Project 2: Modeling Total Utility Bills

Jonathan Fogel

11/13/2020

Author: Jonathan Fogel

Introduction

Each year, a plethora of data comes out to reveal how the world is using an unsustainably high amount of energy. Humans are burning fossil fuels at an alarming rate to produce energy. This energy is then used everywhere from production lines to electrical energy production. An estimated 64.5% of electrical energy worldwide is created by the burning of fossil fuels. All the while, the usage of fossil fuels is harming our environment signifigantly, and polluting the air we breathe.

One metric to measure one's usage of fossil fuels is their monthly bill for electric and natural gas, since electrical energy is often a result of burning fossil fuels, and natural gas is a form of fossil fuels. In this project, an attempt to model a household's utility bill is made, so that future work can be done to determine where energy usage can be reduced.

Modeling Total Utility Bills

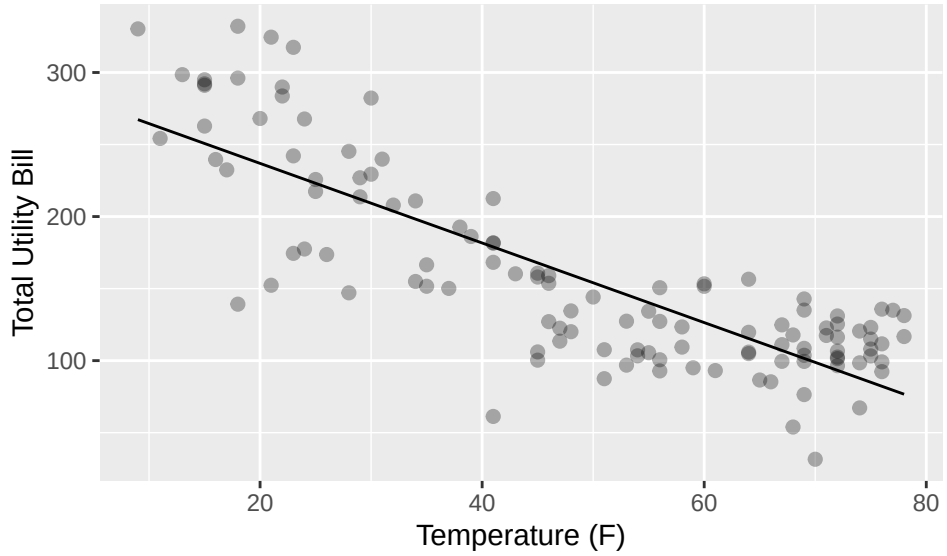
In an effort to understand residential usage of electrical energy and natural gas(which is a fossil fuel), analysis is done on data from utility bills at a private residence. The data used are a built-in set called "Utilities2" for R, and created by Daniel T. Kaplan for "Statistical modeling: A fresh approach" in 2009. Since the data used are for a single home, the energy supplier is unchanged for all points in the dataset, so we will not account for multiple sources/rates.

We expect that signifigant energy will be used to heat a home as it gets colder, so we will first examine the relationship between monthly average temperature and monthly utility bill.

```
cor(totalbill~temp,data=U2)
```

```
## [1] -0.8225802
```

Above, the correlation between temperature and total bill is found to be -0.82, which confirms the relationship between the variables. We now will create a linear model to simply model this relationship.




```
coef(model1)
```

```
## (Intercept)        temp
##  292.155830   -2.762358
```

```
rsquared(model1)
```

```
## [1] 0.6766382
```

The model created is

$Bill = 292.16 - 2.76(Temp)$.

