Household Energy Consumption Estimation

Andrew Norman

29 August, 2025

Data Processing

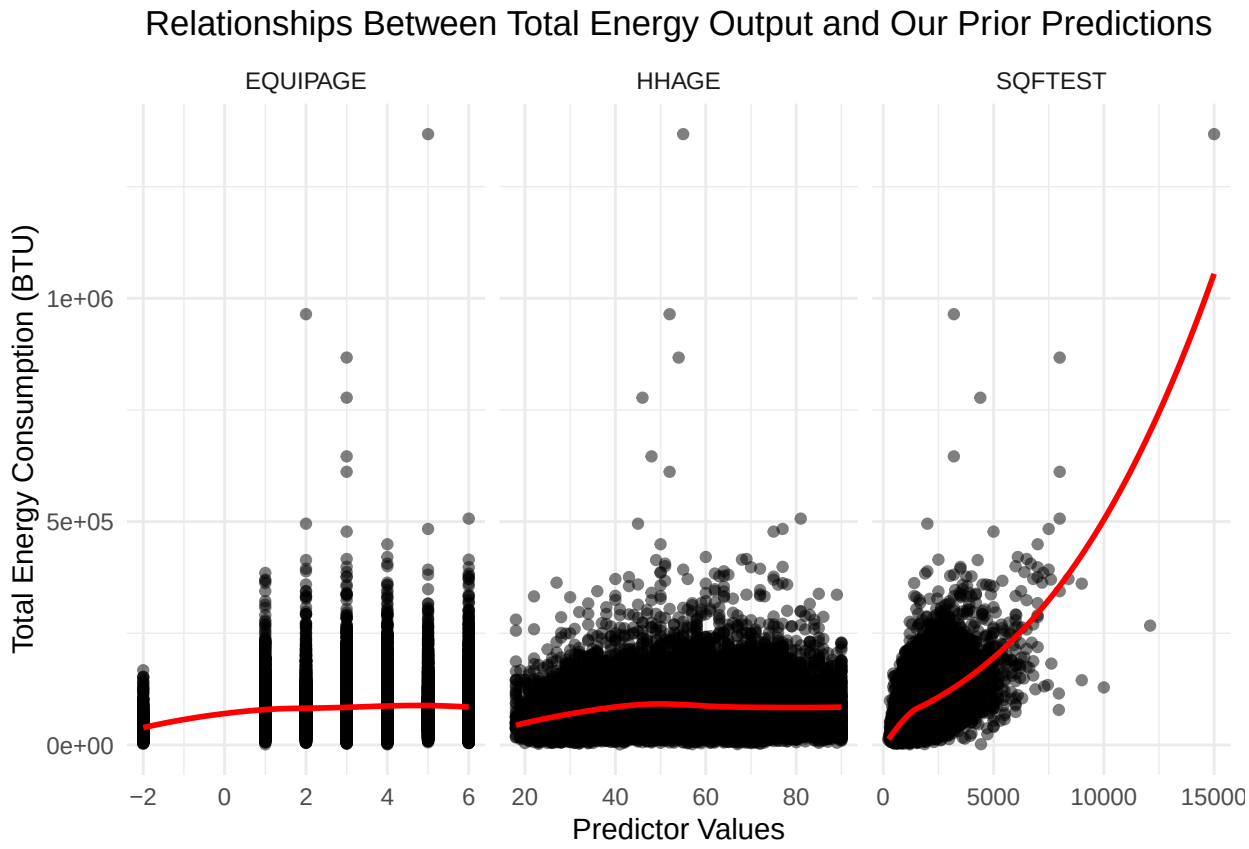
```
recs <- read.csv("recs2020_public_v7.csv", stringsAsFactors = FALSE)

factor_list <- c("TOTALBTU", "SQFTEST", "HDD65", "CDD65", "TYPEHUQ", "STORIES",
                 "TOTROOMS", "WINDOWS", "TYPEGLASS", "WINFRAME", "ADQINSUL",
                 "WASHLOAD", "WASHTEMP", "DRYRUSE", "EQUIPAGE", "FUELHEAT",
                 "AIRCOND", "ACEQUIPAGE", "H2OMAIN", "LGTIN4TO8", "MONEYPY",
                 "ATHOME", "FUELPOOL", "ROOFTYPE", "WALLTYPE", "REGIONC",
                 "YEARMADERANGE", "BEDROOMS", "TOTHSQFT", "TOTCSQFT", "HHAGE")

recs_clean <- recs %>%
  select(-TOTALBTUSPH, -TOTALBTUWTH, -TOTALBTUOTH, -DOEID) %>%
  select(all_of(factor_list)) %>%
  mutate(
    TYPEHUQ = as.factor(TYPEHUQ),
    STORIES = as.factor(STORIES),
    WINDOWS = as.factor(WINDOWS),
    TYPEGLASS = as.factor(TYPEGLASS),
    WINFRAME = as.factor(WINFRAME),
    FUELHEAT = as.factor(FUELHEAT),
    AIRCOND = as.factor(AIRCOND),
    H2OMAIN = as.factor(H2OMAIN),
    LGTIN4TO8 = as.factor(LGTIN4TO8),
