# Electricity Costs in Oregon Households: Multivariate Regression Analysis

Violeta Lio King

3/13/2022

## Exploratory Analysis

The preliminary step is preparing the data in order to be able to carry out the exploratory analysis. One of the predictor variables, BLD, contains information on the housing type but the analysis will only cover houses and apartments. A new column "HA" is created and the rest of the housing types are dropped from the dataframe. Next, it is critical to drop several of the predictors due to collinearity/linear dependence issues later on in part 2. For example, BLD and HA are collinear to one another because HA is derived from BLD so BLD must be dropped. Other variables such as SERIALNO, are unique identifiers and will have no predictive influence so they can be omitted. Using a similar logic, ACR/TYPE/VALP are also dropped from the dataframe.

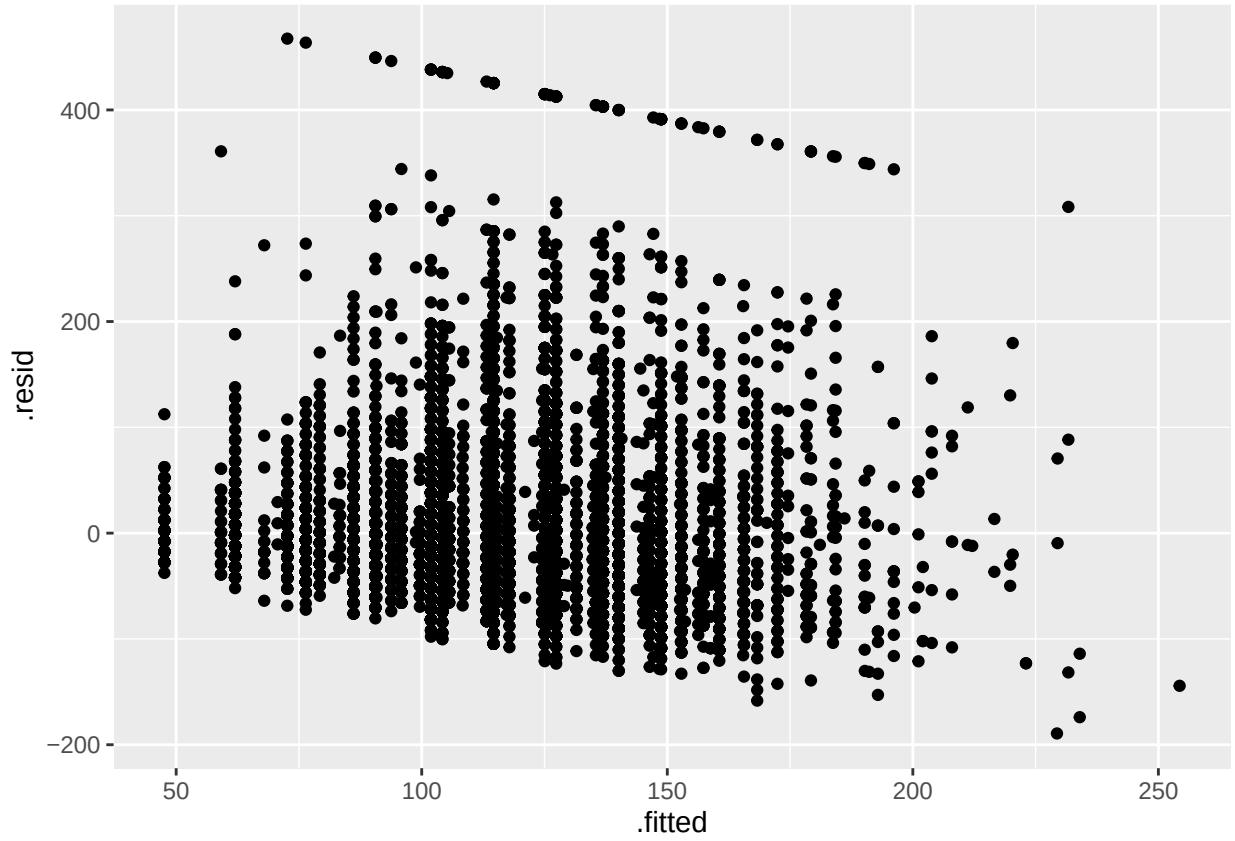
```
# Load data
setwd("~/Downloads/ST517")
housing_df <- read.csv("OR_acs_house_occ.csv", stringsAsFactors=TRUE)

# Create new variable HA
housing_df$HA <- "other" # Initialize everything to "other"
housing_df$HA[which(grepl("house", housing_df$BLD))] <- "house"
housing_df$HA[which(grepl("apartment", housing_df$BLD, ignore.case = TRUE))] <- "apt"
housing_df$HA <- factor(housing_df$HA)

# Data must be scrubbed first. The only housing data we want is from apartments and houses
house_drop <- subset(housing_df, HA == "apt" | HA == "house")
housing_df <- droplevels(house_drop)

# For part 2, it is not necessary to to pass regsubsets() every single variable in the dataframe. We ar
housing_df <- subset(housing_df, select = -c(ACR, BLD, SERIALNO, TYPE, VALP))
```

The first part of this analysis is to create a model with interactions. After that, we run residual diagnostics, and then compare it to a model without interactions by performing the extra SS F-test. Finally, we will perform inference using the preferred model. In this case, the model with interactions is minimal and only uses the HA, NP, and BDSP variables plus the interaction terms. The model with interactions is $\mu(ELEP|HA, NP, BDS) = \beta_0 + HA\beta_1 + \beta_2 NP + \beta_3 BDSP + \beta_4(HA * BDSP) + \beta_5(HA * BDSP) + \beta_6(NP * BDSP)$. If we choose a model without interactions then the model will be $\mu(ELEP|HA, NP, BDS) = \beta_0 + HA\beta_1 + +\beta_2 NP + \beta_3 BDSP$.

