STAT 205 Final Project

University of California, Irvine

April 8, 2019

1

1.A

We will be fitting a linear regression model for daily energy consumption on November 15, 2009 in kWh (V1) using a subset of the six household covariates as predictors. The response is assumed to be normal with a population level precision parameter - we do not yet model the individual household precision because we're only focusing on one day of energy consumption. We plan on using a g-prior for β_j, which we will justify in part 1.E. The household covariates are treated as continuous values rather than categorical values, as this results in simpler models due to less parameters being fit. It is worth noting that this does impose a linear relationship among the levels of the predictors; however, we do not believe that we will suffer a major performance degradation due to our findings in part 1.B. Furthermore, all covariates used will be centered prior to fitting. This is done to gain leverage in parameter interpretability while modeling in problems 3 and 4.

1.B

Figure 1 showcases violin plots of energy consumption on November 15, 2009 vs each predictor. This shows us the trends in energy consumption across predictors as well as the densities at each level. We can see that for number of bedrooms there is an increasing trend in energy consumption, with more spread for households that have more bedrooms. Surprisingly, those with the strongest attitudes about the environment and reducing the bill have higher energy consumption, albeit with higher variability. This trend could potentially be explained when we control for confounders in the regression model. Interestingly the median energy consumption is highest for households with 1 resident, with a distinct tri-modal distribution. Lastly, energy consumption does not seem to be affected by the head of the household's age.

We can see from the correlation plot in Figure 2 that energy consumption on November 15, 2009 (V1) has the highest correlation with number of bedrooms. Surprisingly, the number of bedrooms is not highly correlated with the number of residents - perhaps due to wealthy households having many empty bedrooms. As we would expect, attitude about environment is correlated with attitude about reducing the bill, reflecting the fact that both of these attitudes would likely lead to decreased energy consumption. We can also see that age of head of household is not correlated with energy consumption on November 15, 2009, suggesting that age may not be a helpful predictor for energy consumption.

