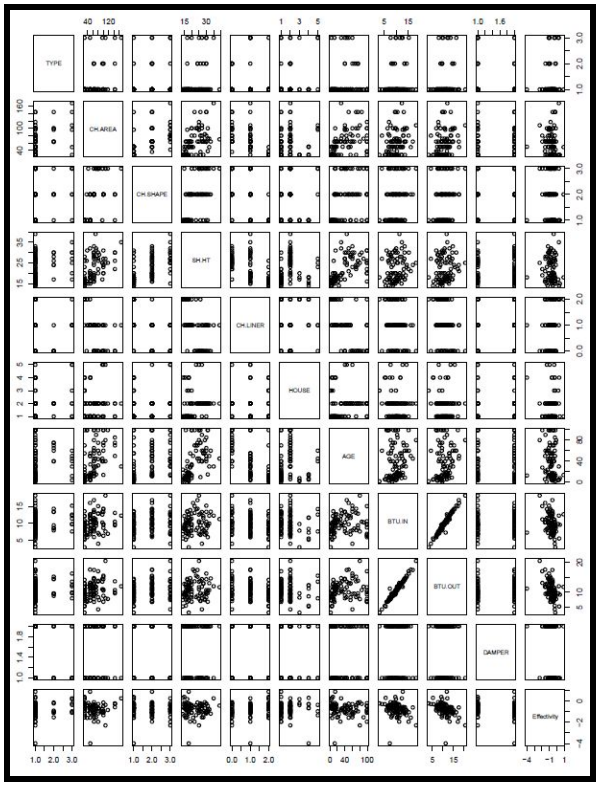
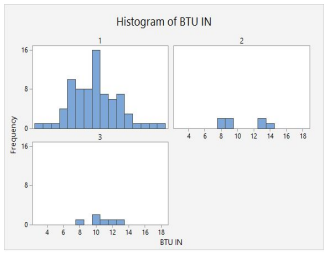


Figure 4

4) Distributions for Independent Variables vs Dependent Variables

a) TYPE (type of furnace)

Most of the furnace samples from the study are (TYPE 1). The gravity furnace (TYPE 2) consumes the most energy and the optimal furnace to minimize consumption is forced air (TYPE 1).



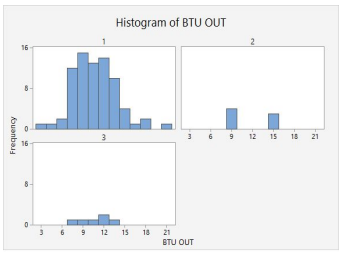


Figure 5

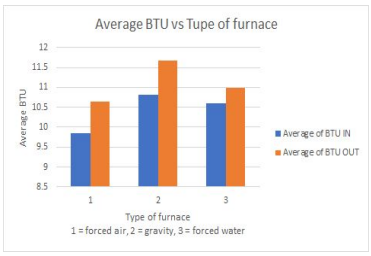
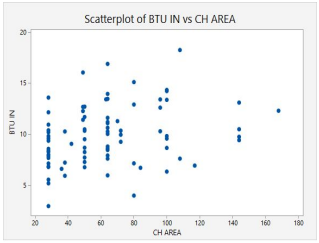


Figure 6

b) CH.AREA (chimney area)

There is no linear association between chimney area and energy consumption. Generally, BTU OUT is greater than BTU IN, except when chimney area is greater than 158. The optimal chimney area to minimize energy consumption is between 38 and 47.



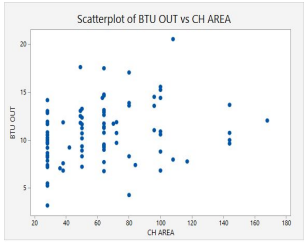


Figure 7

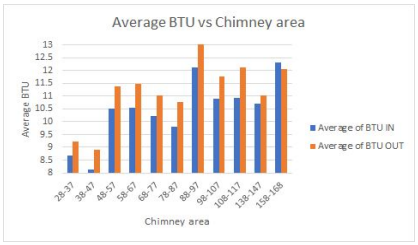
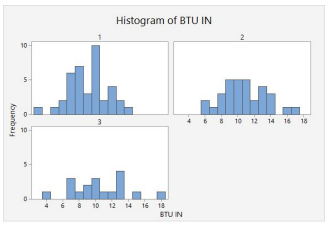


Figure 8

c) CH.SHAPE (chimney shape):

The distribution of BTU IN and BTU out for the respective chimney shapes is normal. The optimal chimney shape to minimize energy consumption is RECTANGULAR.



