MSDS 6372 Project 1 Report

Volodymyr Orlov, Jaime Villanueva, Jason Lin

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 of the data and model future average yearly energy consumption, in US dollars, of a household from various household's features and consumption habit of its tenants. In addition to that we also attempted to answer following questions:

- Q1
- Q2

Data Description

(Why we've chosen 2009 when we could take 2015?!!. Lets either somehow support it or use latest data).

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:

[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

Objective 1

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:

$$y = X\beta + \epsilon$$

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.

Model Selection

Given big number of predictors it made more sense for us to have two models:

1. A model for interpretation and hypothesis testing
2. A model for high predictability

For (1) we used our intuition to choose a best subset of predictors for interpretation. For (2) we have tested multiple combinations of quantitative and categorical predictors and selected the best model using AIC, BIC and CV criteria.

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 where each subsequent model uses increasing number of predictors from this table and 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 model build from predictors marked with a star in Table 1.

Descriptive Model

The methodology used to estimate the interpretive model is a mix of manual and the different model selection techniques i.e. stepwise, forward, backward, and LASSO. Since there are many variables in the data set, as shown earlier, it may take high computational power and time. Therefore, theory and logic are used to subset the variables to what is considered to be relevant in explaining total dollar energy expenditure. The CV press is used as the optimizing criteria to determine the final model for the given model selection techniques. Many iterations of the final subset of variables were considered, table of variables is listed below in the Figures and Tables Section.

With our descriptive model, we wanted to test to see if air and heating contributes most to the total dollar energy expenditures and also what other factors contributes to air and heating usage.The logic used is that from experience and anecdotal evidence heating and cooling uses the most energy especially in Texas during the summer months. Therefore, the indicator for air conditioner (AIRCOND) and heating equipment (EQUIPNOHEAT) are included in the final variable group. Some other factors to consider is of temperature will influence the use of central air and heating. The variables heating (HDD65) and cooling (CDD65) degree days using 65 degrees as the base, and what temp is set when at home, gone, or at night during the winter are good determinants for the behavior of the person when using air and heating. Another factor to consider is the characteristics of the household and building, since these characteristics will influence the efficiency and usage of the air and heating. Therefore, the variables household age (HHAGE), Total Square footage of the building (SQFT_100=TOTSQFT/100), if the person is a home indicator (ATHOME), Number of rooms (TOTROOMS), when the house is built (YEARMADE), Total heating square footage (TOTHSQFT), Total cooling square footage (TOTCSQFT), and any interactions between the variables.

Once it was determined what model is considered of good fit for explaining the factors that affect total energy expenditures based on Adjusted-RSquare and CV press, further refinements are used to put in or take out variables that maybe relevant to the model. 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. The final estimated model that was chosen is the following:

$$logDollar = 11.20 + 0.0097 * sqft100 - 0.001492 * HHAGE - 0.05925 * ATHOME$$
$$+0.003291 * TEMPGONE + 0.09193 * TOTROOMS - 0.002504 * YEARMADE - 0.36051 * AIRCOND$$
$$+0.00011723 * CDD30YR + 0.00004345 * HDD30YR + 0.00005129 * HDD30YR * AIRCOND$$

The interpretation of the variables given in the following (Please See Figures and Tables for Original SAS Estimate):
Note since this is a log-linear relationship, the coefficients are interpreted as

$$e^{\beta}$$

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.

When looking at the fit statistics of the model (Please see the Figures and Tables Section for Model Estimates and Fit Statistics Plots), we can see that all variables are significant at the 99% confidence level. We can also see that the residuals are randomly scattered indicating no issue of correlation of residuals. When looking at the leverage and Cook's D for outliers, Leverage does show some concern at the near 0.004, however, when looking at the Cook's D there does not show of 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 both show very good indication that the residuals are normally distributed where the histogram shows a normal distribution 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.

Some conclusions that can be taken from the interpretive model are that we can see the biggest effect on total dollar energy expenditure is the use of air conditioning in the building. 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 building or house since it follows logic that the more rooms the more people that would be living there increase the probability of use of the 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 to consider is to add region and its interaction with the variables to better parse out the effect of AIRCOND usage since each region has its own climate.