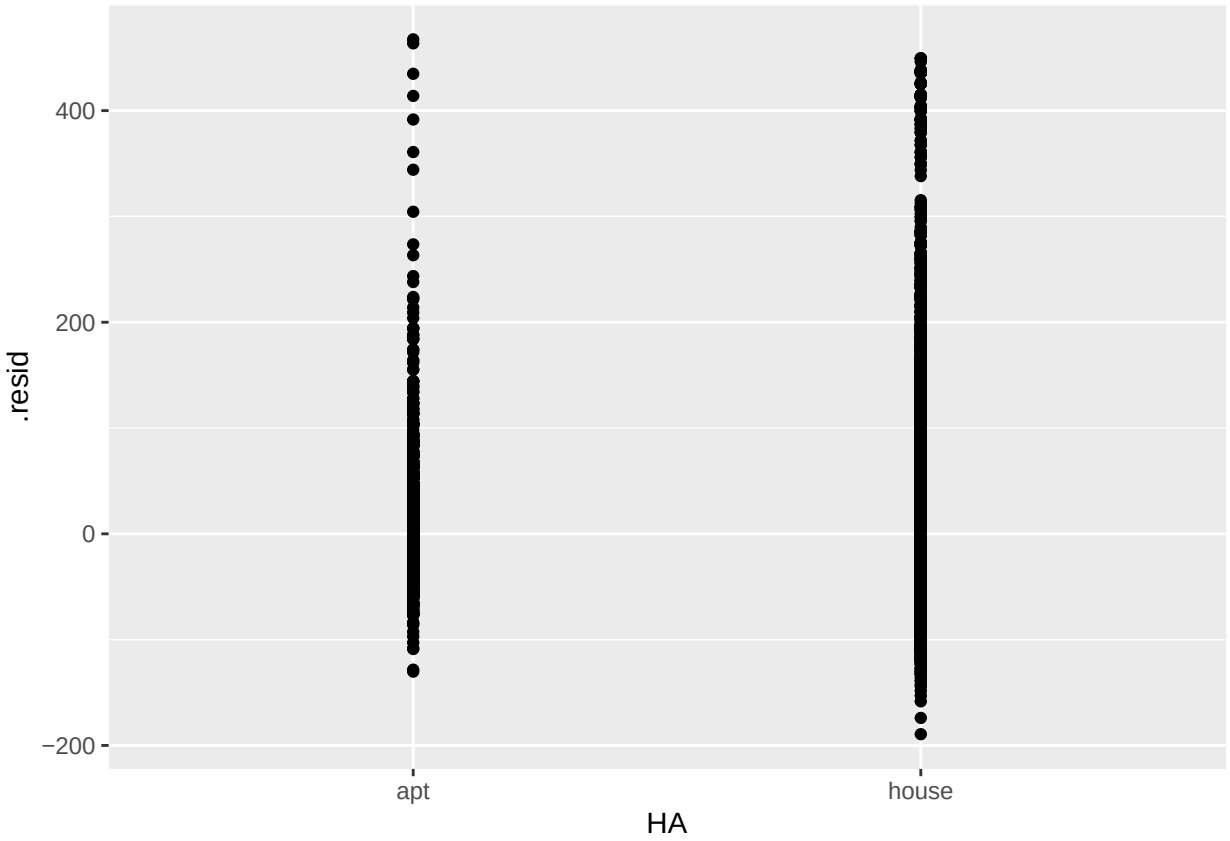
The residual plots show that none of the assumptions are being violated. The first plot shows the residuals vs the fitted values. It shows a near constant variance with the exception of some values. The next three plots are the residuals versus HA, NP, and BDSP. The residuals versus NP plot shows a moderate skew. On the left side of the plot, there are some positive values which are over 400 while the negative values do not

fall below -200. However, the data are quite large and this skew is moderate so there are no assumptions being violated. The rest of the residuals look good. To conclude, the residual plots indicate that the model fit is good and that

