This study explores the determinants of electricity consumption using household-level data from the U.S. Department of Energy 2005 Residential Energy Consumption Survey. This cross-sectional dataset of US dwellings will be analysed using OLS regression models.

1 Model construction

1.1 Variable selection and OVB

Due to the demand-oriented nature of the study, the log-log form will be used for numeric variables in all proposed models. This form is commonly used as it allows us to determine the demand elasticities for the predictor variables. Based on our review of existing literature (Alberini et al. (2011), Barnes et al. (1981), Fell et al. (2014)) as well as common intuition regarding electricity consumption drivers, the following baseline line model is proposed:

$$
\begin{aligned}
log(\text{KWH})_i &= \beta_0 + \beta_1 log(\text{Income})_i \\
&+ \beta_2 \text{ElectricityHeating}_i + \beta_3 \text{Own}_i \\
&+ \beta_4 \text{BuiltPre1970}_i + \beta_5 \text{BuiltPost2000}_i \\
&+ \beta_6 \text{Urban}_i + \beta_7 log(\text{TotalArea})_i + \varepsilon_i, \qquad (1)
\end{aligned}
$$

Where Income is treated as a continuous variable since it reports the midpoints of income groups. As a result, we can apply the log form to this variable to obtain demand elasticities.

To improve our baseline model, we introduce exogenous variables to account for different weather states' impact on electricity consumption (Model 2). Next, to control for regional differences across states, we introduce the geographical divisions (Model 3). This reduces the risk of obtaining inconsistent and biased OLS estimators due to the omitted variable bias. The results reported in Table 2 confirm expectations of OVB in the baseline model. The Wald test proves that controlling for geographical regions results in a statistically significant improvement in the fit of the model (p-value < 0.001).

Next, we add the nonlinear expressions for the heating and cooling days (Model 4) and observe that the coefficient for the square of cooling days is non-significant and exclude it from further models. Model 5 introduces interaction terms to account for the hypothesized relationships between the age of the dwelling, its total area, and the type of heating fuel used. We observe only one significant interaction variable, and so we drop the interactions between dwelling age with its total area and the usage of electricity.

Model 6 incorporates controls for appliances; the variable specifications are presented in Table 3. Finally, Model 7 combines Model 6 with the interaction term from Model 5.

$$
\begin{aligned}
log(\text{KWH})_i &= \alpha + \boldsymbol{x}_i' \boldsymbol{\gamma} + \beta_1 log(\text{Income})_i \\
&+ \beta_2 \text{ElecHeating} : \text{HeatingDays}_i^2 + \varepsilon_i, \qquad (2)
\end{aligned}
$$

where $\boldsymbol{x}_i$ is a vector of dwelling and household characteristics which influence consumption. Tables 2 and 3 presents the results of the regression for the above models. Confidence intervals for the coefficients at a 95% level of confidence are presented in Table 4.

1.2 Heteroskedasticity

The Breusch-Pagan test is used to test for the constant variance of error terms. We arrive at a p-value of 0.001 (for model 7) and thus reject the null hypothesis. As the conditional variance of the error term is not constant, robust standard errors will be reported.

1.3 Normality and collinearity

Overall, the Q-Q plot (Figure 2) shows normality with the exception of the tails, where we note a divergence from normal distribution. However, given the large number of observations, we expect a normal distribution of the error terms based on the Central Limit Theorem.

To ensure that there is no multicollinearity between our predictor variables, we run a Variance Inflation Factor test (Table 5). As no variables besides interaction or nonlinear terms have values above the threshold of 10 (Hair Jr et al., 1995), we report no collinearity.

2 Threats to internal validity

2.1 Measurement error

The data is collected through 45-minute-long interviews with survey participants (N=4382); the veracity of the information, therefore, depends on the participants' ability to provide accurate information. Due to the large size of the sample, the robustness of the collected information should not be an issue should the errors be normally distributed around the real values. It is important to consider whether there occur presistent errors in reporting data such as total square footage, the age of the house, etc, which would affect the obtained coefficients. Nonetheless, we expect minimal levels of inaccuracy in data collection considering the stringent methodology of the study.

