Final Project

Liam Gunn, Delaney Heileman, Elaine Lac

4/26/2022

#Introduction - Background and Motivation

*For this project, we used the 2015 Residential Energy Consumption Survey conducted by the Energy Information Agency. This survey asks a range of questions regarding how much, how often, and what type of energy was used. Our group wanted to examine how various variables impact household electricity usage, in kilowatt-hours.*

#Loading Data

*This data set originally has 759 different variables and 5686 observations. For our project we decide on 21 variables of interest. Below is a table of the variables we selected and what they represent.*

#Data Cleaning

*Our data cleaning was relatively straight forward. For this process we created a data frame called cleanRECS.*

#Removing "not applicable" answers for thermostat questions. N/a for Thermostat likely means they use wood fire or something else  
cleanRECS <- recs1[recs1$THERMAIN != -2 & recs1$THERMAINAC != "-2",]

*We started by removing thermostat values that were “-2” which are coded as “N/A” in the code book. We decided to remove these values because we were only interested in the binary of if someone does or does not have a thermostat.*

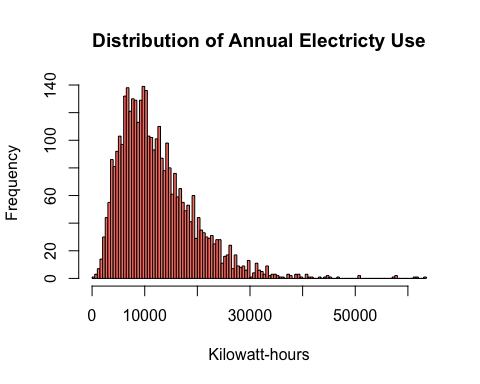
cleanRECS$UrbanRural <- recode(cleanRECS$UATYP10, 'C' = 'U')  
cleanRECS$HeatOrEat <- as.factor(recode(cleanRECS$SCALEB, '2' = '1', '3' = '1', '1' = '1', '0' = '0'))

*We then looked at the UATYP10 variable, which indicated if a home is located in a Urban, Suburban, or Rural setting. We decided to create a new variable called UrbanRural which re-coded all suburban values as urban. We did this because we thought the binary Rural v. Non-rural would be interesting to explore and because the suburban and urban values were very similar. We also created a new category called HeatOrEat to capture instances of energy poverty. The HeatOrEat variable is now binary and shows if a household has gone without necessities at any point during the year in order to afford utility bills.*

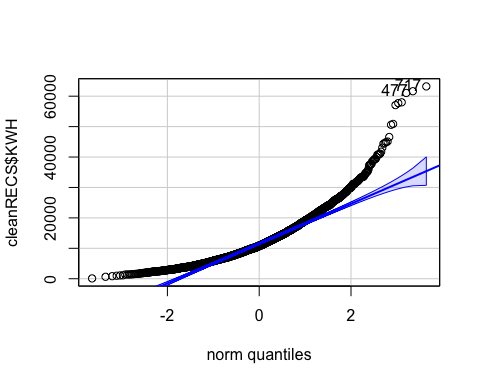
cleanRECS$REGIONC <- as.factor(recode(cleanRECS$REGIONC, '1' = 'Northeast', '2' = 'Midwest', '3' = 'South', '4' = 'West'))  
  
cleanRECS$DIVISION <- as.factor(recode(cleanRECS$DIVISION, '1' = 'New England', '2' = 'Mid Atlantic', '3' = 'East North Central', '4' = 'West North Central', '5' = 'South Atlantic', '6' = 'East South Central', '7' = 'West South Central', '8' = 'Mountain North', '9' = 'Mountain South', '10'= 'Pacific'))  
  
cleanRECS$CENACHP <- recode(cleanRECS$CENACHP, '1' = 'YES', '0' = 'NO')  
  
#recoding energy assistance variable to yes or no value  
cleanRECS$ENERGYASST <- recode(cleanRECS$ENERGYASST, '1' = 'YES', '0' = 'NO')

*We then re-coded some of our values with names to so that future graphics would be more legible*

hist(cleanRECS$KWH, col = "salmon", main = "Distribution of Annual Electricty Use", xlab = "Kilowatt-hours", breaks = 200)



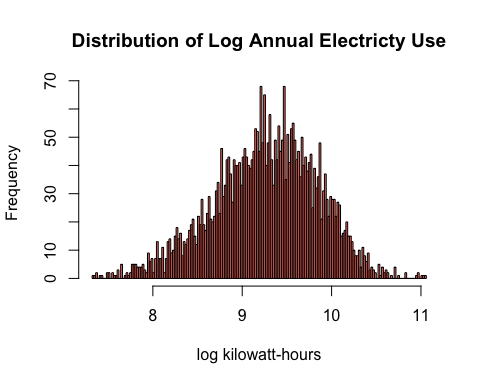
qqPlot(cleanRECS$KWH)