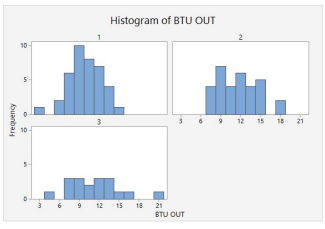


Figure 9

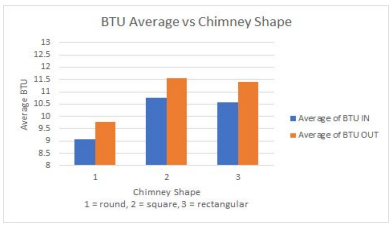
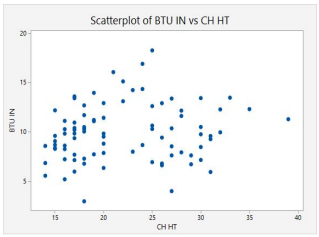


Figure 10

d) CH.HT (chimney height in feet)

There is no linear association between chimney height and energy consumption. The optimal chimney height to minimize energy consumption is between 14 and 16 ft with 29 to 31 ft and 26 to 28 ft close behind it.



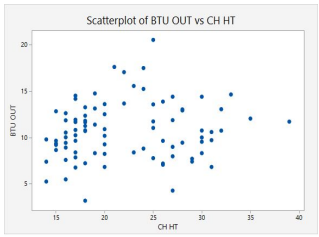


Figure 11

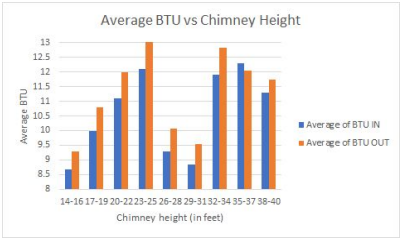
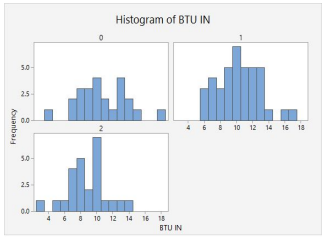


Figure 12

e) CH.LINER (type of chimney liner)

The distribution of BTU IN and BTU out for the respective chimney liner is normal. The optimal chimney liner to minimize energy consumption is METAL.



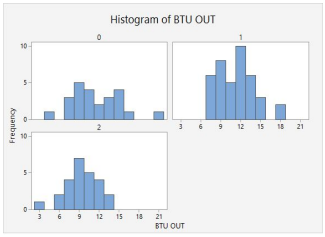


Figure 13

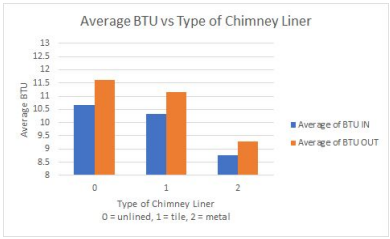
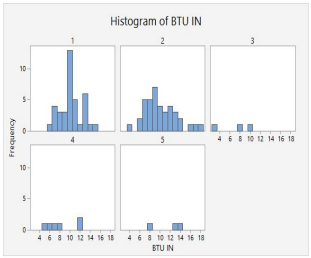


Figure 14

f) HOUSE (type of house)

Most of the samples of house type from the study are RANCH and TWO-STORY. The optimal house to reduce energy consumption is TRI-LEVEL and then BI-LEVEL.