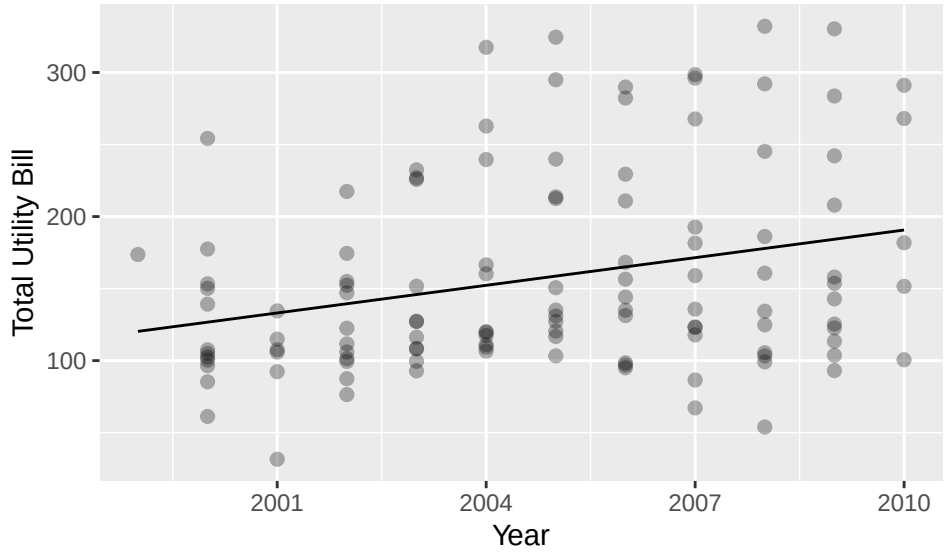
This predicts that for every additional degree (F) of heat, the price of the monthly bill will decrease 2.76 USD. Looking at the plot, it is apparent that this model is suggesting an accurate trend. This is echoed by $R^2 = 0.68$ for the model, which means that almost two thirds of the variability of utility bills can be explained by temperature. To better the model, we will now examine other possible explanitory variables.

One other area that may be important in modeling utility bills is the presence of inflation and the generally increasing energy costs. To do this, we will first examine the possible relationship between total bill and year.

```
cor(totalbill~year,data=U2)
```

```
## [1] 0.2810642
```

The correlation between total bill and year is found to be 0.28. This agrees with the idea of inflation, but it quite a weak correlation.




```
coef(model2)
```

```
##   (Intercept)          year
## -12666.005030      6.396349
```

```
rsquared(model2)
```

```
## [1] 0.07899708
```

The new model, only taking year as an explanatory variable, is

$Bill=-12666.005030 + 6.396349 (Year) $.

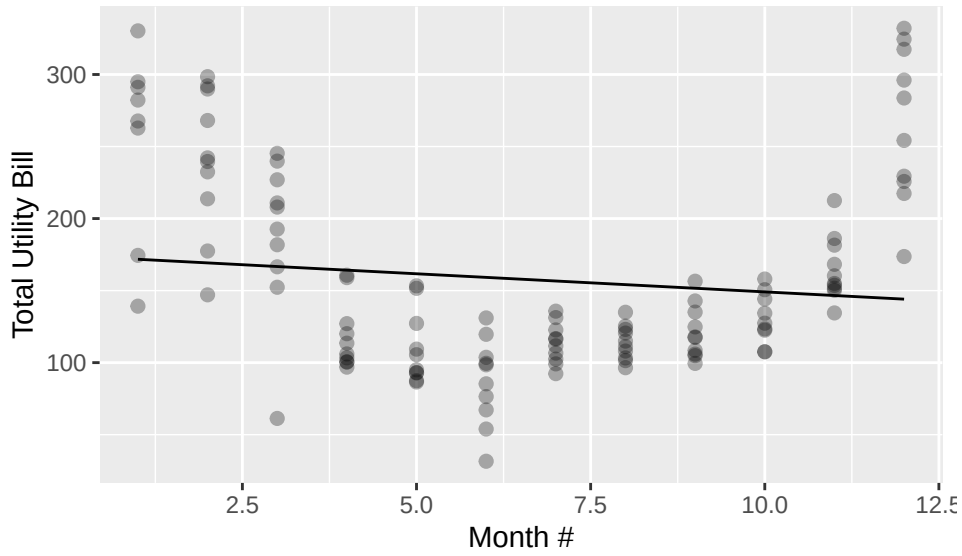
This predicts that, each year, the price of a monthly bill goes up about 6 USD. This model, however, only accounts for 0.08 of the variability. This, along with the very small correlation, leads to the conclusion that, although inflation is likely influencing the price slightly each year, the year does not have a signifigant impact on each month's bill.

