Project 1 - ST595

Inferential vs. Predictive Analysis for 2015 Public Use Microdata Sample (PUMS)

Ruslan Mamedov

2022-05-30

Introduction

We are given a set of of anonymized census data and asked to conduct two types of statistical analysis to answer the following questions:

1. Inferential analysis: Do people living in houses pay more on electricity than those living in apartments? How much? Make sure you adjust for (at least) the number of bedrooms and number of occupants in the household.
2. Predictive analysis: Create a model that could be used to predict electricity costs for a household in Oregon.

We are then asked to compare and contrast the approaches across the two tasks.

Data Description

We are provided with the .csv file 'or_acs_house.csv' which contains household level responses to the American Community Survey for households in Oregon. The dataset is a Public Use Microdata Sample (PUMS) from the 2015 1-year survey obtained from http://www2.census.gov/programs-surveys/acs/data/pums/2015/1-Year/. It contains only households that have at least one person, pay for their electricity, and are not group accommodation, and it may be assumed this is a random sample of all such households in Oregon.

Statistical Modeling

1. Explanatory Problem: Do people living in houses pay more on electricity than those living in apartments, after accounting for the confounding variables? If so, by how much?

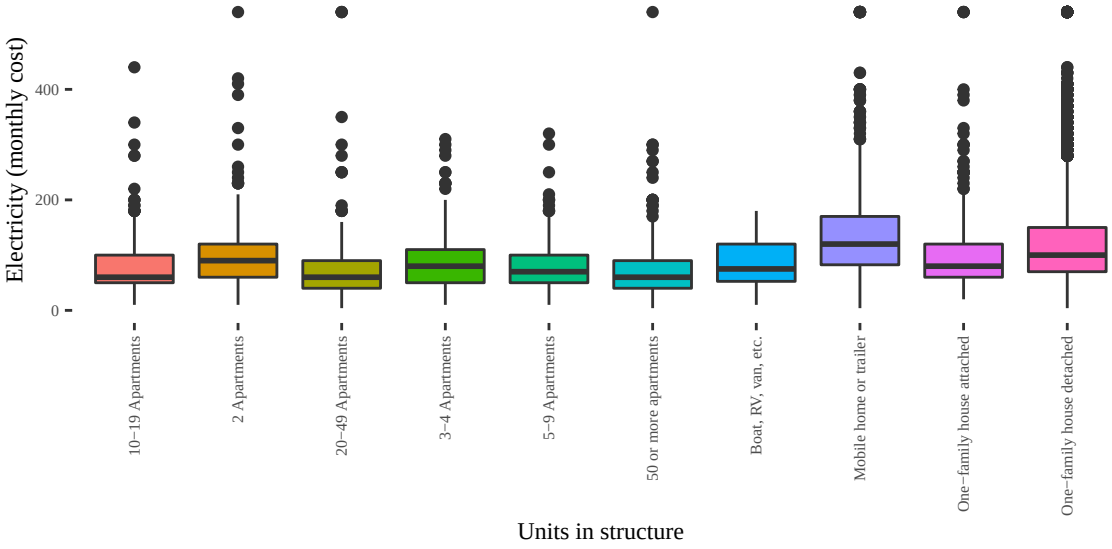
H~0: The mean electricity payments for people living in houses is not greater than for those living in the apartments, after controlling for differences in the other proposed explanatory variables.

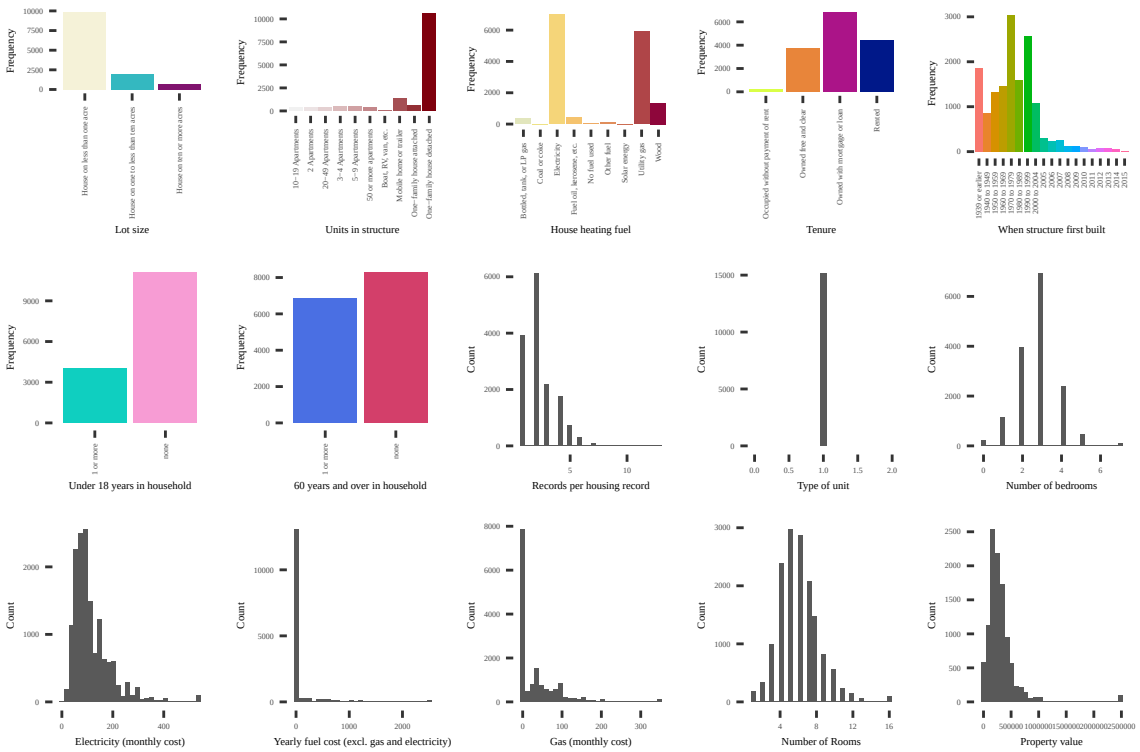
H~A: The mean electricity payments for people living in houses is more than for those living in the apartments after controlling for differences in the other proposed explanatory variables.

Because this is a question of statistical inference, we will fit a model to the data, test the model to see if it satisfies the linear model assumptions, and then interpret it by determining the statistical significance of the coefficients for our proposed explanatory variables. We will determine p-value for the explanatory variable of interest.

