MSDS 6372. Energy Consumption Report

Volodymyr Orlov, Jason Lin, Jaime Villanueva

Introduction

The United States energy consumption is one of the biggest in the world with residential sectors representing approximately 40 percent[1] of it. Here we explore data collected via *Residential Energy Consumption Survey*[2] conducted by the *US Energy Information Administration* in 2009. Using this data we perfom regression analysis and test a hypothesis that there a difference in average annual bill for different types of settlements and year of construction. We build a first model for predicting annual energy consumption, in US dollars, of a household from various household's features and consumption habit of its tenants. In addition to that we test factors that are commonly assumed to affect energy bill with our second model.

Data Description

The *US Residential Energy Consumption Survey* is a comprehensive data collection program that was conducted by the *US Energy Information Administration*[3].

The data file can be obtained from the www.eia.gov website and is formatted as a CSV file. The file contains responses from 12083 randomly selected sample of the population of 113.6 million US households. Response is coded using 940 variables, which can be broken down into following categories:

- 359 imputation flags.
- 85 measurements of various energy consumption metrics.
- 58 various quantitative measurements, collected from a household or its tenants.
- 438 categorical features of a household.

Exploratory Analysis

For our analisys and models we considered only 58 quantitative and 438 categorical features of a household and dropped 359 imputation flags as well as 85 measurements of energy consumption metrics. For our first model we've been primaraly interested in predicting total energy bill per household that has been recorded as a separate variable, in whole US dollars.

To better understand relationship between total yearly cost and household's features we looked at correlation matrix of all quantitative plus our dependent variable. The plot is depicted in Figure 1

By scanning through correlation matrix we've identified several good candidates for strong predictor variables: total square footage, total number of rooms, number of rooms heated, number of ceiling fans used, number of televisions used, heating degree days in 2009 above 65F. From Figure 2 you can see there is a clear linear relashionship between all of these variables and a response we are trying to predict.

In addition to that we've used our intuition to select following variables for further analysis:

Regression Analysis

First goal of our analysis is to model total energy bill of a household using best subset of predictors. All our models are based on linear regression:

[1] https://www.nap.edu/catalog/13360/effective-tracking-of-building-energy-use-improving-the-commercial-buildings
[2] https://www.eia.gov/consumption/residential/about.php
[3] https://en.wikipedia.org/wiki/Energy_Information_Administration

$$y = X\beta + \epsilon \tag{1}$$

Where X is a $x \times k$ matrix with k observation and n independent variables, y is a $n \times 1$ vector of observations, ϵ is a $n \times 1$ irreducible error vector and β be a $k \times 1$ vector of parameters that we want to estimate.

Predictive Model

To select a best predictive model we assembeled an ordered list of qualitative and quantitative predictors using scatterplots and our intuition as a guideline. The final list can be found in Table 1. Next, we have assembled 22 models starting from a simple model with just one predictor and each subsequent model having increasing number of predictors from Table 1. We've measured BIC, AIC and CV metrics for each model. Resulting plots can be found in Figure 3. After looking at these plots we concluded that the best predictive model is the model build from predictors marked with a star in Table 1.

Descriptive Model

With our second model we decided to test predictors that are commonly assumed to have great influence on energy bill. From our experience we know following factors are thought to be important (see Table 4):

- We want to see if air cooling and heating contributes to the total dollar energy consuption. Therefore, the indicator for air conditioner *AIRCOND* and heating equipment *EQUIPNOHEAT* are included.
- Does temperature influence the use of central air and heating? Here we included variables *HDD65* (heating degree days) and *CDD65* (cooling degree days).
- Other factors that characterize a household and a building, since these influence the efficiency of a household as well as usage of the air cooling and heating systems. Therefore, the variables *HHAGE* (household age), *SQFT_100* (total square footage of the building as $TOTSQFT/100$), *ATHOME* (if the person is at home), *TOTROOMS* (number of rooms), *YEARMADE* (when the house was built), *TOTHSQFT* (total heating square footage), *TOTCSQFT* (total cooling square footage) are part of the model.
- Any interactions between these variables.

