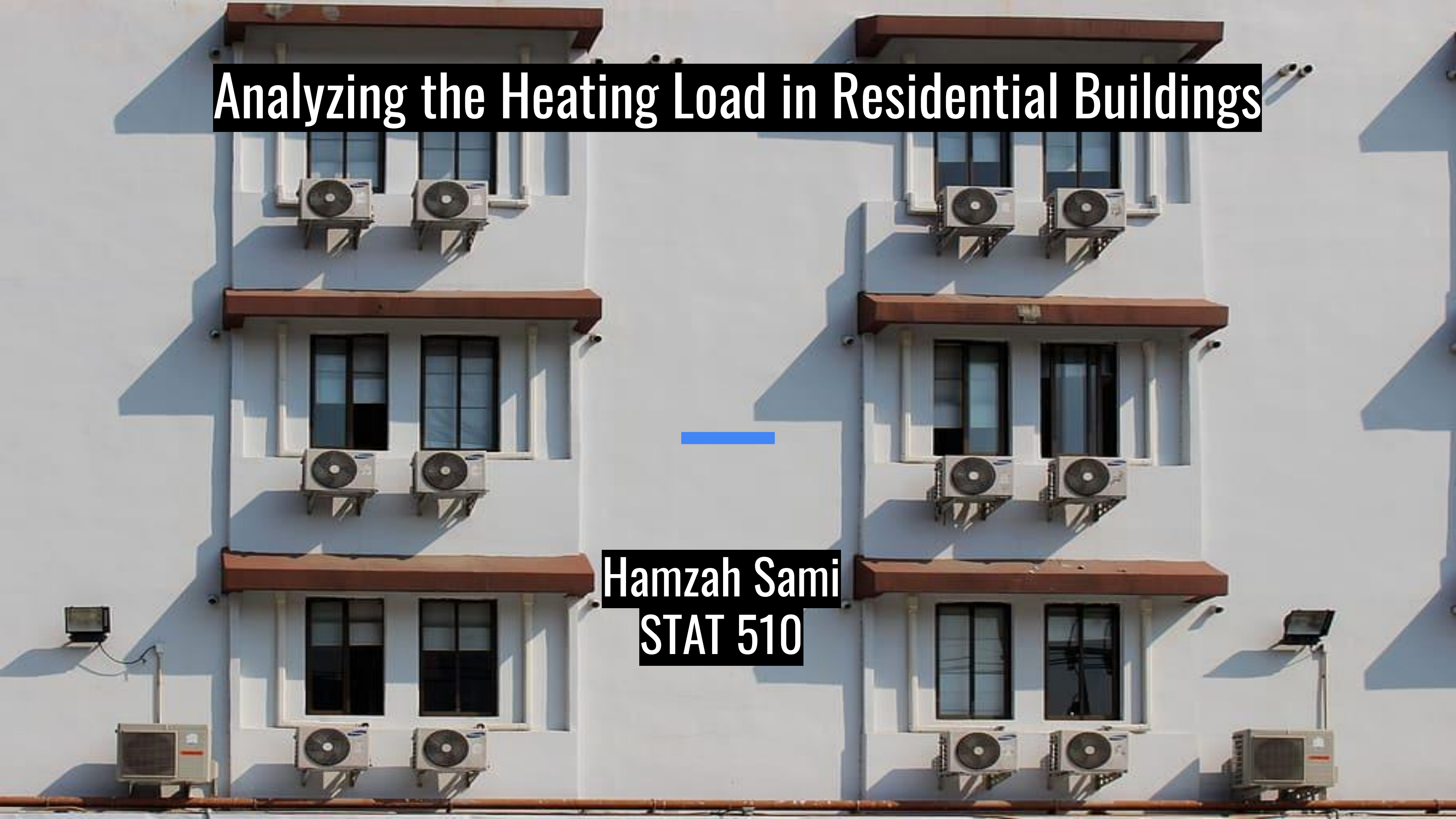
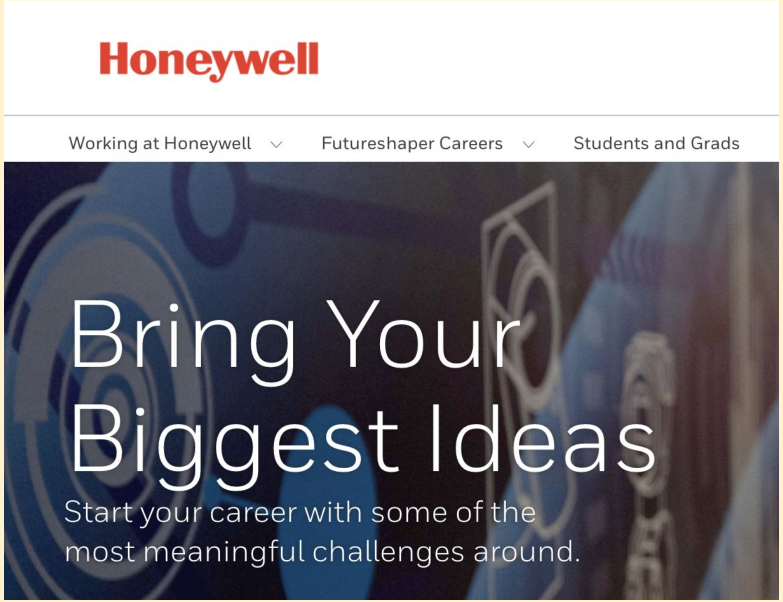