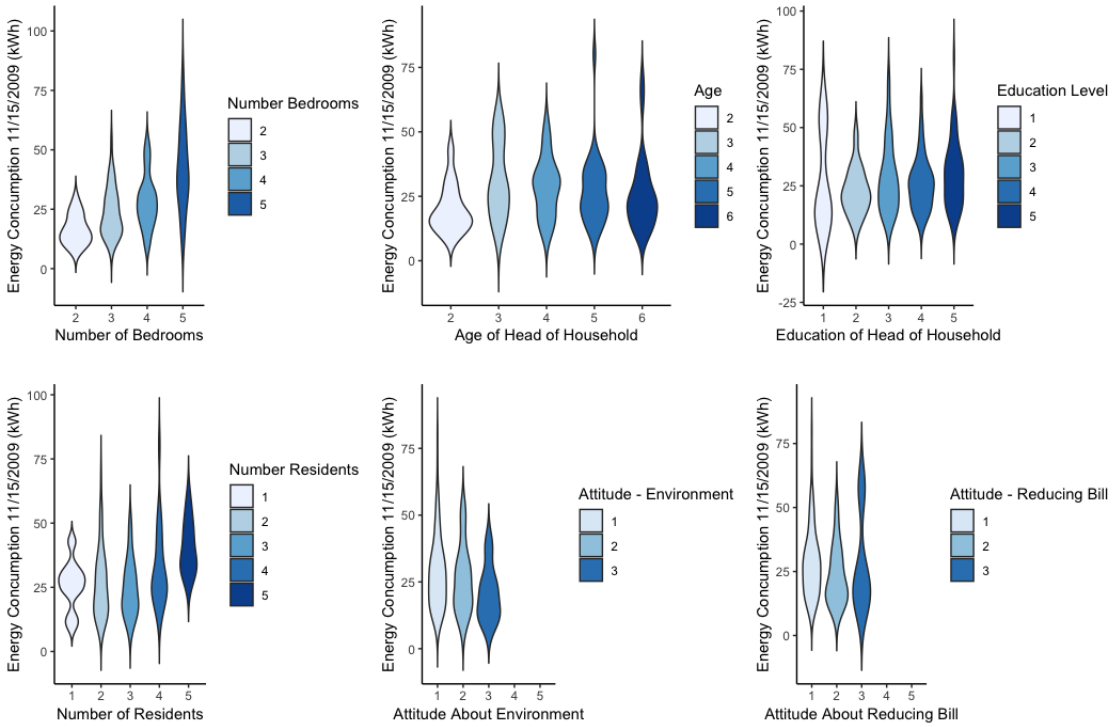

Figure 1: Violin Plots - Predictors vs. Response

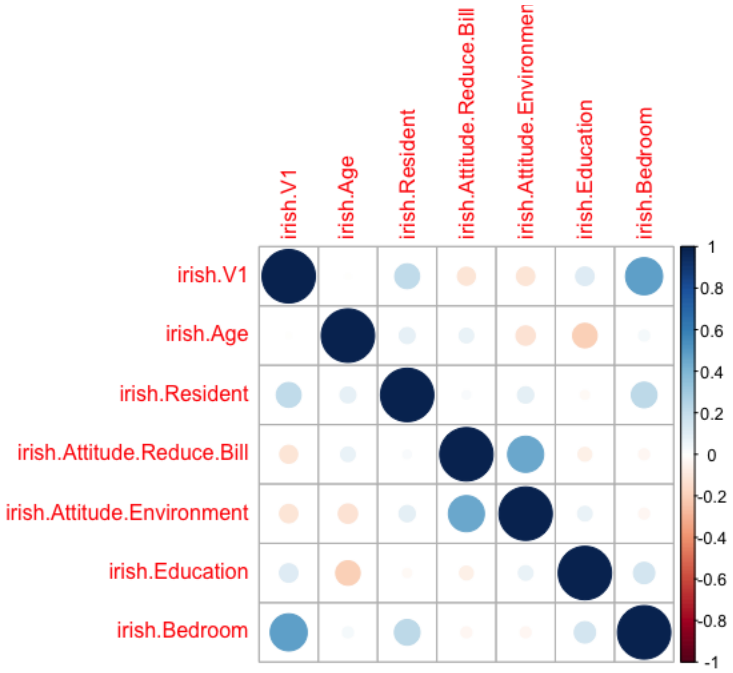

Figure 2: Correlation Plot

1.C

The Standard Improper Reference (SIR) prior, $p(\mu, \tau) = \frac{1}{\tau}$ for normal data, is often considered a "noninformative" prior in the sense that it takes independent flat priors on μ and τ (or rather

$\log \tau$) (pg 71). Essentially, the idea behind using flat priors is that most of the weight will be put on the data so that the posterior results mimic the frequentist results for normal data (pg 111). Similarly, in the linear regression setting the SIR prior $p(\beta, \tau) = \frac{1}{\tau}$ corresponds to the frequentist least squares estimation (pg 226). SIR priors are also used in sensitivity analysis, allowing us to see how our results change when an informative prior is used (pg 230). The downside of using a SIR prior is that it often results in an improper posterior, which does not allow us to use the laws of probability and leads to biased hypothesis tests. Furthermore, there is no such thing as an actual noninformative prior due to the fact that imposing a uniform distribution on a parameter does not mean that subsequent transformations of that parameter will also be uniform.

Proper Vague Reference Priors are an approximation to the SIR prior, but have the benefit of resulting in proper posteriors. This allows for inference in the absence of preexisting beliefs. Similar to SIR priors, proper reference priors can be used in preliminary data analysis when we have little prior information and would like to approximate non-bayesian methods, or in sensitivity analysis if prior information is available (pg 230). Proper reference priors can also be used when we have partial information; we can incorporate the prior information we do have and use reference priors for the rest. Lastly, using independent proper reference priors can help resolve multicollinearity issues.

1.D

Zellner's g prior for linear regression is defined as the following prior:

$$\begin{aligned} Y_i|\underline{X}_i, \underline{\beta}, \tau &\sim N(\underline{X}_i^{'}\underline{\beta}, \tau) \\ \underline{\beta}|\tau &\sim N_r(\underline{\beta}_0, \frac{\tau}{g}(\underline{\underline{X}}^{'}\underline{\underline{X}})) \\ \tau &\sim Gamma(a, b) \end{aligned}$$

where $a, b, \underline{\beta}_0, g$ are hyperparameters. It has the benefit of being simple mathematically (relative to conjugate priors). The g-prior was also designed so that inference is invariant to changes in the scale of the predictors. In addition, the g-prior is linked to frequentist results since the prior variance on the Betas resembles that of the maximum likelihood estimates.

Some argue that the g-prior uses the data to inform the prior, hence leveraging the covariates to inform us of the covariance relationship between the betas. However this should not be an issue since, in the linear regression setting, the covariates are viewed as known and fixed values. The researcher also must decide on what value to use for g; changing g affects how much weight is put on the prior vs. on the data. Lastly, as g approaches infinity the prior becomes uniform, bringing back the Lindley paradox of biased inference.

1.E

Here we consider three values for g: $g = 1$ which puts equal weight to prior and likelihood, $g = n$ which is equivalent to the weight of one observation, and Fernandez's $g = max(n, r^2)$. In this case, we have no prior information available on energy consumption in Ireland, so we would like to be as vague as possible in our prior specification. Therefore, we use $g = max(n, r^2) = 151$.

1.F

While we suspect that age is not a significant predictor for the energy consumption based on the correlation plot, we will still include the variable in our search just to affirm (or rebuke) our suspicions. We will perform an all subsets model search where we fit every possible model consisting of solely the main effects. As we have 6 predictors, each can either be or not be in a potential model, thus resulting in $2^6 = 64$ total possible models (we always include an intercept in every model fitted). The model specification is identical to the one in part 1.D with $a = b = 0.001, g = 151, \underline{\beta}_0 = \underline{0}$. Each model will have the negative of the log psuedomarginal likelihood (nLPML) computed (alongside the DIC and BIC for reference) and the model with the lowest nLPML will be determined as our final choice.

1.G

As mentioned in part F, we chose the model with the smallest nLPML. Coincidentally, this model also yielded the smallest DIC. The covariates in the model include number of residents, number of bedrooms, and attitude about the environment. Our findings are consistent with our hypotheses from the exploratory data analysis - age is not a significant predictor for energy consumption, while covariates that are correlated with energy consumption, such as number of bedrooms and attitude about the environment, are included in the model. The posterior means for the betas can be seen in table 1. As we would expect intuitively, the posterior means for number of bedrooms and number of residents are positive, thus predicting an increase in energy consumption. The posterior mean for attitude about the environment is negative, so that a strong attitude about the environment predicts lower energy consumption.