The final estimates for the model can be found in Table 5. The model can be described using following equation:

$$\begin{aligned} log(Dollar) = 11.20 + 0.0097 * sqft100 - 0.001492 * HHAGE - \\ 0.05925 * ATHOME + 0.003291 * TEMPGONE + 0.09193 * TOTROOMS - \\ 0.002504 * YEARMADE - 0.36051 * AIRCOND + 0.00011723 * CDD30YR + \\ 0.00004345 * HDD30YR + 0.00005129 * HDD30YR * AIRCOND \end{aligned} \tag{2}$$

We have determined that while the model fits data relatively well and all p-values are below threshold of 0.05 cutoff for significance, most of the predictors in this model have surprisingly low practical influence on annual energy bill. Final AIC, BIC and CV metrics for this model were 13854.78, 13943.57 and 0.42 respectively, where our best predictive model gave us 8122.132 (AIC), 10556.48 (BIC) and 0.33 (CV). On the other hand, since we have over 12,000 observations there is a chance that the test statistics could be very sensitive to minor differences, and so practical vs. statistical interpretation is used to determine if the variable is necessary or not. After comparing this model to our best predictive model we realized that, for example, number of windows, gross household income and number of televisions used are far more important predictors of the annual energy bill.

Here is how we interpret final estimates of descriptive model.

- For every 100 square foot increase of the building leads to a multiplicative change of 1.009758 with a 95% confidence range of 1.00896 to 1.01056 in Median value of total energy expenditure.

- For every 1 year increase in the age of the household leads to a multiplicative change of 0.99850911 with a 95% confidence range of 0.998026 to 0.998993 in Median value of total energy expenditure.

- When comparing people at home vs. people not at home, people not a home has a multiplicative change of 0.94246687 with a 95% confidence interval of 0.92728 to 0.95790 in Median value of total energy expenditure when compared to people at home.

- For every 1 degree increase in temperature when not at home leads to a multiplicative change of 1.0032976 with a 95% confidence range of 1.00271018 to 1.00388491 in Median value of total energy expenditure.

- For every 1 room increase in the building leads to a multiplicative change of 1.096288 with a 95% confidence interval range of 1.090605 to 1.1020015 in Median value of total energy expenditure.

- For every 1 year increase in the year the house is made leads to a multiplicative change of 0.9975 with a 95% confidence interval range of 0.99717135 to 0.99782736 in Median value of total energy expenditure.
- When comparing air conditioner used vs NO air conditioner used, people not using an air conditioner has a multiplicative change of 0.69732145 with a 95% confidence interval of 0.66346322 to 0.73290756 in Median value of total energy expenditure when compared to people using an air conditioner.

- For every 1 day increase in the average 30 year cooling degree days leads to a multiplicative change of 1.00011724 with a 95% confidence interval range of 1.00010361 to 1.00013087 in Median value of total energy expenditure.

- For every 1 day increase in the average 30 year heating degree days leads to a multiplicative change of 1.00004345 with a 95% confidence interval range of 1.00003678 to 1.00005012 in Median value of total energy expenditure.

- In the interaction of average 30 year heating degree days and air conditioner used, it shows that the effect of average 30 year heating degree days when air conditioner is not used leads to an increase multiplicative change of 1.00005129 with a 95% confidence interval of 1.00004212 to 1.00006047 in the Median value of energy expenditure when compared to average 30 year heating degree days when air conditioner is used.

Fit characteristics of the model (Figure 4) as well as Table 5 let us conclude that all variables are significant at the 99% confidence level. We can see that the residuals are randomly scattered indicating no issues. The leverage plot is of some concern at the near 0.004, however, when looking at the Cook's D there we do not see any values indicating of concern where the largest Cook's D value is 0.10. When looking at the histogram and QQ plots of the residuals, we can see that the residuals are normally distributed, the histogram show a normally distributed response and the QQ plots show majority of point falling on the 45 degree line. We have also tested for multicollinearity between indepedent variables in the descriptive model and found all variance inflation factors to be below 10.