Analyzing the Heating Load in Residential Buildings

Hamzah Sami
STAT 510

Background

- Applying for internship at Honeywell.
- Intern Program was cancelled but finishing project anyways.
- Learn more about energy data.



Background

- **Heating Load:** amount of heat energy needed to maintain a building’s internal temperature.
- When designing residential buildings, structural engineers need to know the estimated cooling and heating load in a building to build an efficient air conditioner and furnace to maintain the internal temperature.
- For this presentation, we’ll only look into regressing the heating load of a building.

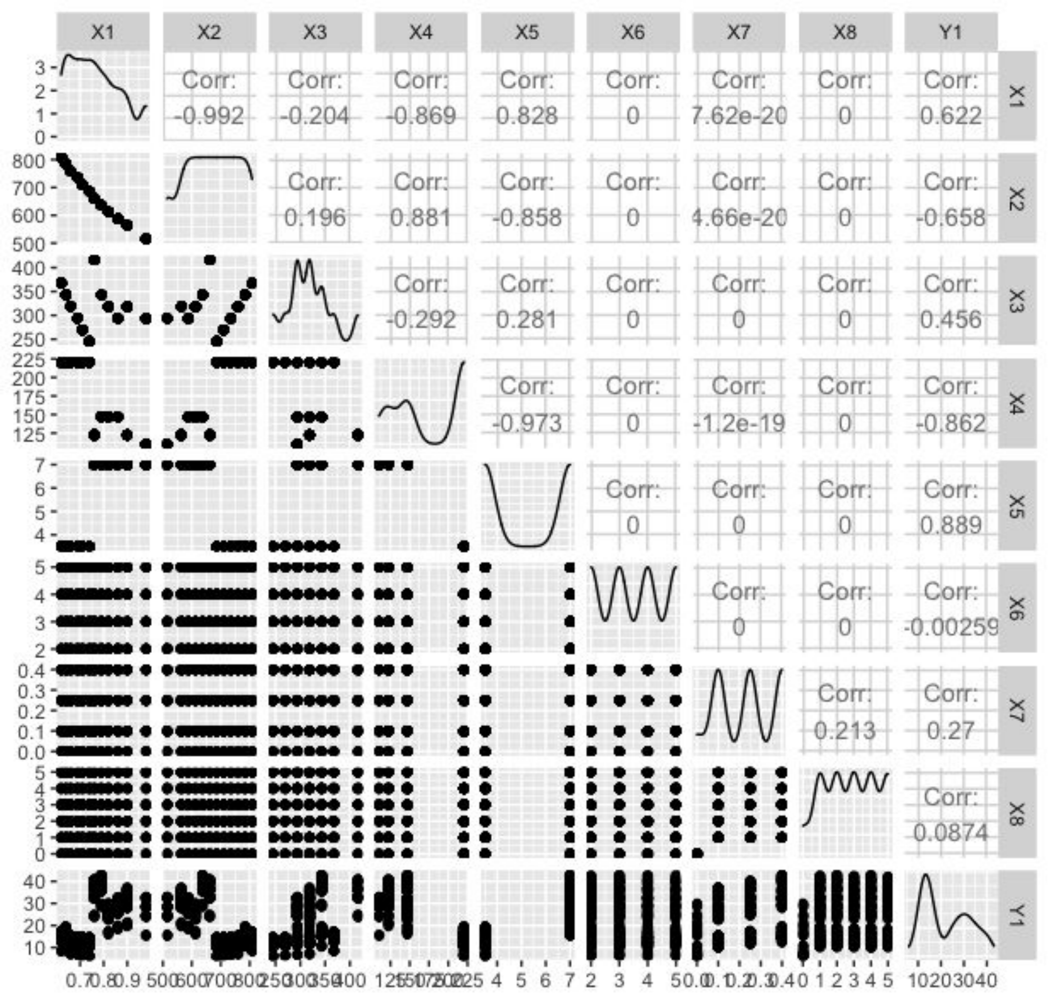
Data

- Goal: Predicting Heating Load (y1) in residential buildings.
- 769 observations
- X1: Relative Compactness: ratio of volume to surface area.
- X2 Surface Area: surface area of building. (yd^2)
- X3 Wall Area: area of building covered by width of wall (yd^2)
- X4 Roof Area: Area covered under roofs (yd^2)
- X5 Overall Height: height of buildings (yd)
- **X6** Orientation direction front of building faces sun.
- {2 =North facing
- 3 = South facing
- 4 = East facing
- 5 = West facing}
- X7 Glazing Area (area of building covered in glass) (yd^2)
- X8 Glazing Area Distribution (percentage of building covered in glass)
- y1 Heating Load : how much heat energy is needed to maintain internal building temperature (units: btu)
- y2 Cooling Load: how much heat energy is removed to maintain internal building temperature (units: btu)

“Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools”- Tsanas & Xifara

Correlation

-x2 and x1 are negatively correlated (surface area and relative compactness)
-x5 and n4 are negatively correlated (overall height and orientation)
-x4 and x2 are positively correlated (orientation and surface area)
-y1 and x5 are positively correlated (heating load and overall height)
-y1 and x8 have poor correlation (heating capacity and glass area distribution)
-x8 has zeros haha



Model Building

Stepwise Regression

- Stepwise Regression with partial F tests
- After 6 iterations,
- Starting Model:
- Y1= x1+x2+x3+x5+x7+x8
- Note we omit x4 and x6. These two predictors correspond to Roof Area and Orientation.

```
Step:  AIC=1659.48
y1 ~ x5 + x7 + x3 + x1 + x2 + x8
```

Stepwise Regression

Summary

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.932873  19.018972   4.413 1.17e-05 ***
x1          -64.773991  10.283093  -6.299 5.06e-10 ***
x2           -0.087290   0.017065  -5.115 3.97e-07 ***
x3            0.060813   0.006644   9.153  < 2e-16 ***
x5            4.169939   0.337781  12.345  < 2e-16 ***
x7           19.932680   0.813483  24.503  < 2e-16 ***
x8            0.203772   0.069875   2.916  0.00365 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.933 on 761 degrees of freedom
Multiple R-squared:  0.9162,    Adjusted R-squared:  0.9155
F-statistic:  1387 on 6 and 761 DF,  p-value: < 2.2e-16
```

ANOVA

```
Analysis of Variance Table

Response: y1
           Df  Sum Sq Mean Sq   F value    Pr(>F)
x1          1 30238.2 30238.2 3516.2320 < 2.2e-16 ***
x2          1  8092.9  8092.9  941.0751 < 2.2e-16 ***
x3          1 26144.8 26144.8 3040.2338 < 2.2e-16 ***
x5          1  1310.6  1310.6  152.4011 < 2.2e-16 ***
x7          1  5686.0  5686.0  661.1997 < 2.2e-16 ***
x8          1    73.1    73.1    8.5045  0.003647 **
Residuals 761  6544.3     8.6
---
```

Best Subsets Regression

- Same as stepwise.
  Use six predictors.
- Omit x4 and x6.

```
> summary.mod$adjr2
[1] 0.7908142 0.8635453 0.9095145 0.9143384 0.9151686 0.9155346 0.9154303
```

```
  (Intercept)    x1    x2    x3    x4   x5    x6    x7    x8
1        TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE
2        TRUE FALSE FALSE FALSE FALSE TRUE FALSE  TRUE FALSE
3        TRUE FALSE FALSE  TRUE FALSE TRUE FALSE  TRUE FALSE
4        TRUE  TRUE FALSE FALSE  TRUE TRUE FALSE  TRUE FALSE
5        TRUE  TRUE FALSE FALSE  TRUE TRUE FALSE  TRUE  TRUE
6        TRUE  TRUE  TRUE  TRUE FALSE TRUE FALSE  TRUE  TRUE
7        TRUE  TRUE  TRUE  TRUE FALSE TRUE  TRUE  TRUE  TRUE
```

Starting model:

$$y_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8$$

Mallows' Cp