The next variable to examine is the bill's month issued. One reason why the month may be important is that some months are colder than others, although there are many other reasons, including children's availability in the summer, increased likelihood to travel in certain months, etc. To ensure that taking both temperature and month into account isn't redundant, their correlation is calculated.

```
cor(temp~monthnumeric,data=U2)
```

```
## [1] 0.2606197
```

As expected, there is a slight correlation between the two. Since the correlation is only 0.26, it seems that month should be examined as well.



Above is a linear model of bill as a function of month, using month as a quantitative variable between 1 and 12. Obviously, this model does not fit the data very well, because of the parabolic trend of the data. To fix this, month will be accounted for as a categorical variable, binning all data for each month.

```
coef(model4)
```

```
## (Intercept)      month2      month3      month4      month5      month6
##    255.3425    -15.2445    -66.7985   -136.5885   -145.1435   -168.6665
##      month7      month8      month9     month10     month11     month12
##   -141.8475   -141.4285   -134.0155   -124.7358    -89.9865     10.0875
```

```
rsquared(model4)
```

```
## [1] 0.75361
```

This model, only taking month into account, accounts for 0.75 of the variability of bills. This model is then

$Bill=255.34 -15.24(Feb) -66.80(Mar) -136.59(Apr) -145.14(May) -168.67(June) -141.84(July) -141.43(Aug) -134.02(Sep) -124.74(Oct) -89.99(Nov) + 10.09(Dec) $

In this equation, you insert a 1 into each variable if it is that month, and insert 0 if it isn't. Now, an attempt to make a more complicated model is made, by accounting for both temperature and month. Since there was such a small correlation between year and bill, year will not be included in the model.

The effect that the temperature has on total bill should not depend on the month, as a cold day would require heating, no matter the month. This leads to the conclusion that there should not be any interaction term between month and temperature.

```
coef(model5)
```

```
## (Intercept)        temp      month2      month3      month4      month5
##  306.123313   -2.726487   -9.587040  -27.605255  -55.134712  -37.515441
##      month6      month7      month8      month9     month10     month11
##  -30.501791   11.040237    7.914804   -1.031115  -33.436400  -30.344605
##     month12
##   19.016744
```

```
rsquared(model5)
```

```
## [1] 0.7816398
```

The final model created is then:

$Bill = 306.12 - 2.73(Temp) - 9.59(Feb) - 27.61(Mar) - 55.13(Apr) - 37.52(May) - 30.50(Jun) + 11.04(Jul) + 7.91(Aug) - 1.03(Sep) - 33.44(Oct) - 30.34(Nov) + 19.01(Dec)$

The variables in this model are the same as they were in the previous iterations. This model now accounts for 0.78 of the variability of utility bills. In the model above, the prediction follows the same downward trend as temperature increases, but there is now an adjustment added/subtracted at every point to adjust for what month it is.

Conclusion

The model created in this paper sugggests that both monthly average temperature, and the month itself are both important explanatory variables when it comes to predicting a monthly utility bill. For temperature, the model suggests that the warmer it is, the less energy used. For each degree increase, the model predicts a decrease of 2.73 USD. Also, it appears that months in the winter have signifigantly higher energy consumption than those in the summer, with spring and fall having intermediate values. Although these seem to be very related results, the data suggested that month and temperature were not strongly correlated.

With this result, it seems like the periods where there is the most potential to lower energy consumption are periods of cold, and the winter. Further studies should be done to further analyze how behavior in these periods could be adjusted to lower energy consumption.