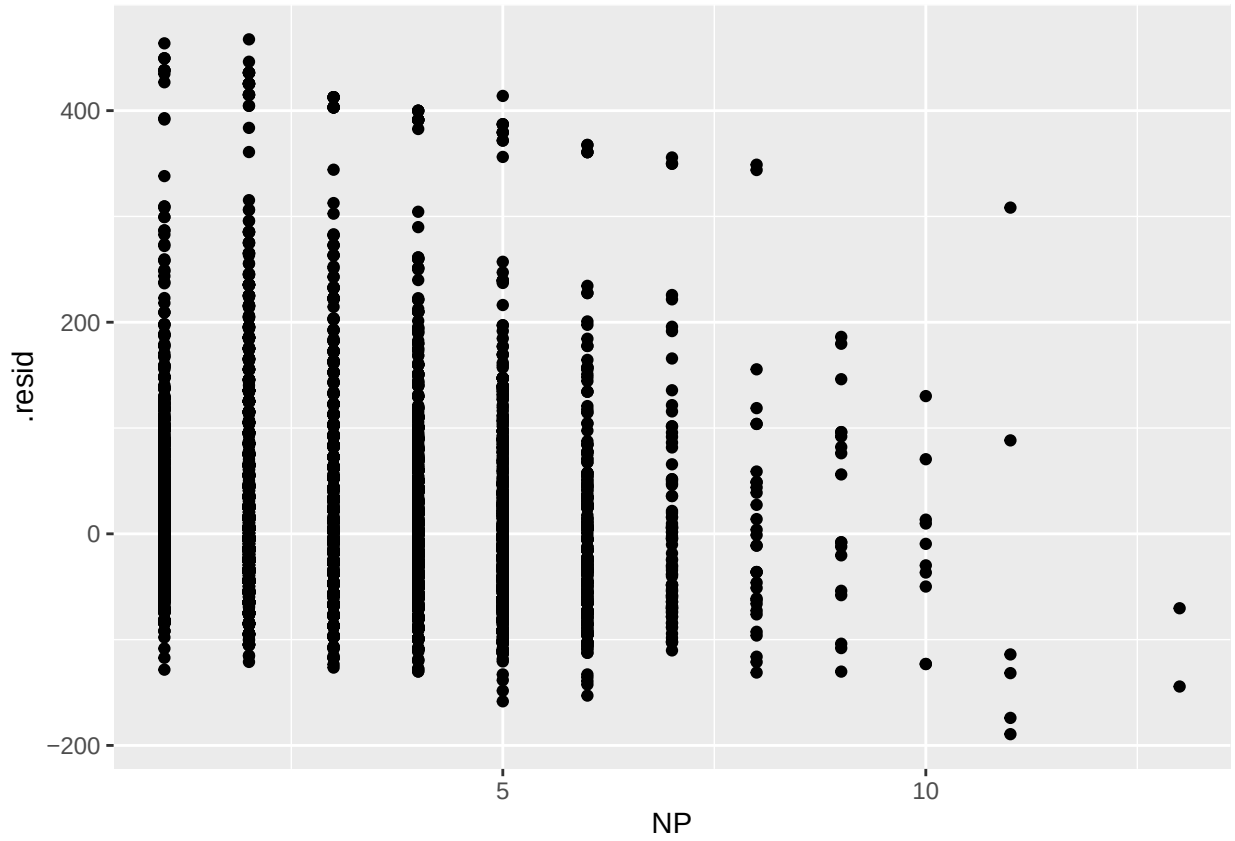
```
# Models with and without interactions
mod_int <- lm(ELEP ~ HA + NP + BDSP + HA:NP + HA:BDSP + NP:BDSP, data = housing_df)
mod_noint <- lm(ELEP ~ HA + NP + BDSP, data = housing_df)

# Residual analysis on model with interactions
mod1_diag <- augment(mod_int, housing_df)
#head(mod1_diag)
qplot(.fitted, .resid, data = mod1_diag)
```



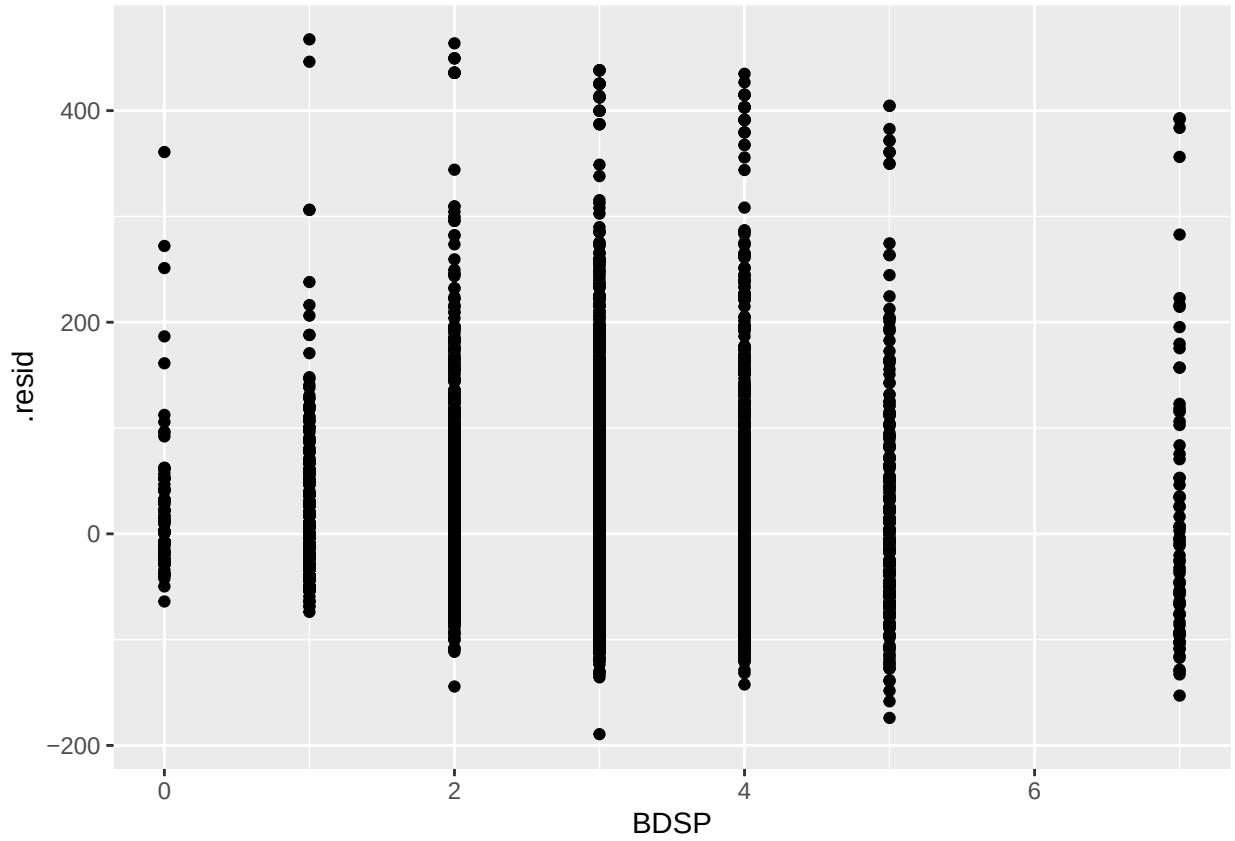

```
qplot(HA, .resid, data = mod1_diag)
```




```
qplot(NP, .resid, data = mod1_diag)
```




```
qplot(BDSP, .resid, data = mod1_diag)
```