```
summary.mod$cp
1] 1128.231082  470.716485   56.367557   13.834344    7.351511    5.060596    7.000000
```

- As we add more predictors, the value drops which is obvious.
- The smallest value corresponds to a model with six predictors but maybe we could have used seven predictors.

Interpretation of Coefficients

All continuous variables

X1: When the relative compactness increases by one unit, the expected heating load decreases by 64.773991.

X2: When the surface area increases by one unit, the expected heating load decreases by 0.087290.

X3: When the wall area increases by one unit, the expected heating load increases by 0.060813.

X5: When the overall height increases by one unit, the expected heating load increases by 4.169939.

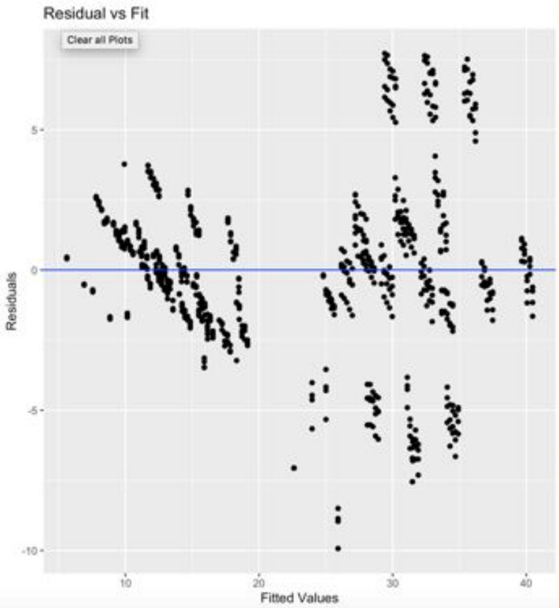
X7: When the glazing area increases by one unit, the expected heating load increases by 19.932680.

X8: When the glazing area distribution increases by one unit, the expected heating load increases by 0.203772.

Residual Analysis

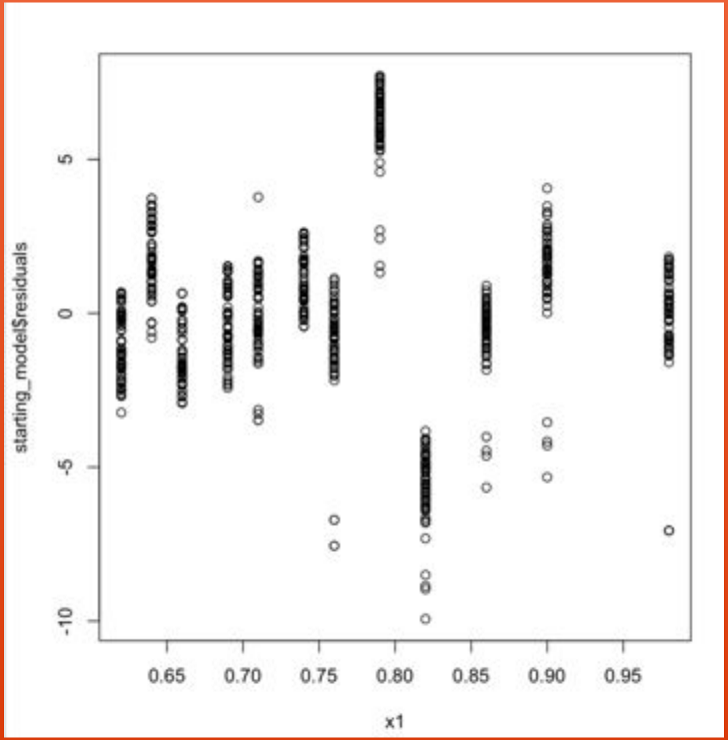
Residual Analysis

- LINE conditions not satisfied
- From graphing starting formula, it appears we need a quadratic transform
- Have to graph residual vs predictor plots
- Transform relevant predictors

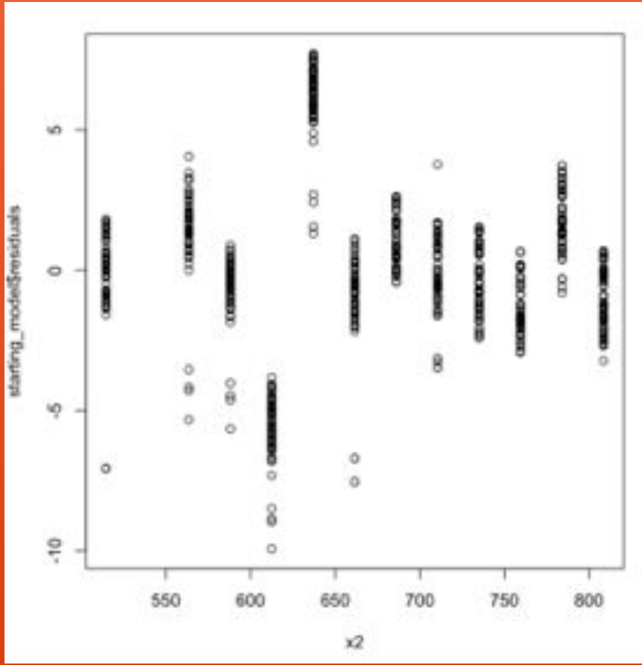