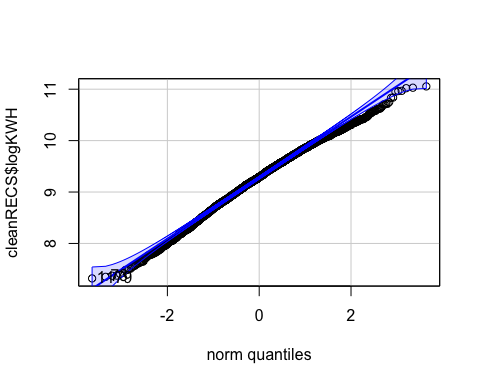
## [1] 717 477

*Since we will mainly be looking at how different variables impact annual energy consumption in KWH we wanted to ensure the KWH variable was normally distributed. Our histogram and QQ-plot both indicate the KWH variable is not normally distributed. To remedy this we removed data points below 1500KWH and took the log.*

#At the advice of the professor, create a cutoff for KWH < 1500 in order to remove influential variables which are affecting normality. Anything households whose annual usage was less than 1500 KWH was re-labeled as NA and then removed. Then KWH is log transformed.   
cleanRECS$KWH[cleanRECS$KWH < 1500] <- NA  
cleanRECS <- na.omit(cleanRECS)  
  
cleanRECS$logKWH <- log(cleanRECS$KWH)  
  
hist(cleanRECS$logKWH, col = "salmon", main = "Distribution of Log Annual Electricty Use", xlab = "log kilowatt-hours", breaks = 150)



qqPlot(cleanRECS$logKWH)



## [1] 1179 73

#Converting the Income Brackets into a numeric variable. Because we do not know the exact variables, we are assuming that each household is at the upper limit of that income bracket, and removing the highest income bracket that has no upper limit to avoid making unbounded assumptions about their incomes.  
cleanRECS$IncBracket <- as.numeric(recode(cleanRECS$MONEYPY, '1' = '20000', '2' = '40000', '3' = '60000', '4' = '80000', '5' = '100000', '6' = '120000', '7' = '140000', '8' = 'NA'))

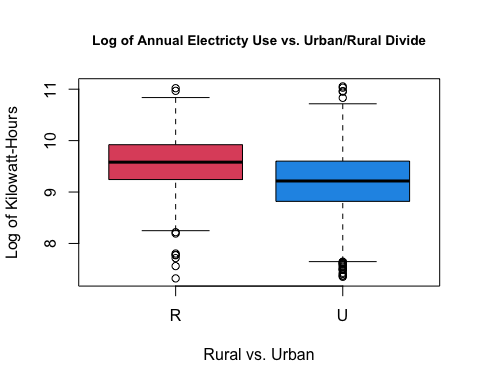
## Warning: NAs introduced by coercion

*After the log transformation the KWH variable was still not perfectly normal but it was much closer and was now more reasonable to work with.*

# T-Test

*After playing around with the data it looked like a t-test based on Kilowatt-Hours based on Rural v. Urban would be interesting. In our data cleaning we already created a logKWH variable, to account for normalcy, so that is what we will be using. We began by looking at a boxplot of the log transformed data*

boxplot(logKWH ~ UrbanRural, main = "Log of Annual Electricty Use vs. Urban/Rural Divide", ylab = "Log of Kilowatt-Hours", data = cleanRECS, col = as.factor(cleanRECS$logKWH), xlab = "Rural vs. Urban", cex.main = 0.85)



*Null hypothesis: There is no difference in KWH usage between the rural and urban groups.*

*Alternative hypothesis: There is a difference in KWH usage between the rural and urban groups.*

#t.test of KWH and region type   
(test1 <- t.test(logKWH ~ UrbanRural, data = cleanRECS, conf.level = .95))

##   
## Welch Two Sample t-test  
##   
## data: logKWH by UrbanRural  
## t = 16.178, df = 1153.2, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group R and group U is not equal to 0  
## 95 percent confidence interval:  
## 0.3173262 0.4049172  
## sample estimates:  
## mean in group R mean in group U   
## 9.554033 9.192912

test1$p.value < 0.05

## [1] TRUE

*The t-test shows us that there is a true difference in means between the rural and urban group with a p-value of 2.2e-16. This small p-value indicated high statistical significant. Thus, WE REJECT THE NULL HYPOTHESIS under a 95% confidence level. (α=0.05).*

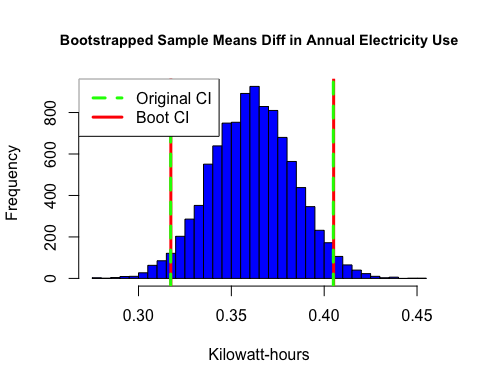
#Bootstrapping *To further test the KWH usage based on Rural v. Urban settings we will perform bootstrapping*

## [1] "Bootstrapped Confidence Intervals"

## 2.5% 97.5%   
## 0.32 0.41

## [1] "Original Confidence Intervals"

## [1] 0.32 0.40  
## attr(,"conf.level")  
## [1] 0.95

 *Our bootstrapped confidence interval is nearly identical to the original confidence interval, further supporting our rejection of the null hypothesis.*

