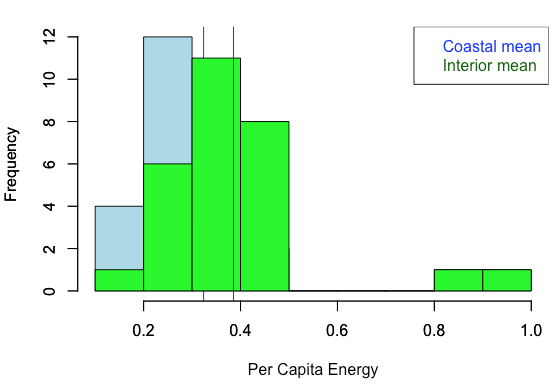
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized by whether they are in the South, West, Midwest, or East of the country (Region) or on the coast (Coast, 0 = no; 1 = yes). Depending on the questions below, you may need to construct your own variable that is a combination of the variables included in the dataset (e.g. when per capita is used). 14 points total.

1. Does ***per capita*** energy consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).

**Ho: Per capita energy consumption does not differ between coastal and non-coastal states**

**Ha: Per capita energy consumption differs between coastal and non-coastal states**

* 1. Please create a visual plot to answer this question (1 point).



* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

**The independent variable is discrete and binary (coastal/non-coastal) and the dependent variable (per-capita energy consumption) is continuous. The data is normal, although with unequal variances and sample sizes, so we run a Welch’s t-test to compare the means of coastal vs non-coastal states**

* 1. Please run the statistical test and interpret the result (1 point).

**P=0.2245**

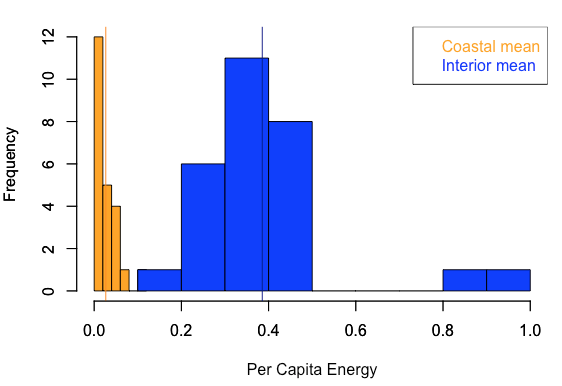
**Therefore, we cannot reject the null hypothesis that per capita energy consumption does not differ between coastal and non-coastal states…**

1. Does ***per capita*** coal consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).

**Ho: Per capita coal consumption does not differ between coastal and non-coastal states**

**Ha: Per capita coal consumption differs between coastal and non-coastal states**

* 1. Please create a visual plot to answer this question (1 point).



* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

**The independent variable is discrete and binary (coastal/non-coastal) and the dependent variable (per-capita coal consumption) is continuous. The data is normal, although with unequal variances and sample sizes, so we run a Welch’s t-test to compare the means of coastal vs non-coastal states**

* 1. Please run the statistical test and interpret the result (1 point).

**P=0.001936**

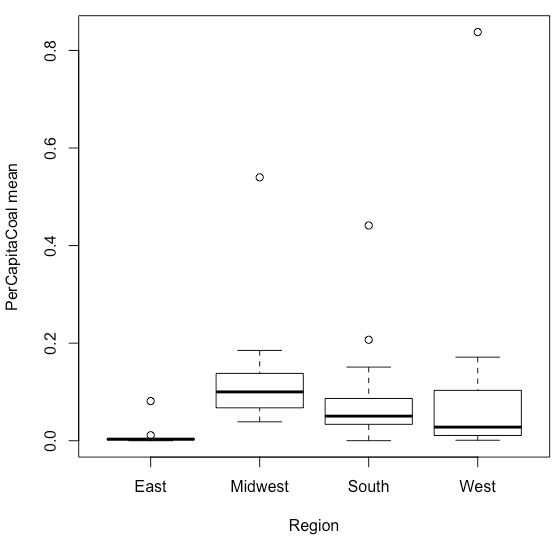
**Therefore, we cannot reject the null hypothesis that per capita energy consumption does not differ between coastal and non-coastal states…**

1. Does ***per capita*** coal consumption differ depending on the region in which a state is found?
   1. Please write the null and alternate hypothesis (1 point).