```
        Shapiro-Wilk normality test

data:  starting_model$residuals
W = 0.9542, p-value = 1.006e-14
```

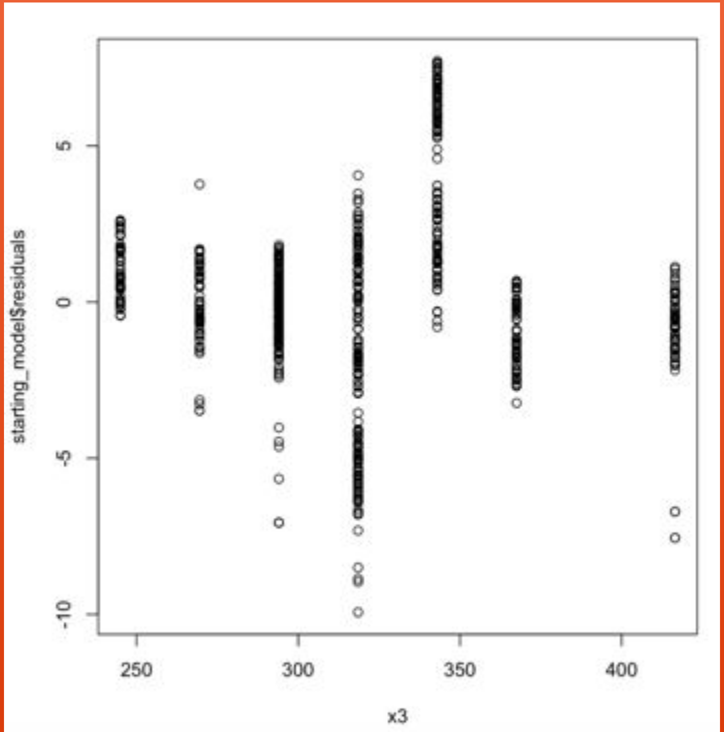
x1 vs Residuals



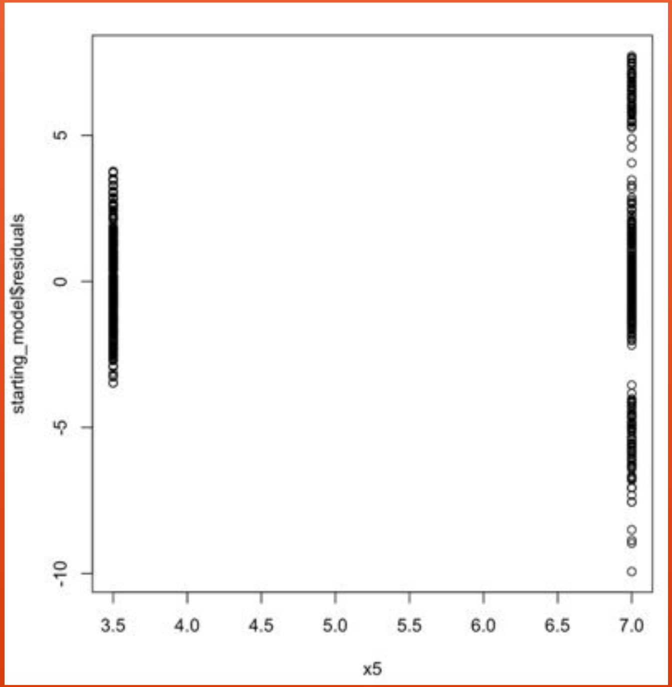
x2 vs Residual



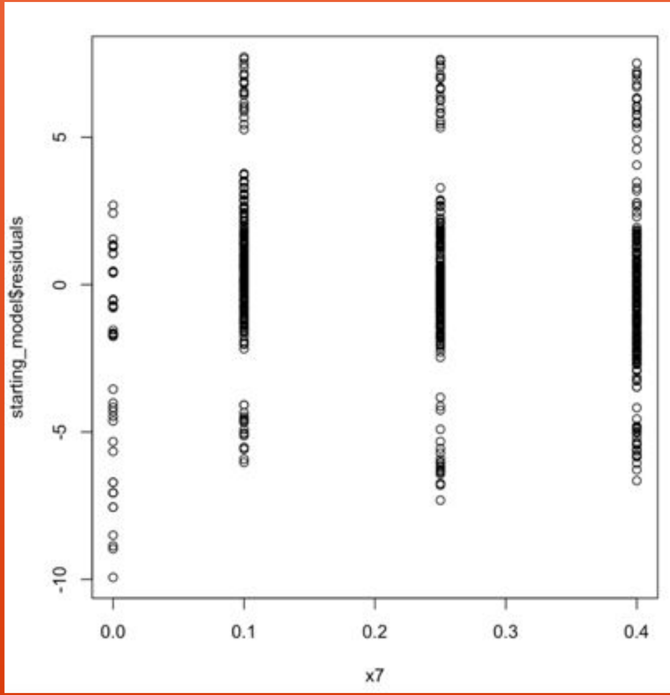
x3 vs Residual



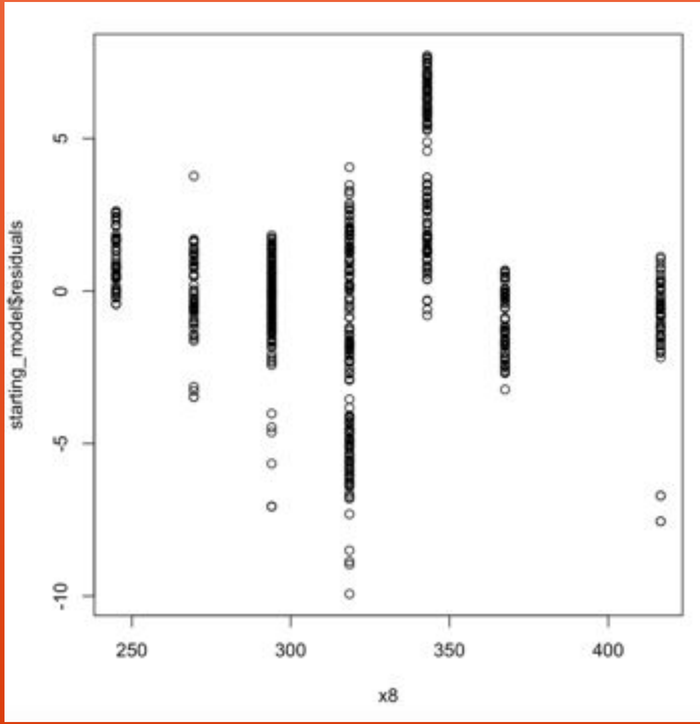
