Maryanne Evans
ST 517

Final Project Report

Data Analytics I
Final Project

Background

This project uses data which contains household level responses to the **American Community Survey for households in Oregon**. It is essentially the 2015 one-year Public Use Microdata Sample (PUMS) survey. I am provided a subset of variables and *only households* that have **at least one person**, pay for their electricity, and are not group accommodation. I may assume this is a *random sample* of all such households in Oregon. There are two questions to address, an explanatory problem and a prediction problem. In this report, I compare the two outcomes.

Explanatory Problem

Do people living in apartments pay less on electricity than those living in houses? How much? Make sure there are adjustments for (at least) the number of bedrooms and number of occupants in the household.

Answer

People living in apartments pay less on electricity than those living in houses in Oregon, based on the American Community Survey for households in Oregon accounted for by the 2015 one-year Public Use Microdata Sample (PUMS) survey. According to this analysis, living in an apartment opposed to a house in Oregon saves $19.47 per month, during the year 2015. With 95% confidence, it was determined that the average difference in electricity bills for apartments versus houses is between $15.96 and $22.99 per month.

Analysis

For these data, the hypothesis question is:

H_O: People living in apartments do not pay less on electricity than those living in houses.
H_A: People living in apartment pay less on electricity than those living in houses.

Immediately, we can see that there appears to be a slight difference between apartments and houses, with apartments paying less, respectively.

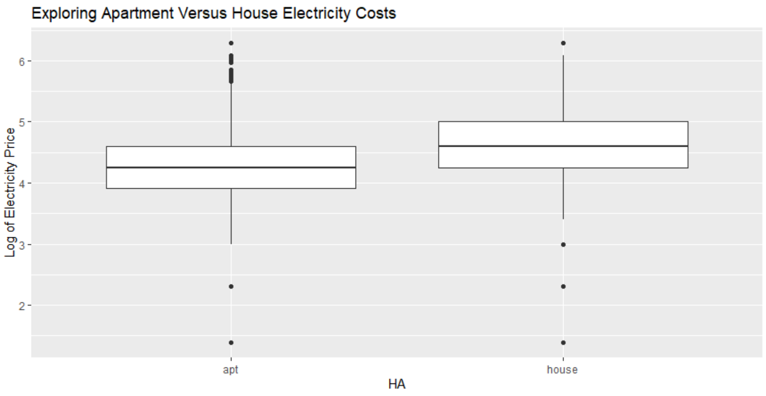


Figure 1: Boxplot exploring differences in Log Electricity Price between Apartments and Houses (HA) in Oregon.

This analysis used a **Multiple Linear Regression**. Though a Generalized Linear Model was originally considered for this analysis in the project strategy, ultimately the analysis was completed using a Multiple Linear Regression. I was concerned with the how much the dependent variable, housing type (apartment or house), was dictated by changes in the independent variables, number of people and number of bedrooms (NP and BDSP), which contributed to my modeling choice. The data was **not** transformed for the analysis. The squared residuals displayed variability when plotted, with a range from above 400 to lower than -200. However, they did appear to centralize around zero as is expected when a population estimate is captured. Furthermore, the sample size for this dataset is so large, that the power of the analysis is robust and the distribution, using the *Law of Large Numbers*, trends towards a normal distribution. The assumption of normality is not violated with these data.

Model comparisons were performed to pare down all candidate models. The models were adjusted to interpret housing options as binary, apartment, or house, and all “other” structures were removed and listed in a generated data column, attached to a subset data frame in R. A full model including all possible independent variables, after removing four problem variables from the original 17 in the dataset, and only those which did not contribute to the analysis were removed (Type, Serialno, BLD and ACR), leaving 13 variables. This was determined after reviewing the impact of ACR and BLD variable on electricity prices. Though the dataset was not subject to a transformation, the exploration boxplots were plotted with log(ELEP) to produce a more appealing visually. Additionally, for variables of interest such as NP, BDSP, ELEP, FULP, GASP, I used the cor() function in R to explore possibility for collinearity.