#Permutation Test

*To further examine the difference in KWH usage between rural and urban groups we performed a permutation test.*

rm(mean)

## Warning in rm(mean): object 'mean' not found

(actualdiff <- by(cleanRECS$logKWH, cleanRECS$UrbanRural, mean))

## cleanRECS$UrbanRural: R  
## [1] 9.554033  
## ------------------------------------------------------------   
## cleanRECS$UrbanRural: U  
## [1] 9.192912

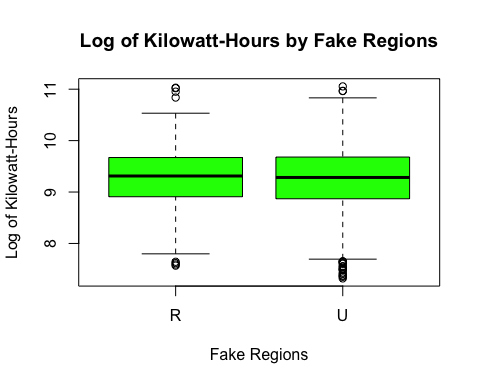
(actualdiff <- actualdiff[1] - actualdiff[2])

## R   
## 0.3611217

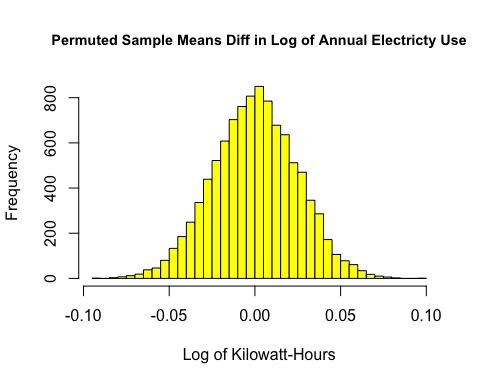
fakeregion <- sample(cleanRECS$UrbanRural) # default is replace = FALSE - this is a PERMUTATION (a reordering) of the age group assignments  
sample(1:10)

## [1] 1 8 9 3 6 5 10 4 2 7

boxplot(cleanRECS$logKWH ~ fakeregion, main = "Log of Kilowatt-Hours by Fake Regions", col = "green", ylab = "Log of Kilowatt-Hours", xlab = "Fake Regions")



diff0 <- mean(cleanRECS$logKWH[fakeregion == "R"]) - mean(cleanRECS$logKWH[fakeregion == "U"])  
  
set.seed(1)  
N <- 10000  
diffvals2 <- rep(NA, N)  
for (i in 1:N) {  
 fakeregion <- sample(cleanRECS$UrbanRural) # default is replace = FALSE  
 diffvals2[i] <- mean(cleanRECS$logKWH[fakeregion == "R"]) - mean(cleanRECS$logKWH[fakeregion == "U"])  
}

 *Our permutation test found no significant difference in the means of two random groupings of households, UNLIKE in our real-world data of urban & suburban households versus rural. This supports reject the null hypothesis of no difference between these two regions.*

#ANCOVA - Analysis of Covariance *The ANCOVA model used three variables. The dependent variable was a continuous log transformed kilowatt-hour variable (logKWH) and the independent variables were a continuous variable of annual gross income (MONEYPY) and a categorical variable of whether a household had a heat pump (CENACHP). The MONEYPY variable is ordered from lowest income (1 - <20,000 USD) to highest income (8 - >140,000 USD). CENACHP records ‘No’ for no heat pump and ‘Yes’ for having a heat pump.*

* KWH (continuous): on a scale of annual kilowatt-hours used by households in the US in the year 2015
* logKWH (continuous): on a log transformed scale of kilowatt-hours used by households in the US in the year 2015
* CENACHP (categorical): the presence of a heat pump “Yes” or “No”
* MONEYPY (continuous): Gross annual income of a household in the year 2015. 8 levels.

*The ANCOVA model aims to fit a model predicting the continuous variable kilowatt-hours ‘KWH’ based on and the categorical variable of whether households have a heat pump or not ‘CENACHP’, and whether there is a significant interaction between the continuous variable of annual gross income ’MONEYPY and having a heat pump.*

# Subset the dataset used for the ANCOVA model, clean the data, and check the characteristics of the dataset.  
ancmod <- cleanRECS[c("KWH", "MONEYPY", "CENACHP")]  
ancmod$CENACHP <- as.factor(ancmod$CENACHP)  
dim(ancmod)

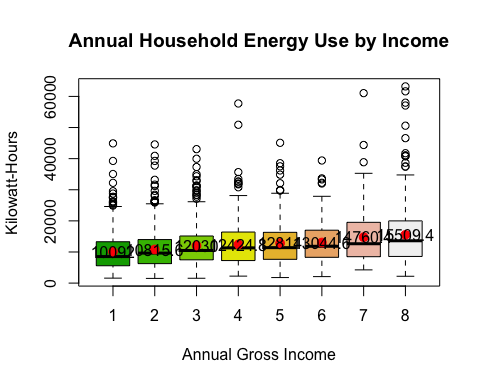
## [1] 3671 3

str(ancmod)