1.H

Since the model that was chosen in part G did not include the education covariate, in order to evaluate if the level of education has an effect on energy consumption, we will fit a model with education as a predictor in addition to the covariates chosen from part G. We will evaluate $P(|\beta_{education}| > \delta|\underline{Y})$ while fitting the model. If the resulting probability is greater than 0.65 than we will deem it to be deviating consistently from 0 enough to claim that it has a significant effect on the energy consumption. δ will be set to be 0.2.

After fitting the model, the resulting probability is $P(|\beta_{education}| > \delta|\underline{Y}) = 0.874$. As such, we deem this covariate to be significant in predicting energy consumption. The mean posterior value for this coefficient is 0.712, meaning that for an increase in education status by 1 level (e.g. from high school to college) there is an associated expected increase of energy consumption for November 15, 2009 of 0.712 kWh.

Table 1: All Subsets Search Results

	Mean Posterior Betas							Criterion		
# Predictors	(Intercept)	Age	# Residents	Bill Att.	Env. Att.	# Bedrooms	Education	BIC	DIC	nLPML
6										
	27.448	-0.375	1.729	-1.094	-1.978	7.032	0.567	1204.607	1180.157	590.995
5										
	27.489	—	1.704	-1.237	-1.855	7.034	0.665	1199.778	1178.553	589.816
	27.485	-0.446	1.729	—	-2.493	7.072	0.659	1199.931	1179.244	590.103
	27.474	-0.487	1.697	-1.170	-1.893	7.182	—	1200.039	1179.607	589.958
4										
	27.464	—	1.684	—	-2.370	7.034	0.722	1195.240	1177.201	589.051
	27.457	-0.584	1.696	—	-2.423	7.175	—	1195.430	1177.610	589.046
	27.449	—	1.628	-1.343	-1.728	7.144	—	1195.432	1177.490	588.852
3										
	27.480	—	1.623	—	-2.198	7.188	—	1190.949	**1176.011**	**588.177**
	27.457	—	1.510	-2.272	—	7.237	—	1191.466	1176.511	588.258
	27.487	—	1.540	—	—	7.190	0.613	1192.793	1177.826	589.264
2										
	27.455	—	1.516	—	—	7.334	—	1188.295	1176.118	588.205
	27.487	—	—	—	-2.013	7.637	—	1188.679	1176.757	588.479
	27.463	—	—	-2.103	—	7.619	—	1188.879	1176.957	588.494
1										
	27.494	—	—	—	—	7.714	—	**1185.569**	1176.688	588.462
	27.487	—	2.920	—	—	—	—	1218.099	1209.223	604.844
	27.440	—	—	—	—	—	1.463	1223.104	1214.035	607.474
0										
	27.504	—	—	—	—	—	—	1220.204	1214.372	607.423

2

2.A

LPML is used to evaluate the predictive accuracy of a model by maximizing the log pseudomarginal likelihood function. It is based on the i^{th} conditional predictive ordinate - the predictive density based on all of the data except the i^{th} observation. The LPML is easily estimated from a posterior sample, and thus can be calculated based on MCMC output (pg 81). An advantage of LPML is that it does not depend directly on the model parameters (pg 83). As for many model selection criteria, LPML violates the likelihood principle and is not appropriate for use in comparing models with different scales of measurement for the predictors and the response. Due to this, the major downside of LPML is that it can only be used to compare models with nested parameters.

When fitting a model with and without the interaction term, we obtain LPML values of -606.595 and -605.7239 respectively. As such, we would declare that the model without the interaction term is superior based on LPML.

2.B

Bayes Factors are computed based on the posterior sampling densities of two competing models. Specifically, the posterior odds are calculated as the prior odds times the Bayes factor, where the Bayes factor is the ratio of posterior sampling densities. Often, we are interested in 2 times the log Bayes factor, as it is more stable for extreme values. Not only can Bayes Factors be used to compare models which are not nested, they can be used to compare two models which have nothing in common (pg 74). Further, Bayes Factors can be used to test two competing hypotheses in general parametric testing. The rule of thumb for determining strength of evidence in favor of one model is based on 2LBF as was proposed by Kass and Raftery in 1985: 0-2 is not really worth considering, 2-6 represents positive strength of evidence, 6-10 denotes strong evidence, and anything greater than 10 is very strong evidence. One downside of Bayes factors is that it requires taking additional samples from the prior associated with each model, which requires more computing effort. It is also important to recognize that Bayes factors are calculated under the assumption that one of the two models is correct.

When fitting a model with and without the interaction term, a Bayes Factor term was computed with the model with the interaction being associated with the numerator of the term. The actual value computed was $BF = 1.062653$ or rather $2 \log BF = 0.1215363$. As such, since $0 < 2 \log BF < 2$ we would declare that there is no evidence suggesting that the model with the interaction term is superior to the model without.