The next step is a comparison of the model with interactions and the model without interactions. For this, a simple ANOVA test was carried out. The p-values for HA, NP and BDSP were all less than 2e-16 while the p-value for the interaction term HA:BDSP was 0.03687. Based on the ANOVA test results, the p-values for the extra SS F-test indicate that the model that includes these four terms is the best model. However, after looking at the summary and seeing how small the coefficient is for the interaction term, I have ultimately decided to go with the simpler model, which contains no interactions. This is mainly a judgement call for the sake of making interpretation more straightforward. The interaction terms can be difficult to interpret and although HA:BDSP is statistically significant, the low magnitude of the coefficient associated with it means that it will not have too much of an impact on inference by omitting it. The final step is inference with the simpler value which is carried out by using confint(). This will provide the answer for the first question. The confidence interval of interest is the one associated with HAhouse which is (15.958141, 22.98812). Both the lower bound and the upper bound values are positive which means that people in houses pay more in electricity than homeowners. To be more precise, people living in houses pay between 15.96 and 22.99 dollars more in electricity than those living in apartments. For the model with no interactions, the RMSE is $69.65.

```
# Model comparison
anova(mod_int)
```

```
## Analysis of Variance Table
##
## Response: ELEP
##              Df   Sum Sq Mean Sq   F value  Pr(>F)
## HA            1  3407308 3407308  702.6011 < 2e-16 ***
## NP            1  5026481 5026481 1036.4814 < 2e-16 ***
## BDSP          1  1002138 1002138  206.6450 < 2e-16 ***
## HA:NP         1    11560   11560    2.3837 0.12263
```

```
## HA:BDSP       1    16451   16451    3.3923 0.06552 .
## NP:BDSP       1    21131   21131    4.3573 0.03687 *
## Residuals 13767 66763926    4850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Inference
confint(mod_noint)
```

```
##                 2.5 %   97.5 %
## (Intercept) 36.335033 43.66073
## HAhouse     15.958141 22.98812
## NP          11.117616 12.95362
## BDSP         8.953356 11.78139
```

```
summ <- summary(mod_noint)
sqrt(mean(summ$residuals^2)) # RMSE for the model with no interactions
```

```
## [1] 69.64672
```

## Prediction Problem

The first step in the prediction problem is to set aside 20% of the the data for validation. This will be the last step in the prediction problem and will be used to determine how well the model fits. The rest of the data will be used to determine the best number of variables for the model.

```
set.seed(42)
n <- nrow(housing_df)

# Set aside 20% of all cases (the dataframe has already been cleaned)
valid <- sample(n, size = floor(0.20*n))
housing_valid <- housing_df[valid, ]
housing_train <- housing_df[-valid, ]
#str(housing_valid)
```