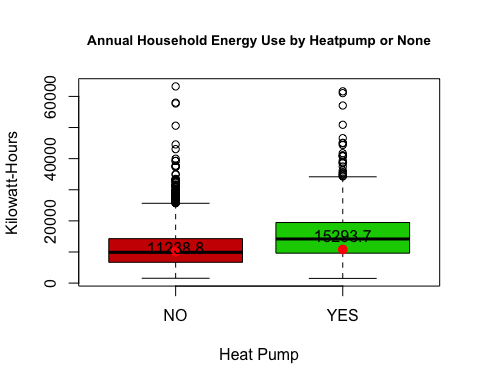
## 'data.frame': 3671 obs. of 3 variables:  
## $ KWH : num 5271 19655 9853 3116 2398 ...  
## $ MONEYPY: int 8 2 3 3 4 5 6 2 6 7 ...  
## $ CENACHP: Factor w/ 2 levels "NO","YES": 1 1 1 1 2 1 1 1 1 1 ...

## Exploratory Data Analysis

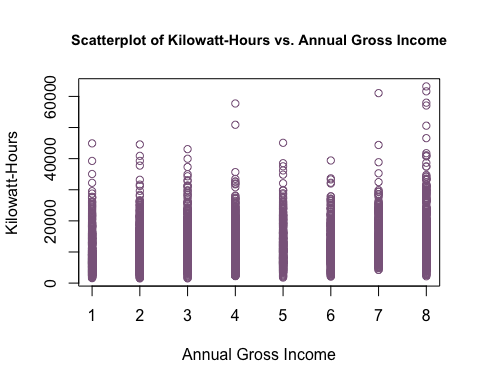
#Create initial exploratory boxplots and scatterplots to determine whether the spread and the means of the data indicate similar variance or evidence of related variables. Only a scatterplot is made for KWH vs. MONEYPY because they are the only continuous variables.   
attach(ancmod)  
boxplot(KWH ~ MONEYPY, data = ancmod, col = terrain.colors(length(unique(ancmod$MONEYPY))), xlab = "Annual Gross Income", ylab = "Kilowatt-Hours", main = "Annual Household Energy Use by Income")  
means1 <- round(tapply(KWH, MONEYPY, mean),1)  
points(means1, col = 'red', pch = 19, cex = 1.2)  
text(x = c(1:8), y = means1, labels = round(means1, 2))



boxplot(KWH ~ CENACHP, data = ancmod, col = c("red3", "green3"), xlab = "Heat Pump", ylab = "Kilowatt-Hours", main = "Annual Household Energy Use by Heatpump or None", cex.main = 0.85)  
means2 <- round(tapply(KWH,CENACHP, mean),1)  
points(means1, col = 'red', pch = 19, cex = 1.2)  
text(x = c(1:length(unique(ancmod$MONEYPY))), y = means2, labels = round(means2, 2))



plot(KWH ~ MONEYPY, data = ancmod, col = "plum4", pch = 21, xlab = "Annual Gross Income", main = "Scatterplot of Kilowatt-Hours vs. Annual Gross Income", ylab = "Kilowatt-Hours", cex.main = 0.9)

 *The initial boxplots indicate that kilowatt-hour means vary across annual gross income, and that kilowatt-hour means may vary between having or not having a heat pump. The spreads of annual gross income and the presence of a heat pump aren’t vastly different. However, boxplots are not a true indication of interactions occurring between variables so other tests need to be completed to verify interactions.*

## Assumption Verification

*The following assumptions must be met in order to proceed with an ANCOVA analysis: the dependent variables should be normally distributed, data should have approximately the same variance.*

#Check Variance Assumption  
stdev2 <- tapply(KWH, CENACHP, sd)  
round(max(stdev2)/min(stdev2),1)

## [1] 1.2

round(max(stdev2)/min(stdev2),1) <= 2

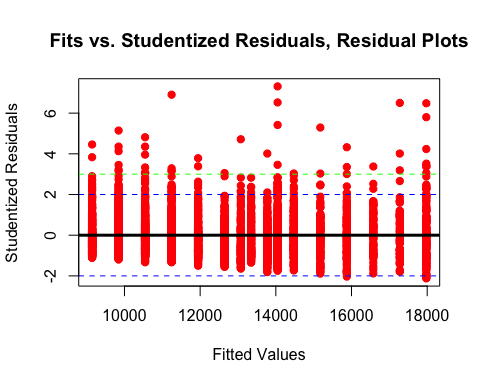
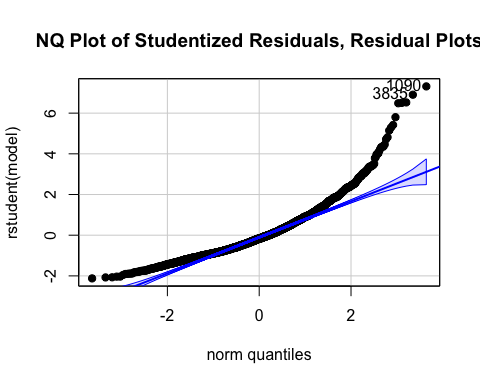
## [1] TRUE

*The variances of KWH and CENACHP are 1.2 so less than 2 thus we satisfy the equal variances assumption to conduct an ANCOVA test. We don’t run this test on MONEYPY because the test would only be testing values of 1-8 and wouldn’t yield any meaningful insights.*

