Xin Zhao: xzhao2
Yingying Gao: yingyin2

Relationship between Energy Consumption for Space Heating and Wall Materials

Background and motivation

Saving energy is always a topic among people and it is really significant in today's world. While we are doing our first assignment, my partner and I found that different factors would influence the energy consumption to some degree such as energy-star of equipment the number of windows. As we all know, exterior wall materials have different insulation properties. We thought maybe this would also have an important influence on the energy consumption for space heating. If we explore their relationship and find they do have a relationship, then we can save energy by change the exterior wall materials when building houses. By searching the Internet, we found that U.S. Department of Energy have a recommendation on total R-Values for new construction houses, by regions and by various parts of the house. The higher latitude the region locates at, the higher R value of material this region is recommended to use when building houses. So this may prove that our assumption is right.

Raw data

Our dataset is 'recs2009_public.csv', '2009 RECS Variable and Response Codebook.csv' and 'public layout.csv'. Firstly we should deal with our raw data. In data table 'recs2009_public.csv', we noticed that column 25 is the major outside wall material. They are brick, wood, siding, stucco, composition, stone and concrete. Because the number of houses with glass wall type is much less than other types, so we just ignore this wall type. Also we use the energy consumption for space heating per heated square footage as our response variable or dependent variable because footage have a direct and apparent relationship with energy consumption. By doing this, we can eliminate the influence of footage. Actually there are many other influences will have an impact on energy consumption for space heating, we decide to calculate their coefficient with energy consumption and choose some factors with big coefficient.

Firstly, we searched the Internet and found the R-value of there materials. Here is the data.

Materials	R-value	Materials	R-value
Common brick[1]	0.8	Composition[2]	sum of each layer's R
Wood	1.25	Stone[3]	0.114
Siding	0.61	Concrete	1.28
Stucco[4]	0.2		

We found that concrete, wood and brick's R-value are relatively higher than other materials. So we may think houses with concrete wall material may have less energy consumption.

Then we did a rough analysis of the relationship between wall materials and energy consumption for space heating without considering other influences except climate zone. We obtained a boxplot in every zone as follows.

[1] http://www.coloradoenergy.org/procorner/stuff/r-values.htm
[2] https://www.e-education.psu.edu/egee102/node/2064
[3] http://www.marble-institute.com/stoneprofessionals/technical-bulletins/rvalue/
[4] https://sizes.com/units/rvalue.htm

