Technical Report

Analysis of Electricity Costs for Households in Oregon

Douglas Fedorczyk

6/3/2021

Introduction

This report is based on data from the American Community Survey for households in Oregon that was conducted in 2015. The intent of this project was to analyze the data from the survey in both an explanatory and predictive nature. The explanatory analysis sought to determine if people living in apartments pay less on electricity than those living in houses; and if so, by how much? The predictive analysis sought to create a model that would predict electricity costs for a household living in Oregon. Lastly, both the explanatory and predictive analyses were compared and contrasted; and a discussion of challenges faced while analyzing the data is also included.

Explanatory Analysis

The explanatory analysis sought to determine if people living in apartments pay less on electricity than people living in houses; and if so, by how much? This section includes the strategy that was used for analyzing the data, a description of the model used to answer the question of interest, followed by the analysis of the model itself, and finally the results from the analysis.

Approach

The strategy used for answering the question of interest begin with fitting a model with interactions. Then a review of the residual diagnostics was conducted to check the residual fit. From there, models with and without interactions were compared using an Extra Sum of Squares (ESS) F-test. Finally, the preferred model as determined from the ESS F-test was used to make an inference to answer the question of interest. These steps are captured in the list below:

Step 1: Fitted model with interactions

Step 2: Performed residual diagnostics to check residual fit

Step 3: Compared models with and without interactions using an Extra SS F-test

Step 4: Performed inference using preferred model

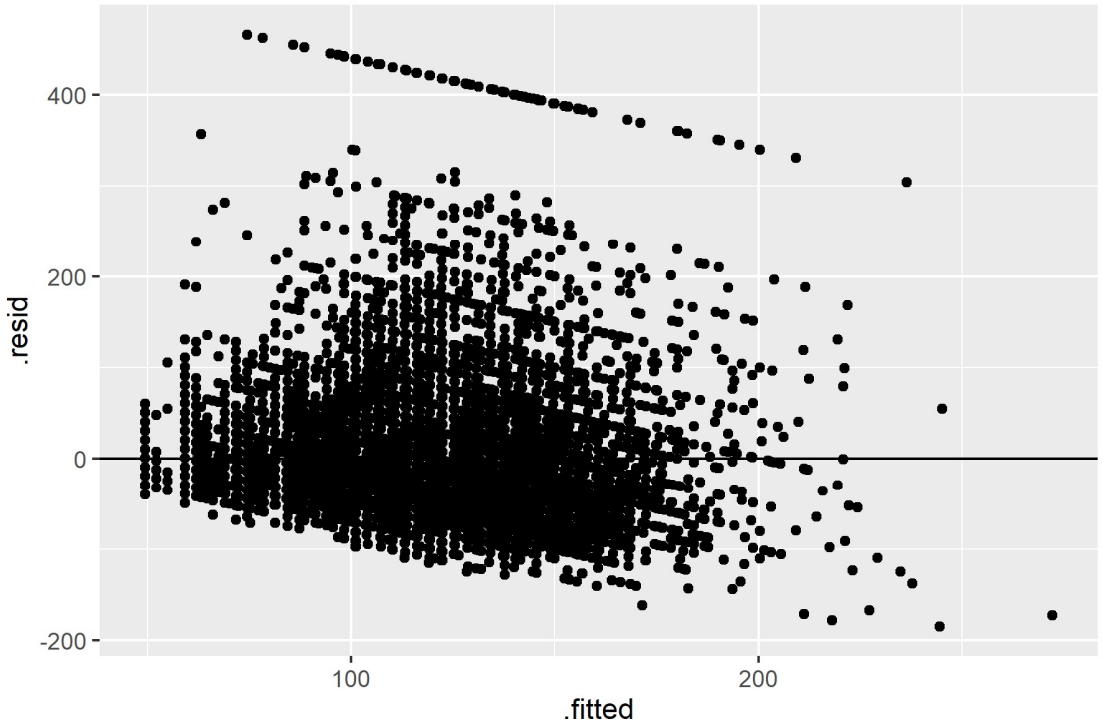
Description of Model

There were two variables that were requested to be included in the model; this included NP, number of persons, and BDSP, number of bedrooms. The variable BLD was included that differentiated between the apartment and house model. Non-apartment and non-house data were removed which included items such as mobile homes, boats, RVs, etc. Various types of apartments and houses were combined so that each type was represented by either a 0 or 1, respectively. The number of rooms variable, RMSP, was also added to capture the square footage of the apartment or house as it was believed this could play a part in the electricity costs. Interactions between variables were also included, specifically the number of persons in a household and the number of rooms or bedrooms with the belief that these have the potential to interact with each other. Based on the aforementioned variables, an initial explanatory model had the following form:

$$ELEP = \beta_0 + \beta_1 NP + \beta_2 BDSP + \beta_3 BLD + \beta_4 RMSP + \beta_5 NP \times BDSP + \beta_6 NP \times RMSP$$

Analysis

When initially checking the model fit, a linear pattern with a negative slope was discovered that can be seen in the residuals versus fitted values plot below. It was determined that this was a underlying issue based on the way the data was inputted, i.e. the response variable (ELEP) was rounded to the nearest 10. It was suggested to determine a transformation and model that looked the best. Various transformations of the response variable were reviewed to include taking the logarithm of the ELEP variable, the squre root, and the inverse. It was determined that taking the logarithm had the best fit overall. This was based on a review of the residuals versus the explanatory variables, namely NP, number of persons, which can be found in the appendix.



