1 Predicting electric energy consumption in the USA

1.1 Problem statement

An energy company wants to develop a model to predict annual electric energy consumption (in dollars) by households in the United States. Energy consumption change depending on a bundle of attributes, such as housing unit attributes (e.g. *BEDROOMS, TOTSQFT_EN, TYPEHUQ, SWIMPOOL)*, location (e.g. *REGIONC, BA_climate*), demographics (e.g. *INCOME, NHSLDMEM*) and consumption habits (e.g. *CWASHER, DISHWASH, ELCOOL, ELWARM, ELWATER, TVCOLOR*). Therefore, for a given household defined on a set of attributes, there is a (conditional) distribution of energy consumption. The management is interested in a prediction interval, defining this range as the difference between the 97.5^{th} and 2.5^{th} quantile of the conditional distribution of interest.

1.2 Dataset

The model is built using the "Residencial Energy Consumption Survey" (RECS) from year 2020 (U.S Energy Information Administration, 2020), which is an American sample poll that collects information related to energy consumption, expenditure and characteristics of housing units occupied as a primary residence and the households that live in them. The dataset is composed by 17.332 observations and 15 variables[1].

Table 1 Description of the energy consumption dataset

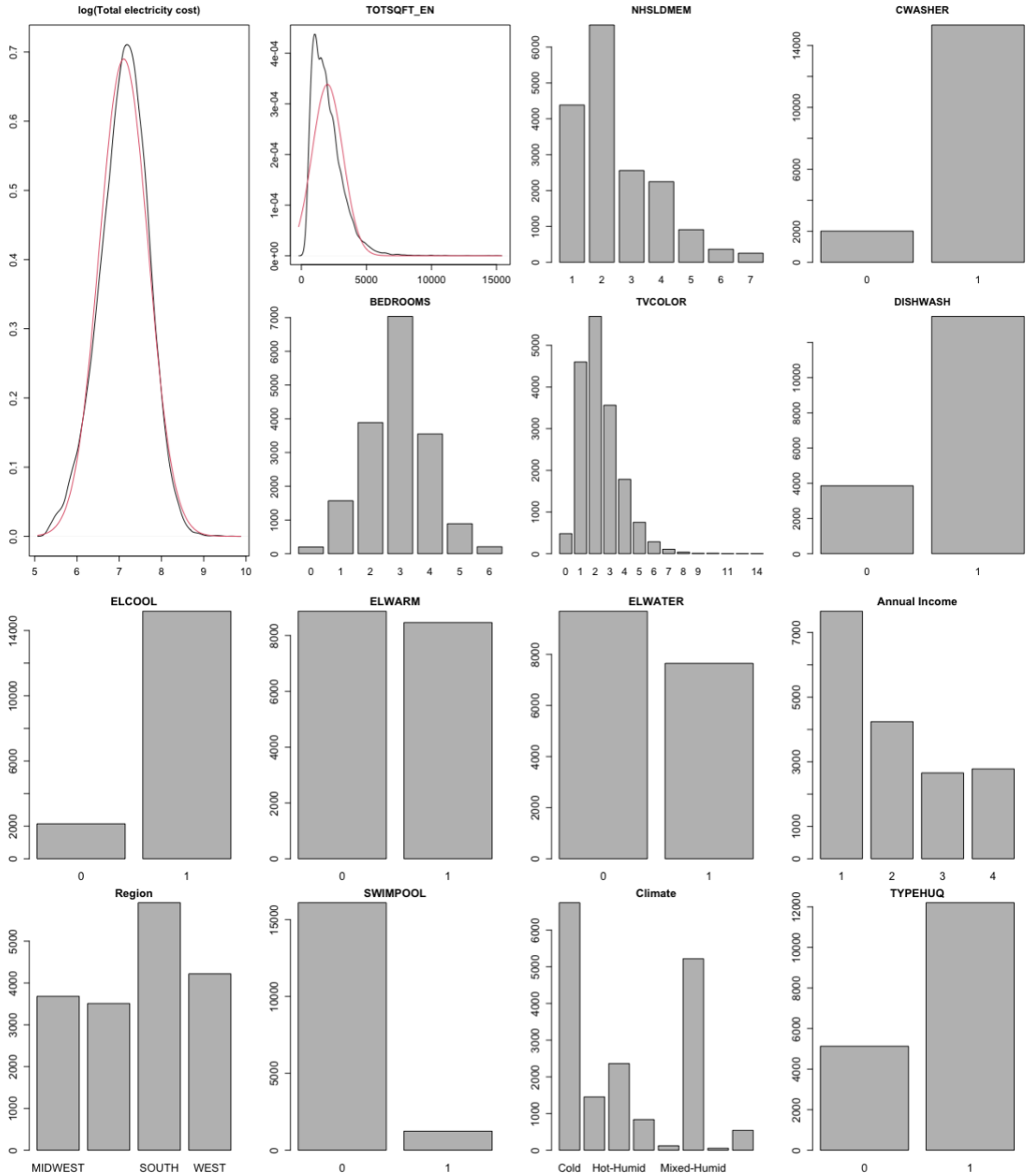
Variable	Description	Type	Outcome
DOLLAREL	Total electricity cost for year 2020 (in USD)	Real	
BA_climate	Building America Climate Zone	Categorical	8 levels: Cold / Hot-Dry / Hot-Humid / Marine / Mixed-Dry / Mixed-Humid / Subarctic / Very-Cold
BEDROOMS	Number of bedrooms	Integer	
CWASHER	Has a clothes washer in home?	Binary	2 levels: 0=No / 1=Yes
DISHWASH	Has a dishwasher?	Binary	2 levels: 0=No / 1=Yes
ELCOOL	Electricity used for air conditioning?	Binary	2 levels: 0=No / 1=Yes
ELWARM	Electricity used for space heating?	Binary	2 levels: 0=No / 1=Yes
ELWATER	Electricity used for water heating?	Binary	2 levels: 0=No / 1=Yes
INCOME	Annual gross household income for the past year (in USD)	Categorical	4 levels: 1= ($0 - $60.000) 2= [$60.000 – $100.000) 3=[$100.000 – $150.000) 4=[$150.000 - ∞)
NHSLDMEM	Number of household members	Integer	
REGIONC	Census Region	Categorical	4 levels: Midwest / Northeast / South / West
SWIMPOOL_	Has a swimming pool?	Binary	2 levels: 0=No / 1=Yes
TOTSQFT_EN	Total energy-consuming area (square footage) of the housing unit.	Real	
TVCOLOR	Number of televisions owned	Integer	
TYPEHUQ_	Is it a Single-family house detached from any other house?	Binary	2 levels: 0=No / 1=Yes