Serial number	Records per housing record	Type of unit	Lot size	Number of bedrooms	Units in structure	Electricity (monthly cost)	Yearly fuel cost (excl. gas and electricity)	Gas (monthly cost)	House heating fuel	Number of Rooms	Tenure	Property value	When structure first built	Under 18 years in household	60 years and over in household
70	4	1	House on less than one acre	2	One-family house detached	70	2	3	Wood	4	Rented	NA	1939 or earlier	1 or more	none
163	2	1	House on less than one acre	2	One-family house detached	100	600	3	Fuel oil, kerosene, etc.	7	Owned with mortgage or loan	225000	1939 or earlier	none	none

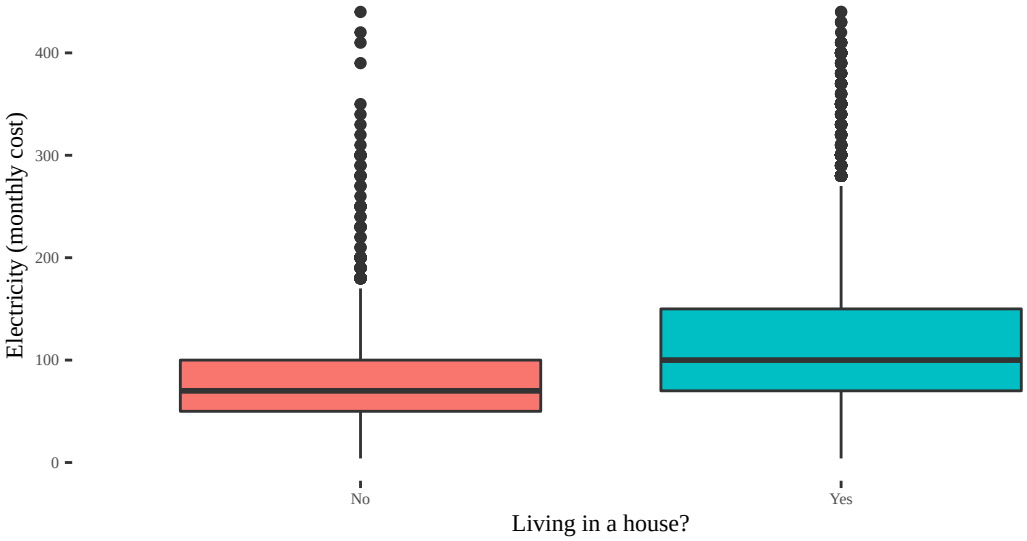
We have quite a large dataset with over 15000 observations and a mix of numerical and categorical values. First, let's have some exploratory analysis done: spread of electricity bills around different building types, distribution of unique values for both numerical and categorical columns, and summary of missing values.





Serial number	Records per housing record	Type of unit	Lot size	Number of bedrooms	Units in structure	Electricity (monthly cost)	Yearly fuel cost (excl. gas and electricity)	Gas (monthly cost)	House heating fuel	Number of Rooms	Tenure	Property value	When structure first built	Under 18 years in household	60 years and over in household
0	0	0	2586	0	0	0	0	0	0	0	0	4632	0	0	0

Some conclusions: 1. Most of the records are for one-family detached houses.
2. Number of bedrooms/rooms have a normal distribution.
3. Property value and Electricity monthly cost have right-skewed disptributions.
4. Type of unit has min=max=1. The dataset includes only housing units and not group quarters which we shall account for when interpreting the results.
5. Electricity monthly cost and couple of other columns seem to have outliers (e.g. ELEP=540) not associated with other variables and most likely representing some cut-off value arbitrarily chosen for very large electric bills. Outliers in electricity column are pretty evenly distributed between housing vs. apartment units, small in numbers and most likely do not represent the actual bills, so we can leave them out of inference analysis since these values are much less accurate than the rest of the data and might skew the model. We'd want to have those outliers back when training the prediction model though.
6. The missing values are only for two columns: "Property value" (roughly 30% of the data) and "Lot size" (about 17% of the rows). We'll have to deal with those later when working on predictive modelling.
7.The boxplots do appear to hint on different electric bills depending on the building type but lets let's group those units on whether they could be considered a house or an apartment. We'll exclude mobile homes, trailers, boats, RVs and vans from the analysis.



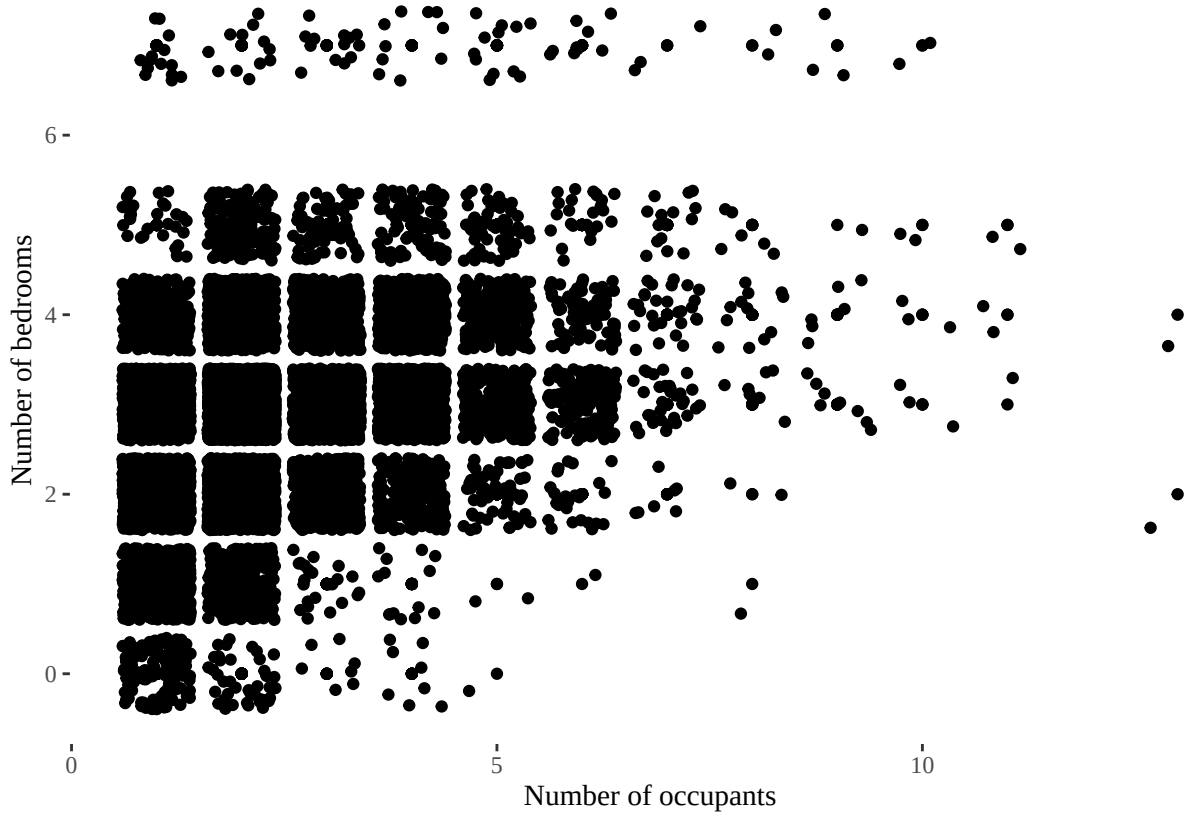
There appear to be some difference in electricity bills depending on whether the people live in a house. Now we'll build inferential model to test that hypothesis and estimate the differences. We'll adjust for the following variables: 1. Number of bedrooms (numerical).
2. Number of occupants (numerical).
3. Year the structure was first built (converted to categorical variable as before or after 2000 to account for the modern energy-efficient home designs).
4. House heating fuel as electricity vs. other (categorical).
5. Presence of persons under 18 years in household (categorical).
6. Interaction term for the number of bedrooms when electricity is used as a heating fuel.