Fitted Values vs Residuals

Results

There were two interactions that were included in the model, the number of persons within the household, NP, and its interaction with both the number of bedrooms, BDSP, and the number of rooms, RMSP. To check the significance of the interactions, the ANOVA function was used to conduct the Extra Sum of Squares F-test; its output is shown in the table below.

Analysis of Variance Table						
Model 1: log(ELEP) ~ NP + BDSP + BLD + RMSP						
Model 2: log(ELEP) ~ NP + BDSP + BLD + RMSP + NP * BDSP + NP * RMSP						
Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	13769	4055.7				
2	13767	4032.4	2	23.266	39.717	< 2.2e-16***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

There is evidence to suggest that the effect of the number of rooms and bedrooms depends on the number of persons in the household (Extra SS F-test, comparing model with interactions to model without interactions, p-value ~ 0).

	Estimate	Std. Error	t value	Pr(>|t|)	95% CI LB	95% CI UB

BLD Coef	0.1744144	0.01430147	12.19556	4.917524e-34	0.1463816	0.2024473

Additionally, there is convincing evidence that people living in apartments pay less for electricity than people living in houses, after accounting for the number of people in the household, number of bedrooms, and number of rooms (p-value ~ 0). It is estimated that the mean amount of money people living in a house pay for electricity is $1.19 more per month than people living in an apartment (95% CI: $1.16 - $1.22). Note that because the logarithm of the response variable was used as part of the model, it was necessary to take the exponentiation of the estimate and the confidence interval for β_3.

Predictive Analysis

The predictive analysis sought to create a model that could predict the electricity costs for a household living in the state of Oregon. This section includes the approach used for the analysis, a description of the model that was constructed for predicting electricity costs, followed by the analysis behind the model, and finally the results.

Approach

The approach used for predicting the electricity costs for a household in the state of Oregon started with determining how to validate the model. For this analysis, it was decided to use a k-folds cross validation approach with k = 5. Prior to validation though, an assessment of appropriate models for prediction purposes was based on comparing MSE (mean squared error) in order to select the best model. Finally, variables selected for the model are outlined in the model description below; the number of variables chosen was based on minimizing the mean squared error as measured by cross validation. The aforementioned steps are listed below:

Step 1: Select means for determining appropriate model => k-folds cross validation with k = 5

Step 2: Determine metric for assessing models => use MSE to compare models

Step 3: Determine method for selecting models for consideration => choose number of variables by minimizing mean squared error as measured by cross validation

Description of Model

A manual approach was initially used for variable selection, followed by the Best Subset method to down-select to a final model. The following two paragraphs go through each variable found in the data dictionary and provides reasoning behind variables that were either included or excluded.

Number of persons in the household (NP) and the number of bedrooms within the household (BDSP) were required to be a part of the initial model per the project instructions. The BLD variable was included as it was shown from the explanatory analysis that building type

affects electricity costs. ELEP, the electricity cost variable, is the item of interest and was included as the response variable. The other two cost variables, FULP and GASP, which stand for fuel costs and gas costs respectively, may be directly or inversely related to electricity costs and were included in the initial model. It should be noted that the majority of the data showed that most people used electricity as their main source of energy so it seemed reasonable that if there was an association that it was probably small. It seemed reasonable that the square footage of the structure would play a part in electricity costs, so the number of rooms variable, RMSP, was included. The year the structure was built, YBL, seemed that it may play a role in the electricity costs, e.g. building materials, building codes, etc., and so it seemed reasonable to include it in the initial model. The last two variables, R18 and R60, which reflect the presence of persons under 18 years of age in the household or the presence of persons over 60 years of age in the household, respectively, seemed that they may have an impact to the electricity costs and so were included in the initial model.

Serial Number was excluded from the model as was Type of unit (TYPE) since all of the data provided were one type. The lot size variable, ACR, was also excluded since it was not readily apparent how it might affect electricity costs. Electricity was listed as a fuel source under the HFL variable (house heating fuel), so it seemed inappropriate to include in the model. It was not readily apparent how the tenure variable (TEN) would affect the electricity cost and so it was excluded as well. Additionally, it was not readily apparent how the value of the property would impact the electricity costs other than as a reflection of the size of the structure, in which case it would not have seemed reasonable to include it since the number of rooms and bedrooms was already captured.

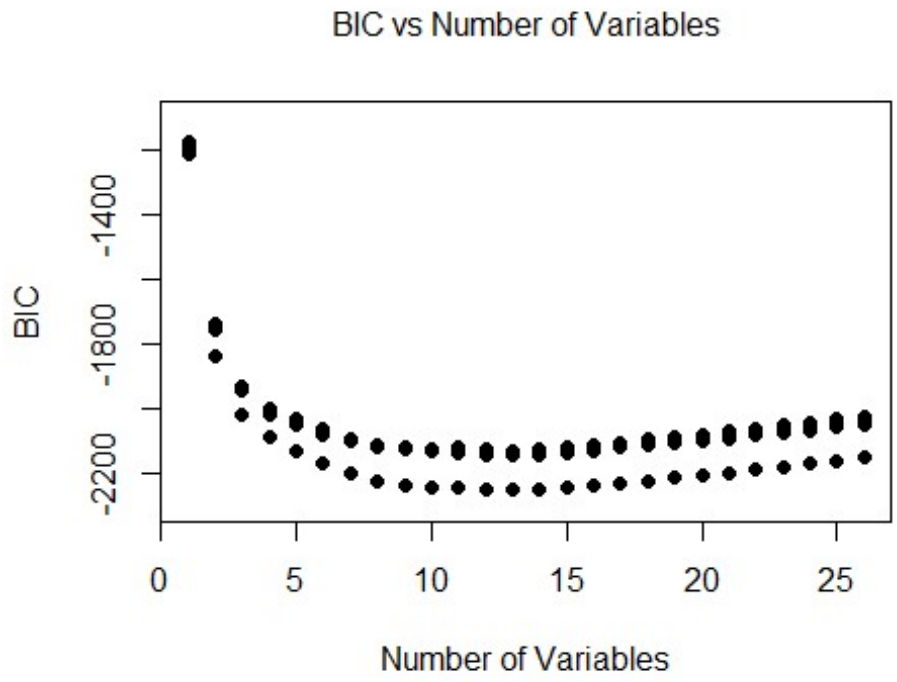
It should be noted that from the explanatory analysis it was discovered that a logarithmic function would be best suited for the ELEP variable; which was included as part of the predictive analysis as well. The initial form of the predictive model took the following form:

$$log(ELEP) =$$

$$\beta_0 + \beta_1 NP + \beta_2 BDSP + \beta_3 BLD + \beta_4 FULP + \beta_5 GASP + \beta_6 RMSP + \beta_7 YBL + \beta_8 R18 + \beta_9 R60$$

Analysis

The figure below shows the BIC values plotted against the number of variables and where a minimum between 12 and 15 variables can be seen.



An example of one of the top few models based on their BIC value is shown below.

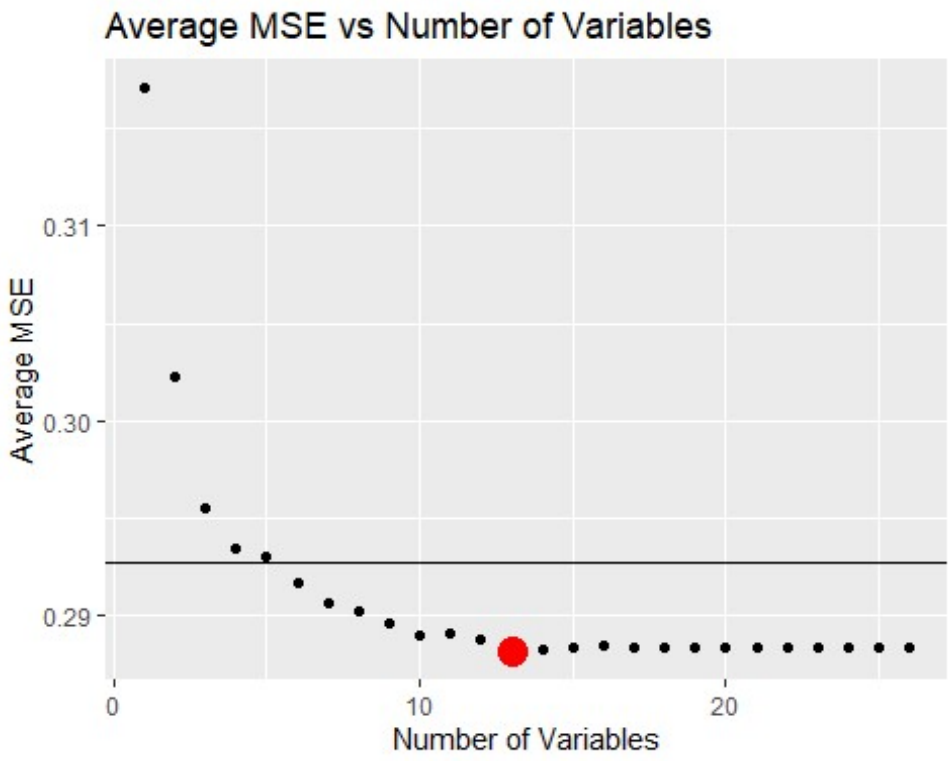
##	(Intercept)	NP	BDSP	BLD1	FULP	GASP	RMSP	YBL1940 to 1949	YBL1950 to 1959
## 13	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
## 12	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
## 14	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
## 11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
## 15	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
## 10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE

The output shows how the factor variables are broken out when applying the regsubsets function. Because of this, an intermediary form starts to emerge:

$log(ELEP) =$

$\beta_0 + \beta_1 NP + \beta_2 BDSP + \beta_3 BLD1 + \beta_4 FULP + \beta_5 GASP + \beta_6 RMSP + \beta_7 YBL\_\{1950 - 1959\}$
$+ \beta_8 YBL\_\{1960 - 1969\} + \beta_9 YBL\_\{1970 - 1979\} + \beta_{10} YBL\_\{1980 - 1989\}$

The plot below shows the average MSE versus the number of variables. The red dot signifies the best model based on average MSE indicating how many variables the model should contain. However, the horizontal line marks the one standard deviation, and using the one standard error rule suggests that a model with seven variables can be used.



Results

A 7-variable model was selected based on the combination of the results from the k-folds cross-validation along with using the one standard error rule. Best subset selection could then be used on the full data set in order to obtain the 7-variable model.