To determine the best model, the function regsubsets() was used on the training data. This is a useful function since it will determine the best number of predictors to include. Unfortunately, it can be difficult to interpret the output since regsubsets() further breaks down the predictors into more variables if they have multiple inputs. So instead of 11 variables, regsubsets() considered 37 of them.

```
# Cross - validation set approach
# Use regsubsets() to determine the best model
regfit_best <- regsubsets(ELEP ~ ., data = housing_train, really.big = TRUE, nvmax = 38)
summary(regfit_best)
```

```
## Subset selection object
## Call: regsubsets.formula(ELEP ~ ., data = housing_train, really.big = TRUE,
##     nvmax = 38)
## 37 Variables  (and intercept)
##                                Forced in Forced out
## NP                                 FALSE      FALSE
```

```
## BDSP                                FALSE      FALSE
## FULP                                FALSE      FALSE
## GASP                                FALSE      FALSE
## HFLCoal or coke                     FALSE      FALSE
## HFLElectricity                      FALSE      FALSE
## HFLFuel oil, kerosene, etc.         FALSE      FALSE
## HFLNo fuel used                     FALSE      FALSE
## HFLOther fuel                       FALSE      FALSE
## HFLSolar energy                     FALSE      FALSE
## HFLUtility gas                      FALSE      FALSE
## HFLWood                             FALSE      FALSE
## RMSP                                FALSE      FALSE
## TENOwned free and clear             FALSE      FALSE
## TENOwned with mortgage or loan      FALSE      FALSE
## TENRented                           FALSE      FALSE
## YBL1940 to 1949                     FALSE      FALSE
## YBL1950 to 1959                     FALSE      FALSE
## YBL1960 to 1969                     FALSE      FALSE
## YBL1970 to 1979                     FALSE      FALSE
## YBL1980 to 1989                     FALSE      FALSE
## YBL1990 to 1999                     FALSE      FALSE
## YBL2000 to 2004                     FALSE      FALSE
## YBL2005                             FALSE      FALSE
## YBL2006                             FALSE      FALSE
## YBL2007                             FALSE      FALSE
## YBL2008                             FALSE      FALSE
## YBL2009                             FALSE      FALSE
## YBL2010                             FALSE      FALSE
## YBL2011                             FALSE      FALSE
## YBL2012                             FALSE      FALSE
## YBL2013                             FALSE      FALSE
## YBL2014                             FALSE      FALSE
## YBL2015                             FALSE      FALSE
## R18none                             FALSE      FALSE
## R60none                             FALSE      FALSE
## HAhouse                             FALSE      FALSE
## 1 subsets of each size up to 37
## Selection Algorithm: exhaustive
##           NP  BDSP FULP GASP HFLCoal or coke HFLElectricity
## 1  ( 1 )  "*" " "  " "  " "  " "             " "
## 2  ( 1 )  " " "*"  " "  " "  " "             " "
## 3  ( 1 )  "*" "*"  " "  " "  " "             " "
## 4  ( 1 )  "*" " "  " "  " "  " "             " "
## 5  ( 1 )  "*" " "  " "  " "  " "             "*"
## 6  ( 1 )  "*" " "  " "  "*"  " "             "*"
## 7  ( 1 )  "*" "*"  " "  "*"  " "             "*"
## 8  ( 1 )  "*" "*"  " "  "*"  " "             "*"
## 9  ( 1 )  "*" "*"  " "  "*"  " "             "*"
## 10  ( 1 ) "*" "*"  " "  "*"  " "             "*"
## 11  ( 1 ) "*" "*"  " "  "*"  " "             "*"
## 12  ( 1 ) "*" "*"  " "  "*"  " "             "*"
## 13  ( 1 ) "*" "*"  " "  "*"  " "             "*"
## 14  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 15  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
```

```
## 16  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 17  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 18  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 19  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 20  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 21  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 22  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 23  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 24  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 25  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 26  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 27  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 28  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 29  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 30  ( 1 ) "*" "*"  "*"  "*"  "*"             "*"
## 31  ( 1 ) "*" "*"  "*"  "*"  "*"             "*"
## 32  ( 1 ) "*" "*"  "*"  "*"  " "             "*"
## 33  ( 1 ) "*" "*"  "*"  "*"  "*"             "*"
## 34  ( 1 ) "*" "*"  "*"  "*"  "*"             "*"
## 35  ( 1 ) "*" "*"  "*"  "*"  "*"             "*"
## 36  ( 1 ) "*" "*"  "*"  "*"  "*"             "*"
## 37  ( 1 ) "*" "*"  "*"  "*"  "*"             "*"
##           HFLFuel oil, kerosene, etc. HFLNo fuel used HFLOther fuel
## 1   ( 1 )  " "                        " "             " "
## 2   ( 1 )  " "                        " "             " "
## 3   ( 1 )  " "                        " "             " "
## 4   ( 1 )  " "                        " "             " "
## 5   ( 1 )  " "                        " "             " "
## 6   ( 1 )  " "                        " "             " "
## 7   ( 1 )  " "                        " "             " "
## 8   ( 1 )  " "                        " "             " "
## 9   ( 1 )  " "                        " "             " "
## 10  ( 1 ) " "                         " "             "*"
## 11  ( 1 ) " "                         " "             "*"
## 12  ( 1 ) " "                         " "             "*"
## 13  ( 1 ) " "                         " "             "*"
## 14  ( 1 ) " "                         " "             "*"
## 15  ( 1 ) " "                         " "             "*"
## 16  ( 1 ) " "                         " "             "*"
## 17  ( 1 ) " "                         " "             "*"
## 18  ( 1 ) " "                         " "             "*"
## 19  ( 1 ) " "                         " "             "*"
## 20  ( 1 ) " "                         " "             "*"
## 21  ( 1 ) " "                         " "             "*"
## 22  ( 1 ) " "                         " "             "*"
## 23  ( 1 ) " "                         " "             "*"
## 24  ( 1 ) " "                         " "             "*"
## 25  ( 1 ) " "                         " "             "*"
## 26  ( 1 ) " "                         " "             "*"
## 27  ( 1 ) " "                         " "             "*"
## 28  ( 1 ) " "                         " "             "*"
## 29  ( 1 ) " "                         " "             "*"
## 30  ( 1 ) " "                         " "             "*"
## 31  ( 1 ) " "                         "*"             "*"
```

```
## 32  ( 1 ) " "                         "*"             "*"
## 33  ( 1 ) " "                         "*"             "*"
## 34  ( 1 ) " "                         "*"             "*"
## 35  ( 1 ) " "                         "*"             "*"
## 36  ( 1 ) " "                         "*"             "*"
## 37  ( 1 ) "*"                         "*"             "*"
##           HFLSolar energy HFLUtility gas HFLWood RMSP TENOwned free and clear
## 1  ( 1 )  " "             " "            " "     " "  " "
## 2  ( 1 )  " "             "*"            " "     " "  " "
## 3  ( 1 )  " "             "*"            " "     " "  " "
## 4  ( 1 )  " "             "*"            " "     "*"  " "
## 5  ( 1 )  " "             "*"            " "     "*"  " "
## 6  ( 1 )  " "             "*"            " "     "*"  " "
## 7  ( 1 )  " "             "*"            " "     "*"  " "
## 8  ( 1 )  " "             "*"            " "     "*"  " "
## 9  ( 1 )  " "             "*"            " "     "*"  " "
## 10  ( 1 ) " "             "*"            "*"     "*"  " "
## 11  ( 1 ) " "             "*"            "*"     "*"  " "
## 12  ( 1 ) " "             "*"            "*"     "*"  " "
## 13  ( 1 ) " "             "*"            "*"     "*"  " "
## 14  ( 1 ) " "             "*"            "*"     "*"  " "
## 15  ( 1 ) " "             "*"            "*"     "*"  "*"
## 16  ( 1 ) " "             "*"            "*"     "*"  "*"
## 17  ( 1 ) " "             "*"            "*"     "*"  "*"
## 18  ( 1 ) " "             "*"            "*"     "*"  "*"
## 19  ( 1 ) " "             "*"            "*"     "*"  "*"
## 20  ( 1 ) " "             "*"            "*"     "*"  "*"
## 21  ( 1 ) " "             "*"            "*"     "*"  "*"
## 22  ( 1 ) " "             "*"            "*"     "*"  "*"
## 23  ( 1 ) " "             "*"            "*"     "*"  "*"
## 24  ( 1 ) " "             "*"            "*"     "*"  "*"
## 25  ( 1 ) " "             "*"            "*"     "*"  "*"
## 26  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 27  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 28  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 29  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 30  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 31  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 32  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 33  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 34  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 35  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 36  ( 1 ) "*"             "*"            "*"     "*"  "*"
## 37  ( 1 ) "*"             "*"            "*"     "*"  "*"
##           TENOwned with mortgage or loan TENRented YBL1940 to 1949
## 1  ( 1 )  " "                            " "       " "
## 2  ( 1 )  " "                            " "       " "
## 3  ( 1 )  " "                            " "       " "
## 4  ( 1 )  " "                            " "       " "
## 5  ( 1 )  " "                            " "       " "
## 6  ( 1 )  " "                            " "       " "
## 7  ( 1 )  " "                            " "       " "
## 8  ( 1 )  " "                            " "       " "
## 9  ( 1 )  " "                            " "       " "
```