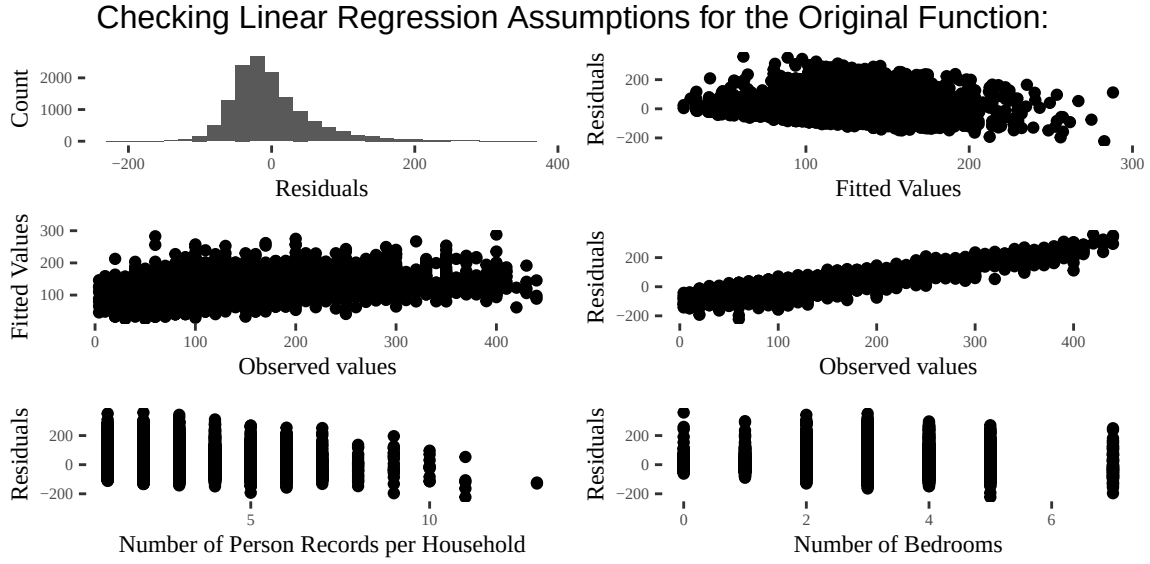
First, let's check if our interaction term is justified. We can compare two models (with vs. without interaction) using extra SS F-test:

```
## Analysis of Variance Table
##
## Model 1: ELEP ~ NP + BDSP + HEATING + YBL + HOUSE + R18
## Model 2: ELEP ~ NP + BDSP + HEATING + YBL + HOUSE + R18 + HEATING:BDSP
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1  13681 48139532
## 2  13680 47949095  1    190437 54.332 1.79e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

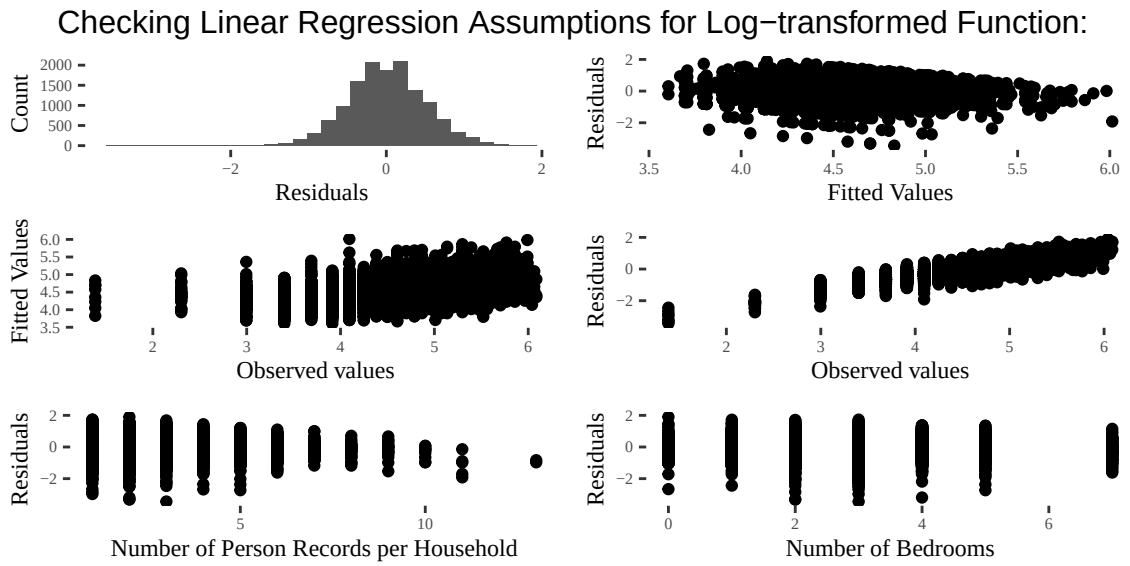
There is convincing evidence that our response variable is associated with at least one of interaction terms, (p = 1.51e-13, extra sum of squares F-test on 6 and 13675 degrees of freedom). We'll keep the interaction term in the model.

Next, let's perform collinearity test between numeric columns and residual diagnostics to check the model satisfies regression assumptions:





There's not much collinearity between the numerical columns. As for the regression assumption, the histogram of the residuals is skewed to the right which violates the normality assumption, although linear regression models are typically robust to this violation, especially for the large sample sizes that we have here. Also, observed vs. fitted values doesn't show linear relationship between two variables. Let's transform the response with log operator and see if it helps with the residuals distribution:



There's still some skewedness in residuals distribution for very high and very low electricity bills but it's much better with log transform and should be good enough for the inference task. Also, the fitted vs. observed value correlation does appear to be linear.