(Intercept)	NP	BDSP	BLD1	GASP	RMSP	YBL1970 to 1979	YBL1980 to 1989
3.7908	0.1101	0.0521	0.2225	-0.0007	0.0280	0.1220	0.1050

It was found that the final predictive model for electricity costs for a household living in Oregon has the following form:

$$log(ELEP) =$$

$$3.79 + 0.11NP + 0.06BDSP + 0.22BLD1 - 0.0007GASP + 0.025RMSP + 0.12YBL\_\{1970 - 1979\} + 0.11YBL\_\{1980 - 1989\}$$

where NP represents the number of persons within the household, BDSP represents the number of bedrooms within the structure, BLD1 indicates if a household resides within a home (equals 1 if yes, 0 otherwise), GASP equals the household gas costs in dollars, RMSP represents the number of rooms within the structure, YBL(1970-1979) indicates if the structure was built between 1970 and 1979 (equals 1 if yes, 0 otherwise), and YBL(1980-

1989) indicates if the structure was built between 1980 and 1989 (equals 1 if yes, 0 otherwise).

Compare and Contrasting Explanatory vs Predictive Analyses

Differences in Approaches

The explanatory analysis intended to keep the analysis as simple as possible and only looked at a few key variables and any interactions between them. The predictive analysis was more expansive in the approach for finding the best overall model for predicting electricity costs. Part of the reason for that, especially for the explanatory analysis, was need to be able to rely upon the use of p-values when determining whether or not to keep interactions in the model. This required not looking at the data ahead of time, i.e. data snooping, and working through the explanatory analysis first before moving onto the predictive analysis.

Why Different Approaches are Required

The different approaches were necessary to ensure that the final models for each respective analysis were valid. When working through an explanatory analysis, it is important not to look through the data so as not to render the p-values meaningless. What may happen is that if the data is looked at, what variables that do end up in the model, the resulting p-values will only reinforce what is already believed to be true and remove any sort of objectivity.

In the case of the predictive model, data snooping is not much of a problem. This is because the end result is to find which variables best predict what the response variable will be. It relies less on p-values and more so on other metrics such as adjusted r-square, root mean square error, etc.

Challenges

The first challenge encountered was figuring out how to consolidate the number of apartment factors and house factors; as well as removing other types of housing structures. This required reassigning values to the BLD variable to turn a 10 factor variable into a two factor variable. The two factors “Mobile home or trailer” and “Boat, RV, van, etc.” were removed since the apartments and the house factors were the only two items under consideration. The various apartment factors and house factors became a binary variable where 0 represented ‘apartment’ and 1 represented ‘house.’

The next challenge that was found was checking the fit for the explanatory model. The linear pattern with the residuals was not expected and it was not entirely clear how to handle it at first. After some discussion, it turned out that this was something to be expected. However, it was advised to try and find a transformation that worked well for the data. This required reviewing multiple transformations to find one that looked the best given the situation with the data.

The last challenge encountered was more of a coding challenge in dealing with the predictive model and also how the analysis was approached. Initially, there was some difficulty plotting

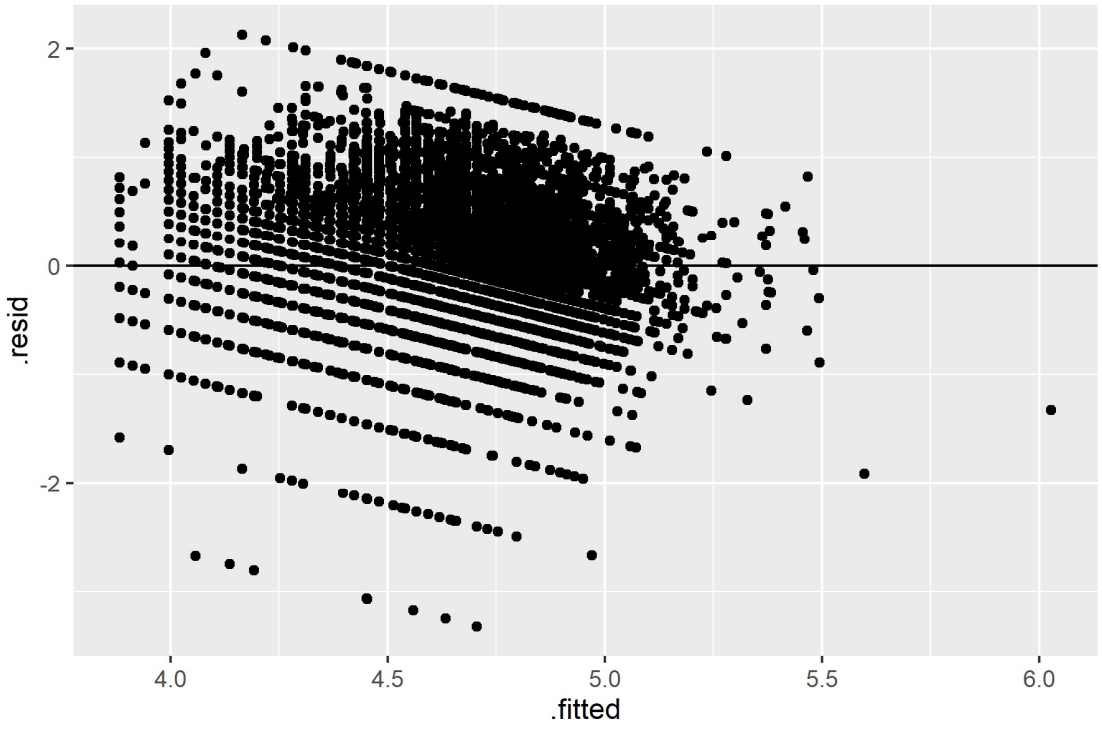
the multiple BIC values. Eventually a solution that worked was found, but was most likely not an elegant solution. Another issue was how the analysis was approached which was not necessarily incorrect, but it did not follow the original plan that had been laid out. Initially, the plan was to split the data into a training and validation set as oppose to using a k-folds cross-validation approach. This was based on a misunderstanding on the author's part in the implementation of k-folds cross validation and that splitting the data was not necessary, i.e. one would either split the data into training and validation sets or use cross validation, not both.