#check normality assumptions for KWH and MONEYPY  
ancmod$KWH[ancmod$KWH < 1500] <- NA  
m1 <- lm(KWH ~ CENACHP + MONEYPY, data = ancmod)  
summary(m1)

##   
## Call:  
## lm(formula = KWH ~ CENACHP + MONEYPY, data = ancmod)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14368 -4766 -1114 3322 49170   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8443.88 238.16 35.45 <2e-16 \*\*\*  
## CENACHPYES 3930.29 248.52 15.81 <2e-16 \*\*\*  
## MONEYPY 700.33 49.67 14.10 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6775 on 3668 degrees of freedom  
## Multiple R-squared: 0.1126, Adjusted R-squared: 0.1121   
## F-statistic: 232.7 on 2 and 3668 DF, p-value: < 2.2e-16

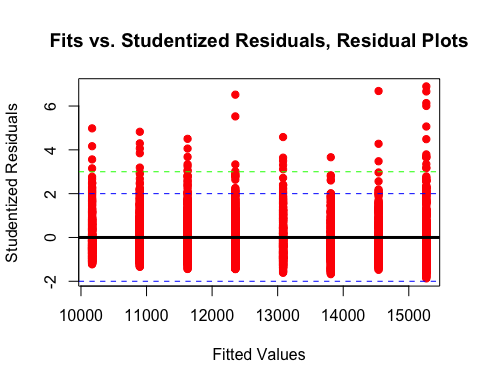
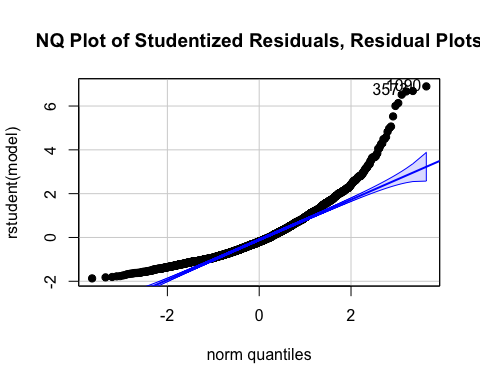
myResPlots2(m1)



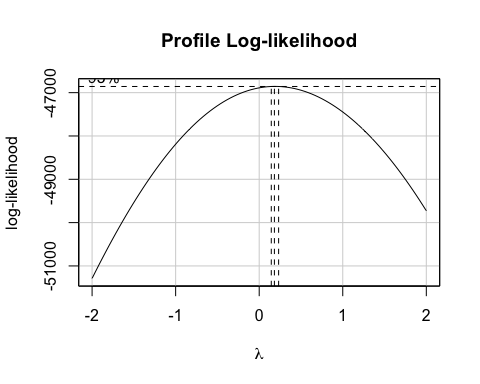
#note anomalies   
m2 <- lm(KWH ~ MONEYPY, data = ancmod)  
summary(m2)

##   
## Call:  
## lm(formula = KWH ~ MONEYPY, data = ancmod)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13071 -5060 -1374 3584 47950   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9441.3 237.3 39.78 <2e-16 \*\*\*  
## MONEYPY 728.2 51.3 14.20 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7001 on 3669 degrees of freedom  
## Multiple R-squared: 0.05207, Adjusted R-squared: 0.05181   
## F-statistic: 201.5 on 1 and 3669 DF, p-value: < 2.2e-16

myResPlots2(m2)



#Normality assumptions have not been met by the residuals so a transformation needs to be made. A box cox test will indicate what transformation is appropriate for the data.   
trans1 <- boxCox(m1)



boxcoxtrans <- trans1$x[which.max(trans1$y)]  
print(boxcoxtrans)

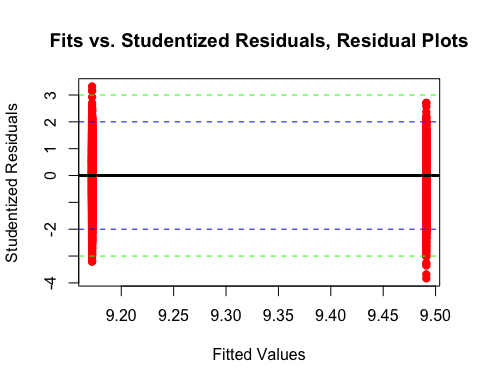
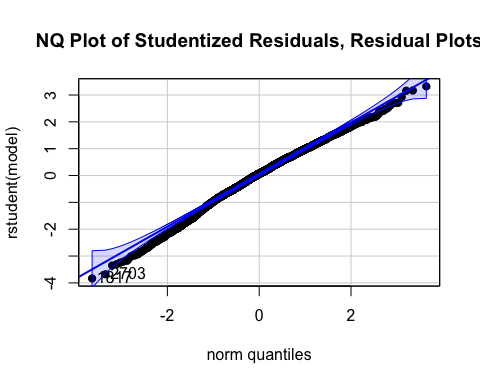
## [1] 0.1818182