```
## 10  ( 1 ) " "                          " "       " "
## 11  ( 1 ) " "                          " "       " "
## 12  ( 1 ) " "                          " "       " "
## 13  ( 1 ) " "                          " "       "*"
## 14  ( 1 ) " "                          " "       "*"
## 15  ( 1 ) " "                          " "       "*"
## 16  ( 1 ) " "                          " "       "*"
## 17  ( 1 ) " "                          " "       "*"
## 18  ( 1 ) " "                          " "       "*"
## 19  ( 1 ) " "                          " "       "*"
## 20  ( 1 ) " "                          " "       "*"
## 21  ( 1 ) " "                          " "       "*"
## 22  ( 1 ) " "                          " "       "*"
## 23  ( 1 ) " "                          " "       "*"
## 24  ( 1 ) "*"                          "*"       "*"
## 25  ( 1 ) "*"                          "*"       "*"
## 26  ( 1 ) "*"                          "*"       "*"
## 27  ( 1 ) "*"                          "*"       "*"
## 28  ( 1 ) "*"                          "*"       "*"
## 29  ( 1 ) "*"                          "*"       "*"
## 30  ( 1 ) "*"                          "*"       "*"
## 31  ( 1 ) "*"                          "*"       "*"
## 32  ( 1 ) "*"                          "*"       "*"
## 33  ( 1 ) "*"                          "*"       "*"
## 34  ( 1 ) "*"                          "*"       "*"
## 35  ( 1 ) "*"                          "*"       "*"
## 36  ( 1 ) "*"                          "*"       "*"
## 37  ( 1 ) "*"                          "*"       "*"
##           YBL1950 to 1959 YBL1960 to 1969 YBL1970 to 1979 YBL1980 to 1989
## 1   ( 1 )  " "              " "             " "             " "
## 2   ( 1 )  " "              " "             " "             " "
## 3   ( 1 )  " "              " "             " "             " "
## 4   ( 1 )  " "              " "             " "             " "
## 5   ( 1 )  " "              " "             " "             " "
## 6   ( 1 )  " "              " "             " "             " "
## 7   ( 1 )  " "              " "             " "             " "
## 8   ( 1 )  " "              " "             " "             " "
## 9   ( 1 )  " "              " "             "*"             " "
## 10  ( 1 ) " "              " "             " "             " "
## 11  ( 1 ) " "              " "             "*"             " "
## 12  ( 1 ) " "              " "             "*"             " "
## 13  ( 1 ) " "              " "             "*"             " "
## 14  ( 1 ) " "              " "             "*"             " "
## 15  ( 1 ) " "              " "             "*"             " "
## 16  ( 1 ) " "              " "             "*"             "*"
## 17  ( 1 ) "*"              " "             "*"             "*"
## 18  ( 1 ) "*"              " "             "*"             "*"
## 19  ( 1 ) "*"              "*"             "*"             "*"
## 20  ( 1 ) "*"              " "             "*"             "*"
## 21  ( 1 ) "*"              " "             "*"             "*"
## 22  ( 1 ) "*"              "*"             "*"             "*"
## 23  ( 1 ) "*"              "*"             "*"             "*"
## 24  ( 1 ) "*"              "*"             "*"             "*"
## 25  ( 1 ) "*"              "*"             "*"             "*"
```

```
## 26  ( 1 ) "*"             "*"             "*"             "*"
## 27  ( 1 ) "*"             "*"             "*"             "*"
## 28  ( 1 ) "*"             "*"             "*"             "*"
## 29  ( 1 ) "*"             "*"             "*"             "*"
## 30  ( 1 ) "*"             "*"             "*"             "*"
## 31  ( 1 ) "*"             "*"             "*"             "*"
## 32  ( 1 ) "*"             "*"             "*"             "*"
## 33  ( 1 ) "*"             "*"             "*"             "*"
## 34  ( 1 ) "*"             "*"             "*"             "*"
## 35  ( 1 ) "*"             "*"             "*"             "*"
## 36  ( 1 ) "*"             "*"             "*"             "*"
## 37  ( 1 ) "*"             "*"             "*"             "*"
##           YBL1990 to 1999 YBL2000 to 2004 YBL2005 YBL2006 YBL2007 YBL2008
## 1  ( 1 )  " "             " "             " "     " "     " "     " "
## 2  ( 1 )  " "             " "             " "     " "     " "     " "
## 3  ( 1 )  " "             " "             " "     " "     " "     " "
## 4  ( 1 )  " "             " "             " "     " "     " "     " "
## 5  ( 1 )  " "             " "             " "     " "     " "     " "
## 6  ( 1 )  " "             " "             " "     " "     " "     " "
## 7  ( 1 )  " "             " "             " "     " "     " "     " "
## 8  ( 1 )  " "             " "             " "     " "     " "     " "
## 9  ( 1 )  " "             " "             " "     " "     " "     " "
## 10  ( 1 ) " "             " "             " "     " "     " "     " "
## 11  ( 1 ) " "             " "             " "     " "     " "     " "
## 12  ( 1 ) " "             " "             " "     " "     " "     " "
## 13  ( 1 ) " "             " "             " "     " "     " "     " "
## 14  ( 1 ) " "             " "             " "     " "     " "     " "
## 15  ( 1 ) " "             " "             " "     " "     " "     " "
## 16  ( 1 ) " "             " "             " "     " "     " "     " "
## 17  ( 1 ) " "             " "             " "     " "     " "     " "
## 18  ( 1 ) " "             " "             " "     " "     " "     " "
## 19  ( 1 ) " "             " "             " "     " "     " "     " "
## 20  ( 1 ) " "             "*"             " "     " "     " "     " "
## 21  ( 1 ) " "             "*"             " "     "*"     " "     " "
## 22  ( 1 ) " "             "*"             " "     "*"     " "     " "
## 23  ( 1 ) " "             "*"             " "     "*"     " "     " "
## 24  ( 1 ) " "             "*"             " "     "*"     " "     " "
## 25  ( 1 ) " "             "*"             " "     "*"     " "     " "
## 26  ( 1 ) " "             "*"             " "     "*"     " "     " "
## 27  ( 1 ) " "             "*"             " "     "*"     " "     " "
## 28  ( 1 ) " "             "*"             "*"     "*"     " "     " "
## 29  ( 1 ) " "             "*"             "*"     "*"     "*"     " "
## 30  ( 1 ) " "             "*"             "*"     "*"     "*"     " "
## 31  ( 1 ) " "             "*"             "*"     "*"     "*"     " "
## 32  ( 1 ) "*"             "*"             "*"     "*"     "*"     "*"
## 33  ( 1 ) "*"             "*"             "*"     "*"     "*"     "*"
## 34  ( 1 ) "*"             "*"             "*"     "*"     "*"     "*"
## 35  ( 1 ) "*"             "*"             "*"     "*"     "*"     "*"
## 36  ( 1 ) "*"             "*"             "*"     "*"     "*"     "*"
## 37  ( 1 ) "*"             "*"             "*"     "*"     "*"     "*"
##           YBL2009 YBL2010 YBL2011 YBL2012 YBL2013 YBL2014 YBL2015 R18none
## 1  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
```

```
## 4  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
## 9  ( 1 )  " "     " "     " "     " "     " "     " "     " "     " "
## 10  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 11  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 12  ( 1 ) " "     " "     " "     " "     " "     " "     " "     "*"
## 13  ( 1 ) " "     " "     " "     " "     " "     " "     " "     "*"
## 14  ( 1 ) " "     " "     " "     " "     " "     " "     " "     "*"
## 15  ( 1 ) " "     " "     " "     " "     " "     " "     " "     "*"
## 16  ( 1 ) " "     " "     " "     " "     " "     " "     " "     "*"
## 17  ( 1 ) " "     " "     " "     " "     " "     " "     " "     "*"
## 18  ( 1 ) " "     " "     " "     " "     " "     "*"     " "     "*"
## 19  ( 1 ) " "     " "     " "     " "     " "     "*"     " "     "*"
## 20  ( 1 ) " "     " "     " "     " "     "*"     "*"     " "     "*"
## 21  ( 1 ) " "     " "     " "     " "     "*"     "*"     " "     "*"
## 22  ( 1 ) " "     " "     " "     " "     "*"     "*"     " "     "*"
## 23  ( 1 ) " "     " "     " "     " "     "*"     "*"     "*"     "*"
## 24  ( 1 ) " "     " "     " "     " "     "*"     "*"     " "     "*"
## 25  ( 1 ) " "     " "     " "     " "     "*"     "*"     "*"     "*"
## 26  ( 1 ) " "     " "     " "     " "     "*"     "*"     "*"     "*"
## 27  ( 1 ) "*"     " "     " "     " "     "*"     "*"     "*"     "*"
## 28  ( 1 ) "*"     " "     " "     " "     "*"     "*"     "*"     "*"
## 29  ( 1 ) "*"     " "     " "     " "     "*"     "*"     "*"     "*"
## 30  ( 1 ) "*"     " "     " "     " "     "*"     "*"     "*"     "*"
## 31  ( 1 ) "*"     " "     " "     " "     "*"     "*"     "*"     "*"
## 32  ( 1 ) "*"     " "     " "     " "     "*"     "*"     "*"     "*"
## 33  ( 1 ) "*"     " "     " "     " "     "*"     "*"     "*"     "*"
## 34  ( 1 ) "*"     " "     "*"     " "     "*"     "*"     "*"     "*"
## 35  ( 1 ) "*"     " "     "*"     "*"     "*"     "*"     "*"     "*"
## 36  ( 1 ) "*"     "*"     "*"     "*"     "*"     "*"     "*"     "*"
## 37  ( 1 ) "*"     "*"     "*"     "*"     "*"     "*"     "*"     "*"
##           R60none HAhouse
## 1  ( 1 )  " "     " "
## 2  ( 1 )  " "     " "
## 3  ( 1 )  " "     " "
## 4  ( 1 )  " "     "*"
## 5  ( 1 )  " "     "*"
## 6  ( 1 )  " "     "*"
## 7  ( 1 )  " "     "*"
## 8  ( 1 )  "*"     "*"
## 9  ( 1 )  "*"     "*"
## 10  ( 1 ) "*"     "*"
## 11  ( 1 ) "*"     "*"
## 12  ( 1 ) "*"     "*"
## 13  ( 1 ) "*"     "*"
## 14  ( 1 ) "*"     "*"
## 15  ( 1 ) "*"     "*"
## 16  ( 1 ) "*"     "*"
## 17  ( 1 ) "*"     "*"
## 18  ( 1 ) "*"     "*"
## 19  ( 1 ) "*"     "*"
```

```
## 20  ( 1 ) "*"     "*"
## 21  ( 1 ) "*"     "*"
## 22  ( 1 ) "*"     "*"
## 23  ( 1 ) "*"     "*"
## 24  ( 1 ) "*"     "*"
## 25  ( 1 ) "*"     "*"
## 26  ( 1 ) "*"     "*"
## 27  ( 1 ) "*"     "*"
## 28  ( 1 ) "*"     "*"
## 29  ( 1 ) "*"     "*"
## 30  ( 1 ) "*"     "*"
## 31  ( 1 ) "*"     "*"
## 32  ( 1 ) "*"     "*"
## 33  ( 1 ) "*"     "*"
## 34  ( 1 ) "*"     "*"
## 35  ( 1 ) "*"     "*"
## 36  ( 1 ) "*"     "*"
## 37  ( 1 ) "*"     "*"
```