```
##
## Call:
## lm(formula = log(ELEP) ~ NP + BDSP + HEATING + YBL + HOUSE +
##     R18 + HEATING:BDSP, data = data <- data.elec)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4503 -0.3112  0.0115  0.3209  1.9581
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.610477   0.029096 124.089  < 2e-16 ***
## NP            0.116781   0.004817  24.241  < 2e-16 ***
## BDSP          0.082659   0.006540  12.639  < 2e-16 ***
## HEATING       0.184639   0.028805   6.410 1.50e-10 ***
## YBL          -0.088912   0.011701  -7.598 3.19e-14 ***
## HOUSE         0.329961   0.014680  22.477  < 2e-16 ***
## R18none       0.053441   0.013993   3.819 0.000135 ***
## BDSP:HEATING  0.049556   0.009438   5.251 1.54e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5086 on 13680 degrees of freedom
## Multiple R-squared:  0.2337, Adjusted R-squared:  0.2333
## F-statistic: 595.9 on 7 and 13680 DF,  p-value: < 2.2e-16
```

```
##                    2.5 %      97.5 %
## (Intercept)   3.55344513  3.66750875
## NP            0.10733804  0.12622382
## BDSP          0.06983961  0.09547748
## HEATING       0.12817805  0.24110026
## YBL          -0.11184774 -0.06597558
```

```
## HOUSE         0.30118650  0.35873455
## R18none       0.02601291  0.08086880
## BDSP:HEATING  0.03105582  0.06805597
```

Since we are using the log-transformation, the coefficient only represented change of the ratio of electricity cost between house and apartment. Thus, we'd need to do some transformation of the response to get the confidence interval.

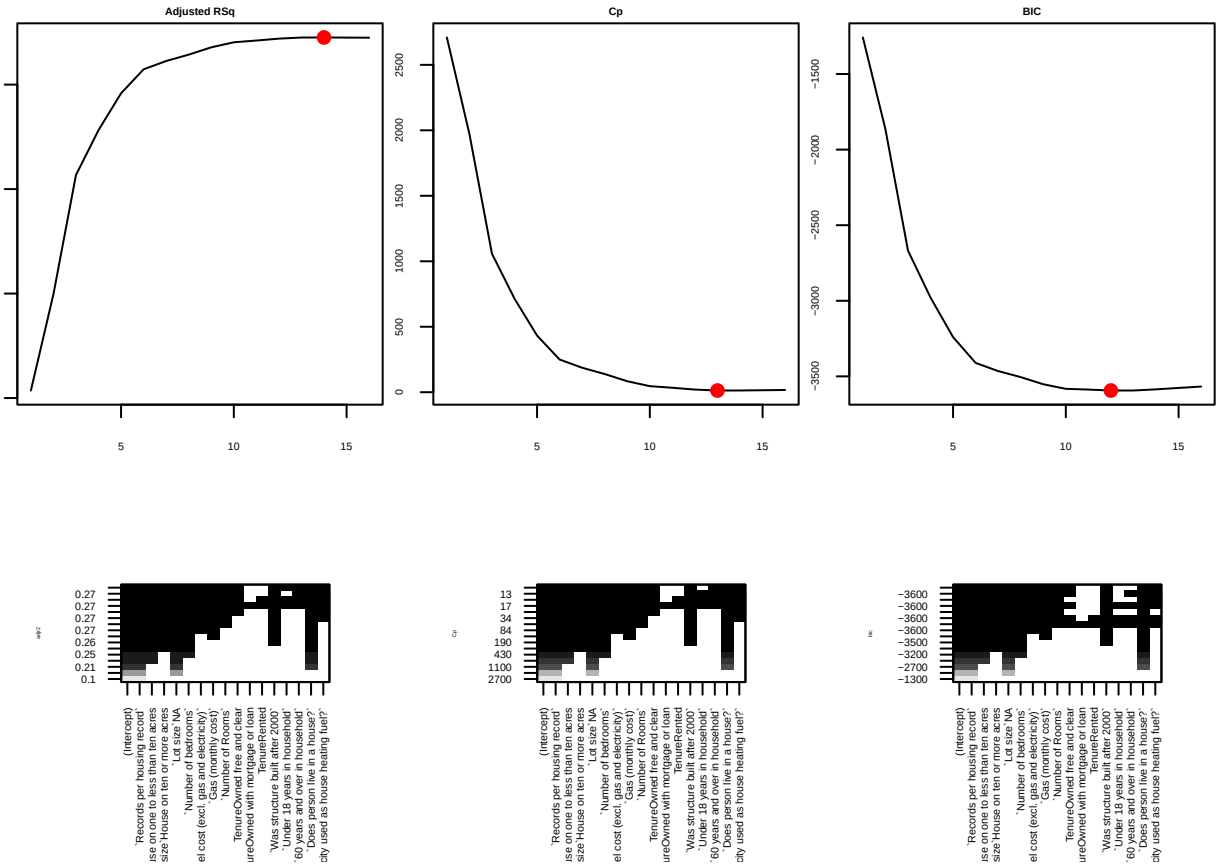
Results There is convincing evidence that people living in houses in Oregon state in 2015 were on average paying more for electricity than those living in apartments (p-value $<$ 2.2e-16), afteraccounting for the number of bedrooms, number of occupants in the household, whether the dwelling was built after 2000, whether it is using electricity for heating, for the presence of persons under 18 years in household, and the number of bedrooms when electricity is used as a heating fuel. It is estimated that living in a house increased the mean electricity monthly cost by 39.09%. With 95% confidence, the mean increase in electricity monthly cost was between 35.15% and 43.15%. The model was based on Public Use Microdata Sample (PUMS) from the 2015 1-year survey in Oregon and only included households that have at least one person, pay for their electricity, and are not group accommodations. The inference shouldn't be extrapolated to the dwellings with very high electricity bills ($540 and above) and institutional/non-institutional group quarters as those were excluded from the analysis.

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

First, we'll clean up the data to get it ready for the modeling. Let's explore that VALP and ACR columns which had missing values. As expected, missing values in "Lot size" column are associated with people who don't live in houses. So we add those NA values as an extra factor to the column. NAs in "Property value" column might correspond to the properties which are either rented or occupied without payment or rent(30%). We can't find a suitable substitute for NAs in this case so we'll omit the whole column. We'll also remove SERIALNO and TYPE columns as irrelevant ones and "Units in structure" and "House heating fuel" columns to their derivatives ("Does person live in a house?" and "Is electricity used as house heating fuel?"). We'll also convert R18 and R60 columns to numeric values.

Let's split the data into train/validate/test (80%/10%/10%) and proceed with the modelling.

We'll use log transformation as it showed to have a better alignment with the regression model assumptions. Let's perform a feature selection analysis on the training set using best subset regression techniques with BIC, CP and Adjusted R2 as the metrics of choice.



Out of 16 variables and intercept, Adjusted Rsq chose 14 variables (excluding "Tenure rented" and "Tenure owned with mortgage and loan"). Cp would exclude "Under 18 years in household" along with two factors of TEN column. BIC model would also exclude "60 years and over in household" (12 variable model).

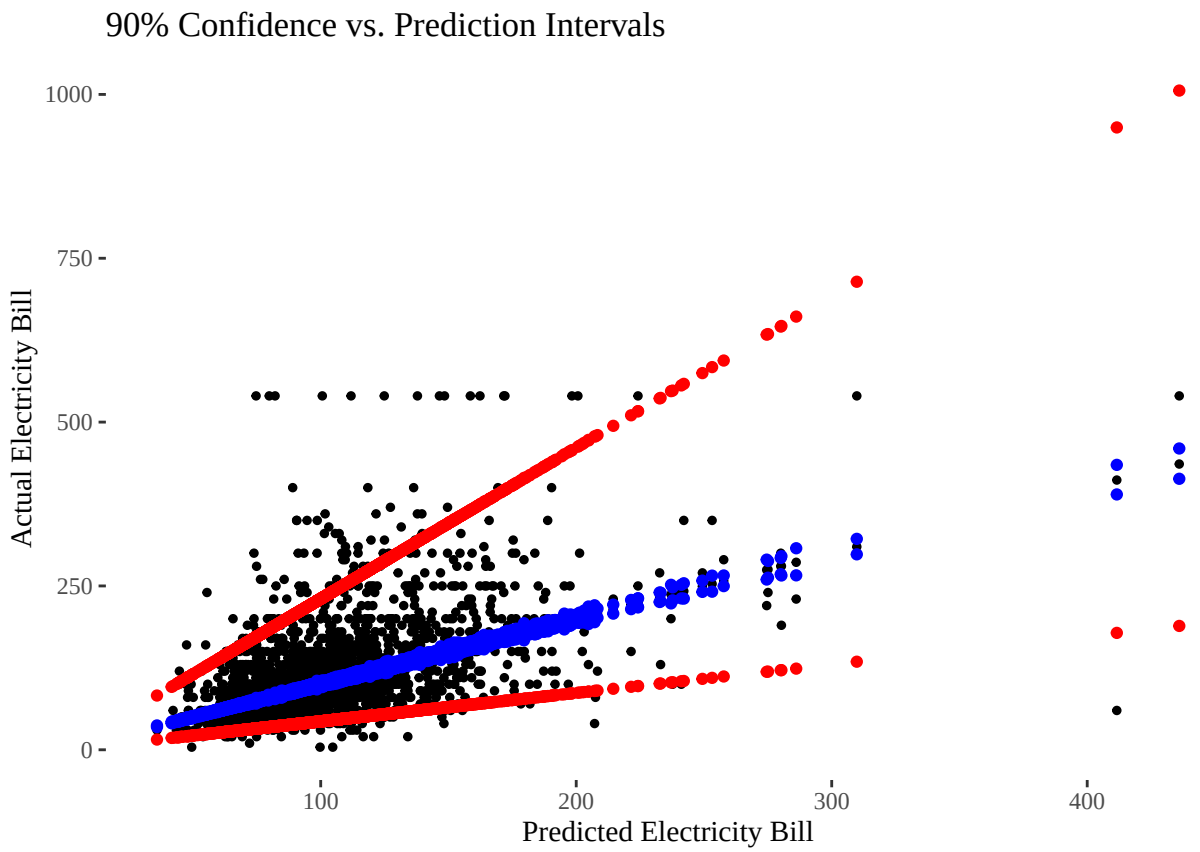
Now let's do a 10-fold cross-validation to check which one of the models produces the least error (MSE) on the validation set and hence is the best model.

Metrics used for variable selection	Number of variables in the model	CV Error on testing data
Adjusted R-squared	14 variables	0.4418096
Cp	13 variables	0.4410038
BIC	12 variables	0.4421923

The model devised with CP metrics shows the smallest CV error.Let's use the full dataset to train the final model and predict electricity costs for a household in Oregon.