2.2 Sample selection bias

This study is not affected by sample selection bias as the participant sample reflects similar trends to the wider US population in regard to factors such as income distribution, the rural-urban ratio, and geographical differences. However, the broad application of "rural" to housing outside of urban centres fails to account for variation within "rural" areas. This applied homogeneity of rural areas may affect perceptions of rural energy consumption.

2.3 Simultaneous causality

In this study, simultaneous causality could stem from the relationship between household income and electricity consumption. However, the income variable is the mid-point of the total combined income of the last 12 months, rather than household disposable income which would have been affected by the level of electricity consumption. As such, this threat can be ignored. Another potential source of simultaneous causality could be the effect of energy consumption on appliances, however, we assume that the adaptation in response would take place through the appliance variables, which is accounted for within the model.

3 Threats to external validity

Generally, the data serves as an appropriate reflection of the US population with regards to geographic divisions. Nevertheless, weak geographic precision in the dataset could result in a failure to account for more region-specific policies and energy provision facilities, limiting effective policy recommendations. Broader generalisations of the data are still possible when applied to the majority of countries. However, the obtained results from such generalisations will depend on a variety of economic aspects (e.g. purchasing power), structural factors (e.g. infrastructure quality), the energy sector characteristics, etc.

4 Interpretation and implications

We find significant income elasticity of 0.075 for our final model. This value is fairly consistent with what was reported by Alberini et al. (2011) who showed at the general value of 0.05, or 0.02 when controlling for dwelling characteristics. Furthermore, our result corresponds with current industry standards of 0.0059 (Fell et al., 2014). Conversely, it is significantly lower than the value of 0.2 observed by Barnes et al. (1981). This difference may be indicative of a change in customer behaviour in the US energy sector between 1981 and 2014.

Electricity heating remains the largest determinant of energy consumption (0.340). As expected, both heating and cooling days are positive and significant determinants of electricity consumption, and the magnitude of their effect increases with more exogenous variables. We also find that the square of heating days is significant, while that of cooling days is insignificant; both squared terms are very close to 0. These results correspond to those of Fell et al. (2014). We also find a significant but small interaction between the square of heating days and electric heating.

We also find that the total area of the house (0.225) and the urban-rural divide (-0.150) are significant in explaining energy consumption, where living in urban areas leads to a reduction in energy consumption. This could be explained by e.g. better electricity infrastructure. We observe that the geographical location is a significant determinant of electricity consumption due to environmental factors, infrastructure, etc.

Juxtaposing Alberini et al. (2011), we find that houses built pre-1970 consume less energy (-0.107) than those built after 2000 (-0.019). Sarkar and Directorate (2011) attribute this to the increase in the total size of dwellings as well as in the number of appliances present in households since 2000. Once we include appliances in our model, however, we observe that houses built after 2000 consume less energy (-0.061) than those built before 1970, which is explained by improved building energy efficiency (Sarkar and Directorate, 2011).

We also find that house ownership leads to increased energy consumption when appliances are not included in the model. This can be attributed to the findings of Fell et al. (2014) who point out that house owners are more likely to fully use heating than renters. However, controlling for appliances sees a lower elasticity of energy consumption (-0.49) by homeowners. This apparent OVB can be attributed to the ability of homeowners to invest in energy-efficient appliances, unlike renters. It would, therefore, be of policy interest to encourage the instalment of more energy-efficient appliances by both landlords and homeowners.

Moreover, we find that the number and type of appliances directly affect electricity consumption: electricity consumption increases by 20.9% when a household uses clothes washer, dishwashers – 9.5%, dryers - 20.7%, microwaves - 8.3%; using an additional fridge increases electricity consumption by 18.7%, while it decreases by 4.9% and 5.4% when a household owns a stove/oven and top-load washers respectively.

Considering these findings, policymakers should legislate to ensure that new houses use efficient electric heating systems and that energy-efficient measures are incorporated in the construction of houses, as indicated by the "New Green Deal". Homeowners and landlords alike should also be encouraged to invest in improving the existing housing stock. Where appliances are concerned, the relative impact of different appliances can be useful for context-specific policymaking. Moreover, new policies should encourage landlords to replace less energy-efficient appliances with more efficient alternatives.