x5 vs Residual



x7 vs residual



x8 vs residual

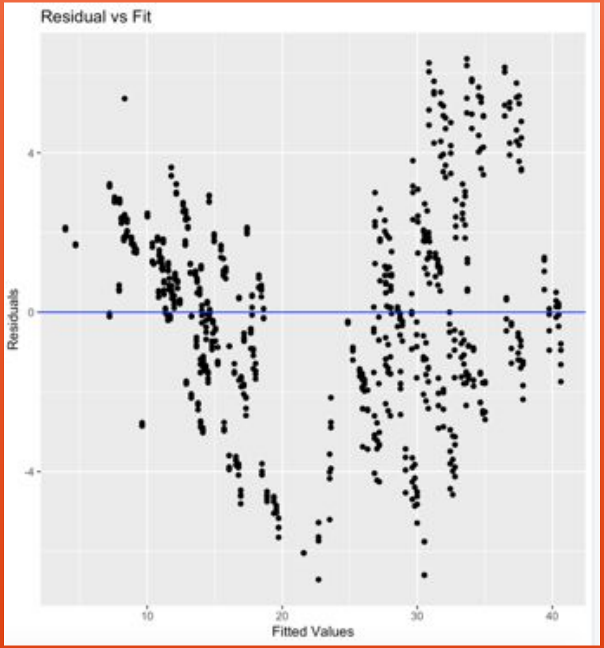


Residual Analysis

- Based off residual plots, I transformed x1, x2, x3, and x8
- Quadratic Model:

$$y_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{88} x_8^2$$

Residual Analysis: Quadratic Model



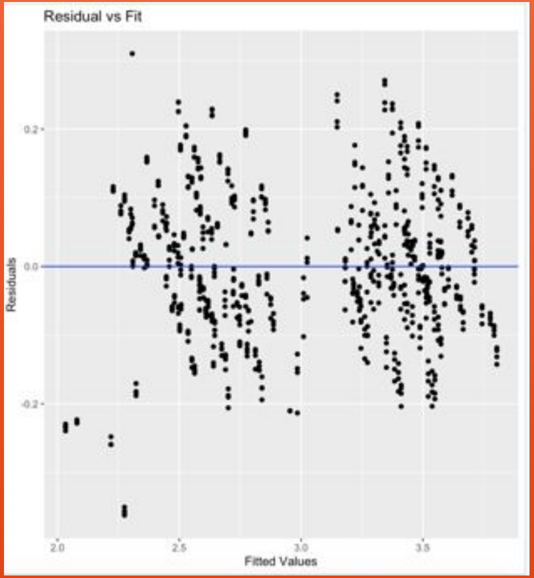