2.C

For sensitivity analysis we compare the estimated coefficients based on a model with g prior with $g = max(n, r^2) = 151$ to a model with a vague prior of $\underline{\beta} \sim N(\underline{0}, 0.01\underline{\underline{I}})$. Figure 3 showcases the posterior distributions for the four beta coefficients fit in the models. As the distributions all look roughly normal and share similar supports among same coefficients, we conclude that utilizing the prior information in the g prior did not significantly effect the results of our analysis.

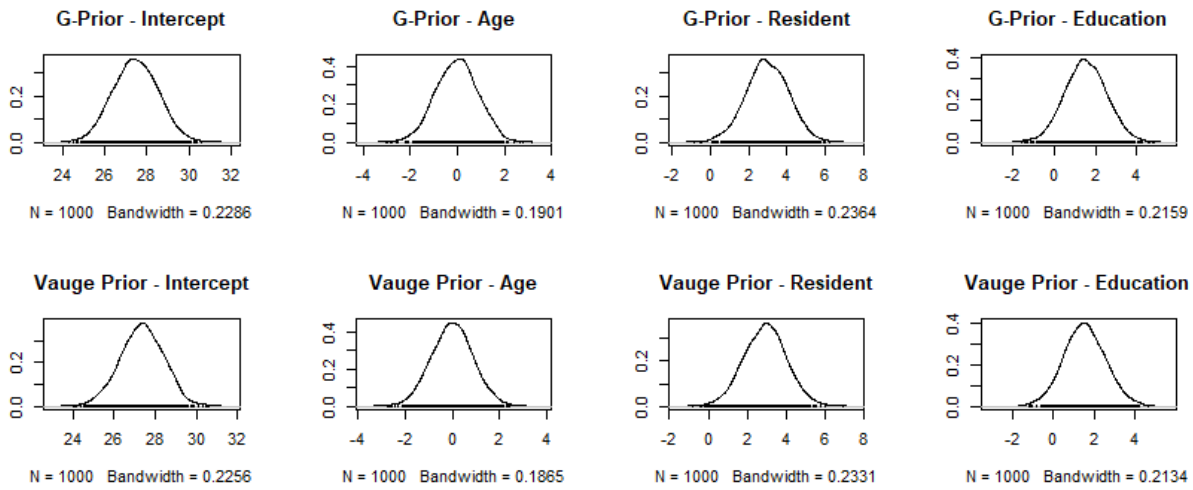

Figure 3: Posterior Distributions for Sensitivity Analysis

2.D

The posterior predictive mean and 95% credible interval for the median household is 25.695 and $(0.048, 51.368)$ respectively. Figure 4 displays the posterior predictive distribution.

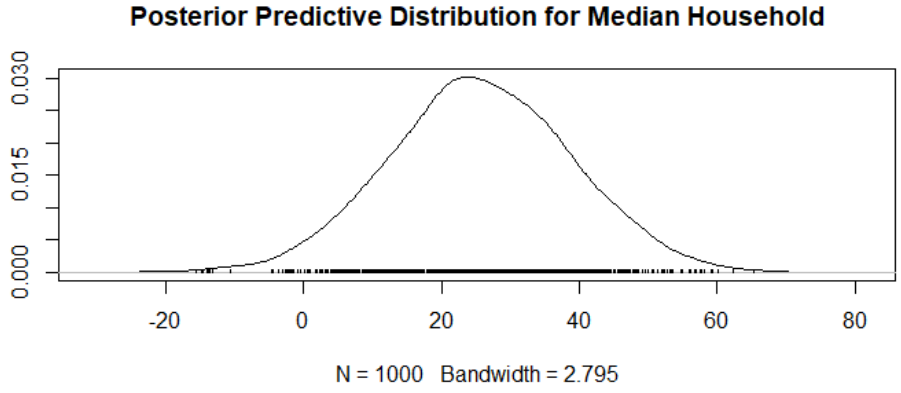

Figure 4: Posterior Predictive Distributions

3

3.A

We propose the following model to describe the daily consumption pattern as a function of the predictors. We used the following notation:

- Y_d^h: Electricity consumption of household h on day d, for $d = 1, \ldots, 121$ and $h = 1, \ldots, 151$

- $\underline{X}^h$: Vector of covariates for household h
- $\underline{\beta}_d$: Vector of regression coefficients for day d
- $\underline{\beta}$: Vector of regression coefficients for population
- μ_d: Mean daily consumption for population
- μ: Mean consumption in population
- τ^h: Precision for household h
- τ_d: Precision for daily consumption
- τ: Population precision

Our model consists of two effects that contribute to mean energy consumption for a house on a specific day: the daily average μ_d and the expected house's deviation from the daily average $\underline{X}^h\underline{\beta}_d$. This is motivated by the fact that we aim at capturing the energy consumption pattern on a daily level in general, and for each house on that day. Both the daily average and the coefficients for the expected deviation have a population level parameter that they are drawn from (μ and $\underline{\beta}$ respectively). This allows for considerable shrinkage in estimating the overall population energy consumption average as well as the population impacts of household covariates on energy consumption.

We do recognize that each house may operate differently than one another on a consistent basis, and so the precision of the energy consumption each house for a specific day is the sum of a daily precision τ_d and the household's precision τ^h. Each of these are drawn from hierarchical parameters to allow for proper shrinkage as well.