Here are our conclusions that can be taken from the interpretive model. We can confirm the effect of air conditioning in the building on total dollar energy expenditure. From the model we can see that individuals without a central air unit pay much less compared to individual with central air by a multiplicative change of 0.69 in the median value of log total dollar energy expenditure. The next biggest effect, with a multiplicative change factor of 1.096, is the number of rooms in the household since bigger number of rooms indicates that more people live there which increases the pressure on air and heating unit.
Things to note in this model, is the factors used in this regression do not have large effects when compared to *AIRCOND* indicating that it may not be really different from zero in a practical sense. Therefore, for future analysis, other variables such as what temperature was set at home, gone, and night during the

summer time since we can see that *AIRCOND* has a significant effect should be included in the model. Other things to consider is maybe finding a better heating unit variable than the one chosen for the subset since the *EQUIPNOHEAT* has many categories which makes it hard to estimate and have little data in some categories. Another thing we have not considered is to add region and its interaction with the variables to better understand the effect of *AIRCOND* on region since each region has its own climate.

Energy cost comparison

In this section we compare total mean energy cost for different types of settlements (urban vs rural) and year of construction (in 10-year ranges). Group differencies are visualized in Figure 5. From plots you can see that there is a clear difference between average energy spendings for households in rural versus urban areas. While both lines on the left side of Figure 5 does not seem to be exactly parallel ANOVA results show that there is no iteraction between type of settlement and construction year. From Table 2 and Table 3 we conclude that:

- The levels of *Construction Year* are associated with statistically significant different energy cost (p-value < 0.00001).
- The levels of *Settlement Type* are associated with statistically significant different energy cost (p-value < 0.00001).
- The interaction between *Construction Year* and *Settlement Type* statistically insignificant, which indicates that the relationships between *Settlement Type* and energy cost does not depends on the *Construction Year* (p-value 0.13).

Model Assumptions

ANOVA assumes that errors are normally distributed and independent and that the variance across groups is homogeneous. To test that both there assumptions holds we looked at residuals versus fit and QQ plots (Figure 6). We haven't found any significant deviations from both assumptions except for one outlier household which turned out to be an error.

Code

All code used to generate models, plots and report related to this work can be found in https://github.com/JaimeVillanueva/Applied-Statistics-Project1