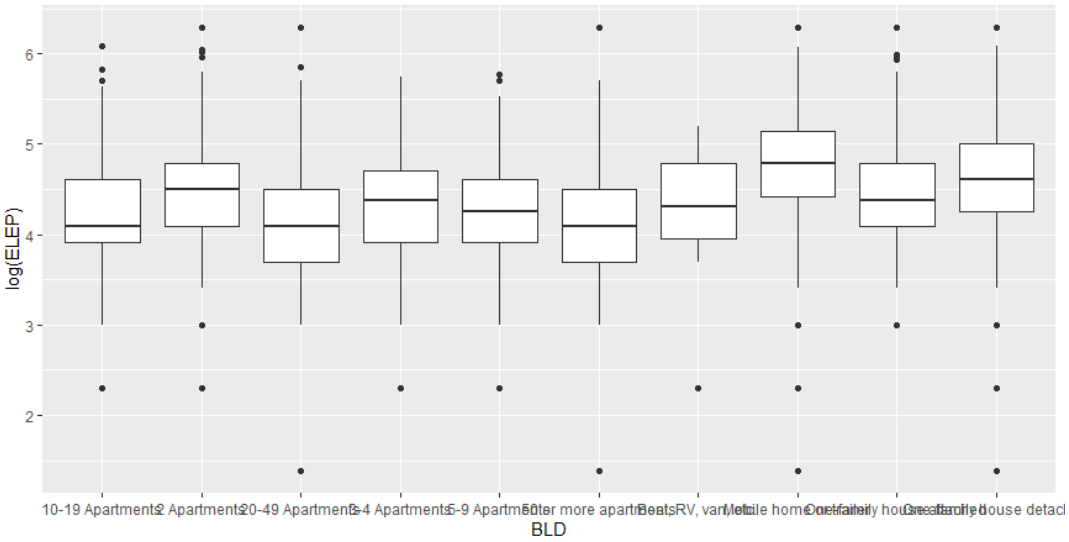


Figure 2: Boxplot exploring differences in Building Type and their impact on Log Electricity Price, an example of comparing the variable BLD to impact on electricity prices. It is not great enough to consider in the final model.

A reduced model considering only the variables of interest, number of people and number of bedrooms:

$$\mu(\text{ELEP}|\text{HA, BDSP, NP,FULP,GASP,HFL,RMSP,TEN,VALP,YBL,R18,R60 }) =$$

$$\beta_o+ \beta_1 \text{HA} + \beta_2\text{BDSP}+\beta_3\text{NP}+ \beta_4\text{FULP} +\beta_5\text{GASP} +\beta_6\text{HFL} +\beta_7\text{RMSP} +\beta_8\text{TEN} +\beta_9\text{VALP} +\beta_{10}\text{YBL} +\beta_{11}\text{R18} +\beta_{12}\text{R60}$$

This model can generate computed results. The problem, however, is that the output is more complex than necessary to answer the question and reduces the robustness of the analysis because it does not minimize the sum of squares residuals

So, a reduced fitted model was made using the fewest possible parameters to answer the hypothesis question:

$$\mu(\text{ELEP}|\text{HA, BDSP, NP}) =$$

$$\beta_o+ \beta_1 \text{HA} + \beta_2\text{BDSP}+\beta_3\text{NP}$$

The reduced fitted model is the chosen model for this analysis

However, exploring potential for interaction is important because dependency in a model can produce misleading results for significant correlations. Therefore a reduced fitted model with interaction terms was made to explore the possible dependency between NP and BDSP.

$$\mu(\text{ELEP}|\text{HA, BDSP, NP,FULP,GASP,HFL,RMSP,TEN,VALP,YBL,R18,R60 }) =$$

$$\beta_o+ \beta_1 \text{HA} + \beta_2\text{BDSP}+\beta_3\text{NP}+ \beta_3(\text{BDSP*NP})$$

When comparing these two models by performing an anova which provided the Extra Sum of Squares test needed for model comparisons, it was found that there was no significant difference between the models (p-value= 0.122). **Meaning, the reduced fitted model with no interactions, the simple model, makes the most sense for this analysis.** Additionally, the variance inflation factor (VIF) was explored for each model, producing modest results that suggest that there is no multicollinearity present in the final model, and interaction terms are not necessary.