```
## [1] "log(ELEP)~ 3.90809+0.11059*`Records per housing record`+0.26096*`Lot size`House on one to "
## [1] "less than ten acres+0.28178*`Lot size`House on ten or more acres+-0.3706*`Lot size`NA+0.06"
## [1] "12*`Number of bedrooms`+0.00013*`Yearly fuel cost (excl. gas and electricity)`+8e-04*`Gas "
## [1] "(monthly cost)`+0.02058*`Number of Rooms`+-0.08984*`Was structure built after 2000`+0.0328"
## [1] "4*R60+0.35008*`Is electricity used as house heating fuel?`+-0.08395*`Does person live in a"
## [1] " house?`+-0.05646*`Tenure owned and free and clear`"
```

Finally, we'll run a prediction on the test dataset and calculate its RMSE.



Results Here we have a model predicting monthly electricity cost for a household in Oregon. The cross-validation estimate of prediction error is 0.441 and RMSE for test set prediction is 0.8473 on a log scale. The model is based on Public Use Microdata Sample (PUMS) from the 2015 1-year survey in Oregon and only includes households that have at least one person, pay for their electricity, and are not group accommodation. The property value column was excluded from the analysis due to missing 30% of the cases which might have affected its accuracy. Adjusted R-squared value for the model is only 0.274 and RMSE is quite large which might indicate that a linear regression model is not the best choice for the data at hand.

Conclusion

Below we tabulated the two approaches employed for the analysis above: drawing inferences vs. making predicitons.

Explanatory Model	Predictive Model
Primary goal is answering the question on difference in electricity bill for houses vs. apartments.	Primary goal is having an accurate estimate of monthly electricity costs as well an estimate for the accuracy of the model.
The strategy is to pick only those variables and interactions which are meaningful/needed for model interpretation (variables of direct concern). Limit the number of models you compare and choose the model to answer the scientific question you care about. The more models you compare the more likely it is that you will find a possibly spurious relationship	The strategy is to choose the set of variables that will lead to a better prediction performance. If your goal is to make predictions of the response (or mean response), it doesn't hurt to fit and compare a large number of models, but you should use some objective way of selecting the final model (e.g. smallest AIC, smallest BIC, lowest prediction error on independent set of data)
No concern for missing values and problematic columns as long as they are not in the selected variables.	Need to clean up the data (missing values etc.) prior to selection.
Need to avoid complicating the model interpretation with too much column transformations.	All column transformations and feature engineering are allowed as long as they improve the accuracy of prediction.
Cannot trust inference after variable selection (using the same data twice). Otherwise, we will be getting small p-values from the sample becuase we selected for those small p-values. That is, the process of variable selection invalidates the properties of the p-values and confidence intervals.	Variable selection is the major step in model optimization.
Typically selecting between only a couple of models (e.g. one with and without interactions) to avoid model selection bias.	Comparing lots of models with best subset selection techniques and crossvalidation on independent dataset.
Have to check for the regression assumptions by residual diagnostics and multicollinearity, Using log transformtion if needed.	For prediction models adjusting for multicollinearity is not important unless it decreases the prediction error. Same is true for checking regression asumptions and transformations of variables.
The explanatory variables for the model need to have significant slope coefficients.	Did not care much if the coefficients for my model were statistically significant.
Interested in p-values for the model and confidence interval for the variable of interest.	Interested in prediction formula with coefficient and the prediction error.
Violation of regression assumptions and collinearity as the main challenges.	Low adjusted R-squared, poor model fit or overfitting as the main challenges.

R code appendix