```
Shapiro-Wilk normality test

data:  q_model$residuals
W = 0.99412, p-value = 0.004347
```

- Model still does not work haha
- We have kind of fixed the linearity, variance conditions
- To correct normality, use log transform on response

Residual Analysis: Updated Model

$$\log(y_1) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{88} x_8^2$$

- Now it works!!!!!!!!!
- We'll call this the updated model




```
        Shapiro-Wilk normality test

data:  updated_model$residuals
W = 0.99632, p-value = 0.07017
```

Recap

Starting model

$$y_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8$$

Quadratic model

$$y_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{88} x_8^2$$

Upgraded Model

$$\log(y_1) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{88} x_8^2$$

Interaction Terms

Interaction Terms

- Pairwise interactions due to quadratic model
- Run the stepwise regression on all possibilities and residual analysis
- Eventually…

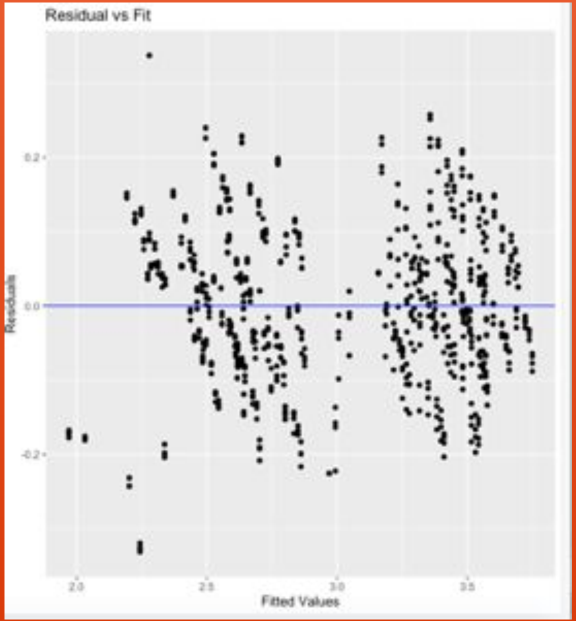
(x1*x7): Interaction of relative compactness and glazing area

(x3*x7): Interaction of wall area and glazing area

```
add1(q_model, ~.+(x1*x2)+(x1*x3)+(x1*x5)+(x1*x7)+(x1*x8)+(x2*x3)+(x2*x5)+(x2*x7)+(x2*x8)+(x3*x5)+(x3*x7)+
        (x3*x8)+(x5*x7)+(x5*x8)+(x7*x8), test = 'F')
```

Final Model

$$\log(y_1) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{88} x_8^2 + \beta_{17} x_1 x_7 + \beta_{37} x_3 x_7$$




```
        Shapiro-Wilk normality test

data:  final_model$residuals
W = 0.99711, p-value = 0.1892
```

Final Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.299e+01  5.265e+00 -10.063  < 2e-16 ***
x1           3.743e+00  4.250e+00   0.881   0.3788    
x2           9.961e-02  1.046e-02   9.519  < 2e-16 ***
x3           3.995e-03  1.834e-03   2.178   0.0297 *  
x5           4.215e-01  2.856e-02  14.760  < 2e-16 ***
x7           2.521e+00  3.273e-01   7.703 4.19e-14 ***
x8           1.225e-01  1.022e-02  11.990  < 2e-16 ***
I(x1^2)      1.169e+01  2.770e+00   4.219 2.75e-05 ***
I(x2^2)     -5.252e-05  5.858e-06  -8.966  < 2e-16 ***
I(x3^2)     -4.705e-06  2.297e-06  -2.048   0.0409 *  
I(x8^2)     -1.911e-02  1.778e-03 -10.748  < 2e-16 ***
x1:x7       -6.835e-01  2.734e-01  -2.500   0.0126 *  
x3:x7       -3.382e-03  6.629e-04  -5.102 4.25e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1044 on 755 degrees of freedom
Multiple R-squared:  0.9528,	Adjusted R-squared:  0.952 
F-statistic:  1270 on 12 and 755 DF,  p-value: < 2.2e-16
```

Final Model-ANOVA

```
Analysis of Variance Table

Response: log(y1)
           Df Sum Sq Mean Sq   F value    Pr(>F)    
x1          1 72.331  72.331 6637.6562 < 2.2e-16 ***
x2          1 11.802  11.802 1083.0379 < 2.2e-16 ***
x3          1 58.255  58.255 5345.9095 < 2.2e-16 ***
x5          1  5.354   5.354  491.3397 < 2.2e-16 ***
x7          1 15.047  15.047 1380.8222 < 2.2e-16 ***
x8          1  0.451   0.451   41.3441 2.263e-10 ***
I(x1^2)     1  0.199   0.199   18.2969 2.133e-05 ***
I(x2^2)     1  0.969   0.969   88.8901 < 2.2e-16 ***
I(x3^2)     1  0.046   0.046    4.1939   0.04091 *  
I(x8^2)     1  1.259   1.259  115.5172 < 2.2e-16 ***
x1:x7       1  0.024   0.024    2.2247   0.13623    
x3:x7       1  0.284   0.284   26.0345 4.247e-07 ***
Residuals 755  8.227   0.011
```