Finally, policymakers should take the geographical regions and the rural-urban divide into account when creating regulations, as certain policies will prove more effective in the urban context due to better energy infrastructure access compared to rural areas. Similarly, with the United States encompassing vastly different environments, policies should be adjusted to reflect climate specificity rather than be applied homogenously across the country or states.

5 Further research

The lack of price information constitutes the key limitation of this study. By failing to address the impact of the variation in price elasticity on electricity consumption, this study will be subject to a degree of OVB. Prices are known to impact heating as well as the number and usage of appliances by households. This would further improve the model as it would then account for the differences in price structures across geographic divisions. Moreover, including measurements over time (i.e. panel data) would allow accounting for state and time fixed effects allowing the model to become more dynamic. Finally, more comprehensive data on the age and energy-efficiency of appliances, heating systems, and buildings would allow for improved evaluation if these require renovating or upgrading.

References

Alberini, A., Gans, W., and Velez-Lopez, D. (2011). Residential consumption of gas and electricity in the us: The role of prices and income. *Energy Economics*, 33(5):870–881.

Barnes, R., Gillingham, R., and Hagemann, R. (1981). The short-run residential demand for electricity. *The review of economics and statistics*, pages 541–552.

Fell, H., Li, S., and Paul, A. (2014). A new look at residential electricity demand using household expenditure data. *International Journal of Industrial Organization*, 33:37–47.

Hair Jr, J. F., Anderson, R. E., Tatham, R. L., and William, C. (1995). Multivariate data analysis (3rd ed). *New York: Macmillan*.

Sarkar, M. and Directorate, D. (2011). How american homes vary by the year they were built. *Housing and household economic statistics working paper: US Census Bureau*.

6 Appendix

Table 1: Summary statistics and description of variables

Variable	Description	Mean	Std.dev
KWH	Kilowatt Hours of electricity used	11326.497	7522.25
household_income	Midpoint of income group of households	47601.552	34679.47
heating_days	Number of heating days	4311.211	2180.79
cooling_days	Number of cooling days	1486.243	966.46
total_area	Total square footage of the dwelling	2284.173	1623.20
electricity_heating	Household uses electricity for heating (yes=1)	0.283	0.45
own	Household owns the dwelling (yes=1)	0.683	0.47
built_pre_1970	Dwelling built before 1970 (yes=1)	0.439	0.50
built_post_2000	Dwelling built after 2000 (yes=1)	0.075	0.26
urban	Dwelling located in an urban area (yes=1)	0.801	0.40
clothes_washer	Household uses a clothes washer (yes=1)	0.824	0.38
dish_washer	Household uses a dishwasher (yes=1)	0.566	0.50
dryer	Household uses a dryer (yes=1)	0.781	0.41
microwave	Household uses a microwave (yes=1)	0.875	0.33
n_fridges	Number of fridges in the household	1.232	0.47
stove_oven	Household uses a stove that has both burners and one or two ovens (yes=1)	0.902	0.30
topload_washer	Household uses a washing machine with load from the top (yes=1)	0.756	0.43