An exploratory data analysis (EDA) is performed on the whole dataset and draw some considerations to be considered in the empirical analysis.

[1] The original dataset is composed by more than 200 variables. For this project, 14 predictors that seemed to be more related with the purpose of this project were taken.

Figure 1 reports the marginal distributions of the variables included in the dataset. Most important, we can note that the DOLLARREL variable displays a positive skewness, asking for caution when using a linear regression model to predict the quantiles.

Figure 1 *Marginal distribution of the variables. A barplot is used for binary/categorical variables while a kernel density (black line) to show continous variables. A normal density (red line) matching the mean and variance of corresponding continous variable is added to help comparison.*



We explore the relationship between each explanatory variable and the dependent variable (DOLLAREL). Figure 2 shows that continuous explanatory variables have a non-linear behavior with the dependent variable. Figure 3 studies the relationship between categorical variables and energy consumption (DOLLAREL).

Figure 2 *Scatterplot of the explanatory variable and DOLLAREL (upper panel). Scatterplot of the fitted values vs residuals from a linear regression of DOLLAREL on the explanatory variable (middel panel). Smoothed regression line in red.*

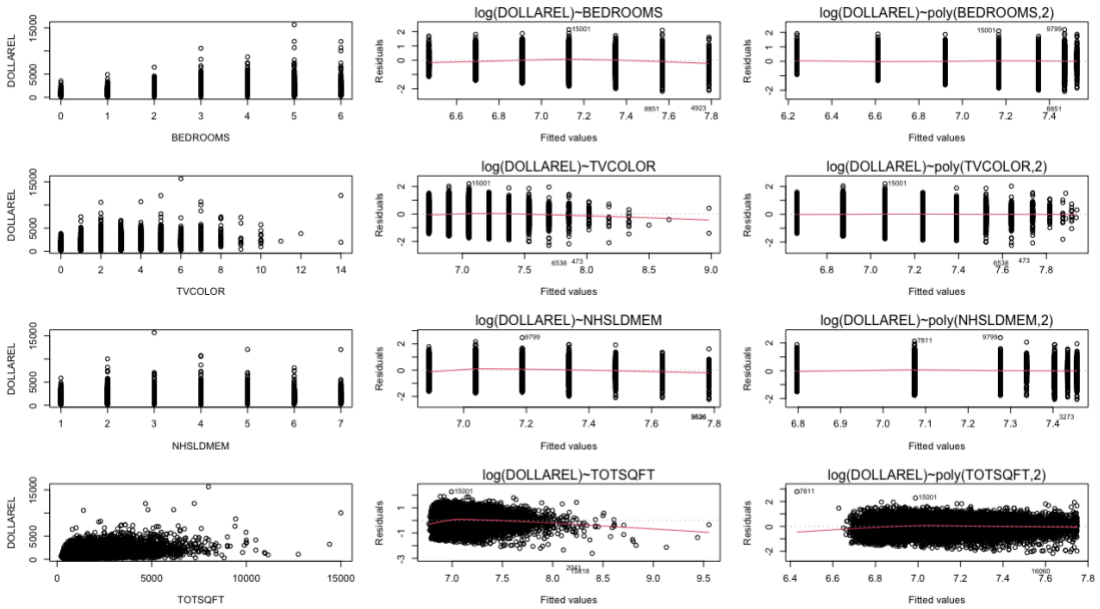


Figure 3 *Boxplot of the energy consumption for each outcome of the categorical explanatory variables.*

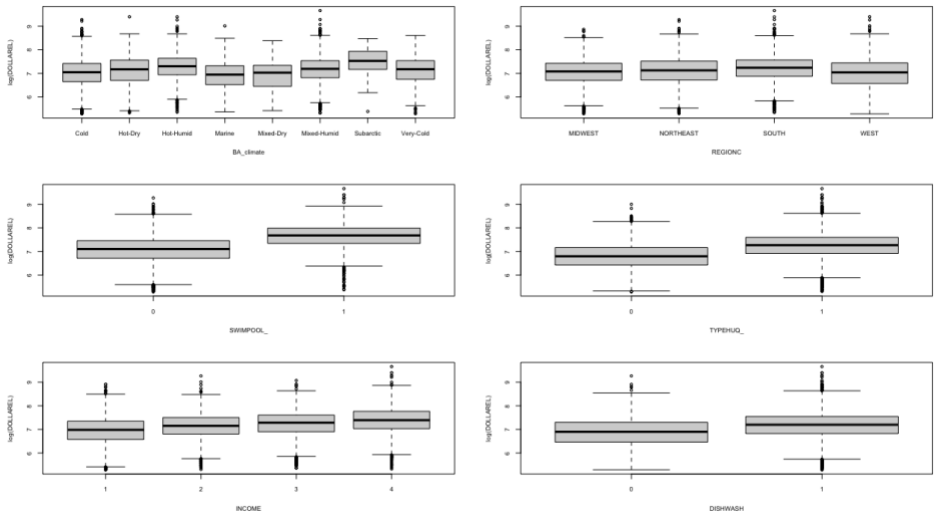
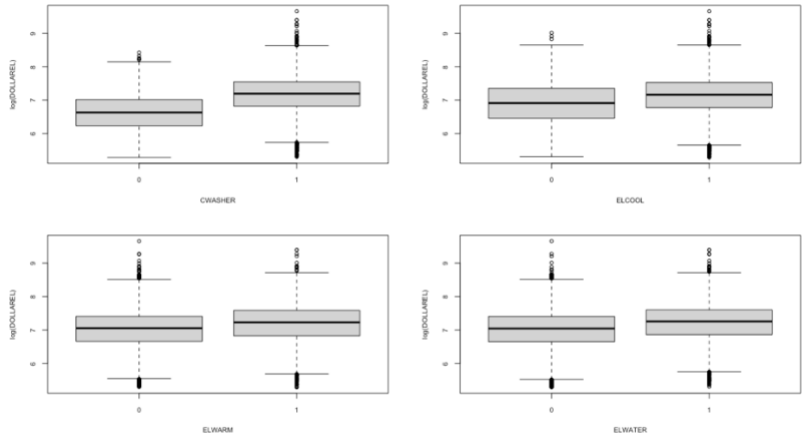


Figure 3 *Boxplot of the energy consumption for each outcome of the categorical explanatory variables.*



1.3 Method

The aim of the predictive model is to return a prediction interval, i.e. a prediction of the energy consumption range for a household of interest. This task demands an accurate predictions of the 2.5th and 97.5th conditional quantiles. Let Y_i be the price of the *i*th transaction and X_i a vector of covariates, we consider the following models:

1.3.1 Gaussian linear regression model.

$$Y_i = X_i' \beta + \boldsymbol{\sigma}\epsilon_i, \tag{1}$$

with $E(\epsilon_i) = 0, V(\epsilon_i) = 1$, estimated by ordinary least squares (OLS). Then, the τth quantile is computed as:

$$\hat{q}_{\tau,i}^{LR} = X_i'\hat{\beta} + \hat{\sigma}\Phi^{-1}(\tau) \tag{2}$$

where Φ^{-1} is the quantile function of a standard normal distribution.