    ATHOME = as.factor(ATHOME),
    FUELPOOL = as.factor(FUELPOOL),
    ROOFTYPE = as.factor(ROOFTYPE),
    WALLTYPE = as.factor(WALLTYPE),
    REGIONC = as.factor(REGIONC),
    YEARMADERANGE = as.factor(YEARMADERANGE),
    MONEYPY = as.factor(MONEYPY))

recs_long <- recs_clean %>%
  select(TOTALBTU, EQUIPAGE, SQFTEST, HHAGE) %>%
  pivot_longer(cols = c(SQFTEST, EQUIPAGE, HHAGE),
               names_to = "vars",
               values_to = "vals")

set.seed(2422024)
```

```r
eval_indices <- sample(1:nrow(recs_clean),
                       size = 1000, replace = FALSE)

eval_sample <- recs_clean[eval_indices, ]

recs_remaining <- recs_clean[-eval_indices, ]

n_remaining <- nrow(recs_remaining)

train_indices <- sample(1:n_remaining,
                        size = n_remaining / 2, replace = FALSE)

train_sample <- recs_remaining[train_indices, ]

test_sample <- recs_remaining[-train_indices, ]

eval_sqft <- eval_sample %>%
  mutate(SQFTEST = SQFTEST * 1.5)

eval_equipage <- eval_sample %>%
  mutate(EQUIPAGE = EQUIPAGE + 10)

eval_hhage <- eval_sample %>%
  mutate(HHAGE = HHAGE + 10)
```