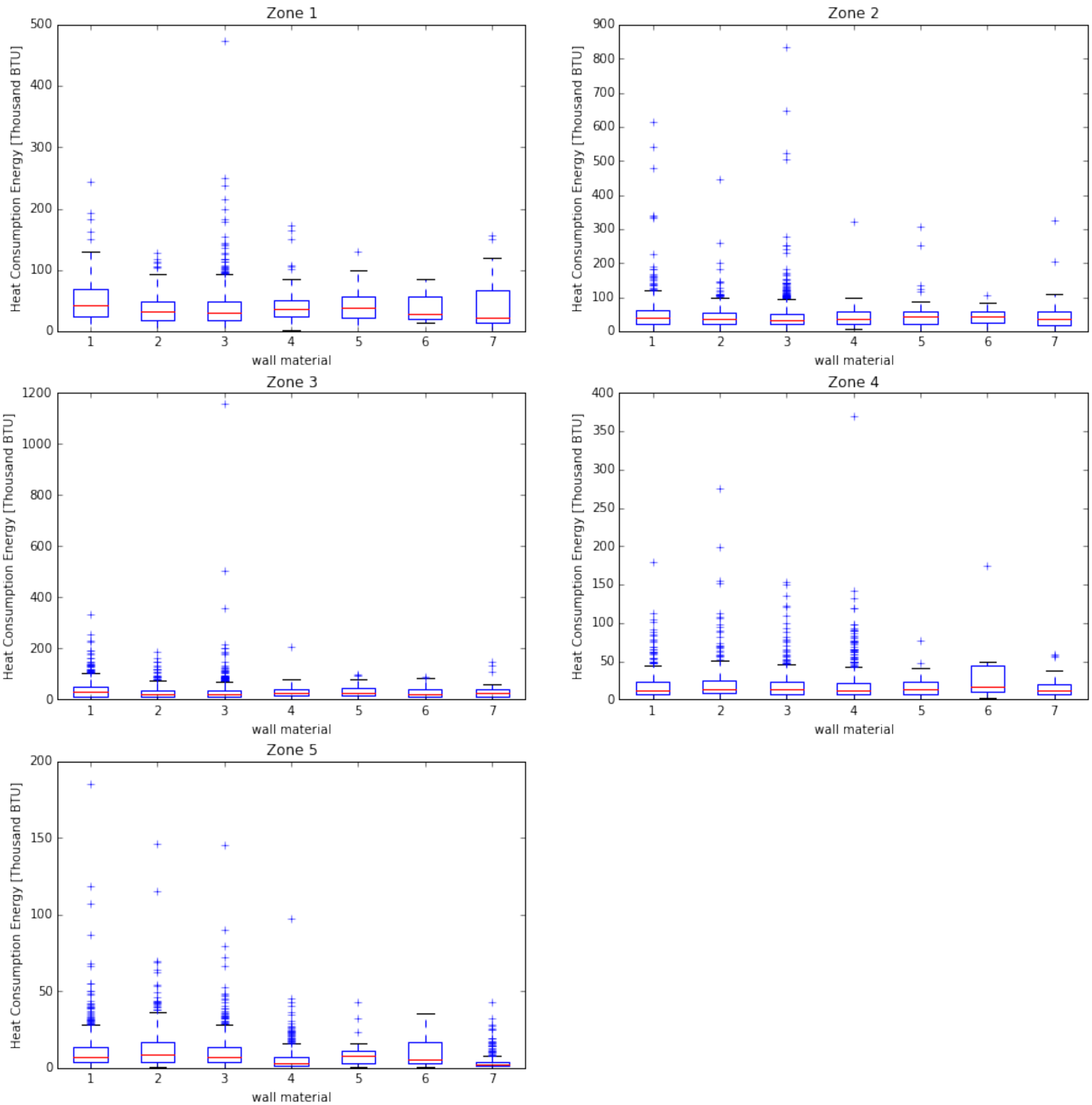

Figure 1 Energy consumption for each wall material in every zone

It seemed that there is no very clear relationship between them. But we did find that houses with concrete wall type have lower energy consumption for space heating compared to other materials in every zone. To obtain a more clear and valid relationship between energy consumption and material type, we need to find some primary variables.

Data selection

As we were looking for the relationship between outer wall materials and energy efficiency for space heating, we needed to consider variables that may have great influences on this energy consumption and construct our model to illustrate the contribution to the space heating energy efficiency for each main variable and variable of outer wall materials if it's not included in the influential ones.

By looking at the 'public_layout.csv' downloaded from EIA, parameters after column of 800 are all about energy consumption and costs. Since we were supposed to extract column of 830 'Total heated square footage' and column of 909 'Total usage for space heating, in thousand BTU, 2009', and then to calculate the usage for space heating per square footage, thus, we didn't need to consider these parameters after column of 800. Indeed, we manually checked these parameters and decided not to take them into consideration for the parameter selection.

To choose variables for our model, we went through each column from 1 to 800 to see their correlation coefficient with the total energy consumption for space heating, and for each variable whose absolute value of correlation coefficient with the heating space energy consumption is greater than 0.35, we created a dictionary to store the position of this variables and the correlation coefficients sorted it according to the value of correlation coefficient. The results are shown in Table 1.