Model Limitations

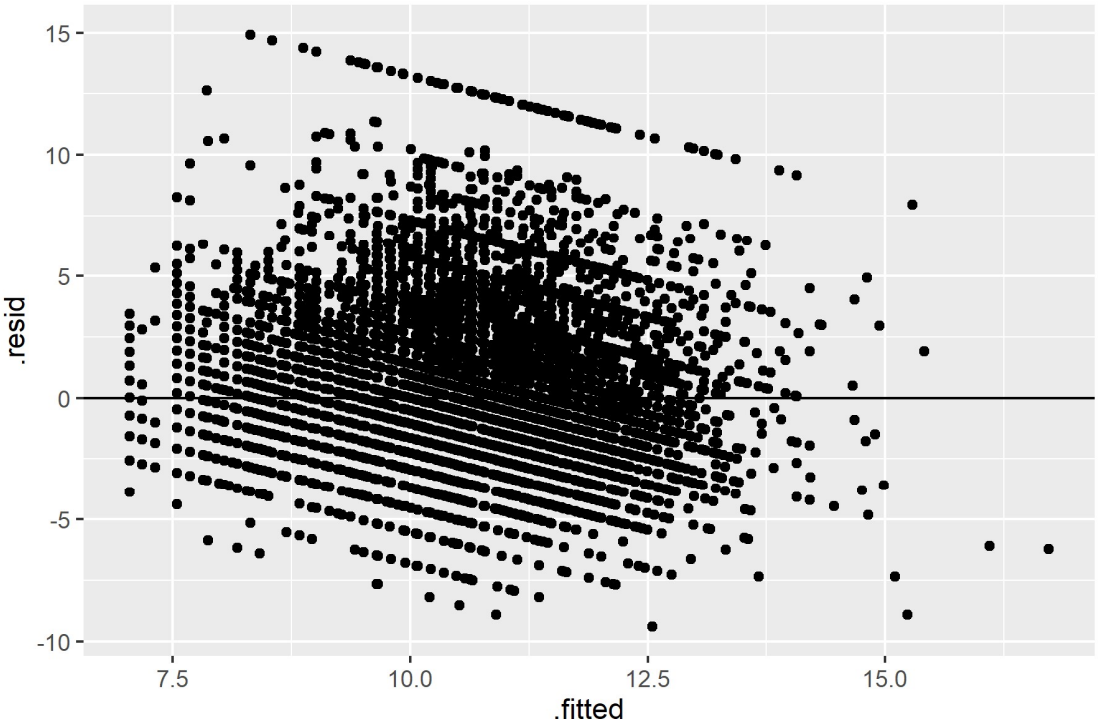
The data is assumed to be a random sample of households in the state of Oregon that have at least one person, pay for electricity, and are not group accommodations; this is assumed to be the case for both the explanatory and predictive analyses. To be clear, both models only take into account households that live in either a house or an apartment. Causal inferences for either model should be limited to either of those populations within the state of Oregon.