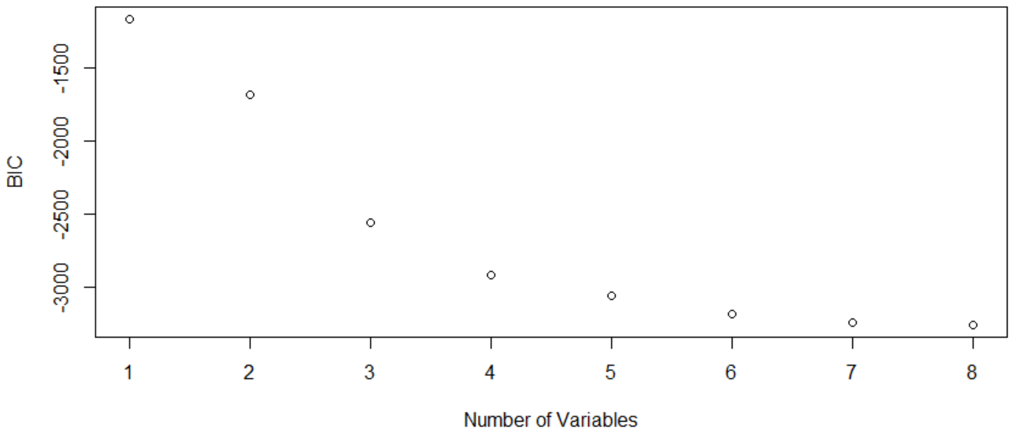
In conclusion, the null hypothesis (H_o) was rejected. The alternative hypothesis (H_A) is supported with a reduced fitted model assessing electricity price as a response variable to the explanatory variables of housing as a binary between a house and an apartment (HA), number of people (NP), and number of bedrooms (BDSP), with no interaction terms because there is no collinearity to address in the final model. It was determined that there is a price difference between living in a house versus an apartment in Oregon during the year of 2015, while adjusting for number of occupants and bedrooms.

Prediction Problem

Create a model that could be used to predict electricity costs for a household in Oregon.

The model ultimately used for the predictive model resembled the Multiple Linear Regression Model used to answer Question 1. The difference was that in addition to removing variables TYPE, SERIALNO, BLD and ACR, VALP was also removed. These were variables that were causing disruptive correlations in the analysis.

After curating the data for computation, I split the data into a training and an observation subset. From there, I calculated the BIC (-1165.115), CP (2440.5468) and RSS (69967607). These were calculated using the regsubset full model, with all considerable variables.



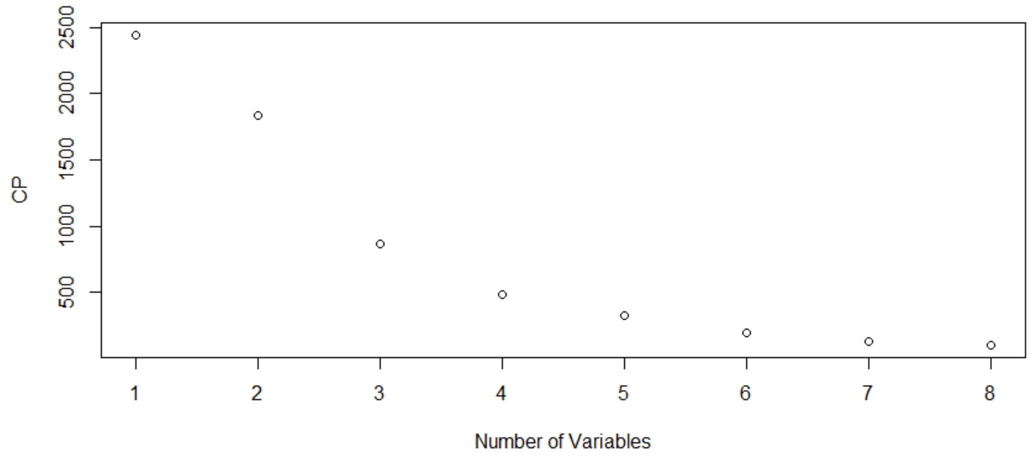


Figure 3: Two plots depicting BIC and CP as the response variables, plotted against the possible explanatory variables.

As discussed in my project summary, I started by curating the data to a manageable data frame. Like Question 1, I used HA as a binary addition to account for house and apartment and removed all "other" shelters. Electricity price remained the response variable. The major difference for this model, as far as I can take it that is, was including all 12 possible variables for model construction.

The reason for this is my R code was throwing errors for the fitting the predictive model with the imputed dataset. Because of this, I was unable to generate multiple predictive models for comparisons using Cross Validation of K-folds of ten, which was my plan. The model that would ultimately be chosen for this analysis would be selected after applying the Cross Validation for model comparison and selecting the model with the smallest possible Mean Squares Errors (MSE), measured using the BIC score, the smallest score suggesting the best fit model. Based on my knowledge of the dataset, these data will still meet the assumptions of normality given the large sample size and thus will not undergo a transformation for the analysis. These data would likely be *brozek*.

The reason it is important to compare the observed data with the training data is to test for overfitting the model. That is one of the largest limitations, aside from computational complexities, regarding predictive models of this nature. However, as we learned in lecture, linear models are predictive, but they cannot necessarily account for new observations, they are limited to set datapoints. It is worth emphasizing that this predictive model aims to generate values for *new* observations, while minimizing MSE, but maximizing fit to represent population level estimates.

Compare and Contrast

Discuss the differences in your approach to Questions 1 and 2. Why are different approaches required? What challenges did you face, and how did they compare across the two tasks?