Prior Assumptions

Prior to testing the relationship between these different variables, and the target variable `TOTALBTU` I can make some predictions about the relationships between these variables. I expect that homes with a larger estimated square footage, `SQFTEST`, will consume more energy overall. These larger homes would require the use of more energy for heating, cooling, and other daily functions. On top of that, I anticipate the age of the main space heating equipment `EQUIPAGE` will likely have an impact on the total energy consumption. I believe that likely the older heating systems aren't as energy efficient, leading to higher energy consumption compared to the newer equipment. `HHAGE` is the measure of the homeowners reported age, I predict that there will be a moderate positive relationship between this and total energy usage since as you get older you may require more or less heat than the average person to be more comfortable. Also, if you are older you are more likely to be retired and may stay at home more thus consuming more energy on average.

```
## `geom_smooth()` using formula = 'y ~ x'
```



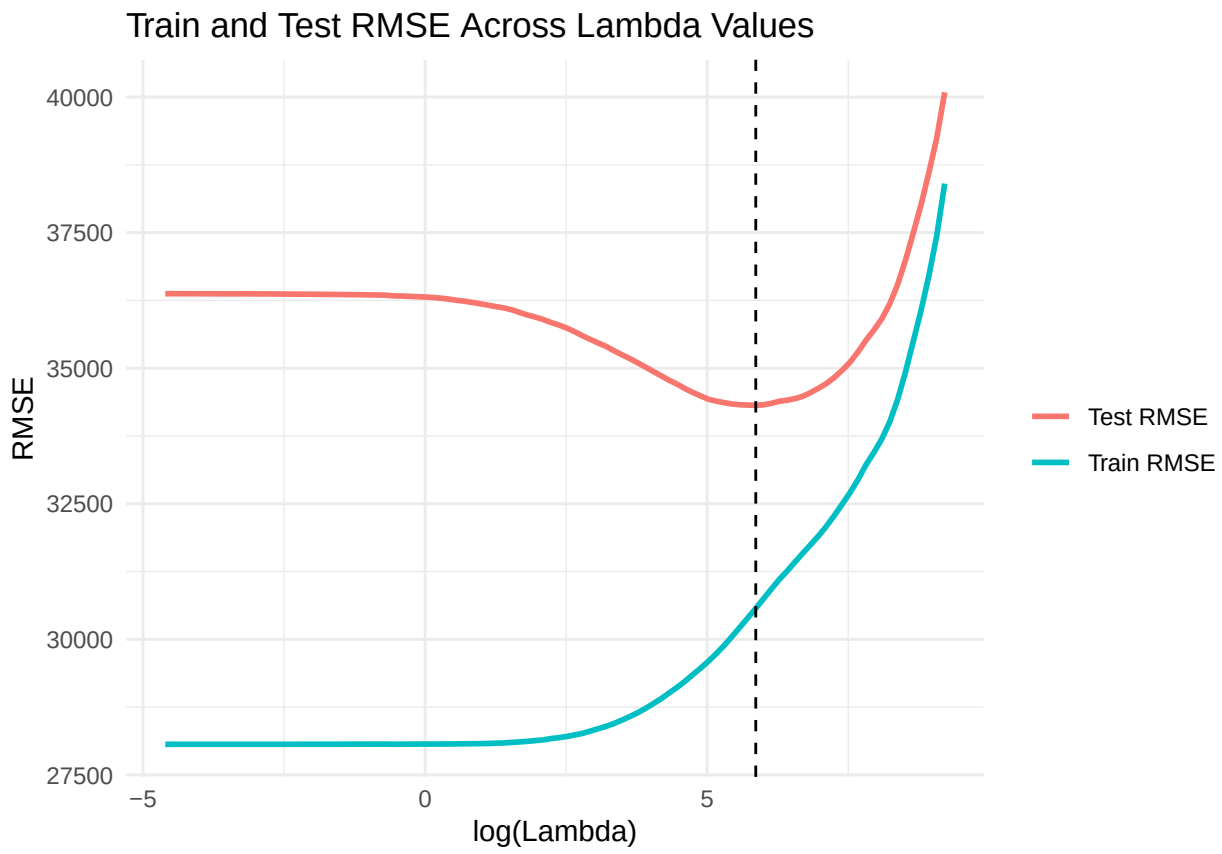
Estimating a Model

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")

## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

Table 1: Train and Test RMSEs for LASSO Across Lambda Values

lambda	train_rmse	test_rmse
9.326	28170.03	35842.22
10.723	28186.38	35795.39
12.328	28208.22	35739.92
14.175	28236.44	35672.32
16.298	28265.77	35596.61
18.738	28308.33	35529.97
21.544	28351.31	35463.27
24.771	28395.06	35401.30
28.480	28447.82	35321.51
32.745	28509.16	35247.77