We are using all six of the available covariates. This is due to the fact that even though our analysis in parts 1 and 2 indicated that some variables were not related to energy consumption, those were for only a single day. We do not know yet if those that were found irrelevant will be significant for later days. As such, we employ proper vague reference priors on all of the variables at the population level to encourage shrinkage towards 0. The model definition is thus as follows:

$$Y_d^h|\underline{\beta}_d, \mu_d, \tau_d, \tau^h \sim \mathrm{N}(\mu_d + \underline{X}^h\underline{\beta}_d, \tau_d + \tau^h)$$

$$\begin{aligned}\underline{\beta}_d|\underline{\beta}, \tau_\beta &\sim \mathrm{N}_r(\underline{\beta}, \tau_\beta \underline{\underline{I_r}}) \\ \underline{\beta} &\sim \mathrm{N}_r(\underline{0}, 0.01\underline{\underline{I_r}})\end{aligned}$$

$$\begin{aligned}\mu_d|\mu, \tau_\mu &\sim \mathrm{N}(\mu, \tau_\mu) \\ \mu &\sim \mathrm{N}(0, 0.01) \\ \tau_d|\gamma, \delta &\sim \mathrm{Gamma}(\gamma, \delta) \\ \tau_h|\epsilon, \eta &\sim \mathrm{Gamma}(\epsilon, \eta)\end{aligned}$$

$$\tau_\beta, \tau_\mu, \gamma, \delta, \epsilon, \eta \sim \mathrm{Gamma}(0.001, 0.001)$$

3.B

In figure 5 we provide a plot that shows the time-varying effect of the number of residents on the daily consumption at the population level (January 1st is marked with a black vertical line). The plots shows the median consumption grouped by the number of residents for each day (We used the medians since the responses were skewed - see Figure 9). We see that households with 2 and 3 residents have very similar electricity consumption. There was a spike in the electricity consumption right around Christmas for all levels of the Resident variable, which makes sense since people tend to use more electricity during holidays especially for decorating their houses. Based on the posterior distribution of the electricity consumption we see that the higher the number of residents in a household the higher the daily electricity consumption. Note that the electricity consumption is very stable for all Resident groups after January 1st compared to the previous year. This might be due to the new tariffs. Lastly, electricity consumption is reduced after January 1st for all Resident groups.

As the data set did not provide any ID numbers for the households, we interpreted households 30, 83, and 91 to be the households corresponding to rows 30, 83, and 91 respectively. We can see the the time-varying daily consumption for households 30, 83, and 91 in figure 6. These households have the following profile:

Table 2: Household Profiles

Household	Age	Resident	Bill	Environment	Bedroom	Education
30	6.00	2.00	1.00	1.00	3.00	4.00
83	4.00	4.00	1.00	1.00	4.00	4.00
91	4.00	3.00	1.00	1.00	3.00	5.00

Household 83 has the highest consumption values and household 30 has the lowest. Similar to the population the electricity consumption reduces after January 1st for all of the three households. The consumption is very steady for the three households after January 1st, but is fluctuating significantly for days before January 1st. Again, there is a peak consumption around the Christmas day for all of the three houses.