Appendix

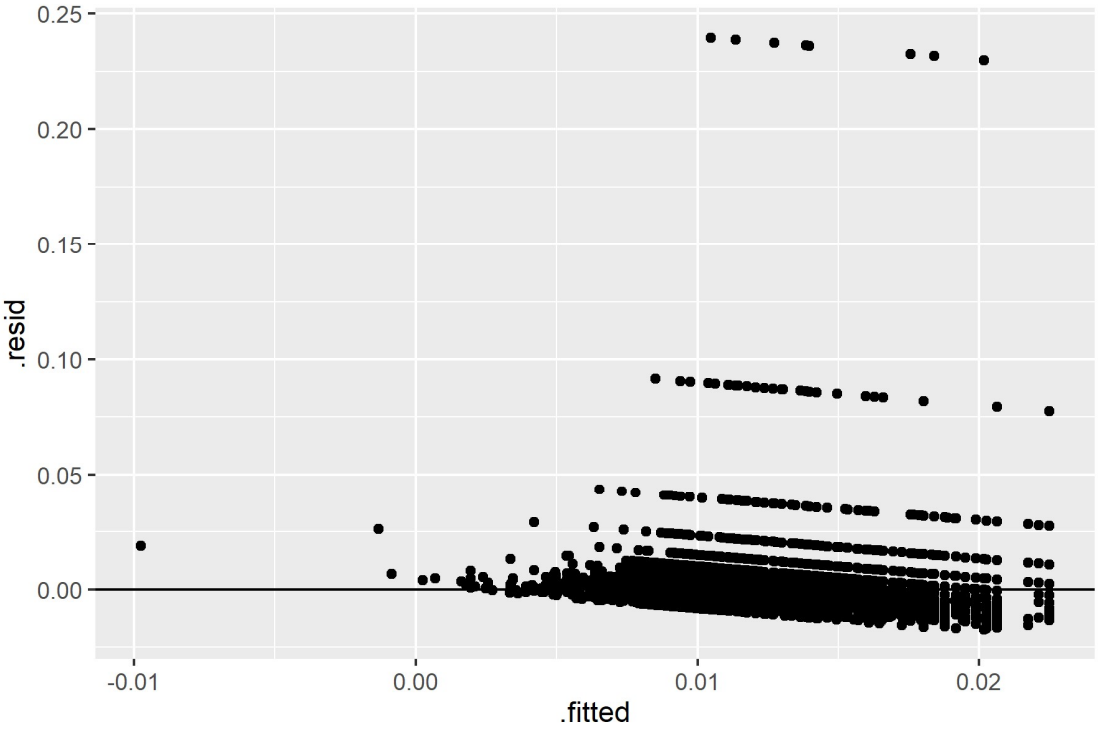
Transformation Figures



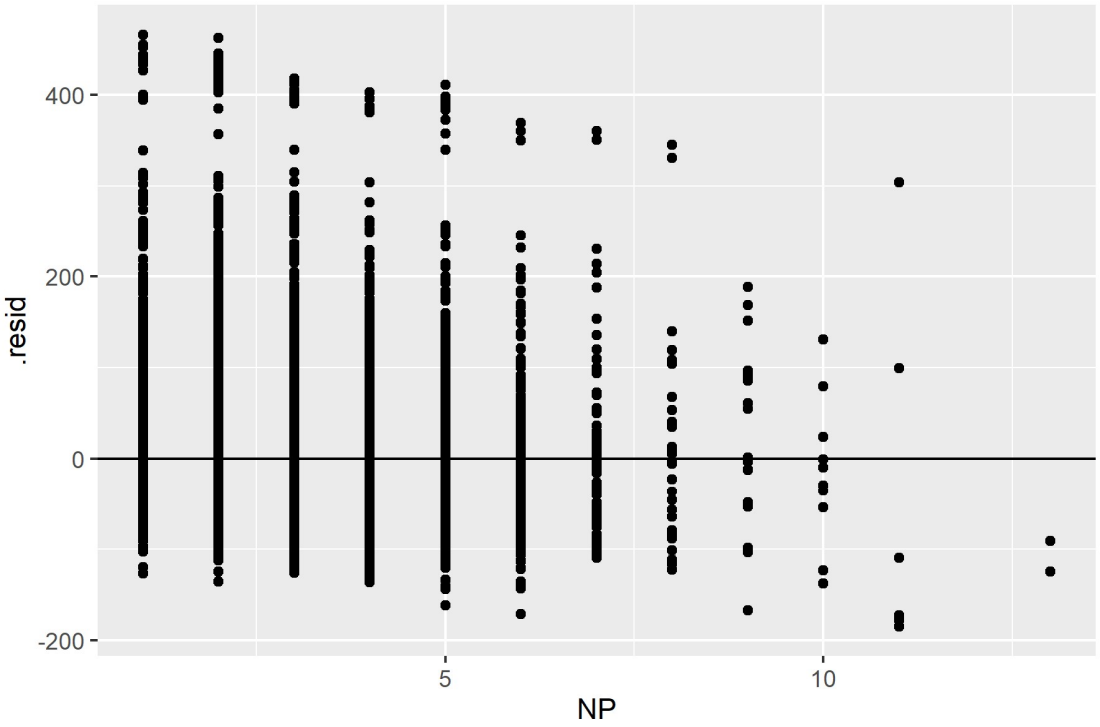
Log Fitted Values vs Residuals



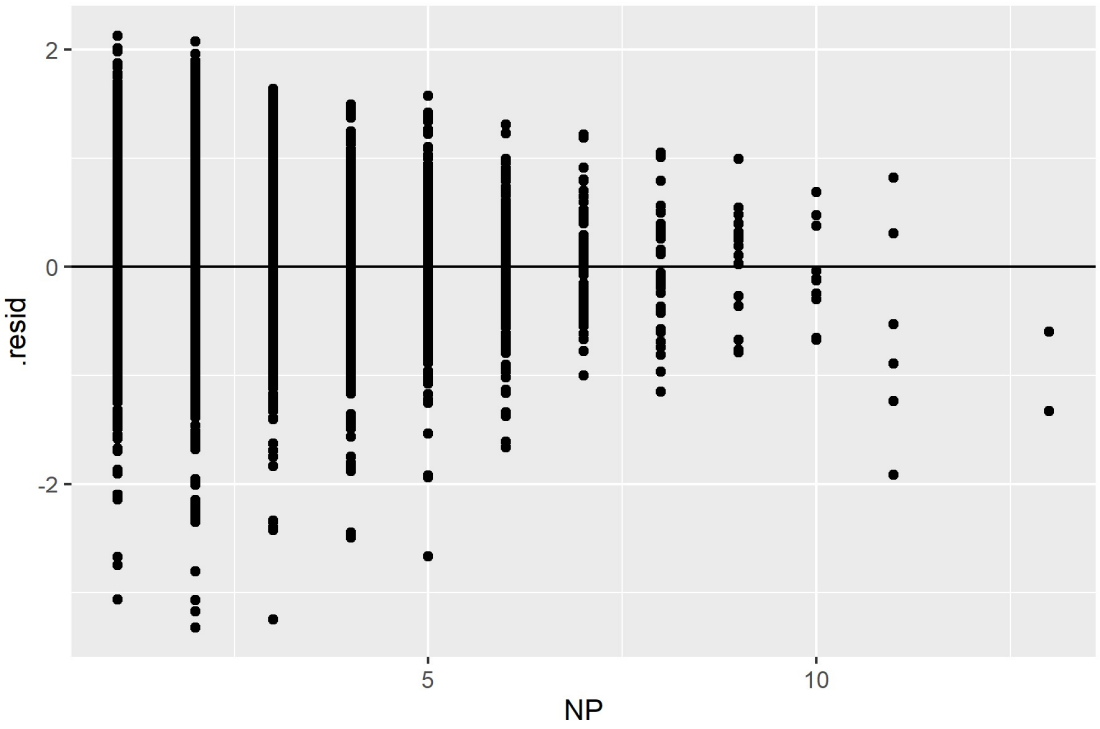
Square Root Fitted Values vs Residuals



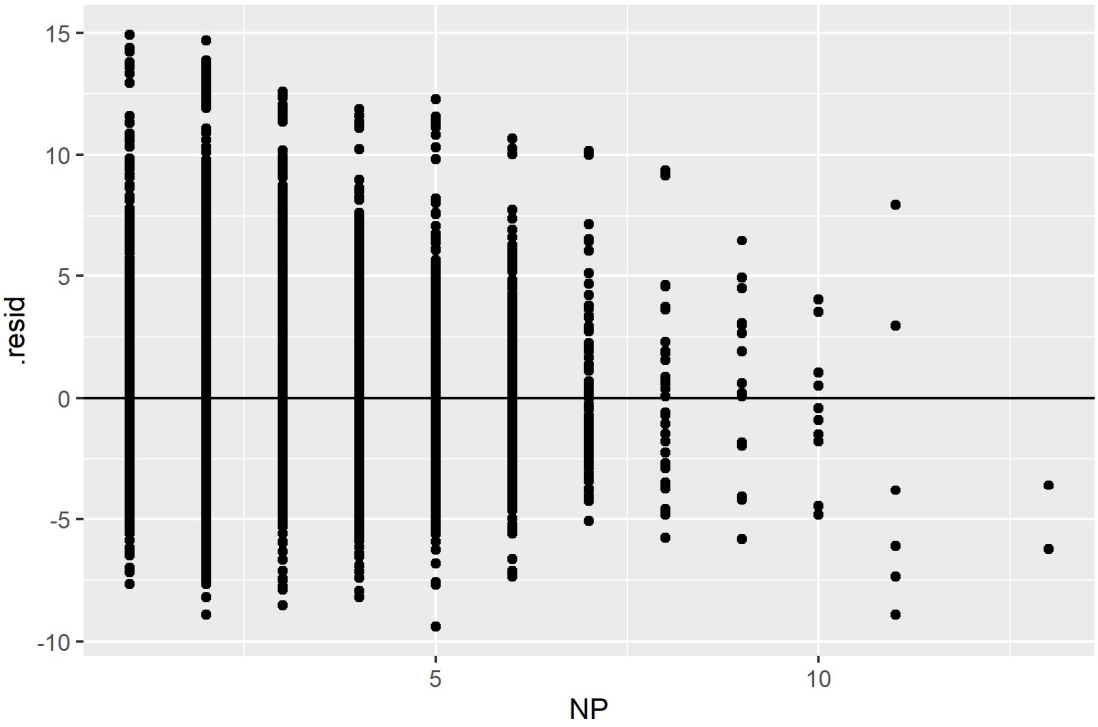
1 / Fitted Values vs Residuals



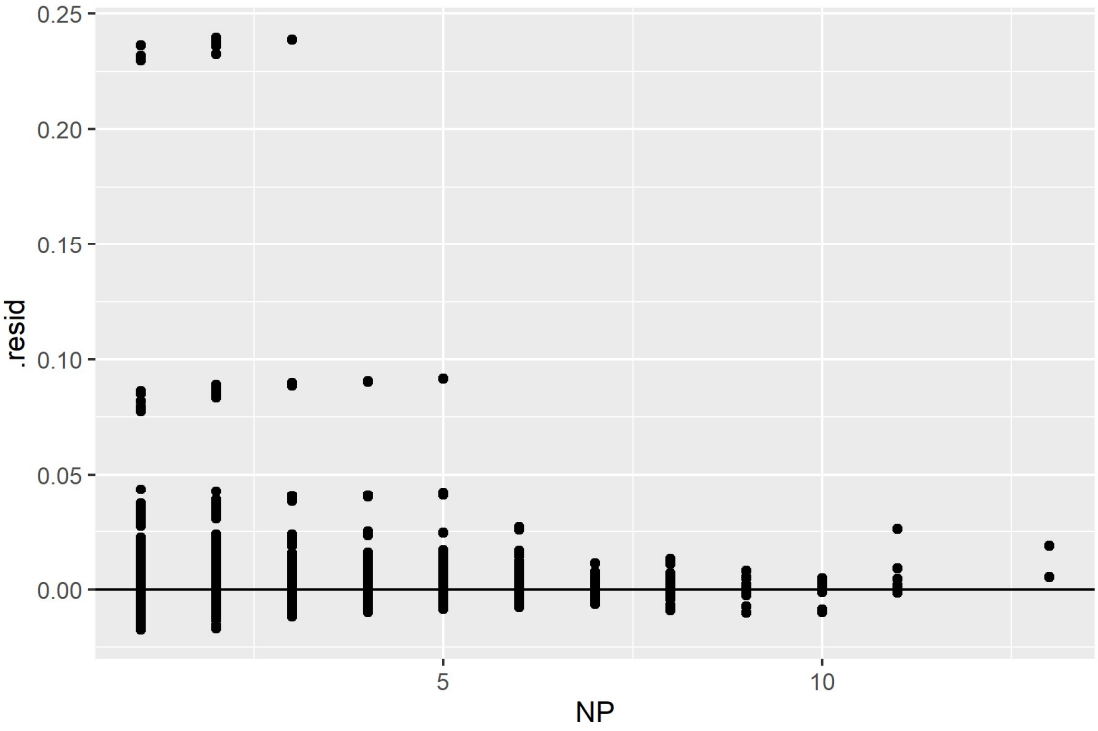
NP vs Residuals



Log NP vs Residuals



Square Root NP vs Residuals



1/NP vs Residuals

R Code

```
library(ggplot2)
library(broom)
library(leaps)
library(ISLR)

# Import data
households <- read.csv("OR_acs_house_occ.csv")

# Remove non-apartments and non-houses
households$BLD[households$BLD == "Mobile home or trailer"] <- NA
households$BLD[households$BLD == "Boat, RV, van, etc."] <- NA
households <- households[!(is.na(households$BLD)),]

# Refactor apartments and houses as 0 and 1, respectively
households$BLD <- as.character(households$BLD)
households$BLD[households$BLD == "2 Apartments"] = 0
households$BLD[households$BLD == "3-4 Apartments"] = 0
households$BLD[households$BLD == "5-9 Apartments"] = 0
households$BLD[households$BLD == "10-19 Apartments"] = 0
households$BLD[households$BLD == "20-49 Apartments"] = 0
households$BLD[households$BLD == "50 or more apartments"] = 0
households$BLD[households$BLD == "One-family house detached"] = 1
households$BLD[households$BLD == "One-family house attached"] = 1
households$BLD <- as.factor(households$BLD)


## Explanatory Analysis ##
# Fit model with interactions
fit_house_int <- lm(log(ELEP) ~ NP + BDSP + BLD + RMSP + NP*BDSP + NP*RMSP, d