After that, it was time to compute the validation set error for the best model corresponding to different sizes. In order to do so, we must set up a test matrix for the test data, run a loop, calculating the predictions through the extracted coefficients and then computing the MSE. They are then stored in the test matrix.

```
# Compute the validation set error for the best model (corresponding to different sizes).
test_mat <- model.matrix(ELEP ~ ., data = housing_valid)

# Run a loop and compute the MSE
val_errors <- rep(NA, 37)
for(i in 1:37){
  coefi <- coef(regfit_best, id = i)
  pred <- test_mat[,names(coefi)] %*% coefi
  val_errors[i] <- mean((housing_df$ELEP[valid] - pred)^2)
}

# Which model is best:
val_errors
```

```
##  [1] 5753.824 5598.843 5327.022 5228.428 5159.786 5105.763 5083.950 5074.089
##  [9] 5068.201 5077.058 5070.788 5071.101 5073.095 5061.987 5058.134 5054.116
## [17] 5048.634 5054.045 5050.790 5053.084 5053.192 5050.835 5048.053 5051.458
## [25] 5048.665 5048.503 5048.539 5047.495 5046.724 5046.685 5044.444 5045.373
## [33] 5045.367 5046.316 5046.504 5046.367 5046.355
```

```
which.min(val_errors)
```

```
## [1] 31
```

```
coef(regfit_best, 31)
```

```
##                   (Intercept)                            NP
##                   1.285689009                 13.173602481
##                          BDSP                          FULP
```

```
##                     7.364917522                     0.005973527
##                            GASP                  HFLCoal or coke
##                     0.151044010                   -18.757587772
##                  HFLElectricity                  HFLNo fuel used
##                    36.024171659                     4.515093130
##                   HFLOther fuel                  HFLSolar energy
##                    33.106073394                   -18.145819853
##                  HFLUtility gas                        HFLWood
##                   -19.681726806                   15.915833299
##                            RMSP          TENOwned free and clear
##                     3.529936641                   -10.125259566
## TENOwned with mortgage or loan                      TENRented
##                    -5.934231381                   -6.581020574
##                  YBL1940 to 1949                 YBL1950 to 1959
##                     8.029600507                     4.055254761
##                  YBL1960 to 1969                 YBL1970 to 1979
##                     2.214030566                     7.691374318
##                  YBL1980 to 1989                 YBL2000 to 2004
##                     5.135691272                    -3.718285831
##                         YBL2005                         YBL2006
##                    -2.169941976                   -6.061412999
##                         YBL2007                         YBL2009
##                    -2.120469310                    4.650257419
##                         YBL2013                         YBL2014
##                   -10.493756019                  -18.599041501
##                         YBL2015                         R18none
##                   -15.710373240                    6.131774672
##                         R60none                         HAhouse
##                    -5.000573632                   31.739588389
```

