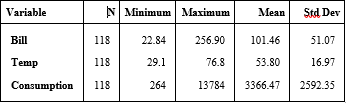
**Dominic Boccaleri**

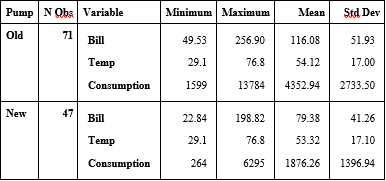
**Project 2:**

**Background**:

The household electric bill varies from house to house. There are many factors that affect this bill. Some factors include age of equipment, consumption, outside temperature, size of house, and number of people who live in the house (JSE). Emerging technologies such as solar are effective ways to decrease energy bills. Solar energy has decreased in cost by 90% and will only cost about $0.03 per kilowatt-hour in the coming years (“Photovoltaics Research and Development”). In this study, we expect to calculate a model predicting the monthly electric bill. We believe that having a new heating pump is an effective way to decrease your electric bill as opposed to having an old, less energy efficient one. This study is important because an electric bill is a huge bill for a household, next to mortgage, and if we can convince people with evidence that having new heat pump will lower their bill a significant amount, then more people might switch to an energy efficient model.

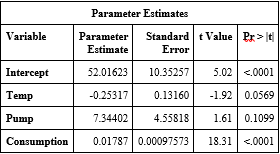
**Description of Statistical Methods:**

 To analyze the data, I used the multiple linear regression model. This method was the most appropriate to analyze if having an old or new pump changes the cost of the electric bill. The population of the study is households. The observational sample is of electric bills taken from houses. In trying to observe electric bills, we had a bill variable that ranged from $22.84 to $256.90. In the analysis, we used monthly average temperature, which was from 29.1 to 76.8 degrees Fahrenheit, consumption, which ranged from 264 to 13784 kilowatt-hours, and a pump, which was either 0 for having an old pump or 1 for a new pump.

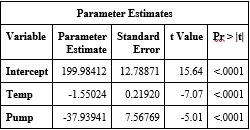
The first table to the right shows the information of each variable with the exception of the “pump” variable since it is comprised of only 0’s and 1’s. To the right is a second table showing the summaries of each variable split on whether the pump is new or old. From these summaries, there are no variables that seem to function as a lurking variable.

For each of the variables, we ran a 2-sample T-test with the null hypothesis being that the means of each variable are equal and the alternative hypothesis being that the means of each variable were different. When conducting our t-test, we met the normality requirement by the Central Limit Theorem. The variances were equal for the bill and the temperature but different for the consumption so as shown in **Figure 1**, we have a p-value less than 0.0001 for the bill and consumption so we reject the null hypothesis and conclude that the means of the bill and consumption between the old and new pump systems are different. The p-value for temperature is 0.8035 so we fail to reject the null hypothesis and conclude that the means are the same for consumption between the old and new heat pump systems.

Based on the data, we need to check there is no collinearity. This is shown in the matrix in **Figure 2**. There is slight collinearity between the temperature and consumption variables and the pump status and consumption variables. We also need to assume need to assume that the residuals are independent of the electric bill and have constant variance and the residuals are in a normal distribution. Based on the plots, the residuals, in **Figure 3**, seem to be relatively heteroscedastic but there are some means for concerns and some possible outliers as well as a possible slight pattern in the residuals. The QQ plot, in **Figure 4,** also seems to be in good standing. The plot is in a relative straight line but there is still means for concern. The point around (2.5, 3.75) seems to be the outlier that may need to be investigated.