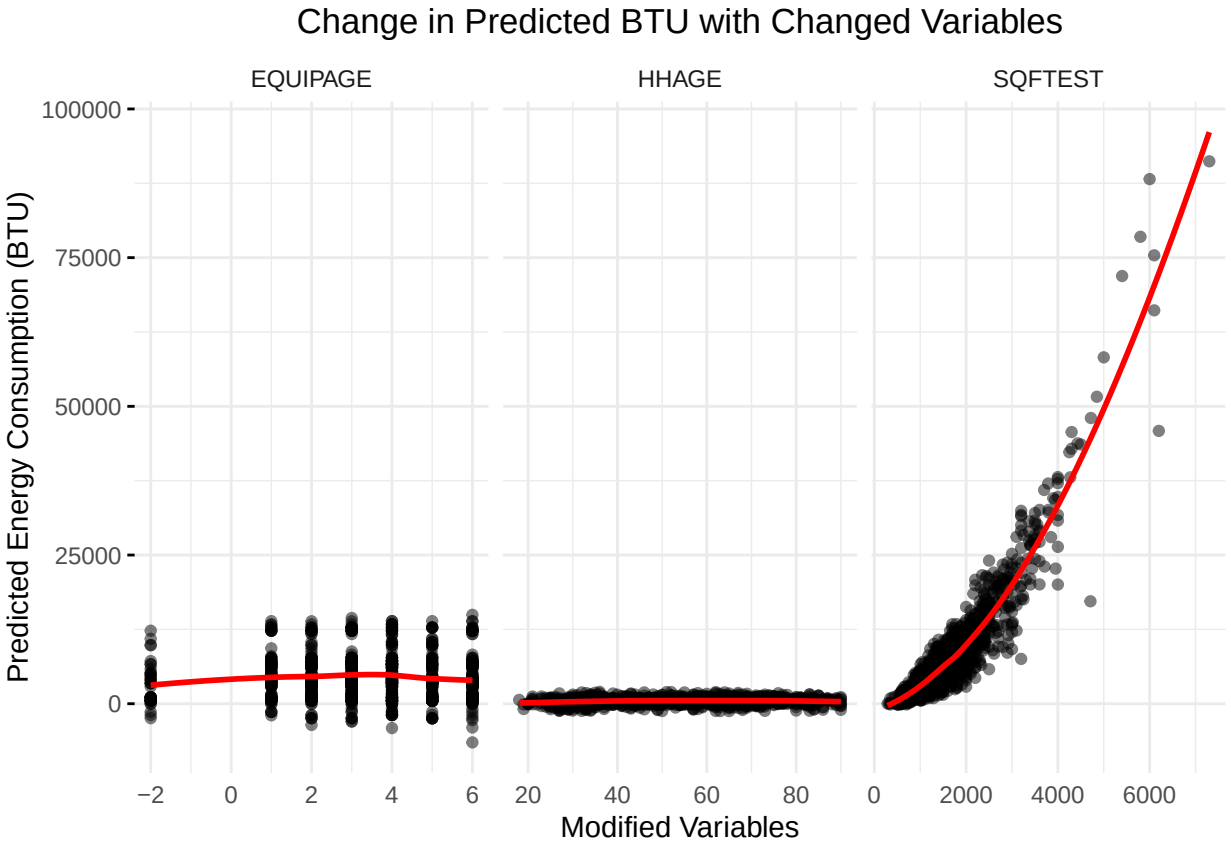
lambda	train_rmse	test_rmse
37.649	28576.09	35177.19
43.288	28645.98	35099.03
49.770	28725.85	35019.99
57.224	28815.05	34938.18
65.793	28906.66	34857.57
75.646	29011.29	34777.40
86.975	29112.98	34705.83
100.000	29224.99	34625.19
114.976	29347.76	34551.73
132.194	29468.73	34488.59
151.991	29597.80	34425.95
174.753	29738.28	34391.39
200.923	29888.26	34363.40
231.013	30054.89	34338.92
265.609	30224.64	34325.73
305.386	30396.17	34318.54
351.119	30569.81	34314.74
403.702	30751.28	34325.75
464.159	30930.80	34351.94
533.670	31105.60	34390.27
613.591	31261.23	34409.93
705.480	31427.91	34439.00
811.131	31591.84	34482.74
932.603	31750.36	34552.52
1072.267	31911.64	34632.40
1232.847	32089.72	34719.54
1417.474	32287.75	34832.01
1629.751	32496.59	34963.72
1873.817	32709.71	35113.23
2154.435	32948.62	35303.20
2477.076	33218.79	35514.64
2848.036	33443.00	35700.64
3274.549	33691.88	35908.66
3764.936	34005.08	36181.53
4328.761	34412.43	36534.16
4977.024	34919.78	36982.04
5722.368	35489.55	37487.75
6579.332	36069.50	38018.68
7564.633	36712.10	38611.68
8697.490	37452.83	39254.60
10000.000	38403.74	40086.47



I chose roughly 20 variables from the 800 that I believe have a large effect on `TOTALBTU`. For my model I decided to use the LASSO given this data sample. I decided to use LASSO because the tuning that is needed is in the lambda value; this valuecan be varied very easily to arrive at the optimal lambda that minimizes the RMSE of the model when I use the evaluation sample. LASSO also works well given a large amount of variables, so even though I am not using all 800 them this method should work well.

Evaluating the Model

```
## `geom_smooth()` using formula = 'y ~ x'
```



After increasing the `EQUIPAGE` and `HHAGE` by 10 years it appears that most of the values for those two charts are positive, meaning that the increase in Equipment Age and Homeowner Age by 10 years increase the predicted energy usage (BTU). When we multiply `SQFTEST` by 1.5 it appears that change in predicted energy usage rapidly increases with an increase in estimated square feet. Since all variables are positive, this means that my prior predictions were correct about these variables having positive relationships with `TOTALBTU`.