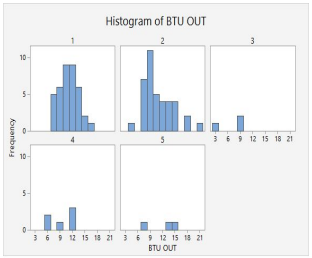


Figure 15

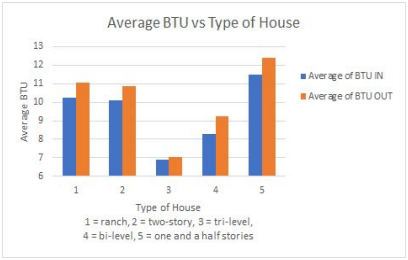
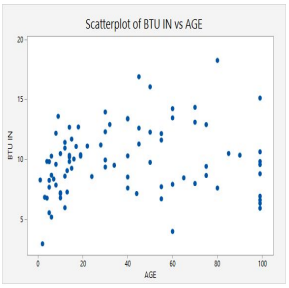


Figure 16

g) AGE (house age in years - 0 to 99+)

There is no linear association between age and energy consumption. The optimal house age in reducing energy consumption is between 0 to 10 years old.



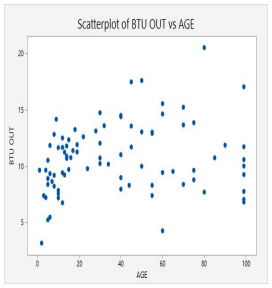


Figure 17

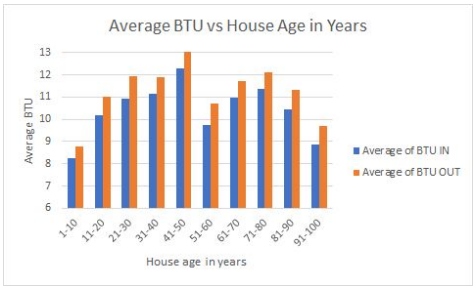


Figure 18

h) Effectiveness

The distribution of Effectivity of each damper is normally distributed. Since BTU IN is greater than BTU OUT—yielding negative values for the mean effectivity—it can be concluded that, both damper 1 and 2 are effective.

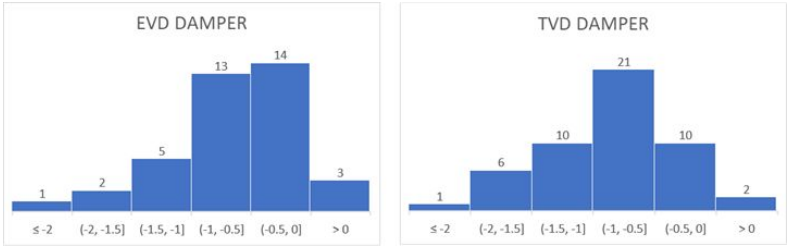


Figure 19

i) Statistical Difference of Dampers

Effectivity	Damper 1	Damper 2
Mean	-0.651282051	-0.8652
Known Variance	0.263322	0.469842
Observations	39	50
Hypothesized Mean Difference	0	
z	1.683366299	
P(Z<=z) two-tail	0.092304203	
z Critical two-tail	1.959963985	

Table 3: The result of Z - Test (Two-Sample for mean)

a. Each sample size is more than 30, sample variance can substitute for population variance. It was assumed that both population variances are known, therefore two-sample z-test for hypothesis testing was used.

c. P-value was determined in Excel and is more than significance level, 0.05, therefore fail to reject H0.

d. Conclusion : There is not enough evidence to support the claim that there is a statistical difference in effectivity between the two types of dampers.

5.2 Part B - Regression Analysis

1) Most Important Variables

The group concluded that there is a difference in effectivity based on what variables are input into our regression model. Based on the multiple linear regression model, below, the group determined that the most influential variable affecting Effectivity is the chimney liner, with a p-value of 0.0071. This p-value allows the group to reject the null hypothesis which states that no variables directly affect overall efficiency. From the same Minitab read out, the damper was the next most influential variable with a p-value of roughly 0.0748, however, the group failed to reject the null hypothesis. The corresponding residual charts for this regression are shown in the Figures below.