**Statistical Results:**

In the model, having a new pump was not a huge factor in deciding the cost of the electric bill when looking at consumption and outside temperature. The p-values, given to the right, show that in this model, temperature and pump status are not statistically appropriate. In the study, after adjusting for pump status and consumption amount, we would predict the electric bill to decrease by $0.25 per each degree increase in temperature. After adjusting for temperature and consumption amount, we would predict the electric bill to increase by $7.34, if the house has a new pump. After adjusting for temperature and pump status, we would predict the electric bill to increase by $0.02 per each kilowatt-hour increase in consumption. The results of this study should not be extended or generalized outside the sample since the p-values are outside of the realm of statistical significance. The time of year, such as the season, could be a lurking variable since the heat pump is not used often in the summer months. Also, humidity changes with the seasons and heat is retained better in warmer temperatures.

After calculating the model using all the variables, we manually removed some variables from the model, keeping the pump status since it was the crux of the hypothesis. When removing the consumption variable, we saw a drop in p-value in both temperature and pump status. Consumption wasn’t the most obvious variable to remove but it did produce the most statistically useful model. The model using consumption and pump status also was not practically useful whereas the model of pump status and temperature is practically useful. In the second model, after adjusting for pump status, the electric bill is predicted to decrease by $1.55 per degree increase in temperature. After adjusting for temperature, the electric bill will decrease by $37.94 if you have a new heat pump, as opposed to an older one. This model is more statistically appropriate to generalize outside the sample. The only part of the population where this study could be extrapolated to is the areas where the sample was taken. This is because the coefficient of determination is very low, r2=0.3901. This means that only 39% of the variance in the electric bills is explained by the temperature and the pump status in the regression model. This shows the model is not practically useful. The seasons could still function as a lurking variable as well as the consumption. Seasons allow heat to travel differently as well as produce natural heat i.e. the summer. Consumption definitely would affect the electric bill but houses in the north would use more energy to combat extreme cold temperatures than the south where it is more moderate all year round. The updated residual plot and QQ plot are show in **Figure 5** and **Figure 6** respectively. The assumptions are much better met for the residual plot by the new model and about the same for the QQ plot. The residual plot is a more heteroscedastic rectangle than before in the first model which, while still semi-heteroscedastic, there was a slight pattern to the residuals.

**Report Summary:**

Using the data, we tried to see if having a new pump was important in lowering the cost of a household’s electric bill. The data was collected over time, surveying the cost of the electric bill, the status of the heat pump, the temperature outside, and the consumption per month. There were 118 months that were documented for analysis. A multiple linear regression was completed on the data to find a model that would effectively predict the electric bill cost based on the temperature outside, the consumption, and the status of the heat pump (new or old). The assumptions of constant variance of the residual plot and linearity among the QQ-plot were checked and confirmed to be appropriate. There were an outlier or two that should be looked into and may be excluded from the data set. The most statistically appropriate model came when we removed the consumption variable. All the p-values were less than 0.0001 for the pump status and the temperature variables. The effect of the temperature was a decrease of $1.55 per degree increase in temperature for the month, after adjusting for pump status and the effect of the pump status, after adjusting for temperature, was a decrease of $37.94 in electric bill for having a new heat pump.

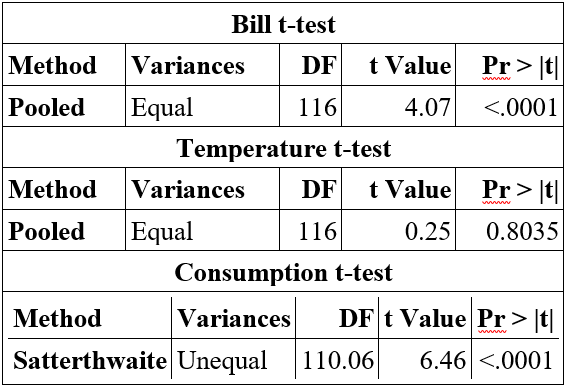
A future study could be on the electric bill and where samples were taken from different areas of the country and put in as a categorical variable to adjust for location as well as ensure that consumption readings were accurate because as consumption increases, we expect the bill to increase but as temperature increases, we expect the bill to decrease and having a new pump should also decrease the bill since it is more energy efficient. We did not find all those things from this study so a more accurate study may be in need.