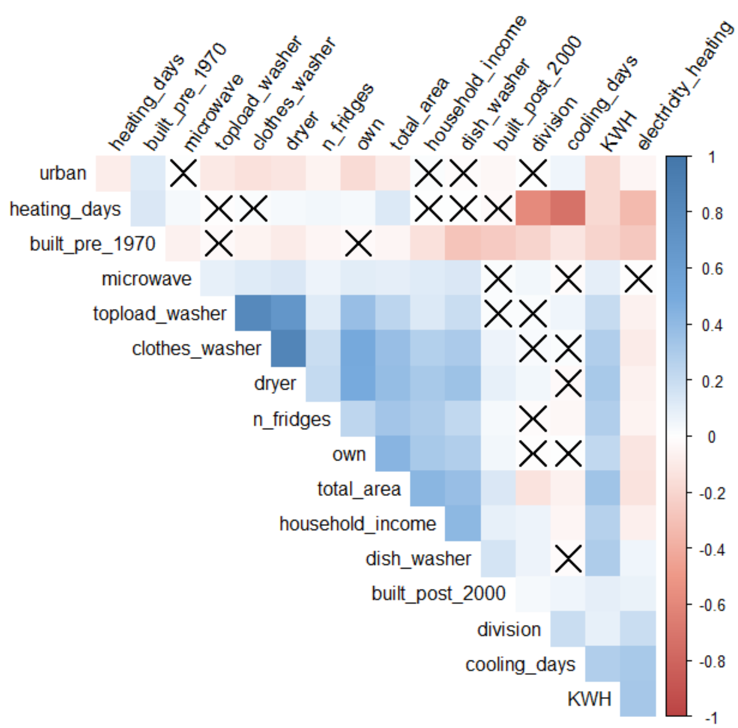

Figure 1: Correlogram - Pearson correlation coefficients, crossed out values of correlations for which the hypothesis that they are significantly different from 0 cannot be rejected. It is observed that the correlation is low between most of the variables of interest, and as none of the correlations are worrying we proceeded to the analysis

Table 2: Regression results, models 1-5

	Dependent Variable: natural logarithm of kilowatthours of electricity consumption				
	log(KWH)				
	(1)	(2)	(3)	(4)	(5)
log(household_income)	0.095[a]	0.106[a]	0.118[a]	0.121[a]	0.121[a]
	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
heating_days		0.00003[a]	0.00005[a]	0.0002[a]	0.0002[a]
		(0.00001)	(0.00001)	(0.00002)	(0.00002)
cooling_days		0.0002[a]	0.0002[a]	0.0002[a]	0.0002[a]
		(0.00001)	(0.00002)	(0.00004)	(0.00002)
electricity_heating	0.564[a]	0.477[a]	0.423[a]	0.423[a]	0.345[a]
	(0.019)	(0.020)	(0.020)	(0.020)	(0.032)
own	0.087[a]	0.062[a]	0.056[a]	0.063[a]	0.064[a]
	(0.022)	(0.022)	(0.021)	(0.021)	(0.021)
built_pre_1970	−0.107[a]	−0.098[a]	−0.072[a]	−0.076[a]	−0.126
	(0.019)	(0.018)	(0.018)	(0.018)	(0.189)
built_post_2000	−0.019	−0.045	−0.061[b]	−0.059[b]	−0.061[b]
	(0.032)	(0.029)	(0.029)	(0.029)	(0.029)
urban	−0.190[a]	−0.213[a]	−0.153[a]	−0.155[a]	−0.156[a]
	(0.022)	(0.021)	(0.021)	(0.020)	(0.020)
log(total_area)	0.355[a]	0.355[a]	0.341[a]	0.339[a]	0.338[a]
	(0.015)	(0.015)	(0.015)	(0.015)	(0.018)
heating_days^2				−0.00000[a]	−0.000[a]
				(0.000)	(0.000)
cooling_days^2				−0.000	
				(0.000)	
electricity_heating:heating_days^2					0.000[a]
					(0.000)
built_pre_1970:log(total_area)					0.006
					(0.025)
built_pre_1970:electricity_heating					0.024
					(0.041)
Constant	5.450[a]	4.960[a]	4.610[a]	4.190[a]	4.260[a]
	(0.131)	(0.130)	(0.149)	(0.158)	(0.169)
Control for divisions			✓	✓	✓
F-test (interactions of electricity_heating)					5.33[c]
					(p=0.05)
Observations	4,382	4,382	4,382	4,382	4,382
R^2	0.344	0.389	0.426	0.435	0.436
Adjusted R^2	0.343	0.388	0.424	0.432	0.434
Residual Std. Error	0.568 (df = 4374)	0.548 (df = 4372)	0.531 (df = 4364)	0.527 (df = 4362)	0.527 (df = 4360)
F Statistic	328.000[a] (df = 7; 4374)	310.000[a] (df = 9; 4372)	191.000[a] (df = 17; 4364)	177.000[a] (df = 19; 4362)	161.000[a] (df = 21; 4360)