ata = households)

# Get summary
summary(fit_house_int)

# Check fit
# Note that the figures below were manipulated multiple times to produce the
transformation figures which were saved to file
fit_house_diag <- augment(fit_house_int, data = households)
qplot(.fitted, .resid, data = fit_house_diag) +
  geom_hline(aes(yintercept = 0))
qplot(NP, .resid, data = fit_house_diag) +
  geom_hline(aes(yintercept = 0))
qplot(BDSP, .resid, data = fit_house_diag) +
  geom_hline(aes(yintercept = 0))
qplot(BLD, .resid, data = fit_house_diag) +
  geom_hline(aes(yintercept = 0))
qplot(RMSP, .resid, data = fit_house_diag) +
  geom_hline(aes(yintercept = 0))
```

```r
# Fit model without interactions
fit_house_noint <- lm(log(ELEP) ~ NP + BDSP + BLD + RMSP, data = households)

# Compare models with and without interactions
anova(fit_house_noint, fit_house_int)

# Get coefficients
summary(fit_house_int)$coefficients["BLD1",]

# Get Confidence Interval
confint(fit_house_int)["BLD1",]


## Predictive Analysis
# Create predict function for regsubsets
predict <- function(object, newdata, id, ...){
  form <- as.formula(object$call[[2]])
  mat <- model.matrix(form, newdata)
  coefi <- coef(object,id=id)
  xvars <- names(coefi)
  mat[,xvars]%*%coefi
}
# Create validation set
set.seed(1)
n <- nrow(households)
valid <- sample(n, size = floor(0.20*n))
households_valid <- households[valid, ]
households_train <- households[-valid, ]

# Set up cross-validation matrix
k <- 5
folds <- sample(1:k, nrow(households), replace = TRUE)
mod_bics <- matrix(NA, k, 26, dimnames = list(NULL, paste(1:26)))