*The values in KWH less than 1500 were creating large outliers in the dataset, and thus large residuals which were affecting the normality of the errors. At the advice of Professor JDRS, a cutoff was imposed on values less than 1500. Normality of the model for the first ‘m1’ model fitting KWH and CENACHP show evidence of non normality from the normal quantile plot which is not linear and evidence of outliers and influential points from the residual plots which show a heavy skew of positive residuals. This model has a r-squared value of 0.1126, meaning 11.3% of the variability of kilowatt-hour usage is explained by the presence of a heat pump. The ‘CENACHP’ variable has a lower p-value of less than 2e-16 meaning it is a statistically significant predictor of kilowatt-hours. The second model ‘m2’ fitting KWH and MONEYPY also show similar evidence of non-normality from a non linear and curved normal quantile plot and residuals heavily positive skewed showing evidence of outliers and influential points. This indicates that the response variable needs some transformation. The ‘m2’ model has a r-squared value of 0.05207 meaning 5.2% of the predictability of kilowatt-hour usage can be explained by the variable annual gross income or ‘MONEYPY’. Like CENACHP, MONEYPY has a low p-value of less than 2e-16 meaning it is a statistically significant predictor. A box cox transformation results in a lambda of 0.18 which is close enough to 0 that we can take a log transformation.*

#Log Transformation   
ancmod$logKWH <- log(ancmod$KWH)  
#check normality assumptions for KWH and CENACHP and KWH and MONEYPY  
m3 <- lm(logKWH ~ CENACHP, data = ancmod)  
summary(m3)

##   
## Call:  
## lm(formula = logKWH ~ CENACHP, data = ancmod)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.17084 -0.35061 0.03312 0.39240 1.88218   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.17214 0.01106 829.22 <2e-16 \*\*\*  
## CENACHPYES 0.31902 0.02081 15.33 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5677 on 3669 degrees of freedom  
## Multiple R-squared: 0.06019, Adjusted R-squared: 0.05993   
## F-statistic: 235 on 1 and 3669 DF, p-value: < 2.2e-16

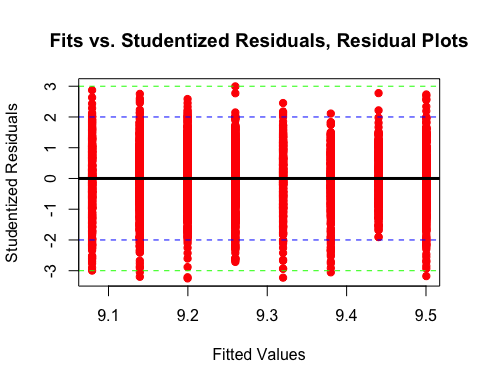
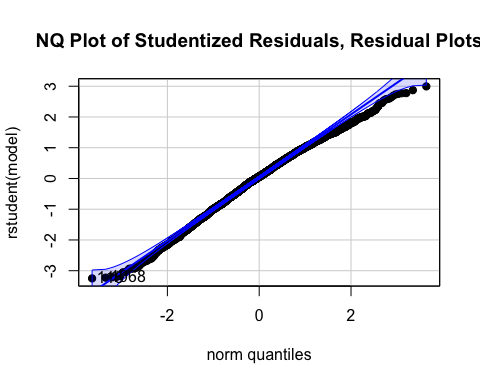
myResPlots2(m3)



m4 <- lm(logKWH ~ MONEYPY, data = ancmod)  
summary(m4)

##   
## Call:  
## lm(formula = logKWH ~ MONEYPY, data = ancmod)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.84880 -0.37763 0.03319 0.41159 1.70369   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.019167 0.019312 467.03 <2e-16 \*\*\*  
## MONEYPY 0.060156 0.004174 14.41 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5697 on 3669 degrees of freedom  
## Multiple R-squared: 0.05358, Adjusted R-squared: 0.05332   
## F-statistic: 207.7 on 1 and 3669 DF, p-value: < 2.2e-16

myResPlots2(m4)



*After transformation, model assumptions have been reasonably met after transformation. We manage to fix some of the heteroskedasticity and the residual plots are approximately normally distributed with equal spread around the 0. The normal quantile plot is approximately linear, which shows that transformation has eliminated some heteroskedasticity. There are a few extreme outliers though. Additionally, the model has over 3000 points which will still yield insight from the data. After the transformation the refitted model ‘m3’ of the log transformed KWH and CENACHP has a lower r-squared value of 0.06019, meaning the predictability of KWH is 6.0% explained by our model. CENACHP is still statistically significant as a predictor. The refitted model ‘m4’ of the log transformed KWH and MONEYPY has an improved r-squared value of 0.05358 meaning our model can predict 5.4% of KWH with the variable MONEYPY. MONEYPY is still a statistically significant predictor.*

##Permutation Test *Our permutation test null hypothesis: there is no significant effect to predicting log kilowatt-hours by our variables heat pump ‘CENACHP’, gross annual income ‘MONEYPY’, and the interaction of heat pumps x gross annual income.*

*The alternative hypothesis: there is a significant effect on predicting log kilowatt-hour by the variable for having a heat pump ‘CENACHP’.*

*The alternative hypothesis: there is a significant effect on predicting log kilowatt-hour by the variable for gross annual income ‘MONEYPY’.*