Note: [c]p<0.1; [b]p<0.05; [a]p<0.01

Table 3: Regression results, model 6 and 7

	Dependent Variable: natural logarithm of kilowatthours of electricity consumption	
	log(KWH)	
	(6)	(7)
log(household_income)	0.074^{a}	0.074^{a}
	(0.011)	(0.011)
heating_days	0.0002^{a}	0.0002^{a}
	(0.00002)	(0.00002)
cooling_days	0.0002^{a}	0.0002^{a}
	(0.00002)	(0.00002)
electricity_heating	0.420^{a}	0.340^{a}
	(0.020)	(0.029)
own	-0.050^{b}	-0.049^{b}
	(0.021)	(0.021)
built_pre_1970	-0.041^{b}	-0.039^{b}
	(0.018)	(0.018)
built_post_2000	-0.061^{b}	-0.065^{b}
	(0.027)	(0.027)
urban	-0.148^{a}	-0.150^{a}
	(0.020)	(0.020)
log(total_area)	0.223^{a}	0.225^{a}
	(0.016)	(0.016)
heating_days^2	-0.000^{a}	-0.000^{a}
	(0.000)	(0.000)
clothes_washer	0.209^{a}	0.209^{a}
	(0.054)	(0.054)
dish_washer	0.093^{a}	0.095^{a}
	(0.019)	(0.019)
dryer	0.204^{a}	0.207^{a}
	(0.041)	(0.041)
microwave	0.082^{a}	0.083^{a}
	(0.026)	(0.026)
n_fridges	0.188^{a}	0.187^{a}
	(0.020)	(0.020)
stove_oven	-0.049^{b}	-0.049^{b}
	(0.024)	(0.024)
topload_washer	-0.051^{c}	-0.054^{c}
	(0.030)	(0.030)
electricity_heating:heating_days^2		0.000^{a}
		(0.000)
Constant	5.070^{a}	5.130^{a}
	(0.169)	(0.169)
Control for divisions	✓	✓
Observations	4,382	4,382
R^2	0.482	0.484
Adjusted R^2	0.479	0.481
Residual Std. Error	0.505 (df = 4356)	0.505 (df = 4355)
F Statistic	162.000^{a} (df = 25; 4356)	157.000^{a} (df = 26; 4355)

Note: ^{c}p<0.1; ^{b}p<0.05; ^{a}p<0.01

Table 4: Regression Results - model 7, 95% confidence interval

	Dependent variable:
	log(KWH
log(household_income)	0.074^{a} (0.053, 0.094)
heating_days	0.0002^{a} (0.0001, 0.0002)
cooling_days	0.0002^{a} (0.0002, 0.0003)
electricity_heating	0.340^{a} (0.284, 0.397)
own	-0.049^{b} (-0.090, -0.008)
built_pre_1970	-0.039^{b} (-0.074, -0.005)
built_post_2000	-0.065^{b} (-0.125, -0.006)
urban	-0.150^{a} (-0.189, -0.110)
log(total_area)	0.225^{a} (0.195, 0.254)
heating_days^2	-0.000^{a} (-0.000, -0.000)
clothes_washer	0.209^{a} (0.111, 0.308)
dish_washer	0.095^{a} (0.058, 0.132)
dryer	0.207^{a} (0.133, 0.281)
microwave	0.083^{a} (0.036, 0.129)
n_fridges	0.187^{a} (0.152, 0.222)
stove_oven	-0.049^{c} (-0.101, 0.003)
topload_washer	-0.054^{c} (-0.115, 0.007)
electricity_heating:heating_days^2	0.000^{a} (0.000, 0.000)
Constant	5.130^{a} (4.820, 5.450)
Observations	4,382
R^2	0.484
Adjusted R^2	0.481
Residual Std. Error	0.505 (df = 4355)
F Statistic	157.000^{a} (df = 26; 4355)
Note:	^{c}p<0.1; ^{b}p<0.05; ^{a}p<0.01

Table 5: VIF analysis of model 7

vif	
log(household_income)	1.5
heating_days	37.1
cooling_days	5.7
electricity_heating	2.9
own	1.6
built_pre_1970	1.3
built_post_2000	1.1
urban	1.1
log(total_area)	2.0
factor(division)	8.5
heating_days^2	20.9
clothes_washer	6.3
dish_washer	1.5
dryer	4.2
microwave	1.1
n_fridges	1.2
stove_oven	1.1
topload_washer	3.1
electricity_heating:heating_days^2	2.5