Why are different approaches needed?

The approaches for Questions 1 and 2 are somewhat similar in theory, but with one important distinguishing objective: *valid observations* collected versus application to *new observations*.

Question 1 is asking for estimates from a limited dataset. The Multiple Linear Regression can compute values within the framework of the available data but cannot make computational predictions for *new* observations.

Question 2 is asking for a model that can predict *new* observations from using observed values. For these types of models, there is a risk of overfitting the data to the observed dataset. This is stemming from the model generating a vector of explanatory values and producing a predictive response. It requires fitting a *training data set* and *observed data set.* Possible limitations can emerge regarding the dataset itself. Data points that are deleted to create the subset test model, if done incorrectly, may challenge the validity of the model's predictive applications, which is another limitation that requires careful consideration.

In short, this means that the **predictive accuracy is different between the models**. Additionally, predictive models that require imputations, validation/test datasets and cross validation approaches are incredible computationally demanding.

Differences in Approach

There were similarities and differences in the approach to both questions. Similarities included the need for model construction, comparison and using a scoring metric to determine best fit. Also, the need to fit residual and reduce error noise is shared. That being said, the major differences in approach were namely the need to generate a training dataset for the predictive model and imputing the values.

What challenges did you face, and how did they compare across the two tasks?

I found the predictive model approach incredibly computationally draining, and time consuming. I felt as though it was much harder to conceptually grasp the steps, and that was reflected in my coding. I still do not fully understand what the code is outputting. For example, I found variable reduction and model construction for the predictive model extremely challenging. It was hard to understand how many variables were being considered, and which ones to focus on for the model. Much of my confusion began at the regsubsets, and then the need to impute the data? I was having a very hard time keeping track of which dataset belong with which coding endeavor. It turns out, once you have a training set and observational set, it is hard to keep track of which is which. This was also reflected in my Lab/HW 8 attempts.

This may have been complicated by my challenges to get the R code to run without developing errors at various stages. Perhaps, I am general unfamiliarity with the code. I am more practiced with linear models within a frequentist framework, so overall, I found Question 2 notably more challenging.