*The alternative hypothesis: there is a significant effect on predicting log kilowatt-hour by the interactions for having a heat pump ‘CENACHP’ and gross annual income ‘MONEYPY’.*

table(MONEYPY, CENACHP)

## CENACHP  
## MONEYPY NO YES  
## 1 316 117  
## 2 559 192  
## 3 409 185  
## 4 383 129  
## 5 266 110  
## 6 235 81  
## 7 149 69  
## 8 317 154

*The table of MONEYPY and CENACHP verifies that there are variables present for every level of the variable.*

fitanco <- lm(logKWH ~ MONEYPY + CENACHP + MONEYPY\*CENACHP, data = ancmod)  
  
Anova(fitanco, type = 3)

## Anova Table (Type III tests)  
##   
## Response: logKWH  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 50482 1 1.6571e+05 < 2.2e-16 \*\*\*  
## MONEYPY 33 1 1.0767e+02 < 2.2e-16 \*\*\*  
## CENACHP 7 1 2.2513e+01 2.167e-06 \*\*\*  
## MONEYPY:CENACHP 3 1 9.1807e+00 0.002463 \*\*   
## Residuals 1117 3667   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

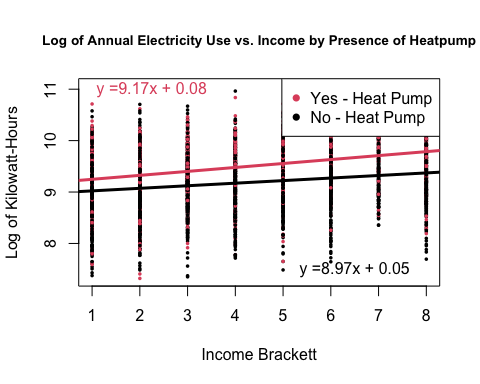
summary(fitanco)

##   
## Call:  
## lm(formula = logKWH ~ MONEYPY + CENACHP + MONEYPY \* CENACHP,   
## data = ancmod)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.00406 -0.35731 0.04563 0.38080 1.79088   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.972508 0.022041 407.078 < 2e-16 \*\*\*  
## MONEYPY 0.050023 0.004821 10.376 < 2e-16 \*\*\*  
## CENACHPYES 0.198088 0.041749 4.745 2.17e-06 \*\*\*  
## MONEYPY:CENACHPYES 0.026876 0.008870 3.030 0.00246 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5519 on 3667 degrees of freedom  
## Multiple R-squared: 0.1121, Adjusted R-squared: 0.1114   
## F-statistic: 154.3 on 3 and 3667 DF, p-value: < 2.2e-16

vif(fitanco)

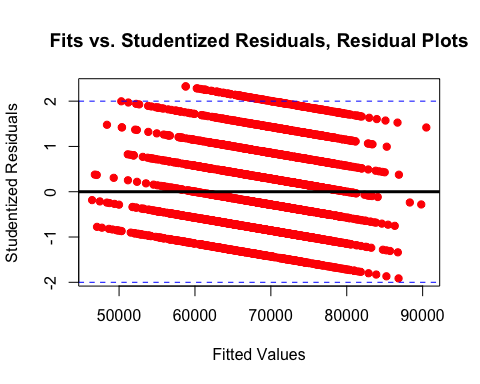
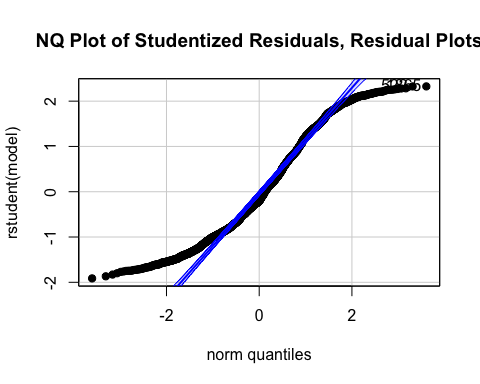
## MONEYPY CENACHP MONEYPY:CENACHP   
## 1.421048 4.257089 4.758726

anccoef <- coef(fitanco)  
plot(logKWH ~ MONEYPY, main = "Log of Annual Electricity Use vs. Income by Presence of Heatpump", ylab = "Log of Kilowatt-Hours", xlab = "Income Brackett", data = ancmod, col = factor(CENACHP), pch = 16, cex = .5, cex.main = 0.85)  
legend("topright", col = 2:1, legend = c("Yes - Heat Pump", "No - Heat Pump"), pch = 16)  
abline(a = anccoef[1], b = anccoef[2], col = 1, lwd = 3)  
text(x = 6.5, y = 7.5, col = 1, paste0("y =", round((anccoef[1]), 2), "x + ", round((anccoef[2]), 2)))  
abline(a = anccoef[1] + anccoef[3], b = anccoef[2] + anccoef[4], col = 2, lwd = 3)  
text(x = 2.25, y = 11, col = 2, paste0("y =", round((anccoef[1] + anccoef[3]), 2), "x + ", round((anccoef[2] + anccoef[4]), 2)))