Since we want to minimize the error, the best model is the one that contains the lowest error which is 31 variables. While this seems like a lot of variables, it is because multiple variables correspond to one predictor variable. After grouping and counting the unique predictor variables this subset of variables falls under, it leads me to conclude that the best model contains all 11 predictors (NP, BDSP, HA, FULP, GASP, HFL, RMSP, TEN, YBL, R18, and R60). This is not the most interesting answer in and of itself, but the cross-validation confirms it. The last thing to so is to compute the Root Mean Squared error (RMSE), which turned out to be $71.04.

```
full_model <- lm(ELEP ~ NP + BDSP + HA + FULP + GASP + HFL + RMSP + TEN + YBL + R18 + R60,
                 data = housing_train)


housing_valid$pred <- predict(full_model, newdata = housing_valid)
(rmse <- with(housing_valid, sqrt(mean((ELEP - pred)^2))))
```

```
## [1] 71.0377
```

## Discussion

The main differences between my approach to the first and second question has to do with the purpose of what I wanted to achieve. For the first question, the purpose was to determine which of the predictor variables in the housing data actually had an impact on electricity prices. Even the magnitude of the

coefficients tell us about the strength of the relationship that an input variable has on the output variable. Once the correct model had been chosen, the confidence intervals gave us a "snapshot" of the current state of the data. The second question wanted to find the correct model, but the ultimate goal was to evaluate the performance of the model with regards to how it fit new (test) data. So this could be used for knowing how good it would be at predicting the electricity costs for any household in Oregon. It was important to split the data in this case (20% test data) because we would have no way of knowing if the model needs to be adjusted to be more flexible or more rigid.