R Code Appendix

```
#Evans Final Project
#American Community Survey for households in Oregon
#set up library packages
library(ggplot2) #viz
library(tidyr) #viz
library(car) #standard package
library(leaps)
library(broom)
library(Amelia)
#-----------------------------------------------------------------------------------------------
#Question 1: Do people living in apartments pay less on electricity than those living in houses?
#How much? Make sure there are adjustments for (at least) the number of bedrooms and number of
#occupants in the household.
#-----------------------------------------------------------------------------------------------
#Import Data
HouseData1 <- read.csv("C:/Users/Marya/OneDrive/Desktop/OSU/ST517 (Data I )/Final Project/OR_house_data.csv",
                    stringsAsFactors=TRUE)
#Modifying the dataset to read as house, apartment and other, with removal of "other" for new column HA
HouseData$HA <- "other"
HouseData$HA[which(grepl("house", HouseData$BLD, ignore.case=TRUE))] <- "house"
HouseData$HA[which(grepl("apartment", HouseData$BLD, ignore.case=TRUE))] <- "apt"
HouseData <- HouseData[!(HouseData$HA == "other"),]
HouseData$HA <- factor(HouseData$HA)
#Looking at House Versus Apartment in Electricity Price
qplot(HA, ELEP, xlab="HA", ylab="Electricity Price", main="Exploring Apartment Versus House Electricity Costs"
      , data = HousePlease, geom = "boxplot")
#Create Multiple Linear Regression Models
#Identify the response variable as "ELEP" electricity, and the explanatory variables need to be explored.
#Full model without interactions
```

```
Mod_Bod <- lm(ELEP ~ ., data = HousePlease)
summary (Mod_Bod)
vif(Mod_Bod)
#Explore the raw data
#Variable ACR
qplot(ACR, log(ELEP), data = HouseData, geom = "boxplot")
#Variable BLD
qplot(BLD, log(ELEP), data = HouseData, geom = "boxplot")
#These plots look extremely similar, at a visual glance. Meaning ACR does not explain a lot about electricity prices.
#Also, there is a presence of leverage points, outliers or otherwise problematic data, which means that the model
#will likely perform better with this variable removed. It is important to find the least complex model,
#and maximize the information added as explanatory variables.
#The pattern here suggests that there may be collinearity or dependence in the data.
#Exploration Continues to look at Collinearity/Multicollinearity
#Subset HouseData to exclude problem variables that do not contribute to the analysis
HousePlease <- subset(HouseData, select = -c(ACR, BLD, SERIALNO, TYPE))
#Look at the remaining variables
str(HousePlease)
ncol(HouseData) #original dataset is 17 variables
ncol(HousePlease)#demonstrates dataset was reduced to 13 variables
#Examine variables for collinarity; NP, BDSP, ELEP, FULP, GASP
cor(HousePlease[,1:5]) #outcomes are modest
#Reduced Model Without Interactions. Knowing the model must account for HA, NP and BDSP
Mod_Fit_Skinny <- lm(ELEP ~ HA + BDSP + NP, data = HousePlease)
summary(Mod_Fit_Skinny) #p-values show a statistically significant relationship, suggesting these variables remain
Skinny <- vif(Mod_Fit_Skinny) #variance inflation factor
barplot(Skinny, main = "VIF Scores", col = "purple") #viz
#Model with Interaction Terms
Mod_Fit_Social <- lm(ELEP ~ HA + BDSP + NP + BDSP*NP, data = HousePlease)
#Residual Plots for Mod Fit Social, Model with Interactions
Fit_Aug <- broom::augment(Mod_Fit_Social, data = )
```

```
qplot(.fitted, .resid, data = Fit_Aug)

qplot(NP, .resid, data = Fit_Aug) +

  geom_hline(aes(yintercept=0))

qplot(BDSP, .resid, data = Fit_Aug) +

  geom_hline(aes(yintercept=0))

nrow(MLB_AugB)

#ANOVA compares models. AIC and BIC values comparison of this proposed model and the reduced model

anova(Mod_Fit_Skinny, Mod_Fit_Social) #large p-value, no significant difference between models

#when there is not a significant difference, choose the simple model

#comparing model without interaction terms to interaction model

AIC(Mod_Fit_Skinny, Mod_Fit_Social)

BIC(Mod_Fit_Skinny, Mod_Fit_Social)

#Estimate calculations and Confidence Intervals for selected model (Mod_Fit_Skinny)

summary(Mod_Fit_Skinny)$coefficients

confint(Mod_Fit_Skinny)

#---------------------------------------------------------------------------------------------------------------------

###QUESTION 2###PREDICITIVE MODEL

#---------------------------------------------------------------------------------------------------------------------

#set up library packages

library(boot)

library(leaps)

library(ggplot2) #viz

library(tidyr) #viz

library(car) #standard package

library(Amelia)

library(broom)

library(faraway)

#import dataset, new name

HousePred <- read.csv("C:/Users/Marya/OneDrive/Desktop/OSU/ST517 (Data I )/Final Project/OR_house_data.csv"
                    ,stringsAsFactors = TRUE, skipNul = TRUE)

#Modifying the dataset to read as house, apartment and other, with removal of "other" for new column HA

HousePred$HA <- "other"

HousePred$HA[which(grepl("house", HousePred$BLD, ignore.case=TRUE))] <- "house"

HousePred$HA[which(grepl("apartment", HousePred$BLD, ignore.case=TRUE))] <- "apt"

HousePred <- HousePred[!(HousePred$HA == "other"),]

HousePred$HA <- factor(HousePred$HA)

#Remove problem variables detected in question 1

HousePlease <- subset(HousePred, select = -c(ACR, BLD, SERIALNO, TYPE, VALP))

#set aside 20% of the data for validation data

set.seed(385)

n <- nrow(HousePlease)

observed <- sample(n, size = floor(0.20*n))

house_obs <- HousePlease[observed, ]

house_train <- HousePlease[-observed,]

#using Amelia to impute data

set.seed(123)

n.imp <- 50 # m=50, Set the number of imputed datasets

imputed <- amelia(house_train, m=n.imp, p2s=0, idvars="HA")

imputed

#subset selection

reg_house_fit <- regsubsets(ELEP ~ ., data= HousePlease, really.big=TRUE, method="forward")
#dependencies detected

reg_sum <- summary(reg_house_fit)

reg_sum

names(reg_sum)

reg_sum$rss #RSS values for best subset

#plotting
```

```
which.min(reg_sum$bic) #9

plot(reg_sum$bic, xlab="Number of Variables", ylab="BIC")

#fitting predictive model with imputed data

betas <- matrix (0, nrow=n.imp, ncol = 12)

ses <- matrix (0, nrow=n.imp, ncol= 12)

for (i in 1:n.imp){

  NewFit <- lm(ELEP ~ . , HousePlease = imputed$imputations[[i]])

  beta[i,] <- coef(NewFit)

  ses[i,] <- coef(summary(NewFit))[,2]

}

#Evaluating models with k-fold (k=10) cross-validation
```