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value
Constant	-1.3942	0.4805	(-2.3504, -0.4380)	-2.90	0.0048
TYPE	0.0671	0.1364	(-0.2043, 0.3385)	0.49	0.6240
CH AREA	0.004325	0.003121	(-0.001886, 0.010537)	1.39	0.1697
CH SHAPE	-0.0848	0.1182	(-0.3200, 0.1503)	-0.72	0.4748
SH HT	0.02014	0.01588	(-0.01146, 0.05173)	1.27	0.2083
CH LINER	0.3952	0.1429	(0.1108, 0.6796)	2.77	0.0071
HOUSE	-0.03170	0.06622	(-0.16349, 0.10008)	-0.48	0.6334
AGE	0.000267	0.002891	(-0.005487, 0.006020)	0.09	0.9268
DAMPER	-0.2379	0.1318	(-0.5001, 0.0243)	-1.81	0.0748

Table 4: The Coefficients value for each independent variables

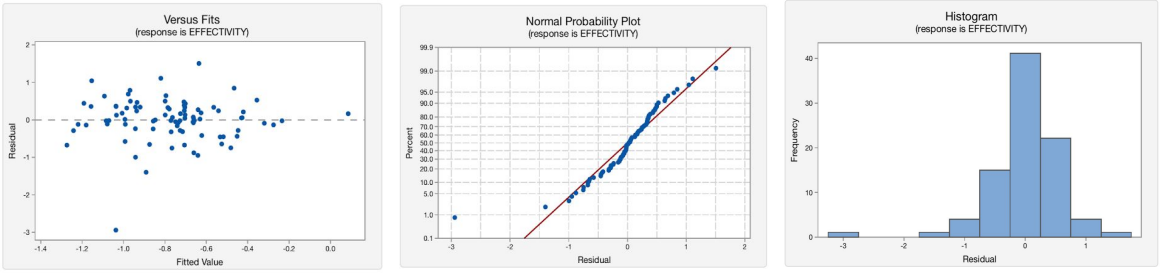


Figure 20

2) Correlated Variables

All variables are correlated, with some having positive-correlation and some having negative-correlation. The selected variables have a p-value that shows they're statistically significant. As shown the CH SHAPE and CH.HT have the strongest correlation

Correlations

	DAMPER	EFFECTIVITY	TYPE	CH AREA	CH SHAPE	SH HT	CH LINER	HOUSE
EFFECTIVITY	-0.171645 0.1078							
TYPE	-0.151219 0.1572	0.134995 0.2072						
CH AREA	-0.036467 0.7344	-0.020344 0.8499	0.396662 0.0001					
CH SHAPE	-0.124443 0.2453	-0.093398 0.3840	0.226536 0.0328	0.605850 <0.0001				
SH HT	-0.007030 0.9479	-0.018737 0.8616	0.248496 0.0189	0.544089 <0.0001	0.607458 <0.0001			
CH LINER	0.043898 0.6829	0.250828 0.0177	-0.115975 0.2791	-0.673512 <0.0001	-0.589460 <0.0001	-0.596343 <0.0001		
HOUSE	-0.093219 0.3849	-0.018460 0.8637	0.243732 0.0214	0.092629 0.3879	0.097957 0.3611	0.197018 0.0642	-0.103168 0.3360	
AGE	-0.061437 0.5674	-0.117338 0.2735	0.039603 0.7125	0.486248 <0.0001	0.424894 <0.0001	0.571342 <0.0001	-0.662990 <0.0001	0.097655 0.3626

Cell Contents: Pearson correlation
P-Value

Table 5: The correlation value from R

3) Outlying and Influential Observations

It was found that some variables had outliers while others were normal with right skew. For instance, BTU.OUT had an outlier, and CH.AREA was right skewed with 2 outliers. Although some variables were not normally distributed and outliers were present, the group decided to keep the outliers for the regression analysis because reason for such variance is unknown. This will be described in more detail in a later section.