Leverage Points

- Very small leverage values.
- H > 3(p/n)
- P = 13
- N = 769

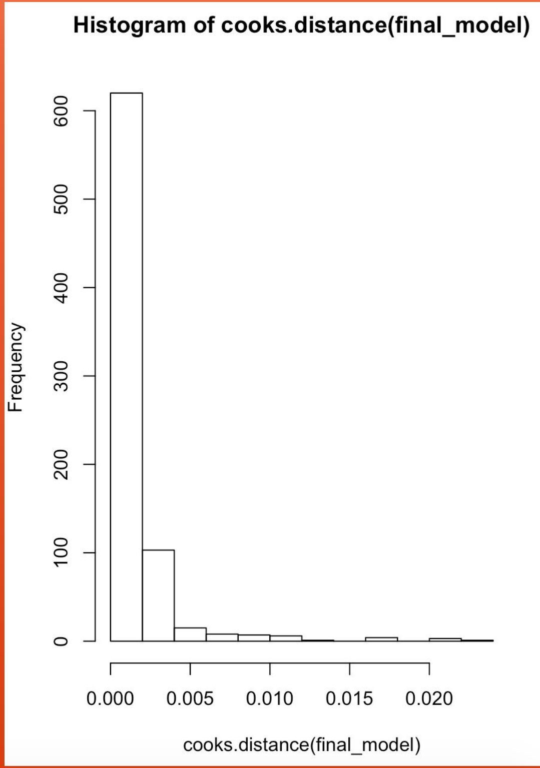
```
> hv = hatvalues(final_model)
> which(hv == max(hv))
21 22 23 24
21 22 23 24
```

```
> max(hv)
[1] 0.04822085
```

```
> which(hv > 39/769)
named integer(0)
```

Cook’s Distance

- All 769 points have a small Cook’s Distance
- No influential points…




```
> hist(cooks.distance(final_model))
```

Research Questions

Research Question 1

- What is the 95% confidence interval for the expected change in heating load when the surface area (x2) is increased by an additional unit?

Answer to Question 1

	2.5 %	97.5 %
(Intercept)	-6.332306e+01	-4.264965e+01
x1	-4.600371e+00	1.208675e+01
x2	7.906682e-02	1.201511e-01
x3	3.939043e-04	7.595460e-03
x5	3.654359e-01	4.775516e-01
x7	1.878389e+00	3.163321e+00
x8	1.024670e-01	1.425912e-01
I(x1^2)	6.248324e+00	1.712268e+01
I(x2^2)	-6.401770e-05	-4.101979e-05
I(x3^2)	-9.214231e-06	-1.947901e-07
I(x8^2)	-2.260265e-02	-1.562106e-02
x1:x7	-1.220184e+00	-1.467885e-01
x3:x7	-4.683510e-03	-2.080937e-03

- Confint(final_model)
- Proper Explanation: We are 95% confident that the true value of beta2 lies between 0.0791 btu/m^2 and 0.1202 btu/m^2.

Research Question 2

- Does a model containing the interaction term x1*x7 useful in predicting heating after controlling for other predictors?

$$H_0: \beta_{17} = 0$$

$$H_1: \beta_{17} \neq 0$$

Decision Rule: Reject H0 if p-value is less than 0.05.

```
Analysis of Variance Table

Response: log(y1)
           Df Sum Sq Mean Sq   F value    Pr(>F)    
x1          1 72.331  72.331 6637.6562 < 2.2e-16 ***
x2          1 11.802  11.802 1083.0379 < 2.2e-16 ***
x3          1 58.255  58.255 5345.9095 < 2.2e-16 ***
x5          1  5.354   5.354  491.3397 < 2.2e-16 ***
x7          1 15.047  15.047 1380.8222 < 2.2e-16 ***
x8          1  0.451   0.451   41.3441 2.263e-10 ***
I(x1^2)     1  0.199   0.199   18.2969 2.133e-05 ***
I(x2^2)     1  0.969   0.969   88.8901 < 2.2e-16 ***
I(x3^2)     1  0.046   0.046    4.1939   0.04091 *  
I(x8^2)     1  1.259   1.259  115.5172 < 2.2e-16 ***
x1:x7       1  0.024   0.024    2.2247   0.13623    
x3:x7       1  0.284   0.284   26.0345 4.247e-07 ***
Residuals 755  8.227   0.011                        
```

Answer to Research Question 2

```
Analysis of Variance Table

Model 1: log(y1) ~ x1 + x2 + x3 + x5 + x7 + x8 + I(x1^2) + I(x2^2) + I(x3^2) +
    I(x8^2) + (x3 * x7)
Model 2: log(y1) ~ x1 + x2 + x3 + x5 + x7 + x8 + I(x1^2) + I(x2^2) + I(x3^2) +
    I(x8^2) + (x1 * x7) + (x3 * x7)
  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1    758 10.0047
2    755  8.2273  3    1.7774 54.37 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The p-value of the partial F test is less than 0.05. Thus, we reject the null hypothesis and conclude that the full model consisting of the additional interaction terms is the best choice of fitting the data.
- Thus, beta17 slope parameter is nonzero.

Summary

- Goal was to fit a regression model of predictors in a structural engineering dataset to predict the heating load in residential buildings.
- Used quadratic transform, logarithmic transform, and added interaction terms.
- High R squared value. Ideally low error.

Conclusion

- Find a less complex model. More interaction terms.
- Look at cooling data to see if there is any kind of connection with heating data
- Predicting using model.
- Periodicity in correlation matrix. Sine transform could potentially be used.

Thank You!