cv_errors <- matrix(NA, k, 26, dimnames = list(NULL, paste(1:26)))

# for loop for calculating MSE
for(j in 1:k){
  best_fit <- regsubsets(log(ELEP) ~ NP + BDSP + BLD + FULP + GASP + RMSP + Y
BL + R18 + R60,
                         data = households[folds != j,], nvmax = 26)
  reg_summary <- summary(best_fit)
  mod_bics[j,] <- reg_summary$bic
  for(i in 1:26){
    pred <- predict(best_fit, households[folds == j,], id = i)
    cv_errors[j,i] <- mean((log(households$ELEP[folds == j]) - pred)^2)
  }
}
```

```r
# Review BICs to help determine best model
plot(mod_bics[1,], xlab = "Number of Variables", ylab = "BIC", font.main = 1,
cex.main = 0.75, main = "Figure 9: BIC vs Number of Variables", pch = 16, yli
m = c(-2300, -1100))
points(mod_bics[2,], pch = 16)
points(mod_bics[3,], pch = 16)
points(mod_bics[4,], pch = 16)
points(mod_bics[5,], pch = 16)

head(reg_summary$which[order(reg_summary$bic),])

# Calculate MSE values and plot
mean_cv_errors <- apply(cv_errors, 2, mean)
best_error <- which.min(mean_cv_errors)
se_cv_errors <- apply(cv_errors, 2, sd)/sqrt(10)
qplot(1:26, mean_cv_errors) +
   annotate("point", x = best_error, y = mean_cv_errors[best_error],
     size = 5, color = "red") +
   geom_hline(yintercept = mean_cv_errors[best_error] + se_cv_errors[best_err
or]) +
   labs(x = "Number of variables", y = "Average MSE", title = "Average MSE vs
Number of Variables")

reg_best <- regsubsets(log(ELEP) ~ NP + BDSP + BLD + FULP + GASP + RMSP + YBL
,
                       data = households, nvmax = 26)
coef(reg_best, 7)
```