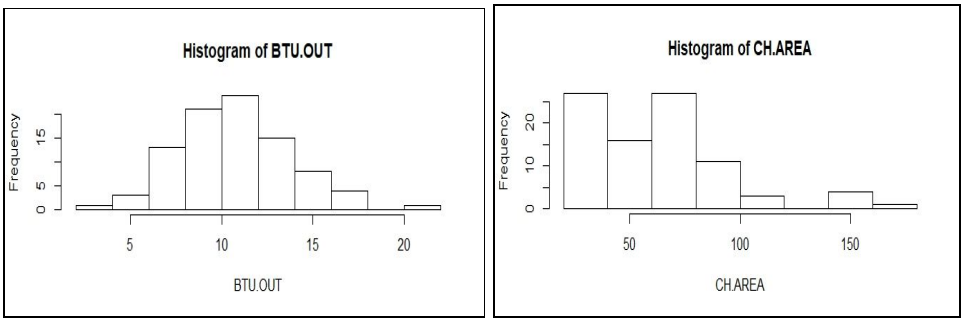


Figure 21

4) Transformation of Variables

Generally, there was no need to transform most variables because all the categorical variables were assigned a number in order to be processed in the statistical software. However, from Part B #1 and #3, the group noticed outliers that were potentially affecting our results. With that observation, 4 rows of outliers were removed and multiple regression was reran. After doing this, the residual plots no longer showed no outliers and the data appeared normal and the dampers effectivity yielded a p-value of 0.02 (ie. statically effective). However, since it was unknown the reason for these outliers, the group elected not to remove the outliers from the remaining report. The Minitab analysis and corresponding residual charts are seen below.

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value
Constant	-0.9597	0.3654	(-1.6875, -0.2320)	-2.63	0.0104
TYPE	-0.01706	0.09965	(-0.21553, 0.18141)	-0.17	0.8645
CH AREA	0.004268	0.002418	(-0.000547, 0.009083)	1.77	0.0815
CH SHAPE	-0.15773	0.08683	(-0.33067, 0.01522)	-1.82	0.0733
SH HT	0.01192	0.01183	(-0.01165, 0.03549)	1.01	0.3172
CH LINER	0.2241	0.1089	(0.0072, 0.4411)	2.06	0.0431
HOUSE	0.08343	0.04970	(-0.01556, 0.18242)	1.68	0.0974
AGE	-0.000814	0.002115	(-0.005027, 0.003398)	-0.38	0.7013
DAMPER	-0.23018	0.09715	(-0.42366, -0.03670)	-2.37	0.0204

Table 6: The Coefficients value with removing outliers

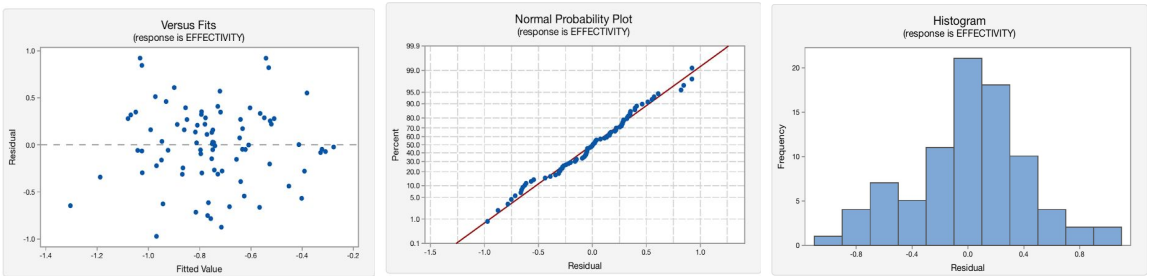


Figure 22

5) Difference in Effectivity between Two Dampers

The pivot table below was created in Excel to show the relationship between the house type and each of the two dampers. In order to determine the best damper for each house type the group compared the average effectivity value between these two dampers. In regards to the effectivity rating, the lower the value the more effective the damper was. For house 1, it was observed that damper 1 had an average effectivity of -0.60 and damper 2 had an average effectivity of -0.88 for the same house. This means that on average, damper 2 is more effective than damper 1 for house 1.

TYPE	**Average of Effectivity**	**TYPE**	**Average of Effectivity**
EVD	**-0.651282051**	**TVD**	**-0.8652**
Ranch	-0.603571429	Ranch	-0.877916667
two-story	-0.759473684	two-story	-0.798
tri-level	-0.105	tri-level	-0.23
bi-level	-0.485	bi-level	-1.165
one & half stories	-0.67	one & half stories	-1.34

