ST517-Final Project

Nora Quick

```
housing = read.csv("OR_acs_house_occ.csv", header=TRUE);
#str(OR);
#head(housing);
```

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

```
new_housing <- filter(housing, housing$NP >= 1 & housing$BDSP == 5 & housing$BLD != "Mobile home or tra
```

I first filtered down the data as required to the building having one or more people living in it, 5 bedrooms (based on the information I gave in the Project Strategy), and removing all buildings that weren't a housing unit or apartment unit.

```
m1 <- lm(log(ELEP) ~ BLD + NP, data = new_housing)
summary(m1)
```

```
##
## Call:
## lm(formula = log(ELEP) ~ BLD + NP, data = new_housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66198 -0.34192 -0.01682  0.35123  1.56762
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.91825    0.33623  14.627  < 2e-16 ***
## BLD3-4 Apartments              0.19901    0.66628   0.299    0.765
## BLDOne-family house attached -0.28895    0.52711  -0.548    0.584
## BLDOne-family house detached -0.35471    0.33423  -1.061    0.289
## NP                             0.08021    0.01538   5.215 2.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5769 on 453 degrees of freedom
## Multiple R-squared:  0.06073,    Adjusted R-squared:  0.05243
## F-statistic: 7.322 on 4 and 453 DF,  p-value: 1.014e-05
```

Here I don't add in number of bedrooms because I have already narrowed it down to only 5 bedrooms so all data seen here should be 5 bedrooms.

```r
confint(m1)
```

```
##                                    2.5 %    97.5 %
## (Intercept)                   4.25747455 5.5790209
## BLD3-4 Apartments            -1.11038148 1.5084015
## BLDOne-family house attached -1.32482946 0.7469341
## BLDOne-family house detached -1.01154523 0.3021233
## NP                            0.04997826 0.1104325
```

For the next part we want to look at apartment vs. housing. This is similar to the gender comparisons in module 5 where we want to look at $\hat{\beta}_1$ which there was males. However, here we have two housing I chose to make the $\hat{\beta}_1$ equal to apartments (3-4 Apartments).

```r
exp(0.19901)
```

```
## [1] 1.220194
```

Based on this outcome our answer to the question would be that we have a 95% confidence that people living in apartments pay more on electricity than than people living in houses. They pay about 22% more on electricity based on the outcome above.

This does seem a bit wrong as logically I would say that people living in houses pay more than people living in apartments. However, I did narrow down the data quite a bit which skews the data. In addition to that we can see that the number of people is quite significant in determining how much is spent on electricity every month.

```r
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: log(ELEP)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## BLD         3   0.698  0.2326  0.6989    0.5531
## NP          1   9.049  9.0486 27.1915 2.808e-07 ***
## Residuals 453 150.746  0.3328
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From this ANOVA test we can see that the number of people in actually the most determening factor about how much a household/apartment for electricity not building.

```r
det_mean <- mean(new_housing$ELEP[new_housing$BLD == ’One-family house detached’])
att_mean <- mean(new_housing$ELEP[new_housing$BLD == ’One-family house attached’])
house_mean <- det_mean + att_mean
house_mean
```

```
## [1] 300.177
```

```r
two_mean <- mean(new_housing$ELEP[new_housing$BLD == ’2 Apartments’])
tf_mean <- mean(new_housing$ELEP[new_housing$BLD == ’3-4 Apartments’])
apt_mean <- two_mean + tf_mean
apt_mean
```

```
## [1] 440
```

Because I didn't believe my answers I decided to look through my whole filtered table and found all the houses and apartments and found the mean of all the electic bills. I found that based on the filtered data I have apartments do pay more for electicity than houses.

```
diff <- 440 - 300.177
diff
```

```
## [1] 139.823
```

How much? Apartments pay about $139.8 more than houses on average.

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

```
pred_data <- within(new_housing, rm(TYPE, BDSP, BLD))
pred_data <- na.omit(pred_data)

apply(pred_data, 2, function(x) length(unique(x)))
```

```
## SERIALNO       NP      ACR     ELEP     FULP     GASP      HFL     RMSP
##      409       10        3       39       31       25        6        8
##      TEN     VALP      YBL      R18      R60
##        2      108       18        2        2
```

For the first part of this question I had quite a large issue with variables that were not working with the data. With some diffing I found that anything with only 1 unique value needed to be removed. After this I needed to remove any NAs. This caused more variables to only have 1 unique value so I removed that as well and was finally left with a set of data I could work with.

I decided to use 39 for my nvmax because of the length of unique variables for ELEP.

```
set.seed(1)

train <- pred_data %>% sample_frac(0.5)
test <- pred_data %>% setdiff(train)

regfit.best_train <- regsubsets(ELEP ~ ., data = train, nvmax = 39)
regfit.best_train_summary <- summary(regfit.best_train)

test_mat <- model.matrix(ELEP ~ ., data = test)

val_errors <- rep(NA, 39)

for(i in 1:39){
  coefi <- coef(regfit.best_train, id = 1)
  pred <- test_mat[,names(coefi)] %*% coefi
  val_errors[i] <- mean((test$ELEP - pred)^2)
}

#val_errors
```

This model created finds the predicted electrcity cost by splitting the prediction dataset into two halves. The first half is for prediction and the second half is for testing. It was found that it predicted about $8821.6.

3. 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?

My approach for question one required quite a bit of editing down data due to the descibed necessity for it. I needed to make sure that I had the number of occupants and rooms that I wanted and then narrowed down the living units to what I wanted to look at. This causes a limitiation to the data shortly descibed in question 1. To explain again, the limitations narrow down the outcome we see and it may not truly represent the real (unedited) data.

My approach for question two relied upon the data I had edited down in question one. In addition to that I needed to delete more data (descibed why in question 2). Once the data was severly limited down I was able to split the data in half to predict and test the data. With the data split in half it was even more limited to what it could predict. In addition to that I found that the limited initial data caused issues with the additional limited data and it would have been best not to use the same data.

The difference between the two was based upon the data we looked at. Qusetion one had limitations for a reason and question two had limitations that required elimination. In addition to the approaches being different because of quality of data, they were different in a way of finding one based on all data and one based on halving the data.

The different approaches were required because they were needed to answer the questions. The first question required limitations because of desired narrowing and the second because of required narrowing. I couldn't predict the model with all the data given as some of it didn't work in the formatting.

I had challenges for each of the questions but they were similar in design. For question one I had challenges because I needed to limit the data and I did it based on research I did about the Oregon housing market. I wanted to find exact numbers which I had to find a way to reasonably narrow down to still work within the data. Question two had a similar challenge because I needed to narrow down the variables but this time due to the regsubsets function not liking my variables. After quite a bit of looking I found that I needed to eliminate all vairables that only had one unique variable and all rows that were NA. This finally allowed me to predict the data I wanted but the removal process took some time. Overall, both tasks has challenges based on data removal/narrowing so that I was always looking at the data I wanted.