**Bibliography:**

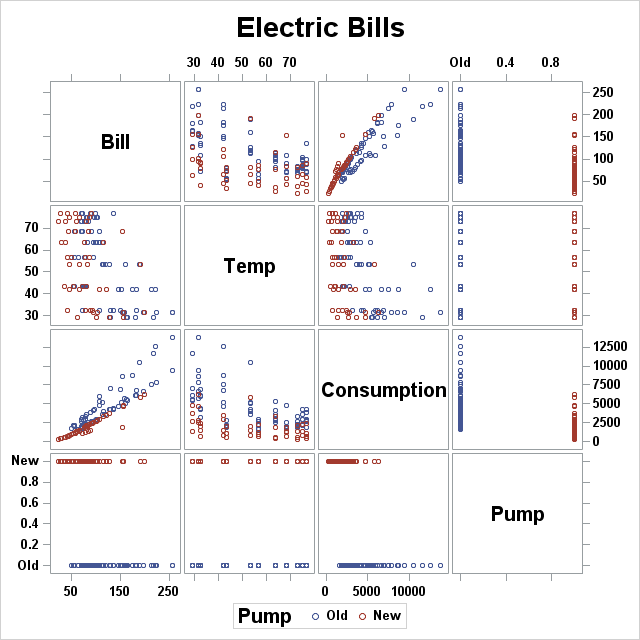
Journal of Statistics Education (JSE), jse.amstat.org/datasets/electricbill.txt.

Journal of Statistics Education (JSE), jse.amstat.org/datasets/electricbill.dat.txt.

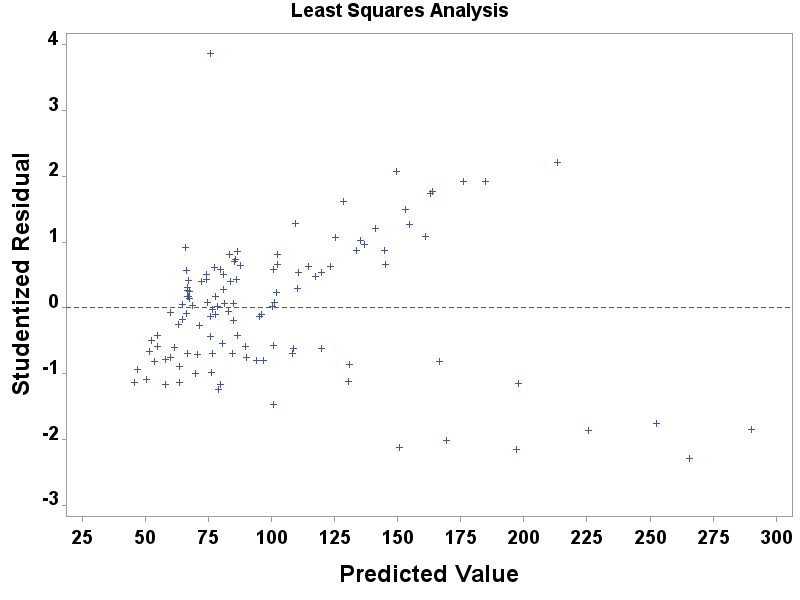
“Photovoltaics Research and Development.” *Energy.gov*, www.energy.gov/eere/solar/photovoltaics-research-and-development.