Table 7: Average effectivity for each type of house and damper

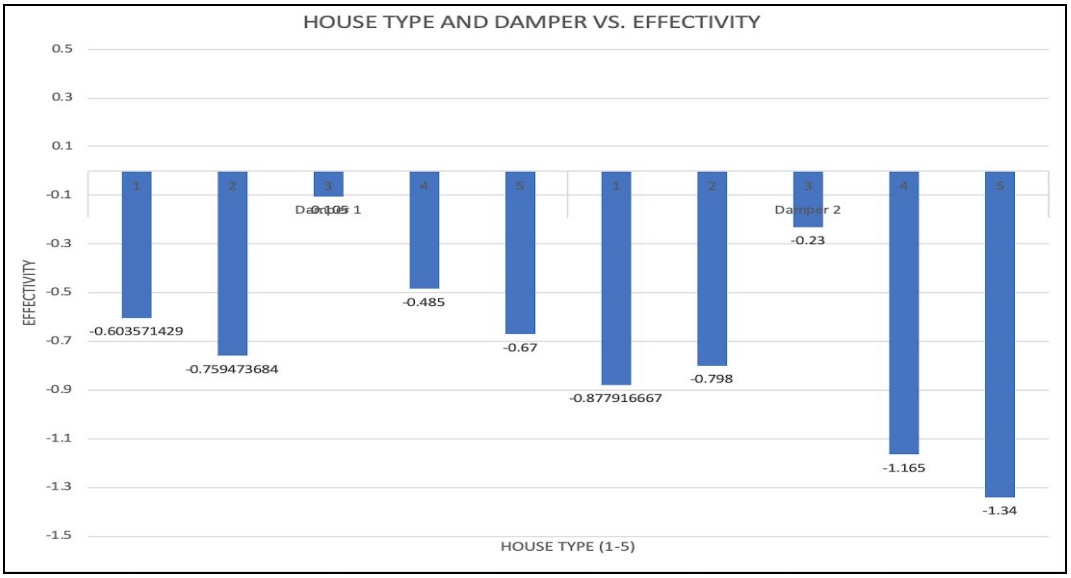


Figure 23

ANOVA analysis showed no statistical difference for Effectivity depending on DAMPER and depending on the HOUSE type as shown in the figures below.

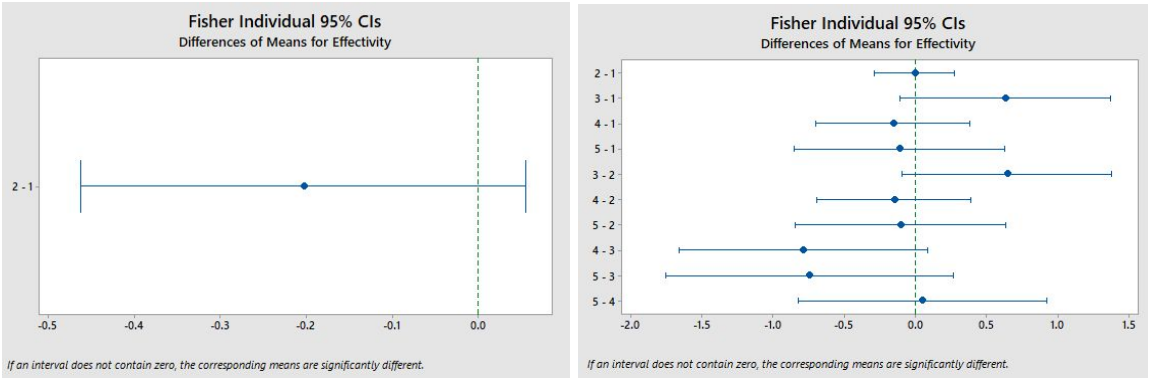


Figure 24

6) Difference between Chimney Area, Shape, Liner, Height

According to the first linear regression model, the group was able to determine that the only variable affecting the overall efficiency is the chimney liner. The p-value of the chimney liner has been reported as .0071 which allowed the group to reject the null

hypothesis that no variables determine the overall efficiency. Instead, there was enough evidence, at a 5% significance level, to show that the chimney liner had a linear relationship with effectivity. Figure 25 below depicts these findings.