```
knitr::opts_chunk$set(include=FALSE)
library(broom)
library(knitr)
library(kableExtra)
library(ggplot2)
library(gridExtra)
library(plyr)
library(leaps)
```

```
library(caTools)
library(NLP)
library(nlme)
library(dplyr)
library(boot)
library(formatR)
library(ggthemes)
library(DAAG)
library(Metrics)
library(sjmisc)
PUMS_2015 <- read.csv('OR_acs_house_occ.csv')
PUMS_2015.nice<-PUMS_2015
colnames(PUMS_2015.nice)<-c("Serial number","Records per housing record","Type of unit","Lot size","Numb
kable(head(PUMS_2015.nice, 2), format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options=c("scale_down","HOLD_position") , font_size = 5, full_width = TRUE)
str(PUMS_2015)
summary(PUMS_2015)
PUMS_2015.int = as.data.frame(PUMS_2015.nice[,1])
PUMS_2015.chr = as.data.frame(PUMS_2015.nice[,1])
for (i in 1:ncol(PUMS_2015.nice)){
  if (!is.numeric(PUMS_2015.nice[,i])){
      PUMS_2015.chr<-cbind(PUMS_2015.chr,PUMS_2015.nice[,i])
      colnames(PUMS_2015.chr)[ncol(PUMS_2015.chr)]<-colnames(PUMS_2015.nice[i])
  }
  else{
      PUMS_2015.int<-cbind(PUMS_2015.int,PUMS_2015.nice[,i])
      colnames(PUMS_2015.int)[ncol(PUMS_2015.int)]<-colnames(PUMS_2015.nice[i])
  }
}
PUMS_2015.int <- subset (PUMS_2015.int, select = -c(1))
PUMS_2015.chr <- subset (PUMS_2015.chr, select = -c(1))
#Boxplots on electricity bills for different building types
ggplot(data=PUMS_2015, aes(factor(BLD), ELEP, fill = factor(BLD))) +
  geom_boxplot()+theme_tufte()+theme(text = element_text(size=rel(3)),axis.text.x = element_text(angle =
#summary of non-numeric column values:
#knitr::kable(list(as.data.frame(table(PUMS_2015.chr[,1])),as.data.frame(table(PUMS_2015.chr[,2])),as.d
library(colorspace)
p1<-ggplot(as.data.frame(table(PUMS_2015.chr[,1])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnam
p2<-ggplot(as.data.frame(table(PUMS_2015.chr[,2])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnam
p3<-ggplot(as.data.frame(table(PUMS_2015.chr[,3])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnam
p4<-ggplot(as.data.frame(table(PUMS_2015.chr[,4])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnam
p5<-ggplot(as.data.frame(table(PUMS_2015.chr[,5])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnam
p6<-ggplot(as.data.frame(table(PUMS_2015.chr[,6])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnam
p7<-ggplot(as.data.frame(table(PUMS_2015.chr[,7])), aes(Var1, Freq, fill = Var1))+geom_col()+xlab(colnam

#summary of numeric columns
#knitr::kable(summary(PUMS_2015.int))%>% kable_styling(latex_options="scale_down", font_size = 4, full_
p8<-ggplot(PUMS_2015.int, aes(x=PUMS_2015.int[,2])) + geom_histogram()+xlab(colnames(PUMS_2015.int[2]))+
p9<-ggplot(PUMS_2015.int, aes(x=PUMS_2015.int[,3])) + geom_bar(width =0.1)+xlab(colnames(PUMS_2015.int[3
p10<-ggplot(PUMS_2015.int, aes(x=PUMS_2015.int[,4])) + geom_histogram()+xlab(colnames(PUMS_2015.int[4]))
p11<-ggplot(PUMS_2015.int, aes(x=PUMS_2015.int[,5])) + geom_histogram()+xlab(colnames(PUMS_2015.int[5]))
p12<-ggplot(PUMS_2015.int, aes(x=PUMS_2015.int[,6])) + geom_histogram()+xlab(colnames(PUMS_2015.int[6]))
p13<-ggplot(PUMS_2015.int, aes(x=PUMS_2015.int[,7])) + geom_histogram()+xlab(colnames(PUMS_2015.int[7]))
```

```
p14<-ggplot(PUMS_2015.int, aes(x=PUMS_2015.int[,8])) + geom_histogram()+xlab(colnames(PUMS_2015.int[8]))
p15<-ggplot(PUMS_2015.int, aes(x=PUMS_2015.int[,9])) + geom_histogram()+xlab(colnames(PUMS_2015.int[9]))


grid.arrange(p1,p2, p3, p4, p5, p6, p7, p8, p9, p10, p11, p12, p13, p14,p15, nrow = 3)

#N/A values
PUMS_2015.na<-PUMS_2015.nice %>% summarise_all(~ sum(is.na(.)))
kable(PUMS_2015.na)%>%kable_styling(latex_options=c("scale_down","HOLD_position"), font_size = 5, full_
PUMS_2015$YBL<-revalue(PUMS_2015$YBL, c("1939 or earlier"=1939,"1950 to 1959"=1959, "1990 to 1999"=1999
PUMS_2015$YBL<-ifelse(PUMS_2015$YBL>2000, 1, 0)
PUMS_2015$HEATING<-ifelse(PUMS_2015$HFL=='Electricity', 1 , 0)
PUMS_2015$HOUSE<-ifelse(PUMS_2015$BLD=="One-family house detached" | PUMS_2015$BLD=="One-family house a