Figures and Tables

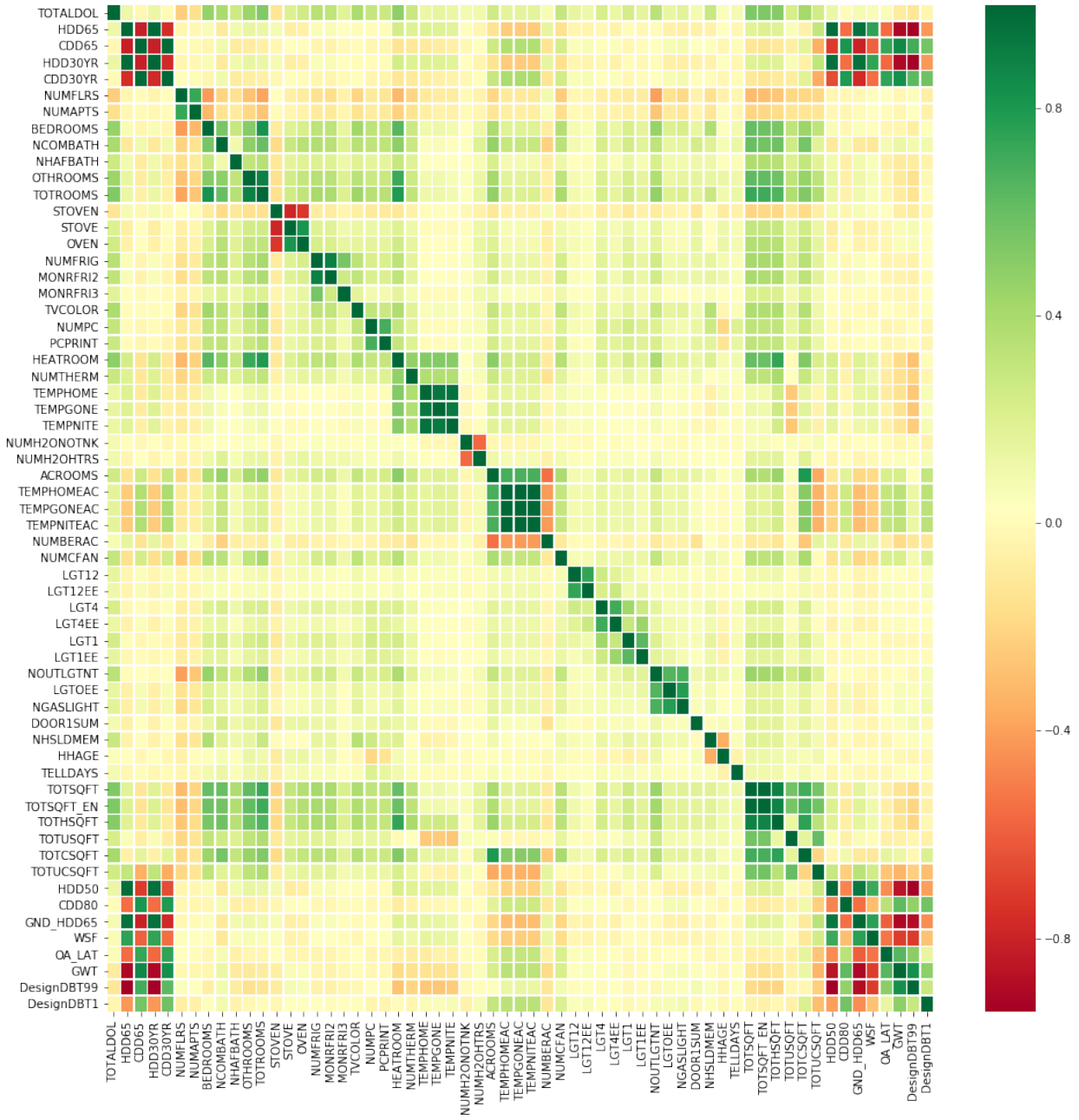

Figure 1: Correlation matrix of all quantitative household measurements, plus dependent variable.