**Ho: Per capita coal consumption does not differ depending on region**

**Ha: Per capita coal consumption differs depending on region**

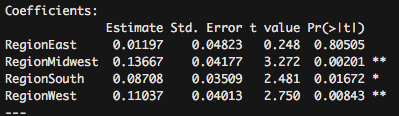
* 1. Please create a visual plot to answer this question (1 point).



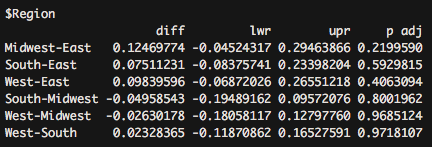
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

**The independent variable is discrete and multiple (Region) and the dependent variable (per-capita coal consumption) is continuous. The Levene test here ensures equal variances, so we run an ANOVA to compare the means of different regions.**

Please run the statistical test and interpret the result (1 point).



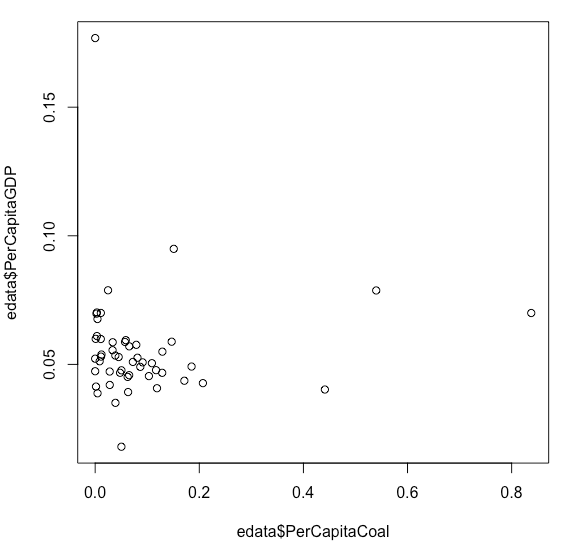
**Here, the P-values simply indicate whether a region’s per-capita coal consumption is significantly different than 0. A post-hoc test is needed to compare these means. The TukeyHSD test yields the following results:**



**Here, we can see that none of the variables are significantly different from eachother. The Midwest-East relationship has the lowest P-value, which makes sense given that their means are the farthest apart.**

1. What is the correlation between ***per capita*** coal use and ***per capita*** GDP? Does this seem like a strong correlation to you? Why or why not? (2 points)

**The correlation coefficient between those terms is extremely weak (0.03598), indicating that there is almost no correlation between the two variables. A scatter plot appears to confirm this:**

****

For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple linear regression model to predict housing value prices (medv). Please do not use an interaction term (unless stated in the question) since they can be challenging to interpret! 14 points + 2 bonus points.

1. Please select three covariates that you will include in your model as independent variables. Please check if these variables are highly correlated with one another to make sure you do not run into problems of multi-collinearity. Check if this model has issues with multi-collinearity using the variance inflation factor. (3 points).