Two Way ANOVA

The two way ANOVA is chosen for the analysis of seeing whether region and urban indicators influence the log total dollar expenditures.Different regions have different climates and therefore energy usage could be very different say from the North and South Region.The North may use heaters more often where as the South uses Air conditioning, which could influence how much dollar expenditure is used since each appliance may have its own energy efficiency.In concern for urban indicators, there is a possibility that living areas are smaller in urban vs. rural and so the amount of dollar expenditure to heat or cool an area is small.However, urban areas may have more energy expenditures because they are more likely to have a night scene such as New York and Las Vegas, which tend to use a lot of energy to keep the businesses running.When running the two way anova, the graph of the log_dollar and regionc and ur, we can see that there maybe interaction effect since the lines do not seem to be parallel. When looking at the TypeIII SSS for the ANOVA test, we can clearly see that the interaction effect is significant. Therefore, the interaction between Urban and Region indicators does have explanatory meaning for log total dollar expenditures.Note, we do have to keep in mind the practical vs. statistical significance since we are using over 12,000 observations. From a theory and logic stand point, it does make sense that these two variables have an interaction, since there is more concentration of urban areas near the northeast and west coast compared to the south and midwest

Figures and Tables

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 *

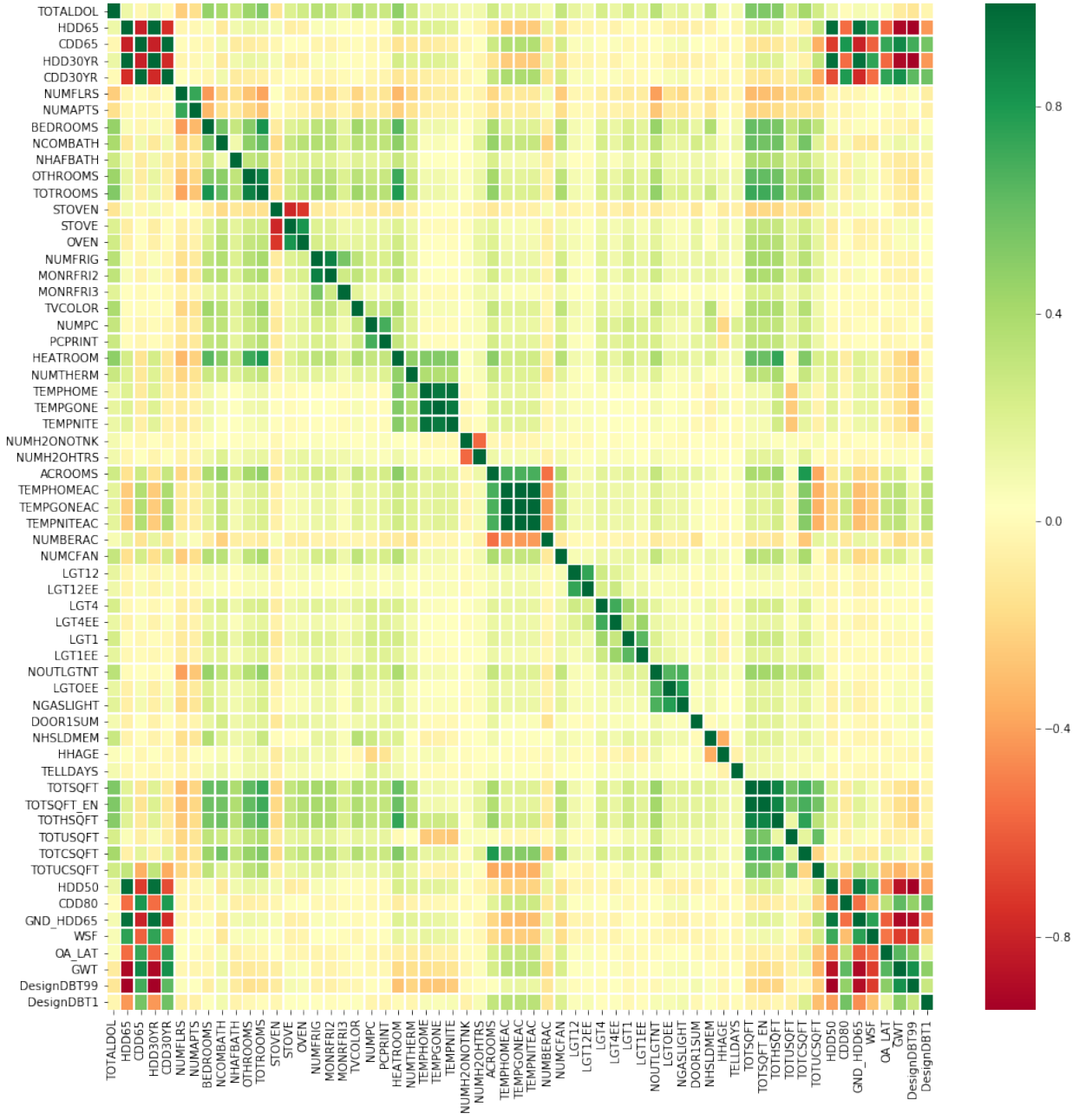

Figure 1: Correlation matrix of all quantitative household measurements, plus dependent variable.

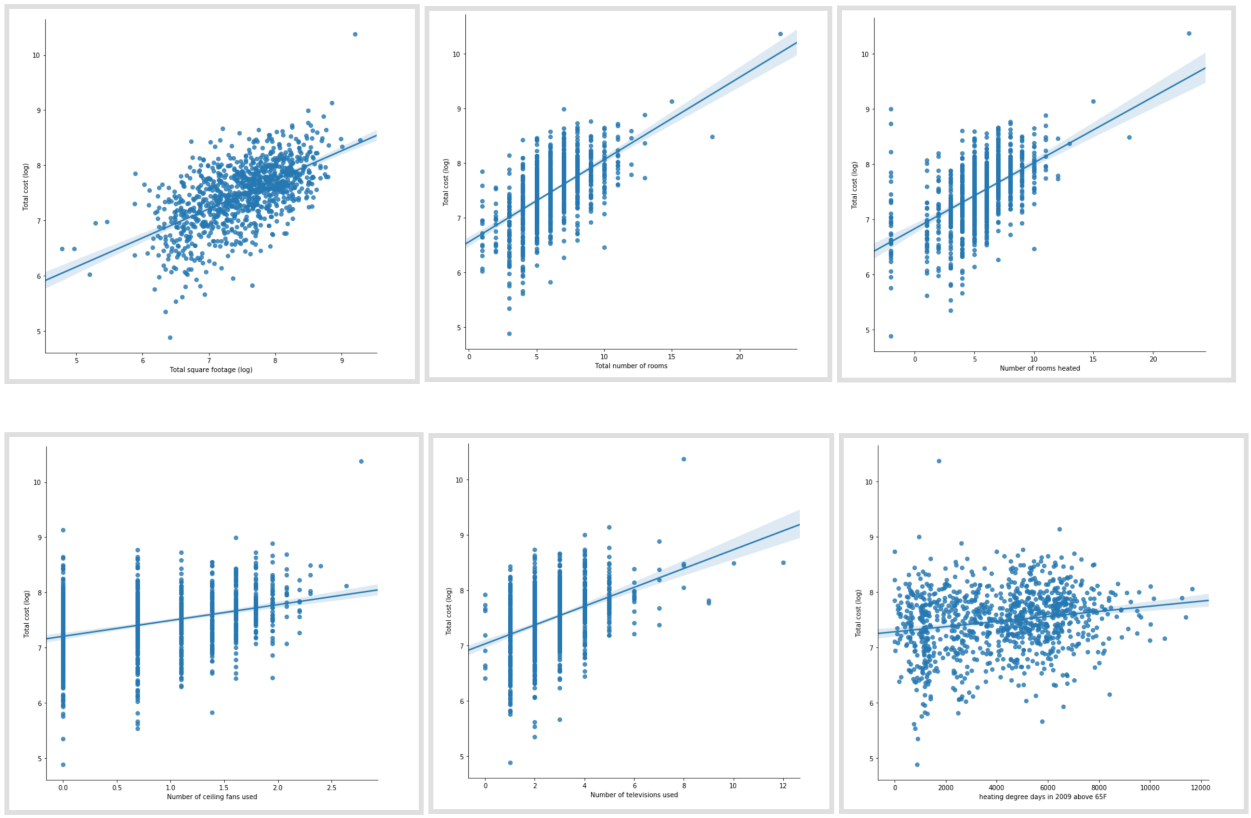

Figure 2: Scatterplots of selected predictors vs log(total cost).