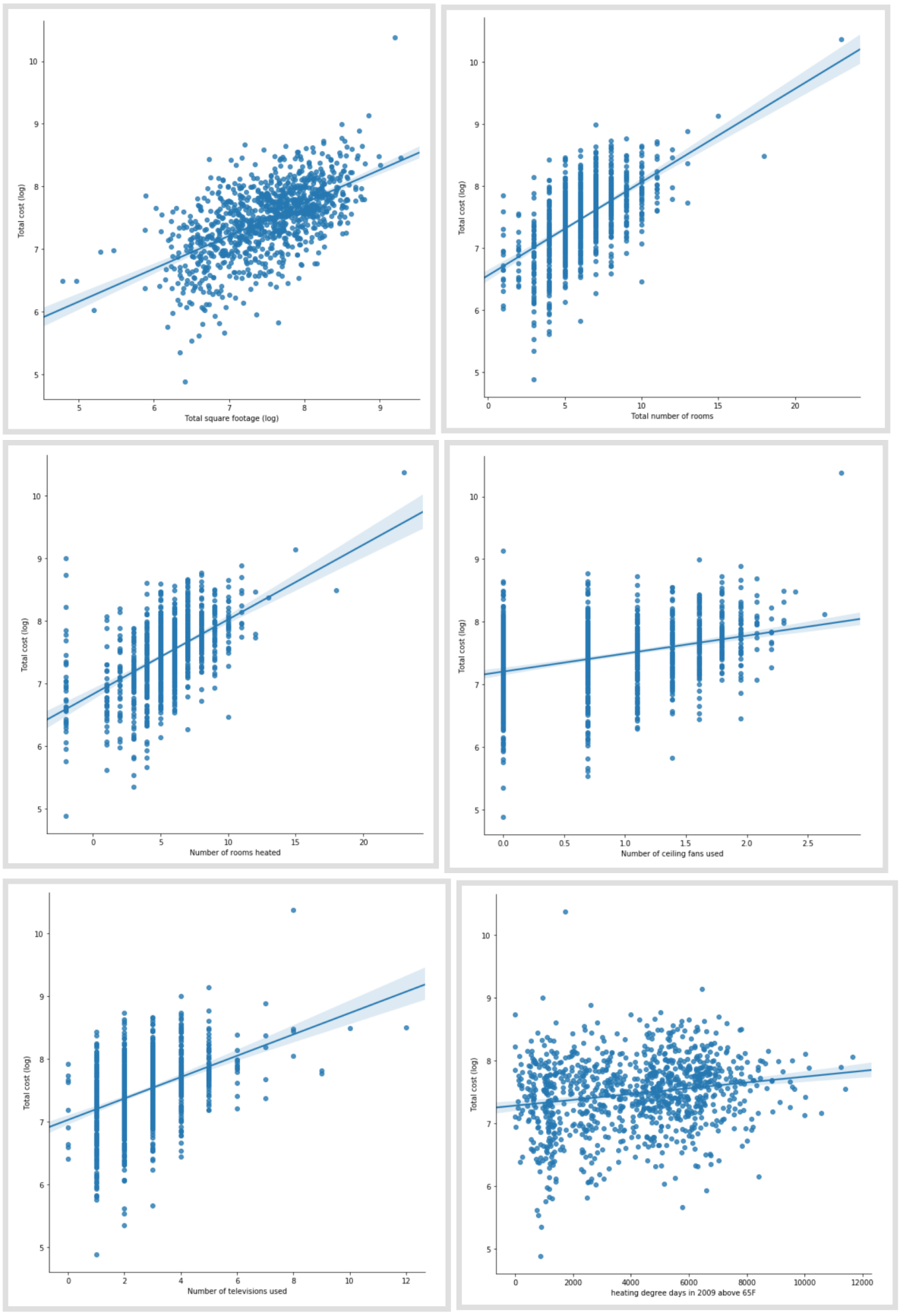

Figure 2: Scatterplots of selected predictors vs log(total cost).

Predictor	Description	Used in Best Model
TOTSQFT	total square footage	*
HEATROOM	number of rooms heated	*
TOTROOMS	total number of rooms	*
NUMCFAN	number of ceiling fans used	*
TVCOLOR	number of televisions used	*
WINDOWS	Number of windows in heated areas	*
MONEYPY	2009 gross household income	*
DIVISION	Census Division	*
YEARMADE	Year housing unit was built	*
POOL	Heated swimming pool	*
AGEHHMEMCAT2	Age category of second household member	*
STUDIO	Studio apartment	*
AGEHHMEMCAT3	Age category of third household member	*
REPORTABLE_DOMAIN	Reportable states and groups of states	*
LPGDELV	Propane delivered	*
PUGCOOK	Who pays for natural gas for cooking	*
SWIMPOOL	Swimming pool	*
USECENAC	Frequency central air conditioner used in summer 2009	*
PCSLEEP2	Sleep or standby mode for second computer when not in use	*
STORIES	Number of stories in a single-family home	*
AGEHHMEMCAT4	Age category of fourth household member	*
ROOFTYPE	Major roofing material	*
FUELPOOL	Fuel used for heating swimming pool	*
USECFAN	Frequency most-used ceiling fan used in summer 2009	*
INSTLWS	Caulking or weather stripping by this household	*
TIMEON2	Daily usage of second most-used computer	*
OVENUSE	Frequency of oven use	*
SIZRFRI2	Size of second most-used refrigerator	*
OVENFUEL	Fuel used by most-used stove	*
WHEATSIZ2	Secondary water heater size (if storage tank)	*
WASHLOAD	Frequency clothes washer used	*
REGIONC	Census Region	*
PCTYPE3	Third most-used computer - desktop or laptop	*
FUELHEAT	Main space heating fuel	*
REFRIGT2	Defrosting type of second most-used refrigerator	*
BATTOOLS	Number of rechargeable tools and appliances used	*
EDUCATION	Highest education completed by householder	*
AIA_Zone	AIA Climate Zone, based on average temperatures from 1981-2010	*
PERIODLP	Number of days covered by Energy Supplier Survey LPG	*
COMBODVR3	DVR built into the cable box or satellite box	*
TVONWE3	Third most-used TV usage on weekends	*
DRAFTY	Is home too drafty in the winter? (respondent reported)	*
DWASHUSE	Frequency dishwasher used	*
PELLIGHT	Who pays for electricity used for lighting and other appliances	*
SIZRFRI3	Size of third most-used refrigerator	*
WHEATAGE2	Secondary water heater age	*
DRYRUSE	Frequency clothes dryer used	*
AGECDRYER	Age of clothes dryer	
TYPEHUQ	Type of housing unit	