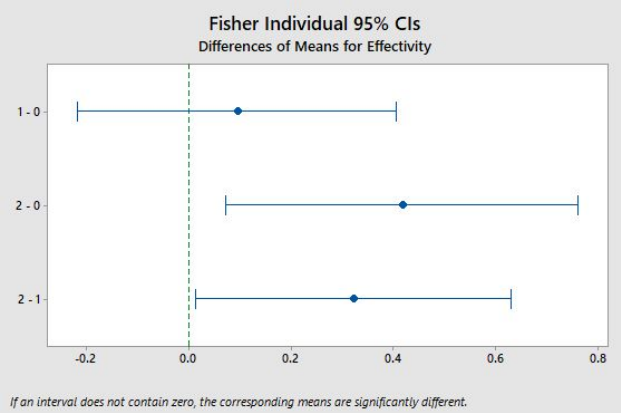


Figure 25

7) Difference between Furnace Types

One-way ANOVA was used in Minitab to determine the difference in effectivity for each type of furnace. The null hypothesis was that all means of the types of furnace are equal and the alternative hypothesis is that at least the mean of one type of furnace are different. The ANOVA output that the P-value of the test was 0.2729, so it can be concluded to fail to reject the null hypothesis. The types of furnace are in same group at the 95% confidence level, as seen in Figure 26 below, allows one to conclude that there was no difference in energy consumption between the three types of furnaces based on the ANOVA alone.

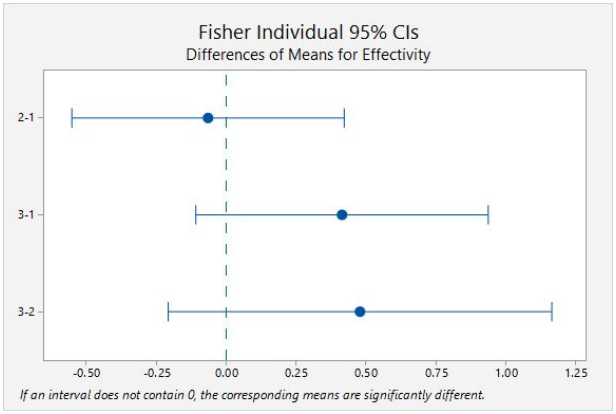


Figure 26

8) Difference between House, Chimney Area, Shape, Height, Liner

There was a statistical difference for energy consumption with Damper in (BTU.IN) depending on House Type, there were differences between types (3,1) and (5,3) as per Figure 27 below. Additionally, there was a difference for BTU.OUT between types (3,1), (3,2), and (5,3) as seen in the Figure.

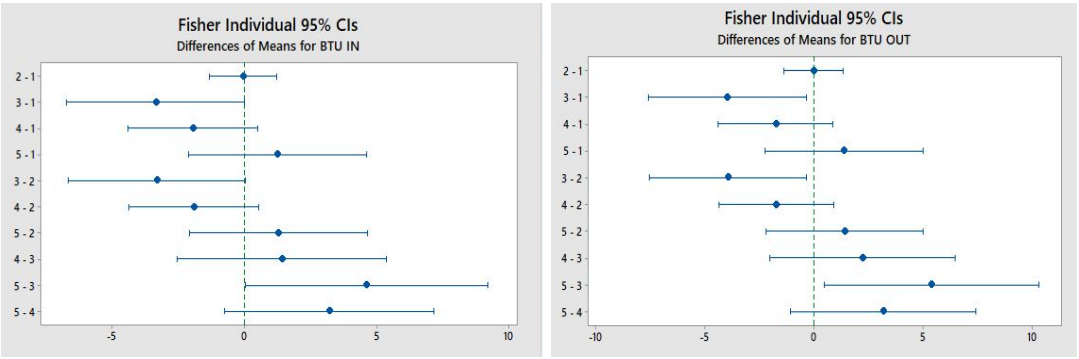


Figure 27

There was also difference in BTU.IN and BTU.OUT due to CH.SHAPE between types (2,1) as seen in Figure 28.