1.3.2 Quantile regression model.

$$\boldsymbol{Y_i} = \boldsymbol{X_i'}\,\boldsymbol{\beta_\tau} + \boldsymbol{\varepsilon_i}, \tag{3}$$

With $Q_\tau(\epsilon_i) = 0$, estimated with the quantile regression estimator minimizing the check loss function. Then, the τth quantile is computed as:

$$\hat{\boldsymbol{q}}_{\boldsymbol{\tau},\boldsymbol{i}}^{\boldsymbol{QR}} = \boldsymbol{X_i'}\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} \tag{4}$$

1.4 Empirical Analysis

We split the available sample in an in-sample, used to estimate the models, and an out-of-sample, used to evaluate the predictive accuracy of both models. Following the discussion of the EDA, the analyst decided to use the following predictors for the model specification:

$$\begin{aligned} X_i = \{&\text{BEDROOMS}, \text{BEDROOMS}^2, \text{TVCOLOR}, \text{TVCOLOR}^2, \text{NHSLDMEM}, \text{NHSLDMEM}^2, \\ &\text{TOTSQFT\_EN}, \text{TOTSQFT\_EN}^2, SWIMPOOL, INCOME, \text{TYPEHUQ}, \text{DISHWASH}, \\ &CWASHER, \text{ELWARM}, \text{REGIONC}, \text{BA\_climate}\ \} \end{aligned} \tag{5}$$

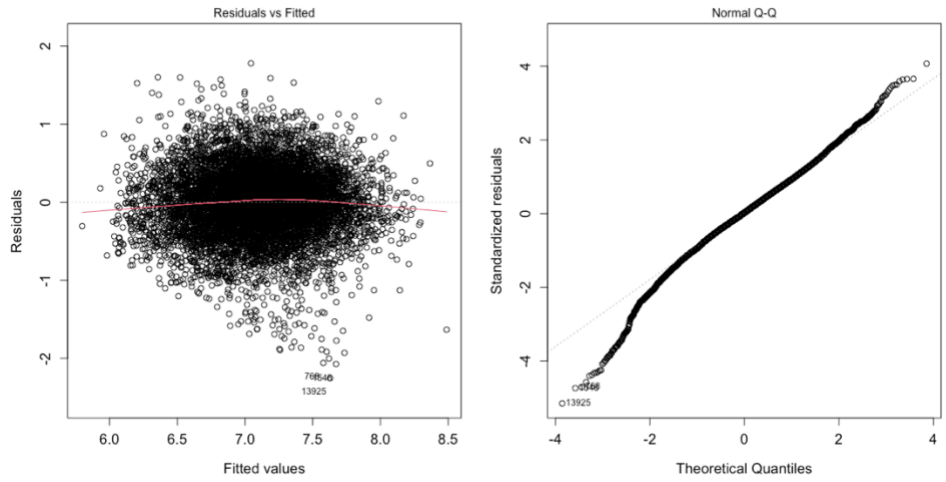
A log transformation was also required for variable DOLLAREL in order to stabilize the variance of the residuals (heteroscedasticity).

Table 2 reports the estimates for the linear regression and quantile regression at the 2.5th and 97.5th quantiles using the predictors stated in (5). In the linear regression, all covariates have a statistically significant relationship with DOLLAREL. Reviewing Figure 4, on the left the Scatterplot of the fitted values and residuals, it seems in general that there is a good adjustment for a linear regression model as there is not dependance left in the residuals to be exploited to improve the fit. On the right, Figure 4 shows a normal QQ-plot of the residuals., which displays a light left-tail.

Table 2 In-sample parameter estimates (Est.) and standard errors (Std.err.) for the linear regression (OLS) and quantile regressions at the 2.5th quantile (LQR) and at the 97.5th quantile (UQR).