Table 1: Ordered list of variables considered for our predictive model. Predictors, selected for the best model are marked with *

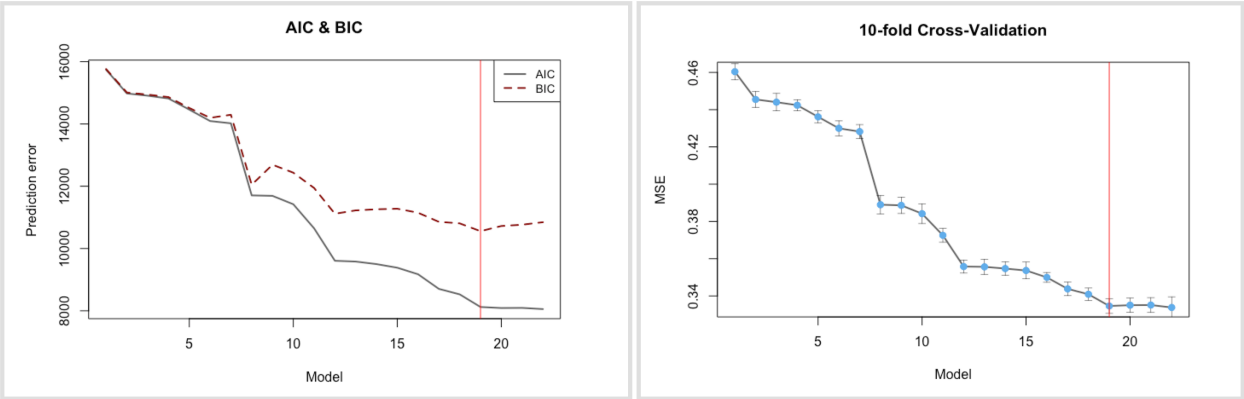

Figure 3: AIC, BIC and CV metrics for 22 predictive models. Vertical red line indicates selected model.

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	18348.9	1	58869.6379	< 0.00001
Construction Year	6.6	7	3.0346	0.003447
Settlement Type	13.5	1	43.2629	< 0.00001
Construction Year $\times$ Settlement Type	3.5	7	1.6047	0.128907

Table 2: Type III ANOVA table for the non-additive model.

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	68199	1	218730.729	< 0.00001
Construction Year	55	7	25.195	< 0.00001
Settlement Type	99	1	318.063	< 0.00001

Table 3: Type III ANOVA table for the additive model.

Predictor	Description
Sqft_100	Total square footage (in 100 sq ft increments)
ATHOME	Household member at home on typical week days
TEMPHOME	Temperature when someone is home during the day
TEMPGONE	Temperature when no on is home during the day
TEMPNIGHT	Temperature at night
YEARMADE	Year housing unit was built
AIRCOND	Air conditioning equipment used
EQUIPNOHEAT	Unused space heating equipment type
HDD30YR	Heating degree days, 30-year average 1981- 2010, base 65F
CDD30YR	Cooling degree days, 30-year average 1981- 2010, base 65F

Table 4: Variables used in descriptive model.