PUMS_2015_inferential<-PUMS_2015[PUMS_2015$BLD!= 'Mobile home or trailer' & PUMS_2015$BLD!='Boat, RV, v


data.elec<-subset(PUMS_2015_inferential, ELEP<500)

data.elec$HOUSING<-ifelse(data.elec$HOUSE==1,"Yes","No")

ggplot(data.elec, aes(factor(HOUSING), ELEP, fill = factor(HOUSING))) +
  geom_boxplot()+theme_tufte()+theme(text = element_text(size=rel(3)),legend.position = "none")+xlab("L
lm.inference.int<-lm(ELEP~NP+BDSP+HEATING+YBL+HOUSE+R18+HEATING:BDSP, data<-data.elec)
model<-lm.inference.int%>%tidy()
#kable(model)%>%
 # kable_styling(latex_options = c("striped","scale_down"), stripe_color = "gray!15", font_size = 7, fu
summary(lm.inference.int)
lm.inference.noint<-lm(ELEP~NP+BDSP+HEATING+YBL+HOUSE+R18, data<-data.elec)
anova(lm.inference.noint, lm.inference.int)
ggplot(data.elec, aes(NP, BDSP))+geom_point()+theme_tufte()+geom_jitter()+labs(x="Number of occupants",
data.additional<- fortify(lm.inference.int, data = data.elec)
p1<-ggplot(data.additional, aes(x=.resid)) + geom_histogram()+xlab("Residuals")+ylab("Count")+theme_tuf
p2<-ggplot(data.additional, aes(y=.resid, x=.fitted))+ geom_point()+xlab("Fitted Values")+ylab("Residua
p3<-ggplot(data.additional, aes(y=.fitted, x=ELEP))+ geom_point()+xlab("Observed values")+ylab("Fitted
p4<-ggplot(data.additional, aes(y=.resid, x=ELEP))+ geom_point()+xlab("Observed values")+ylab("Residuals
p5<-ggplot(data.additional, aes(y=.resid, x=NP))+ geom_point()+xlab("Number of Person Records per Housel
p6<-ggplot(data.additional, aes(y=.resid, x=BDSP))+ geom_point()+xlab("Number of Bedrooms")+ylab("Residu
grid.arrange(p1,p2,p3,p4,p5,p6, nrow=3, top = 'Checking Linear Regression Assumptions for the Original
lm.inference.int<-lm(log(ELEP)~NP+BDSP+HEATING+YBL+HOUSE+R18, data<-data.elec)
data.additional<- fortify(lm.inference.int, data = data.elec)
p1<-ggplot(data.additional, aes(x=.resid)) + geom_histogram()+theme_tufte()+theme(text = element_text(s
p2<-ggplot(data.additional, aes(y=.resid, x=.fitted))+ geom_point()+theme_tufte()+theme(text = element_
p3<-ggplot(data.additional, aes(y=.fitted, x=log(ELEP)))+ geom_point()+xlab("Observed values")+ylab("Fi
p4<-ggplot(data.additional, aes(y=.resid, x=log(ELEP)))+ geom_point()+xlab("Observed values")+ylab("Resi
p5<-ggplot(data.additional, aes(y=.resid, x=NP))+ geom_point()+xlab("Number of Person Records per Housel
p6<-ggplot(data.additional, aes(y=.resid, x=BDSP))+ geom_point()+xlab("Number of Bedrooms")+ylab("Residu
grid.arrange(p1,p2,p3,p4,p5,p6, nrow=3, top = 'Checking Linear Regression Assumptions for Log-transform
lm.inference.int<-lm(log(ELEP)~NP+BDSP+HEATING+YBL+HOUSE+R18+HEATING:BDSP, data<-data.elec)
summary(lm.inference.int)
ci<-confint(lm.inference.int)
(ci)
log.ci.convert<-function(n){(exp(n)-1) * 100}
```

```
(ci.low<-round(log.ci.convert(ci[6,1]),2))
(ci.high<-round(log.ci.convert(ci[6,2]),2))
(point.est<-round(log.ci.convert(lm.inference.int$coefficients[6]),2))
data.na.VALP<- subset(PUMS_2015, is.na(VALP))
head(data.na.VALP)
summary(data.na.VALP)
table(data.na.VALP$TEN)

data.na.ACR<- subset(PUMS_2015, is.na(ACR))
head(data.na.ACR)
summary(data.na.ACR)
kable(table(data.na.ACR$BLD))%>%
  kable_styling(latex_options = c("striped","scale_down"), stripe_color = "gray!15", font_size = 10, ful
data.elec.predict<-PUMS_2015
data.elec.predict$R18<-ifelse(data.elec.predict$R18=='1 or more', 1, 0)
data.elec.predict$R60<-ifelse(data.elec.predict$R60=='1 or more', 1, 0)
data.elec.predict<-subset(data.elec.predict, select=-c(SERIALNO,TYPE,BLD,HFL,VALP))
data.elec.predict$ACR <- addNA(factor(data.elec.predict$ACR))
#pairs(select_if(data.elec.predict, is.numeric))
set.seed(123)
split = sample.split(data.elec.predict, SplitRatio = 0.8)
training_set = subset(data.elec.predict, split == TRUE)

training_set.named<-training_set
colnames(training_set.named)<-c("Records per housing record","Lot size","Number of bedrooms", "ELEP", "

testing_set = subset(data.elec.predict, split == FALSE)

set.seed(321)
split = sample.split(testing_set, SplitRatio = 0.5)
validation_set = subset(testing_set, split == TRUE)
test_set = subset(testing_set, split == FALSE)
str(training_set.named)
regfit.full<-regsubsets(log(ELEP)~.,training_set.named, nvmax = 20)
regfit.summary<-summary(regfit.full)
#regfit.summary

par(mfrow=c(2,3), mar=c(1,1,1,1),cex.axis = 0.5, cex.lab = 0.2, cex.main=0.5)
#plot(regfit.summary$rss, xlab="Number of Variables", ylab="RSS", type="l")
plot(regfit.summary$adjr2, xlab="Number of Variables", main="Adjusted RSq", type="l")
max.adjr2<-which.max(regfit.summary$adjr2)
points(max.adjr2, regfit.summary$adjr2[max.adjr2], col="red", cex=2, pch=20)

plot(regfit.summary$cp, xlab="Number of Variables", main ="Cp", type="l")
min.cp<-which.min(regfit.summary$cp)
points(min.cp, regfit.summary$cp[min.cp], col="red", cex=2, pch=20)

min.bic<-which.min(regfit.summary$bic)
plot(regfit.summary$bic, xlab="Number of Variables", main="BIC", type="l")
points(min.bic, regfit.summary$bic[min.bic], col="red", cex=2, pch=20)
plot(regfit.full, scale="adjr2")
plot(regfit.full, scale="Cp")
plot(regfit.full, scale="bic")
```