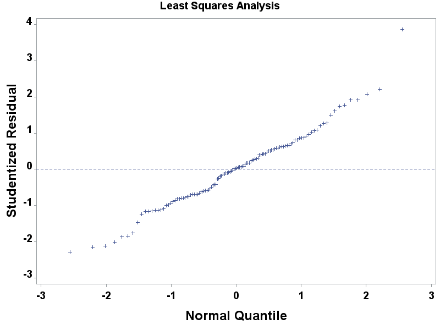
**Appendix A:**

**Figure 1:**

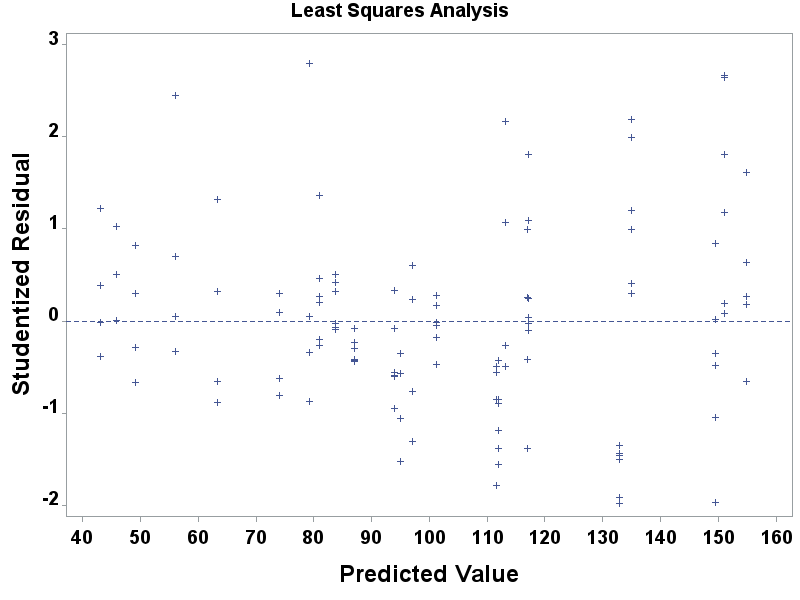
****

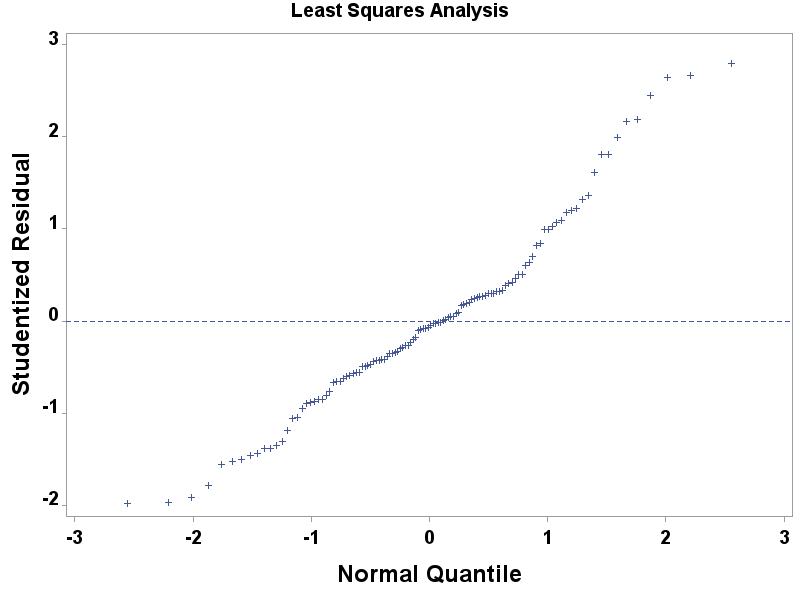
**Figure 2:**

**Figure 3:**



**Figure 4:**

**Figure 5:**



**Figure 6:**

**Appendix B:**

\*\*\*Multiple Linear Regression On Project 2\*\*\*

\*\*\*\*\*\*Electric Bills\*\*\*\*\*;

**PROC** **TEMPLATE**;

DEFINE STYLE STYLES.BIGFONT;

PARENT = STYLES.HTMLBLUE;

STYLE GRAPHFONTS FROM GRAPHFONTS /

'GraphTitleFont' = ("<sans-serif>, <MTsans-serif>",**20**pt,bold)