 *The final fitted ANCOVA model is the log transformed kilowatt-hour variable (logKWH) predicted by annual gross income (MONEYPY), the presence of a heat pump (CENACHP), and the interaction of MONEYPY x CENACHP. The r-squared of this model improved from our separated predictors and our model explains 11.2% of the predictability of transformed kilowatt-hours. Each of the predictors have p-values less than our alpha 0.05 meaning they are statistically significant predictors. Additionally, the interaction term between MONEYPY and CENACHP is statistically significant with a p-value of 0.00246. The model indicates that with there is a higher kilowatt usage with a higher gross annual income at a rate of (8.97). The presence of a heat pump means an increase in kilowatt-hours at a rate of (8.97 + 0.20 = 9.17). The presence of a heat pump and a higher annual gross income have a increase in kilowatt-hour usage at the rate of (8.97 + 0.02 = 8.99) so the effect of the two together is fairly modest. The model has two equations fitted. One for the presence of a heat pump and an increase in gross annual income which is y = 9.17x + 0.08. The base equation without a heat pump and increasing gross annual income is y = 8.97x + 0.05. This model rejects the null hypothesis that log transformed KWH is not significantly predicted by MONEYPY, CENACHP, and MONEYPY x CENACHP and we can conclude that all the variables are significant predictors of log KWH.*

##Generalized Linear Model — Best Subsets Regression

myResPlots2(lm(cleanRECS$IncBracket ~ cleanRECS$logKWH))



cleanRECS$logIncBracket <- log(cleanRECS$IncBracket)

*We see in the below residual plots that neither income nor the residuals of income for predicting the log of energy use are normally distributed. We are not surprised regarding the distribution of income, because it is generally expected to be normally distributed only when transformed in a log scale, and in this case we only have single lines representing income brackets. We therefore go with the standard linear transformation generally used for income variables. After the log transformation, the income variable was still not normal (due to aforementioned single lines representing income brackets), but it was much closer and was now more reasonable to work with.*

mod1 <- regsubsets(logKWH ~ logIncBracket + DIVISION + UATYP10 + KOWNRENT + DIVISION\*UATYP10, data = cleanRECS, nvmax = 7)  
mod1sum <- summary(mod1)  
mod1sum$which[1, ]

## (Intercept) logIncBracket   
## TRUE FALSE   
## DIVISIONEast South Central DIVISIONMid Atlantic   
## FALSE FALSE   
## DIVISIONMountain North DIVISIONMountain South   
## FALSE FALSE   
## DIVISIONNew England DIVISIONPacific   
## FALSE FALSE   
## DIVISIONSouth Atlantic DIVISIONWest North Central   
## FALSE FALSE   
## DIVISIONWest South Central UATYP10R   
## FALSE TRUE   
## UATYP10U KOWNRENT   
## FALSE FALSE   
## DIVISIONEast South Central:UATYP10R DIVISIONMid Atlantic:UATYP10R   
## FALSE FALSE   
## DIVISIONMountain North:UATYP10R DIVISIONMountain South:UATYP10R   
## FALSE FALSE   
## DIVISIONNew England:UATYP10R DIVISIONPacific:UATYP10R   
## FALSE FALSE   
## DIVISIONSouth Atlantic:UATYP10R DIVISIONWest North Central:UATYP10R   
## FALSE FALSE   
## DIVISIONWest South Central:UATYP10R DIVISIONEast South Central:UATYP10U   
## FALSE FALSE   
## DIVISIONMid Atlantic:UATYP10U DIVISIONMountain North:UATYP10U   
## FALSE FALSE   
## DIVISIONMountain South:UATYP10U DIVISIONNew England:UATYP10U   
## FALSE FALSE   
## DIVISIONPacific:UATYP10U DIVISIONSouth Atlantic:UATYP10U   
## FALSE FALSE   
## DIVISIONWest North Central:UATYP10U DIVISIONWest South Central:UATYP10U   
## FALSE FALSE

*This model finds that the household’s location in a rural area, compared to an urban or suburban areas, was the only significant predictor of household energy use.*

*To show this visually, the following figure plots median income vs. household energy use on a log-log scale. The trend lines illustrate the OLS best fit lines for urban and suburban (combined) households vs. rural households differ significantly only in their y-intercept, but not in their slopes. In other words, being in an urban or suburban vs. rural area has a significant effect on household energy use, and there is not an interaction effect with median household income.*

##   
## Call:  
## lm(formula = cleanRECS$logKWH ~ cleanRECS$logIncBracket \* cleanRECS$UrbanRural)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.13828 -0.36151 0.02805 0.38699 1.74603   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 7.75443 0.42602 18.202  
## cleanRECS$logIncBracket 0.16082 0.03867 4.158  
## cleanRECS$UrbanRuralU -0.79503 0.47164 -1.686  
## cleanRECS$logIncBracket:cleanRECS$UrbanRuralU 0.03919 0.04282 0.915  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## cleanRECS$logIncBracket 3.29e-05 \*\*\*  
## cleanRECS$UrbanRuralU 0.092 .   
## cleanRECS$logIncBracket:cleanRECS$UrbanRuralU 0.360   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5494 on 3196 degrees of freedom  
## (471 observations deleted due to missingness)  
## Multiple R-squared: 0.1014, Adjusted R-squared: 0.1005   
## F-statistic: 120.2 on 3 and 3196 DF, p-value: < 2.2e-16