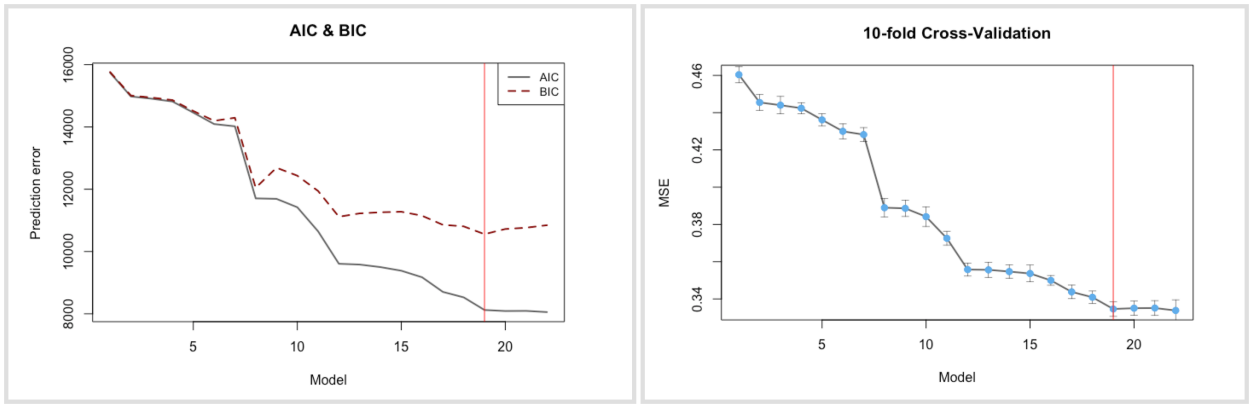

Figure 3: AIC, BIC and CV metrics for 22 predictive models. Vertical red line indicates selected model.

Variable Names
Sqft_100 = TOTSQFT /100
HHAGE
ATHOME: 0=No, 1=Yes
TEMPHOME: 0=No, 1=Yes
TEMPGONE: 0=No, 1=Yes
TEMPNITE: 0=No, 1=Yes
TOTROOMS
YEARMADE
AIRCOND: 0=No, 1=Yes
EQUIPNOHEAT
HDD65
CDD65
HDD30YR
CDD30YR
AIRCOND*HDD30YR
EQUIPNOHEAT*CDD30YR
AIRCOND*HDD65
EQUIPNOHEAT*CDD65
TOTHSQFT
TOTCSQFT
AIRCOND*TOTCSQFT
EQUIPNOHEAT*TOTHSQFT

Figure 4: Variable Subset Used in Descriptive Model

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	11.20483227	B	0.33316270	33.63	<.0001	10.55177990	11.85788463
sqft_100	0.00971098		0.00040432	24.02	<.0001	0.00891846	0.01050351
HHAGE	-0.00149200		0.00024712	-6.04	<.0001	-0.00197640	-0.00100759
ATHOME 0	-0.05925451	B	0.00828651	-7.15	<.0001	-0.07549740	-0.04301163
ATHOME 1	0.00000000	B	.	.	.	.	.
TEMPGONE	0.00329194		0.00029867	11.02	<.0001	0.00270651	0.00387738
TOTROOMS	0.09193033		0.00265169	34.67	<.0001	0.08673258	0.09712807
YEARMADE	-0.00250383		0.00016776	-14.93	<.0001	-0.00283266	-0.00217500
AIRCOND 0	-0.36050878	B	0.02539235	-14.20	<.0001	-0.41028186	-0.31073570
AIRCOND 1	0.00000000	B	.	.	.	.	.
CDD30YR	0.00011723		0.00000695	16.86	<.0001	0.00010360	0.00013086
HDD30YR	0.00004345	B	0.00000340	12.77	<.0001	0.00003678	0.00005012
HDD30YR*AIRCOND 0	0.00005129	B	0.00000468	10.96	<.0001	0.00004212	0.00006047
HDD30YR*AIRCOND 1	0.00000000	B	.	.	.	.	.