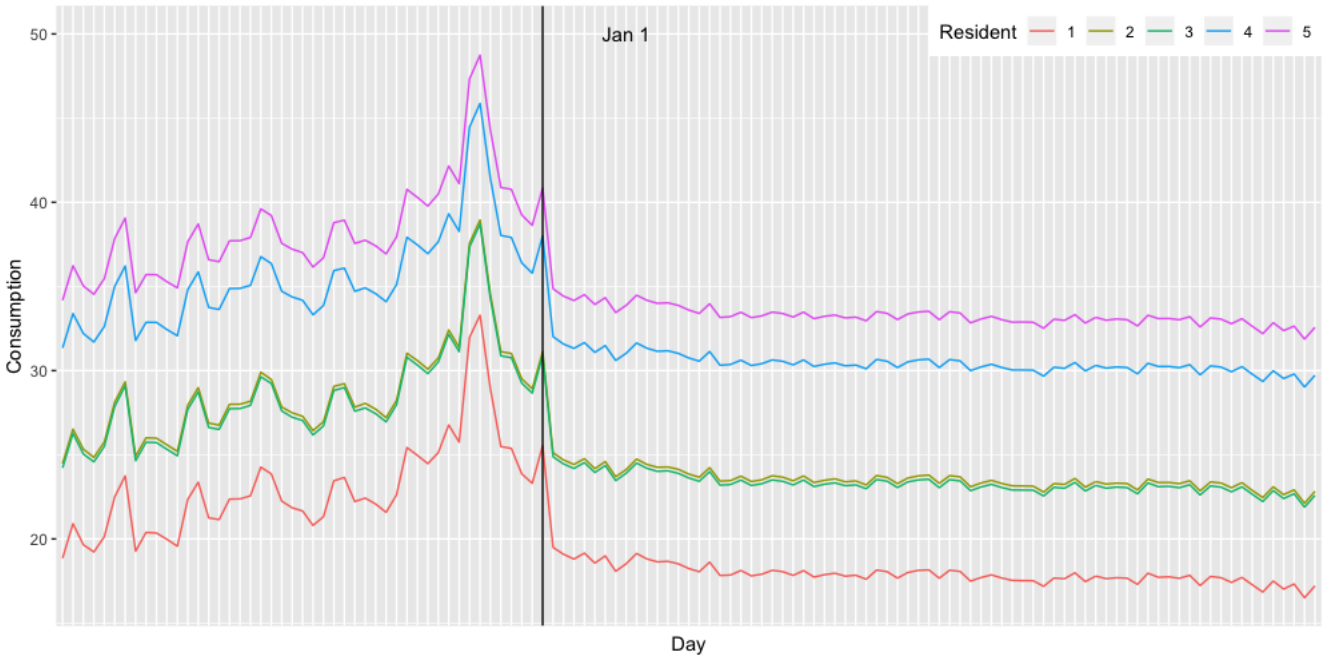

Figure 5: Time Series Plots for Daily Consumption for Different Number of Residents

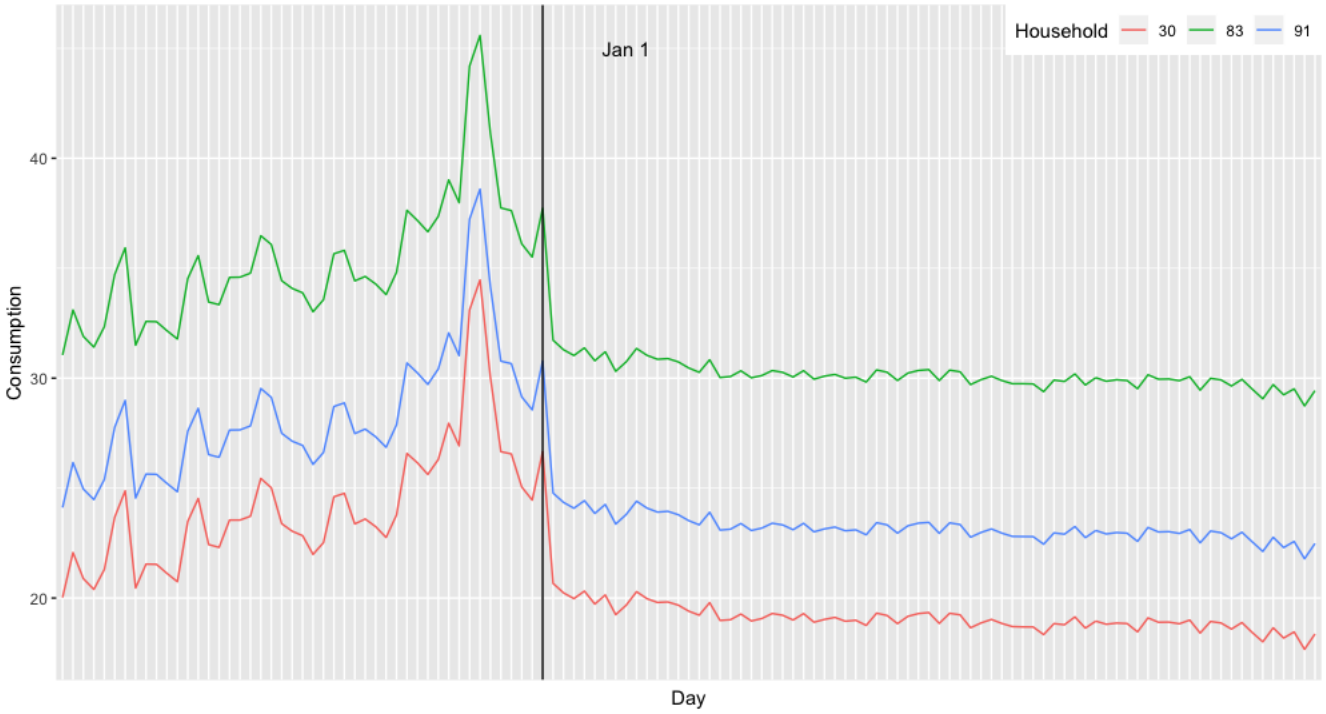

Figure 6: Daily Consumption for Households 30, 83, and 91

4

To evaluate if there was an impact on the daily energy consumption before and after the policy was set in place, we will fit the following model:

$$
\begin{aligned}
Y_d^h|\underline{\beta}_{\underline{d}}, \mu_d, \tau_d, \tau^h &\sim \mathrm{N}(\mu_d + \underline{\underline{X}}^h \underline{\beta}_{\underline{d}}, \tau_d + \tau^h) \\
\\
\underline{\beta}_{\underline{d}}|\underline{\beta}^p, \tau_\beta &\sim \mathrm{N}_r(\underline{\beta}^p, \tau_\beta^p \underline{\underline{I_r}}) \qquad \text{where } p = 1 \text{ if } d \text{ is before Jan 1,} \\
\underline{\beta}^p &\sim \mathrm{N}_r(\underline{0}, 0.01\underline{\underline{I_r}}) \qquad \text{otherwise it is 0} \\
\\
\mu_d|\mu^p, \tau_\mu^p &\sim \mathrm{N}(\mu^p, \tau_\mu^p) \\
\mu^p &\sim \mathrm{N}(0, 0.01) \\
\tau_d^p|\gamma^p, \delta^p &\sim \mathrm{Gamma}(\gamma^p, \delta^p) \\
\tau^h|\epsilon, \eta &\sim \mathrm{Gamma}(\epsilon, \eta) \\
\\
\tau_\beta^p, \tau_\mu^p, \gamma^p, \delta^p, \epsilon, \eta &\sim \mathrm{Gamma}(0.001, 0.001)
\end{aligned}
$$