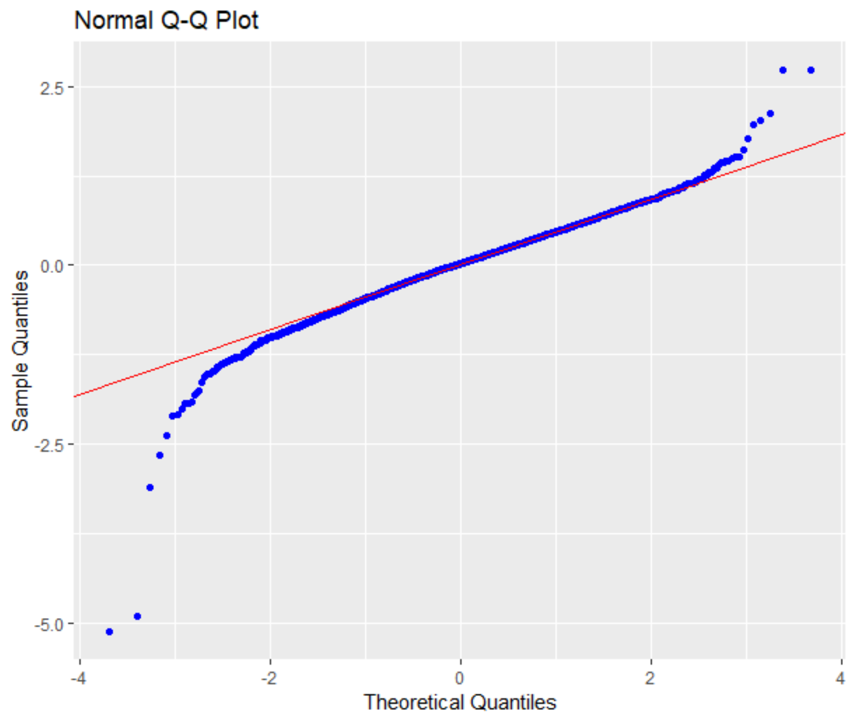

Figure 2: QQplot of residuals from model 7

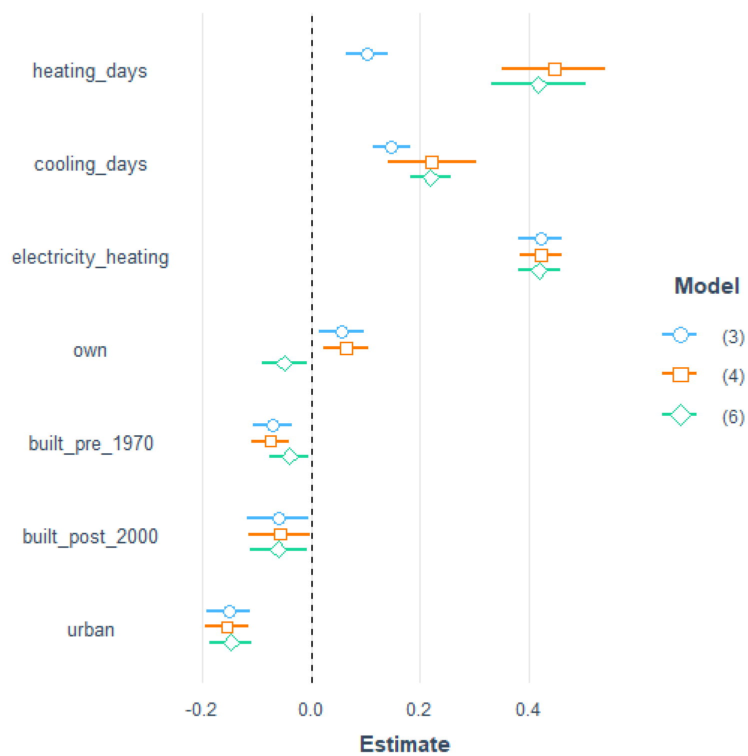

Figure 3: Comparison of coefficients of selected dummies and models without interaction terms, reported together with their 95% confidence intervals