Then we manually selected variables we needed for our model from the results in Table 1. We excluded those related to squares or having nothing to do with energy consumed for heating space firstly. For the rest, since the first two variables in the table are pointing to similar thing, we chose 'HDD65'which tells us information about the current year. For 'UGWARM', since both this variable and 'FUELHEAT' means the way or whether some way is used for heating and the latter can include more ways used by different households, we only chose the latter for the model. For 'PGASHEAT', we thought it's strongly related to the way used for heating, so we excluded that as well. For the last one 'AIA_Zone', it might have similar effects on energy consumption as 'HDD65', and we chose 'HDD65' as it has greater correlation coefficient with energy consumption.

Variable Name	**Position (start counting columns from column 0)**	**Correlation coefficient with total energy consumption for space heating**
HDD30YR, heating degree days, 30-year average 1981-2010, base 65F	Column 8	0.5781
HDD65, heating degree days in 2009, base temperature 65F	**Column 6**	**0.5735**
BASEUSE, portion of basement exclusively used by housing unit in apartment building with 2-4 unit	Column 46	0.5134
BASEHEAT, heating used in basement	Column 40	0.4988
BASEFIN, finished basement	Column 38	0.4973
BASECOOL, cooling used in basement	Column 43	0.4888
UGWARM, natural gas used for space heating	Column 667	0.4184
PGASHEAT, who pays for natural gas for space heating	Column 705	0.4115
CELLAR, basement in housing unit	Column 35	0.4083
DIPSTICK, automotive block or engine heater or battery blanket used	**Column 315**	**0.3793**

HEATROOM, number of rooms heated	Column 461	0.3657
FUELHEAT, main space heating fuel	**Column 430**	**-0.3709**
DIVISION, census Division	Column 2	-0.4191
CDD30YR, cooling degree days, 30-year average 1981-2010, base 65F	Column 9	-0.4419
REGIONC, census Region	Column 1	-0.4443
REPORTABLE_DOMAIN, reportable states and groups of states	Column 3	-0.4676
CDD65, cooling degree days in 2009, base temperature 65F	Column 7	-0.4728
AIA_Zone, AIA Climate Zone, based on average temperatures from 1981 - 2010	Column 11	-0.5549

Table 1 Correlation coefficient

The variable 'WALLTYPE' which means materials for outer walls is not among the most influential ones, but since we want to see the relationship between it and energy consumption for space heating, we also took this as a variable in our model. Therefore, the variables we finally chose for the model and their details as indicated in the 'recs2009_public.csv' file are as follows.

WALLTYPE was explained just now. HDD65 is the heating degree days in 2009, base temperature 65 F. DIPSTICK means automotive block or engine heater or battery blanket used and there are three choices: 0 means no, 1 means yes and -2 means not available. We didn't delete data labeled with '-2' in this variable because it would result in a rather small sample. Finally, the FUELHEAT means the main space heating fuel and there are 9 kinds of fuels: Natural gas, LPG, Fuel oil, Kerosene, Electricity, Wood, Solar, District steam and other fuel.

Build model

Firstly, our model will have 7 parameters $\theta_1, \theta_2, \dots, \theta_7$ for every material. Similarly, there are 3 parameters $\alpha_1, \alpha_2, \alpha_3$ for dipstick because there are 3 choices. Also there are 9 parameters $\gamma_1, \gamma_2, \dots, \gamma_9$ for 9 main space heating fuels respectively. To HDD65, because the range is really large, from 53 to 12525, and hard to give every value a parameter so we decided to use the same method as we used in assignment 2. This HDD65 effect can be modeled with a piecewise linear and continuous HDD65-dependent load model. For each facility, we divide HDD65 into twenty-five equally sized intervals. Base on these parameters, we built our model like this:

$$Y(m_i, h_i, d_i, f_i) = \theta_i + \sum_{j=1}^{25} \varepsilon_j \, h_j + \alpha_i + \gamma_i$$

Our parameter matrix is like this:

$$\beta = \left[\theta_{1,}\theta_2, \dots, \theta_7, \varepsilon_{1,}\varepsilon_2, \dots, \varepsilon_{25}, \alpha_1, \alpha_2, \alpha_3, \gamma_1, \gamma_2, \dots, \gamma_9\right]^T$$