	Linear Regression		0.025 Quantile Regression		0.975 Quantile Regression	
Variable	Estimate	Std.Error.	Estimate	Std.Error.	Estimate	Std.Error.
(Intercept)	6,471	0,026	5,686	0,035	7,398	0,044
poly(BEDROOMS, 2)1	4,713	0,943	0,077	2,033	7,323	1,398
poly(BEDROOMS, 2)2	-2,554	0,706	-3,439	1,235	-0,678	0,968
poly(TVCOLOR, 2)1	11,714	0,716	16,643	2,498	9,242	1,898
poly(TVCOLOR, 2)2	-2,361	0,667	-3,494	3,041	-1,088	2,390
poly(NHSLDMEM, 2)1	14,270	0,715	12,106	2,866	12,112	1,037
poly(NHSLDMEM, 2)2	-4,644	0,657	-4,093	2,487	-1,789	0,850
poly(TOTSQFT_EN, 2)1	12,208	0,882	10,788	1,984	14,221	1,486
poly(TOTSQFT_EN, 2)2	-2,654	0,714	-5,743	2,706	-2,998	1,775
SWIMPOOL_1	0,240	0,019	0,148	0,078	0,218	0,020
INCOME2	0,039	0,012	0,056	0,026	0,004	0,021
INCOME3	0,050	0,015	0,042	0,049	0,023	0,036
INCOME4	0,087	0,016	0,019	0,033	0,092	0,030
TYPEHUQ_1	0,175	0,014	0,158	0,028	0,097	0,028
DISHWASH1	0,045	0,013	0,079	0,022	-0,045	0,022
CWASHER1	0,125	0,018	0,156	0,025	0,178	0,030
ELWARM1	0,127	0,010	0,129	0,020	0,175	0,019
ELCOOL1	0,077	0,016	0,103	0,023	0,002	0,025
ELWATER1	0,237	0,011	0,169	0,024	0,183	0,020
REGIONCNORTHEAST	0,094	0,015	-0,067	0,032	0,224	0,028
REGIONCSOUTH	-0,087	0,019	0,047	0,048	-0,165	0,036
REGIONCWEST	-0,087	0,018	-0,309	0,036	0,118	0,043
BA_climateHot-Dry	0,221	0,022	-0,024	0,132	0,114	0,043
BA_climateHot-Humid	0,260	0,021	0,024	0,070	0,345	0,043
BA_climateMarine	0,034	0,027	0,180	0,071	-0,143	0,068
BA_climateMixed-Dry	0,094	0,057	0,148	0,065	-0,038	0,040
BA_climateMixed-Humid	0,160	0,016	0,027	0,046	0,179	0,021
BA_climateSubarctic	0,659	0,086	0,150	0,090	0,347	0,077
BA_climateVery-Cold	0,190	0,029	0,018	0,031	0,171	0,048

Figure 4 *Scatterplot of the fitted values and residuals (Left) and normal QQ-plot of the residuals (Right) from the linear regression model (OLS) with covariates in (5).*



In order to evaluate the accuracy prediction of both models, a check in the unconditional coverage of predictions was done. There is an indicator variable define as

$$I_{\mathrm{i}} = 1\,(Y_i < q_{\tau,i}), \tag{6}$$

if $q_{\tau,i}$ is the τth conditional quantile of Y_i given X_i, then $I_i \sim Be(\tau)$. To test the unconditional coverage, the analyst performs a likelihood ratio test on the hypothesis:

$$\begin{aligned} \mathrm{H_0} &= \mathrm{E(I_i)} = \tau \\ \mathrm{H_A} &= \mathrm{E(I_i)} \neq \tau \end{aligned} \tag{7}$$

Table 3 reports the number of out-of-sample violations, $V = \sum_{i=1}^{n.oos} I_i$, and the corresponding p-values. Predictions obtained from the quantile regression outperform those of the Gaussian linear regression model. Moreover, the hypothesis test of accurate unconditional coverage is rejected under a confidence interval 95%, only for the case of lower quantile of the linear regression model (2.5^{th}).

Table 3 *Number of out-of-sample violations and the p-value from the likelihood ratio test stated in* ***(7)****. The number of expected violations under the null hypothesis is* $\boldsymbol{V_0 = 217}$

	Linear Regression		**Quantile Regression**	
	q 0.025	**q 0.975**	**q 0.025**	**q 0.975**
V	252	196	216	218
P-value	0,018	0,149	0,964	0,926

2 Bibliography

U.S Energy Information Administration. (2020). Retrieved June 15, 2023, from Residential Energy Consumption Survey (RECS): https://www.eia.gov/consumption/residential/data/2020/index.php?view=microdata