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Interval		VIF
Intercept	11.20483227	0.33316270	33.63	< 0.0001	10.55177990	11.85788463	0
sqft_100	0.00971098	0.00040432	24.02	< 0.0001	0.00891846	0.01050351	2.19059
HHAGE	-0.00149200	0.00024712	-6.04	< 0.0001	-0.00197640	-0.00100759	1.08281
ATHOME	-0.05925451	0.00828651	-7.15	< 0.0001	-0.07549740	-0.04301163	1.06757
TEMPGONE	0.00329194	0.00029867	11.02	< 0.0001	0.00270651	0.00387738	1.15319
TOTROOMS	0.09193033	0.00265169	34.67	< 0.0001	0.08673258	0.09712807	2.11673
YEARMADE	-0.00250383	0.00016776	-14.93	< 0.0001	-0.00283266	-0.00217500	1.09912
AIRCOND	-0.36050878	0.02539235	-14.20	< 0.0001	-0.41028186	-0.031073570	5.96546
CDD30YR	0.00011723	0.00000695	16.86	< 0.0001	0.00010360	0.00013086	3.20138
HDD30YR	0.00004345	0.00000340	12.77	< 0.0001	0.00003678	0.00005012	5.52045
HDD30YR$\times$AIRCOND	0.00005129	0.00000468	10.96	< 0.0001	0.00004212	0.00006047	8.87914

Table 5: Parameter estimates of the descriptive model.

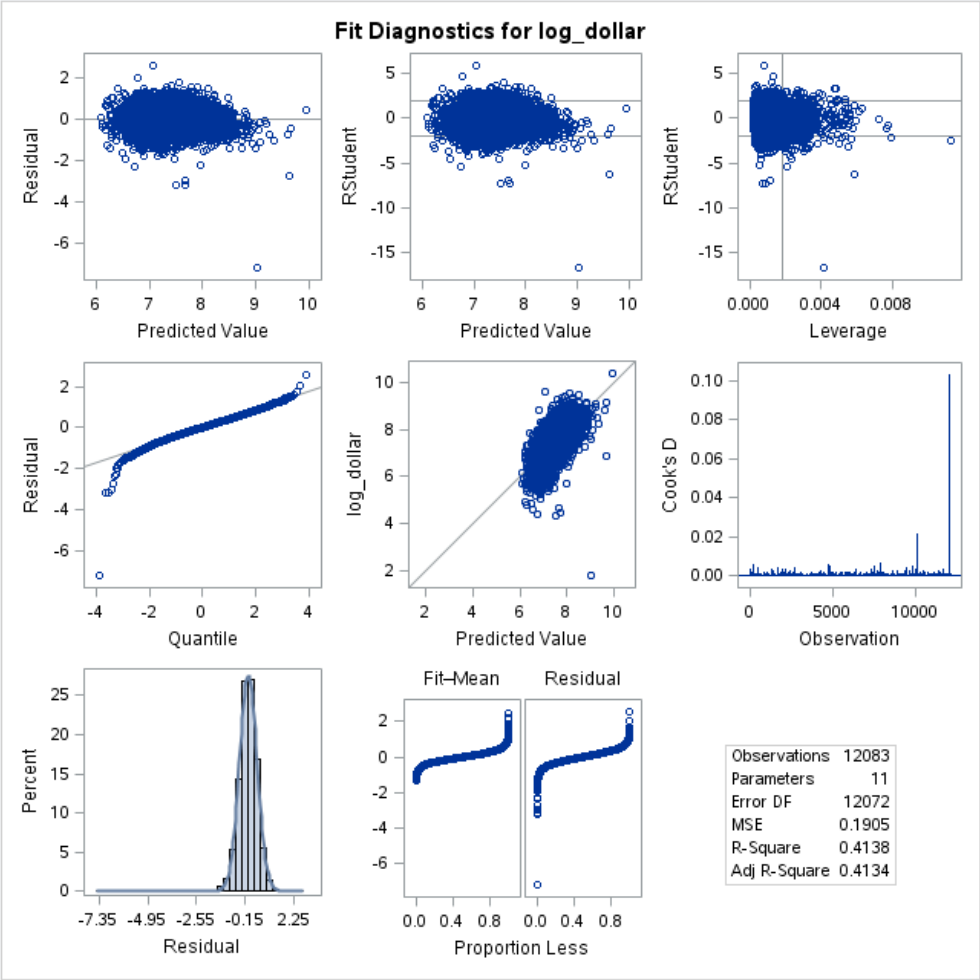

Figure 4: Fit statistics and residual plots of the descriptive model.