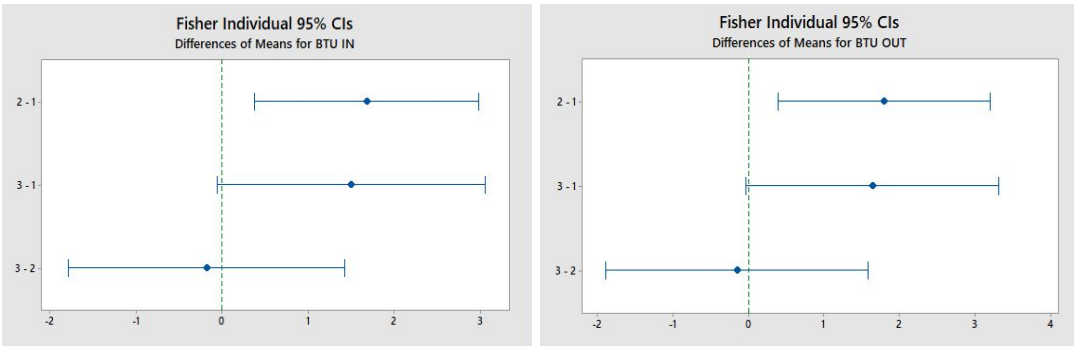


Figure 28

There also was a difference in BTU.OUT between the types of CH.LINER: (2,0) as well as (2,1).

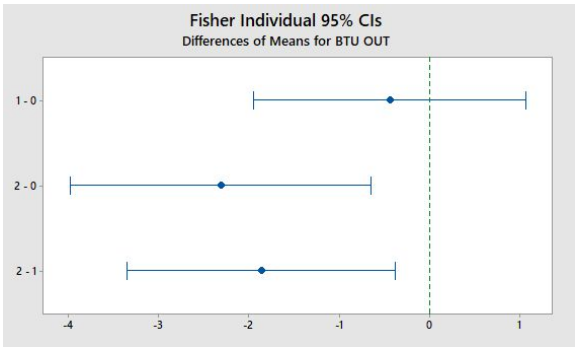


Figure 29

9) Adjusting for Age of House

The group created two different linear regression models to determine the effect that the age of the house had on the overall effectivity. The Table 5 has a p-value of 0.258 based on our given data. Therefore, our group cannot reject the null based on this information and we conclude that the age of the house is not linearly related to the overall effectivity of the house. Table 6 contains adjusted data wherein the ages of the house were changed to years 1 and 2. The p-value with this adjustment changes to 0.04

which allows one to reject the null hypothesis and then hypothetically concluded that the age of the house has an affect on effectivity. Therefore, the group determined that age potentially has an affect on the effectivity level but it cannot be certain without a larger sample size. This is most likely due to the large range of house ages and with a larger sample size the group could gain a better understanding of the houses age and its effect on effectivity.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.4953	0.495269	1.30	0.2580
Error	88	33.6182	0.382025		
Total	89	34.1134			

Table 8: Regression of original data

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1.5115	1.51153	4.08	0.0464
Error	88	32.6019	0.37048		
Total	89	34.1134			

Table 9: Regression of adjusting age data

10) Potentially Important Variables

Our group was able to determine correlation between the different variables and energy consumption, ultimately answering the problem statement, yet also identified other factors that could have an additional impact on our results. One variable that would be weather conditions around the houses location, how many people occupied each home during the experiments, and potentially the age of the people as well. Currently, these items are assumed to be non-impacting to the data, but having this information will allow the group to cover its bases in determining the amount of energy consumption a house would use more accurately.

11) Assumptions

The group assumed that the dependent variables are normally distributed and all values for each independent variable were taken in the same manner in order to reduce

variability during the data collection process. Also, all the potential useful variables above are assumed to be negligible in this analysis.

6. Conclusions and Recommendations

In conclusion, the group determined that the most significant factor that affects the energy consumption was the chimney liner type with a p-value of 0.0071. In regression #5, it was determined that there is no statistical significance between the dampers and the house types with regards to overall efficiency, however, damper 2 showed a higher efficiency ratings for all house types, on average. Therefore, the group concludes that either damper could be used for any type of house. Based on the one way ANOVA in regression #7, the group determined that there is no statistical difference in effectivity based on the furnace types. In regression #9, it was determined that it would be difficult to determine the effectivity due to age for the following reason: the analysis of the original dataset had no effect on the overall effectivity, but after removing years 1 and 2 years (the outliers determined by the team) and comparing the new p-value, 0.0464, to the significance level showed that age hypothetically could affect the overall effectivity. However, without a larger sample size to determine the validity of the outliers, the group cannot say with absolute certainty at this point in time and would recommend looking into it during future analyses.

Best fitting linear model with highest Rsq:

Effectivity = -1.3945071 + 0.0043245 CH.AREA + 0.0201177 CH.HT + 0.3952814 CH.LINER + -0.2377852 DAMPER