Figure 5: Parameters with 95 Percent Confidence Interval

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.78507	0.33081	32.60	<.0001	0
sqft_100	1	0.00971	0.00040432	24.02	<.0001	2.19059
HHAGE	1	-0.00149	0.00024712	-6.04	<.0001	1.08281
ATHOME	1	0.05925	0.00829	7.15	<.0001	1.06757
TEMPGONE	1	0.00329	0.00029867	11.02	<.0001	1.15319
TOTROOMS	1	0.09193	0.00265	34.67	<.0001	2.11673
YEARMADE	1	-0.00250	0.00016776	-14.93	<.0001	1.09912
AIRCOND	1	0.36051	0.02539	14.20	<.0001	5.96546
CDD30YR	1	0.00011723	0.00000695	16.86	<.0001	3.20138
HDD30YR	1	0.00009474	0.00000413	22.95	<.0001	5.52045
hdd30_aircond	1	-0.00005129	0.00000468	-10.96	<.0001	8.87914

Figure 6: Variance Inflation Factor Plots

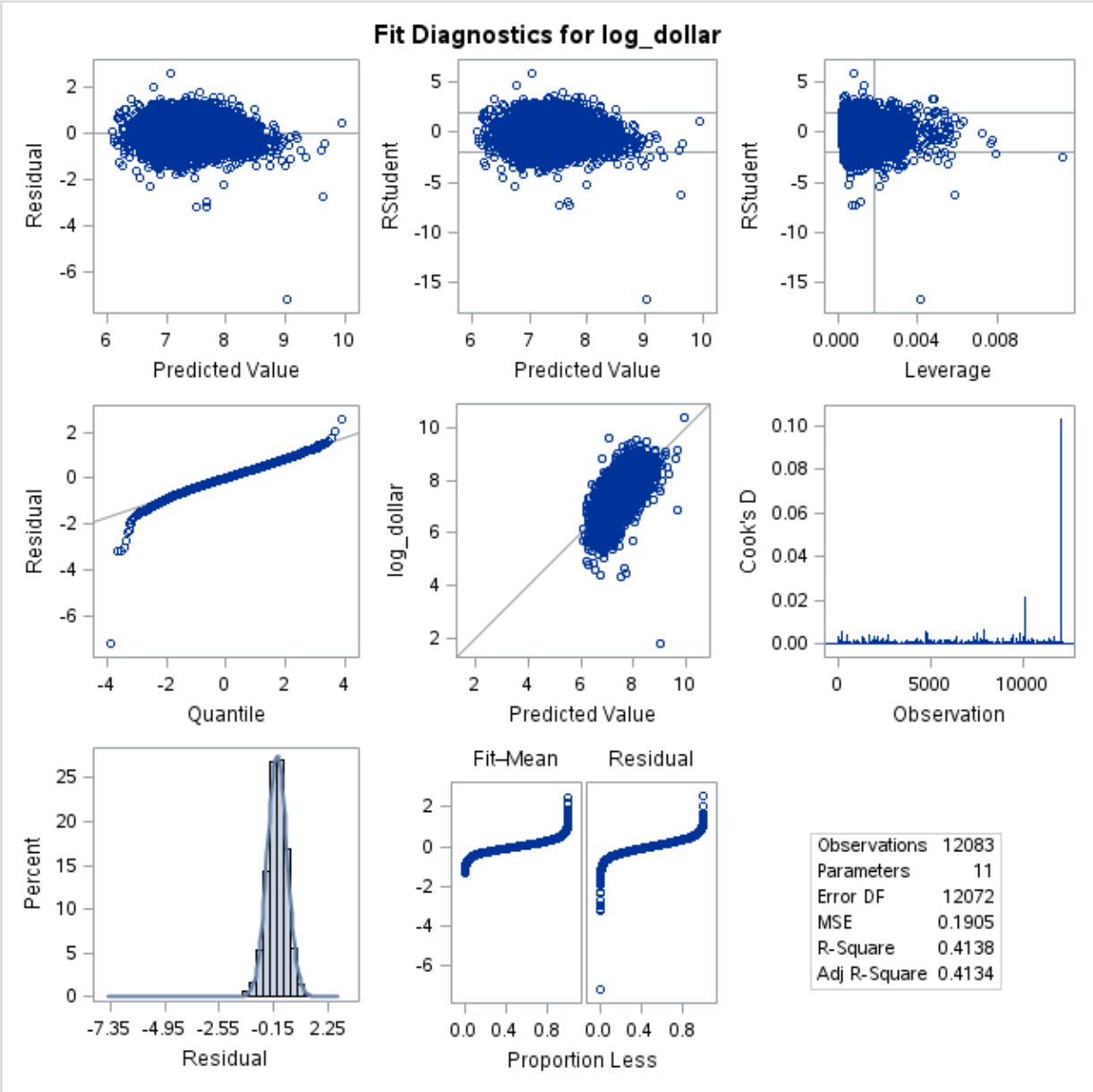

Figure 7: Fit Statistics and Residual Plots

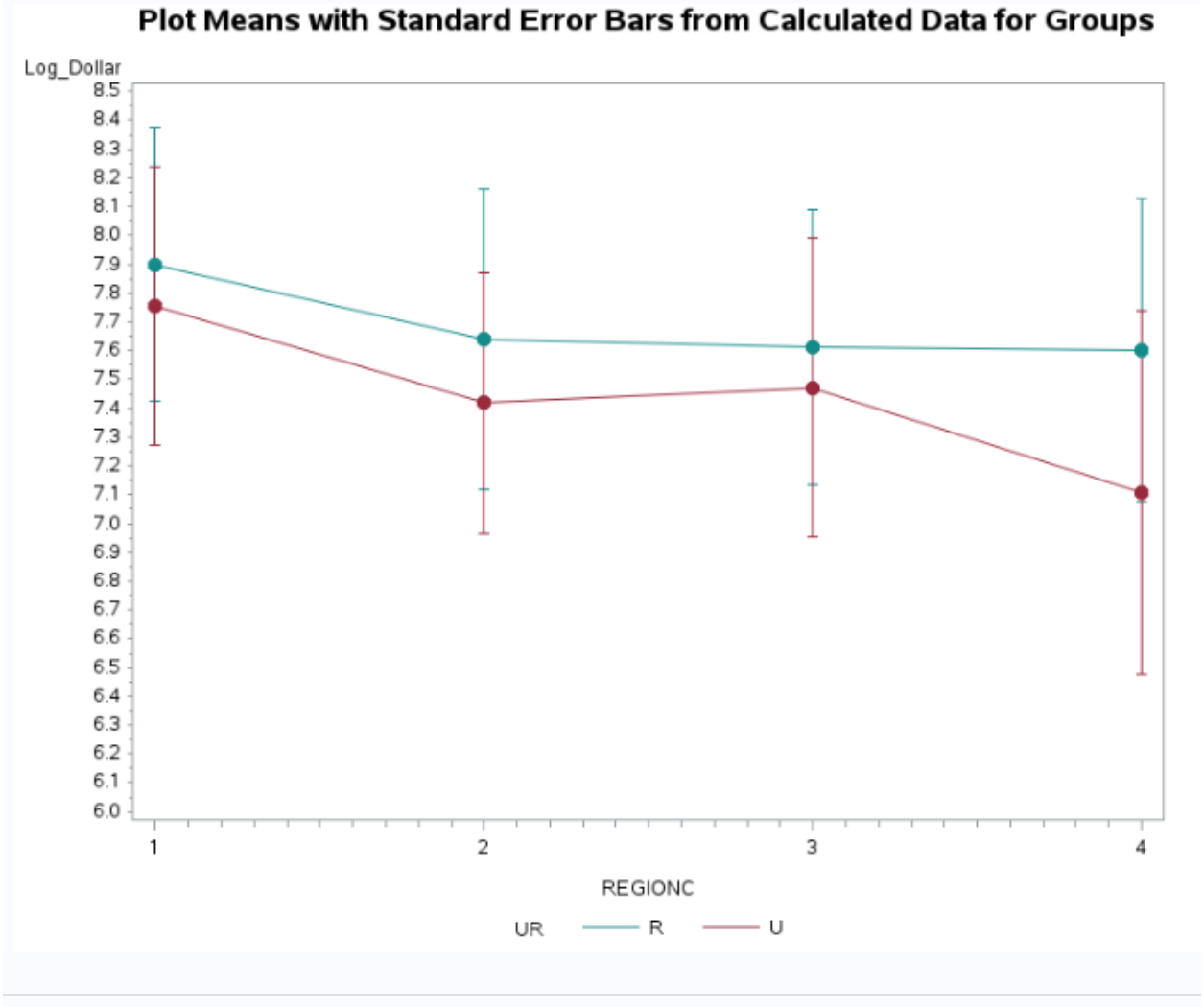

Figure 8: Plot Means with Standard Error Bars from Calculated Data for Groups

Source	DF	Type III SS	Mean Square	F Value	Pr > F
REGIONC	3	124.0722365	41.3574122	150.37	<.0001
UR	1	94.4891978	94.4891978	343.56	<.0001
REGIONC*UR	3	27.6078062	9.2026021	33.46	<.0001

Figure 9: Two Way ANOVA F Tests