Our design matrix is the combination of separate matrix for each factor. For example, to get the matrix of materials, we will do like this:

```
In [834]: N = len(Data_2)
          DM_material = np.zeros((N,7))
          for i in range(N):
              DM_material[i][Data_2['WALLTYPE'][i]-1] = 1
          DM_material

Out[834]: array([[ 1.,  0.,  0., ...,  0.,  0.,  0.],
                 [ 0.,  1.,  0., ...,  0.,  0.,  0.],
                 [ 1.,  0.,  0., ...,  0.,  0.,  0.],
                 ...,
                 [ 0.,  0.,  1., ...,  0.,  0.,  0.],
                 [ 1.,  0.,  0., ...,  0.,  0.,  0.],
                 [ 0.,  1.,  0., ...,  0.,  0.,  0.]])
```

If i_{th} material is brick, then we will put value 1 in the ith row and in the first column of matrix and the other column's values are 0 in this row. So basically we just need to put value 1 in the i_{th} column where i is the respond number od material.

The other separate matrixes are designed by the same method. The final matrix is like this:

```
In [1297]: X = np.hstack((DM_material,DM_fuel,DM_hdd,DM_dipstick))
           X

Out[1297]: array([[ 1.,  0.,  0., ...,  0.,  0.,  1.],
                  [ 0.,  1.,  0., ...,  0.,  0.,  1.],
                  [ 1.,  0.,  0., ...,  0.,  0.,  1.],
                  ...,
                  [ 0.,  0.,  1., ...,  1.,  0.,  0.],
                  [ 1.,  0.,  0., ...,  1.,  0.,  0.],
                  [ 0.,  1.,  0., ...,  0.,  0.,  1.]])
```

According to the formula of beta hat, we can calculate the beta hat.

$$\hat{\beta} = (X^T * X)^{-1} * X^T * Y$$

Validation

To validate our linear model, we would like to how well our model can fit the actual situation by calculating the value of R-square. We took our generated design matrix X multiplied by matrix of beta hat containing coefficients for each variable in X, and then got our predicted matrix Y which indicates the energy consumption for space heating predicted by our model. By calculating SSres and SStot, we got the R-square value with the formula $\mathrm{R}^2 = 1 - \frac{\mathrm{SSres}}{\mathrm{SStot}}$ and the value is -5.82. The value is negative and is less than -5, thus we considered this model doesn't have a very desirable imitative effects. However, considering there are actually too many variables and many not applicable data, we supposed this model is still validated and it can indicate the contribution to energy efficiency for space heating of some main variables.

We also created intervals for each beta based on a 95% confidence level with calculating the MSE, standard deviation for beta and t-test value with α = 5%. Each interval is as $[\beta_i - t_{\frac{\alpha}{2}} \times \sigma, \beta_i + t_{\frac{\alpha}{2}} \times \sigma]$. It means there is a 95% chance that the confidence interval we calculated contains the true mean value of β_i.

Finally we assume that we have a null hypothesis H0 : βk = 0 for each one of the coefficients in β. At level α=0.01, we found that we can reject all null hypothesis. So we can say that these variables are all significant.

So according to each material's β value, 22, -26, 26, -80, -19, -12 and -38, we can know that houses with material stucco, concrete, wood, composition and stone have relatively lower energy consumption and brick and siding have higher energy consumption for space heating. Actually this does not consist to the R-value results.

Future work

The R-square value in our model is less than negative 5 and it remains to be improved. There are more than 10 thousands data in the dataset, so there must be complex dependence among different variables and it's hard to analyze each variable. To deal with that, in the following work we may analyze more variables in the dataset and also analyze the correlation among those variables themselves rather than mainly focus on their correlation with energy consumption. If the dependence among the main variables can be largely weaken, the model can be more accurate.