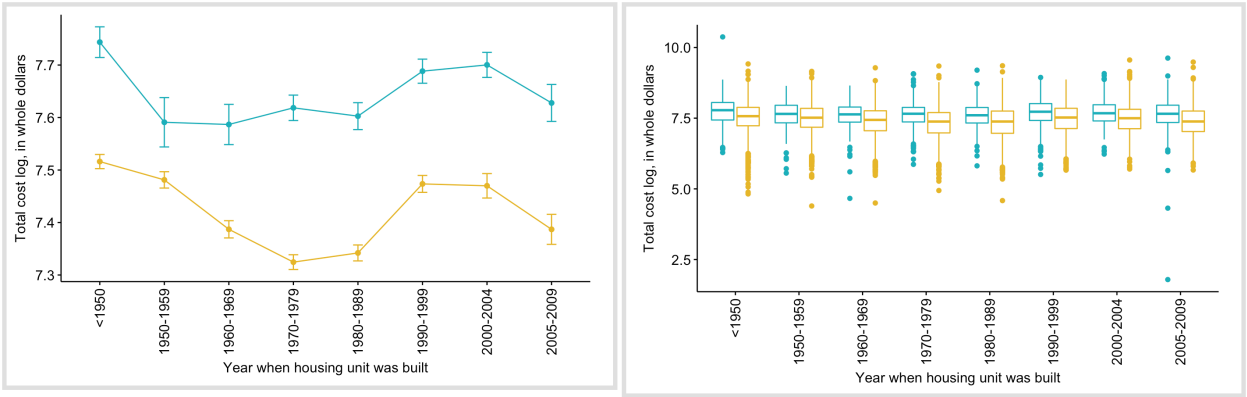

Figure 5: **On the right:** box plot of energy cost grouped by types of settlements and year of construction. **On the left:** two-way interaction plot of the response for combinations of factors. On both plots energy cost for rural areas is coded with blue color and yellow color represents energy cost for urban areas.

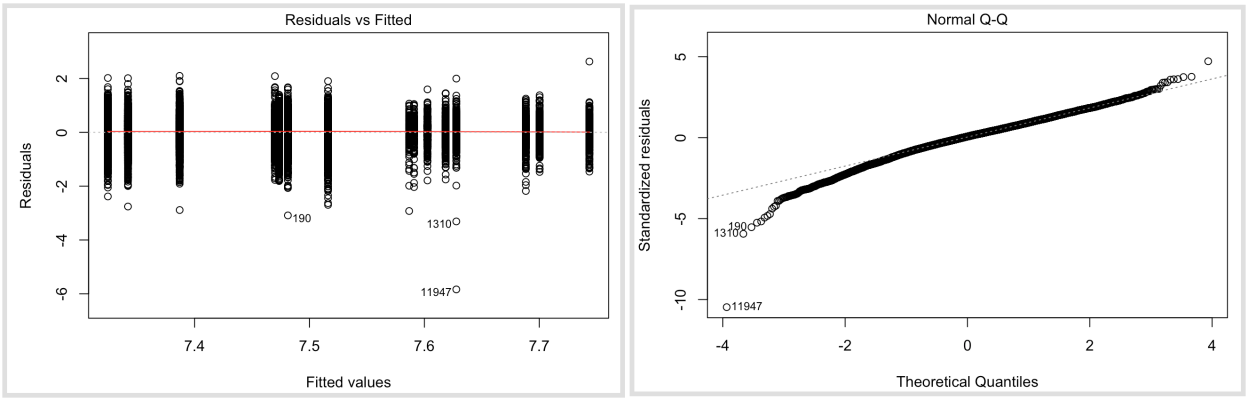

Figure 6: ANOVA QQ (right) and residual plots (left).