and evaluate the probability statement: $P(\mu^1 > \mu^0|\underline{Y})$. This will tell us if the policy had a decrease on the energy consumption for the population. The motivation behind this model is to allow completely different consumption patterns before and after the policy. The result, should they actually be different, will have the coefficients shrink towards different population level values for before and after the policy was enacted.
After fitting the model, the result is as follows: $P(\mu^1 > \mu^0|\underline{Y}) = 1.000$. As such, we conclude that the policy definitely did accomplish its goal of decreasing energy consumption.

In order to evaluate if the energy consumption pattern at the population level tends to return to before the policy was enacted, we will compare the previously fitted model using Bayes Factor to the following model:

$$
\begin{aligned}
Y_d^h|\underline{\beta}_{\underline{d}}, \mu_d, \tau_d, \tau^h &\sim \mathrm{N}(\mu_d + \underline{\underline{X}}^h \underline{\beta}_{\underline{d}}, \tau_d + \tau^h) \\
\\
\underline{\beta}_{\underline{d}}|\underline{\beta}^p, \tau_\beta &\sim \mathrm{N}_r(\underline{\beta}^p, \tau_\beta^p \underline{\underline{I_r}}) \qquad \text{where } p = 1 \text{ if } d \text{ is before Jan 1 or} \\
\underline{\beta}^p &\sim \mathrm{N}_r(\underline{0}, 0.01\underline{\underline{I_r}}) \qquad \text{after March 8, otherwise it is 0} \\
\\
\mu_d|\mu^p, \tau_\mu^p &\sim \mathrm{N}(\mu^p, \tau_\mu^p) \\
\mu^p &\sim \mathrm{N}(0, 0.01) \\
\tau_d^p|\gamma^p, \delta^p &\sim \mathrm{Gamma}(\gamma^p, \delta^p) \\
\tau^h|\epsilon, \eta &\sim \mathrm{Gamma}(\epsilon, \eta) \\
\\
\tau_\beta^p, \tau_\mu^p, \gamma^p, \delta^p, \epsilon, \eta &\sim \mathrm{Gamma}(0.001, 0.001)
\end{aligned}
$$

The idea is that if the overall population trend returns to how it was prior to the policy, then the data will be better modeled by having observations near the end of the study be drawn from the same parameters that modeled observations prior to the policy being enacted.
Doing so resulted in a $2 * \log BF = 70.45$ which indicates extremely strong evidence in favor of the

first model. In other words, there is extremely strong evidence to suggest that the population level energy consumption patterns do not return to how they were before the policy was enacted.