```r
#We'll need to split TENR to its factor columns
training_set$TEN_O<-if_else(training_set$TEN=="Owned free and clear", 1, 0)
training_set$TEN_L<-if_else(training_set$TEN=="Owned with mortgage or loan", 1, 0)
training_set$TEN_R<-if_else(training_set$TEN=="Rented", 1, 0)
validation_set$TEN_O<-if_else(validation_set$TEN=="Owned free and clear", 1, 0)
validation_set$TEN_L<-if_else(validation_set$TEN=="Owned with mortgage or loan", 1, 0)
validation_set$TEN_R<-if_else(validation_set$TEN=="Rented", 1, 0)
test_set$TEN_O<-if_else(test_set$TEN=="Owned free and clear", 1, 0)
test_set$TEN_L<-if_else(test_set$TEN=="Owned with mortgage or loan", 1, 0)
test_set$TEN_R<-if_else(test_set$TEN=="Rented", 1, 0)
data.elec.predict$TEN_O<-if_else(data.elec.predict$TEN=="Owned free and clear", 1, 0)
data.elec.predict$TEN_L<-if_else(data.elec.predict$TEN=="Owned with mortgage or loan", 1, 0)
data.elec.predict$TEN_R<-if_else(data.elec.predict$TEN=="Rented", 1, 0)

cv.new.error<-rep(0,3)

glm.fit.AdjR2 = glm(formula = log(ELEP) ~ .-TEN_L-TEN_R-TEN, data = training_set)
cv.new.error[1]<-cv.glm(validation_set, glm.fit.AdjR2, K=10)$delta[1]
glm.fit.Cp = glm(formula = log(ELEP) ~ .-R18-TEN_L-TEN_R-TEN, data = training_set)
cv.new.error[2]<-cv.glm(validation_set, glm.fit.Cp, K=10)$delta[1]
glm.fit.BIC = glm(formula = log(ELEP) ~ .-R18-R60-TEN_L-TEN_R-TEN, data = training_set)
cv.new.error[3]<-cv.glm(validation_set, glm.fit.BIC, K=10)$delta[1]
best.models <- data.frame(c('Adjusted R-squared', 'Cp','BIC'),c('14 variables', '13 variables','12 varia
colnames(best.models)<- c('Metrics used for variable selection', 'Number of variables in the model', 'CV
kable(best.models, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options = c("striped","HOLD_position"), stripe_color = "gray!15")
data.elec.predict<-rbind(training_set, validation_set)
data.elec.predict.copy<-data.elec.predict
colnames(data.elec.predict)<-c("Records per housing record","Lot size","Number of bedrooms", "ELEP", "Ye
best.model.prediction<-lm(formula = log(ELEP) ~ . - R18 - TEN_L- TEN_R - TEN,
    data = data.elec.predict)
summary(best.model.prediction)
coefficients(best.model.prediction)
prediction.formula<-paste0("log(ELEP)~ ", round(as.vector(coefficients(best.model.prediction)[1]), 5))
for (i in 2:length(coefficients(best.model.prediction))){
      prediction.formula<-paste0(prediction.formula,"+",round(as.vector(coefficients(best.model.predicti
}
j<-0
while (j<nchar(prediction.formula)){
  print(substr(prediction.formula, j+1, j+90))
  j<-j+90}
colnames(test_set)<-colnames(data.elec.predict)
predictions <- best.model.prediction %>% predict(test_set, interval="prediction", level=0.90)
predictions.ci <- best.model.prediction %>% predict(test_set, interval="confidence", level=0.90)
colnames(predictions.ci)<-c("fit", "lower.ci","upper.ci")
# Model performance
# (a) Prediction error, RMSE
(pred.error.test<-rmse(predictions, log(test_set$ELEP)))
pred.interval<-cbind(test_set, predictions, predictions.ci[,2],predictions.ci[,3])
ggplot(pred.interval, aes(x=exp(fit)))+geom_point(aes(y=ELEP), size=1)+geom_point(aes(y=exp(fit)), size=
discussion<-as.data.frame(matrix(nrow=9,ncol=2))
colnames(discussion)<-c('Explanatory Model', 'Predictive Model')
discussion[1,1]<-c('Primary goal is answering the question on difference in electricity bill for houses
```

```
discussion[1,2]<-c('Primary goal is having an accurate estimate of monthly electricity costs as well an

discussion[2,1]<-c('The strategy is to pick only those variables and interactions which are meaningful/
discussion[2,2]<-c('The strategy is to choose the set of variables that will lead to a better predictio


discussion[3,1]<-c('No concern for missing values and problematic columns as long as they are not in th
discussion[3,2]<-c('Need to clean up the data (missing values etc.) prior to selection.')

discussion[4,1]<-c('Need to avoid complicating the model interpretation with too much column transforma
discussion[4,2]<-c('All column transformations and feature engineering are allowed as long as they impr

discussion[5,1]<-c('Cannot trust inference after variable selection (using the same data twice). Otherw
discussion[5,2]<-c('Variable selection is the major step in model optimization.')

discussion[6,1]<-c('Typically selecting between only a couple of models (e.g. one with and without inte
discussion[6,2]<-c('Comparing lots of models with best subset selection techniques and crossvalidation

discussion[7,1]<-c('Have to check for the regression assumptions by residual diagnostics and multicolli
discussion[7,2]<-c('For prediction models adjusting for multicollinearity is not important unless it de

discussion[8,1]<-c('The explanatory variables for the model need to have significant slope coefficients
discussion[8,2]<-c('Did not care much if the coefficients for my model were statistically significant.'

discussion[9,1]<-c('Interested in p-values for the model and confidence interval for the variable of in
discussion[9,2]<-c('Interested in prediction formula with coefficient and the prediction error.')

discussion[10,1]<-c('Violation of regression assumptions and collinearity as the main challenges.')
discussion[10,2]<-c('Low adjusted R-squared, poor model fit or overfitting as the main challenges.')

kable(discussion, format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options = c("striped","scale_down"), stripe_color = "gray!15", font_size = 7, full
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=I(60)), tidy=TRUE)
```