'GraphLabelFont' = ("<sans-serif>, <MTsans-serif>",**14**pt,bold)

'GraphValueFont' = ("<sans-serif>, <MTsans-serif>",**10**pt,bold)

'GraphDataFont' = ("<sans-serif>, <MTsans-serif>",**10**pt);

END;

ODS HTML STYLE=BigFont;

**RUN**;

**PROC** **IMPORT** DATAFILE='H:\Desktop\stat530\Data Sets\project2.formatted.data.xlsx'

OUT = proj2

DBMS=XLSX

REPLACE;

GETNAMES = yes;

**RUN**;

**PROC** **PRINT** DATA = proj2 (obs = **5**);

**RUN**;

\*Format Pump Variable;

**PROC** **FORMAT**;

VALUE pumpF

**0** = "Old" **1** = "New";

\*Numerical Summaries;

**PROC** **MEANS** DATA = proj2 n min max mean std MAXDEC = **2**;

VAR bill temp consumption;

**RUN**;

\*Numerical Summaries;

**PROC** **MEANS** DATA = proj2 n min max mean std MAXDEC = **2**;

VAR bill temp consumption;

CLASS pump;

FORMAT pump pumpF.;

**RUN**;

\*Make the matrix scatterplot;

TITLE 'Electric Bills';

**PROC** **SGSCATTER** DATA = proj2;

MATRIX bill temp consumption pump / group = pump;

**RUN**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

USE ALL THE VARIABLES

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*Correlation;

**PROC** **CORR** DATA = proj2;

VAR bill pump temp consumption;

**RUN**;

**PROC** **TEMPLATE**;

DEFINE STYLE STYLES.BIGFONT;

PARENT = STYLES.HTMLBLUE;

STYLE GRAPHFONTS FROM GRAPHFONTS /

'GraphTitleFont' = ("<sans-serif>, <MTsans-serif>",**24**pt,bold)

'GraphLabelFont' = ("<sans-serif>, <MTsans-serif>",**18**pt,bold)

'GraphValueFont' = ("<sans-serif>, <MTsans-serif>",**14**pt,bold)

'GraphDataFont' = ("<sans-serif>, <MTsans-serif>",**14**pt);

END;

ODS HTML STYLE=BigFont;

**RUN**;

\*Compute Regression;

**PROC** **REG** DATA = proj2;

TITLE "Least Squares Analysis";

MODEL bill = temp pump consumption;

PLOT STUDENT. \* PREDICTED. / NOMODEL NOSTAT;

PLOT STUDENT. \* NQQ. / NOMODEL NOSTAT MODELlab= "QQ PLOT";

**RUN**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

REMOVE SOME VARIABLES------DOES NOT MATTER

TEMP OR CONSUMPTION, SAME EFFECT.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*Correlation;

**PROC** **CORR** DATA = proj2;

VAR bill pump consumption ;

**RUN**;

**PROC** **TEMPLATE**;

DEFINE STYLE STYLES.BIGFONT;

PARENT = STYLES.HTMLBLUE;

STYLE GRAPHFONTS FROM GRAPHFONTS /

'GraphTitleFont' = ("<sans-serif>, <MTsans-serif>",**24**pt,bold)

'GraphLabelFont' = ("<sans-serif>, <MTsans-serif>",**18**pt,bold)

'GraphValueFont' = ("<sans-serif>, <MTsans-serif>",**14**pt,bold)

'GraphDataFont' = ("<sans-serif>, <MTsans-serif>",**14**pt)

;

END;

ODS HTML STYLE=BigFont;

**RUN**;

\*Compute Regression;

**PROC** **REG** DATA = proj2;

TITLE "Least Squares Analysis";

MODEL bill = consumption pump ;

PLOT STUDENT. \* PREDICTED. / NOMODEL NOSTAT;

PLOT STUDENT. \* NQQ. / NOMODEL NOSTAT MODELlab= "QQ PLOT";

**RUN**;