Appendix

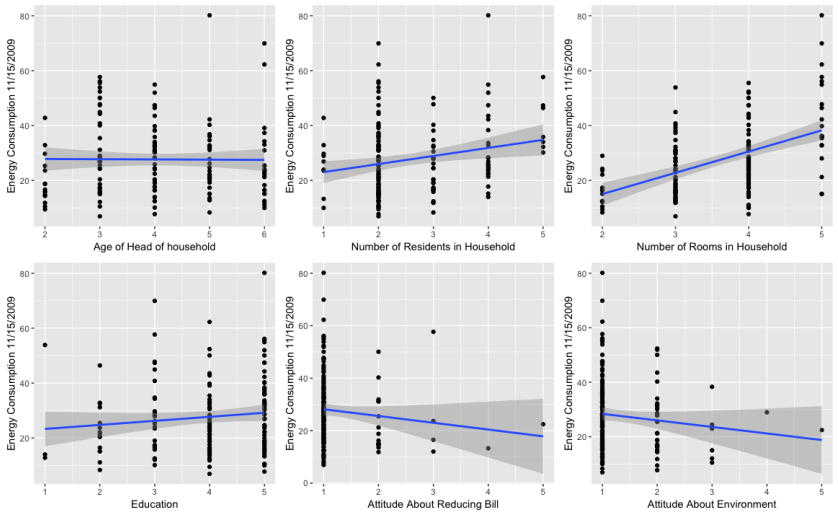

Figure 7: Scatterplots: Predictors vs. Response

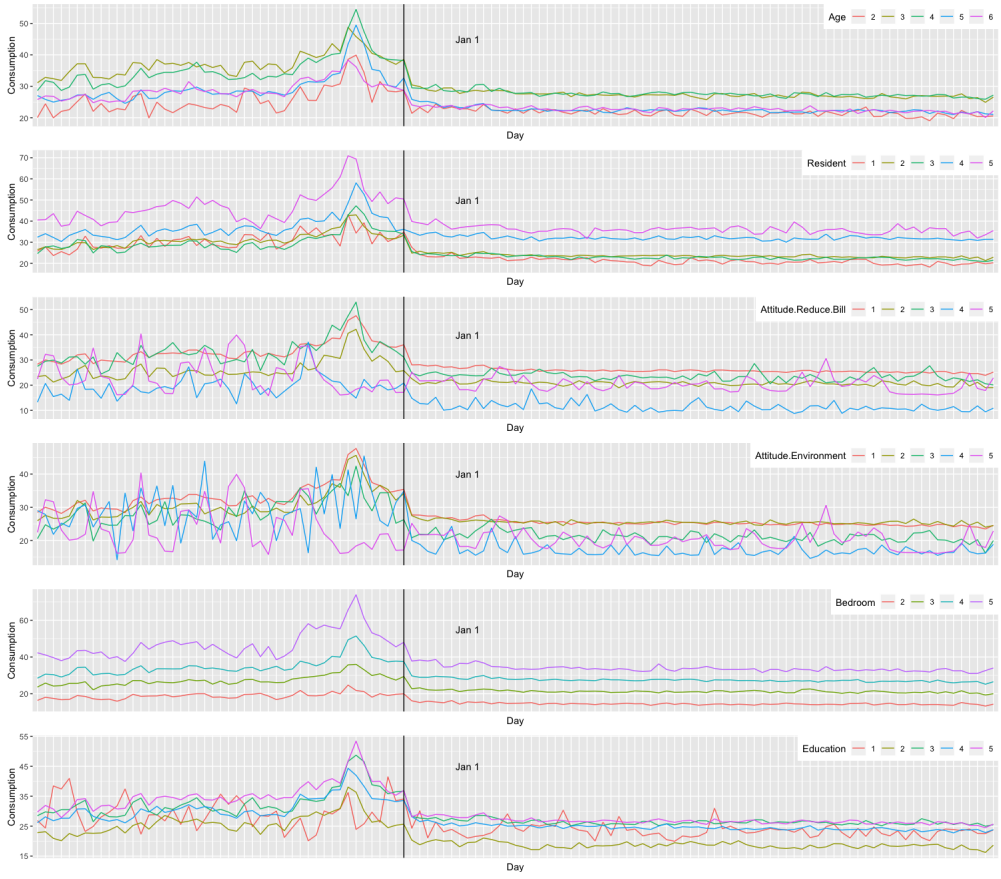

Figure 8: Time Series Plot: Predictors vs. Time

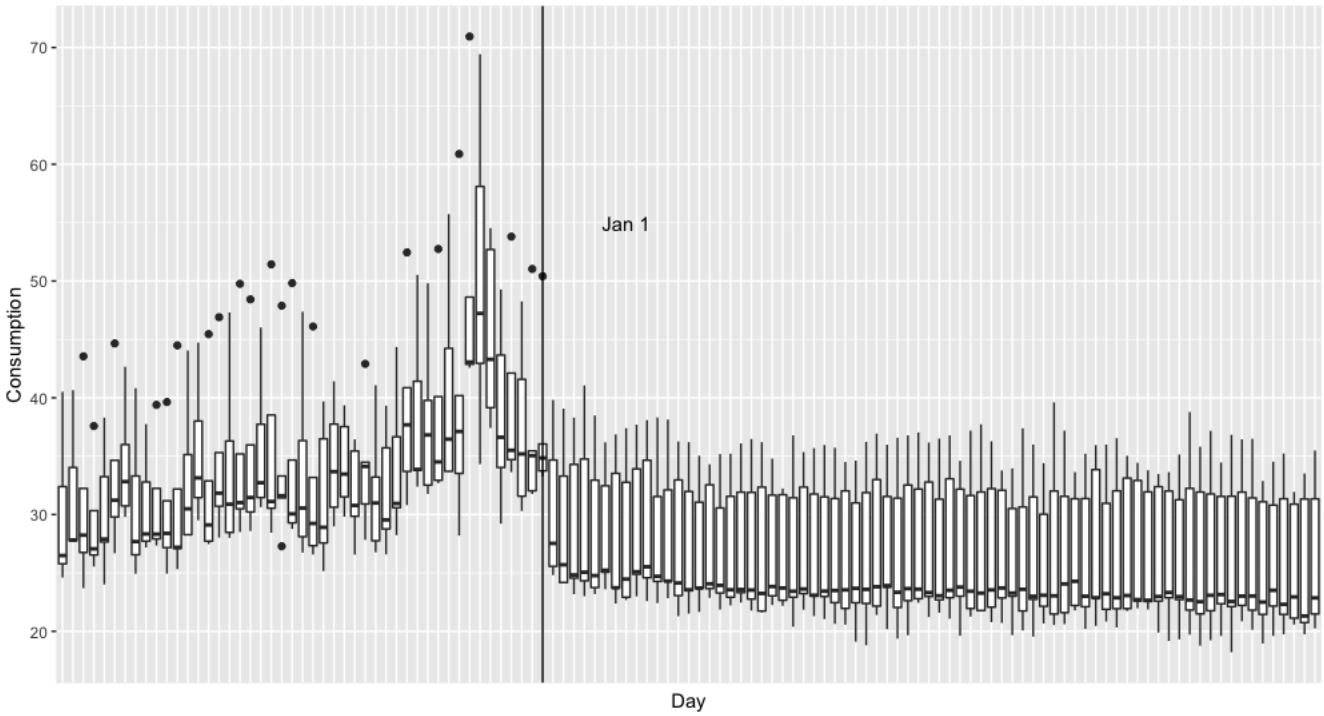

Figure 9: Distribution of the Daily Electricity Consumption