**Correlation values:**

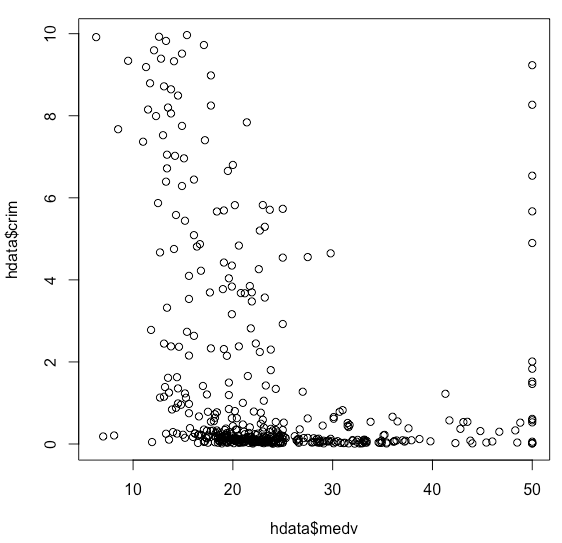
**crim-rm: -0.142**

**crim-ptratio: 0.3194**

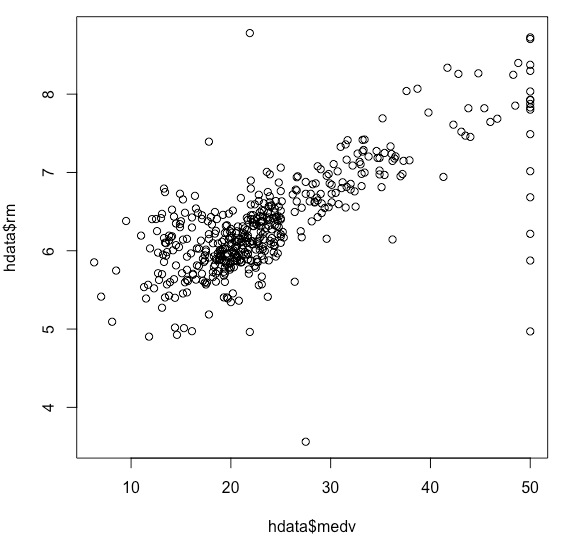
**rm-ptratio: -.3341**

1. Plot the relationship between each of your three independent variables and the dependent variable (medv). (medv; 3 points).

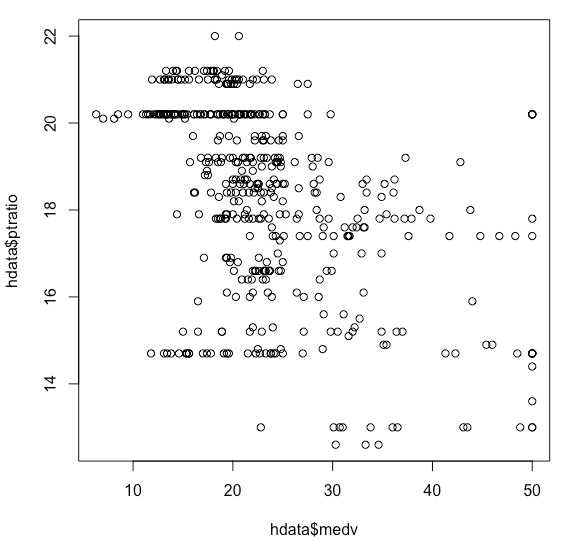
**Crime rate and housing prices: The relationship appears to be negative, perhaps logarithmic, as in lower crimes rates yield higher housing prices. The crime spike > 50 may be a result of those data points being “squished.” If not, the relationship could be parabolic.**



**Average number of rooms per dwelling and housing price: No surprise, the relationship appears to be positive, and linear. As the rooms per dwelling increases, so does the housing price.**

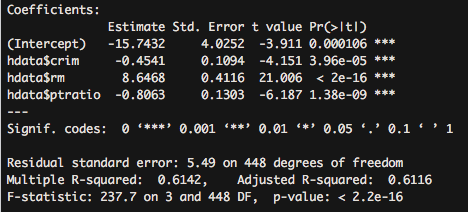


**Pupil-teacher ratio and housing price: The relationship appears to be loosely negative and linear, suggesting that less students per teacher correlate with higher housing prices.**



1. Run your multiple linear regression model. Check whether any assumptions are violated. Please state **which assumptions** you checked, **whether they were violated**, and **how you know** whether or not they were violated. If any assumptions are violated (e.g. normality), we will give you bonus points if you are able to identify a way to overcome this problem (3 points, plus additional 1 point bonus).

**Both independent and dependent variables are continuous and the data is relatively homoscedastic. The data is not normal, however. I did not transform the data and will therefore be analyzing the regression knowing that this assumption is violated.**



1. Interpret the results of the linear regression model. State **what the coefficient and its significance means** for the intercept and each of your three independent variables. Please explain what each regression coefficient means and do not just state that the coefficient is significant or not significant. For 1 bonus point, add in an interaction term, rerun the model, and interpret the result (3 points plus additional 1 point bonus).

**The intercept value (-15.74) refers to the house price if all other variables are zero. It is a somewhat specious value, because houses cannot have zero rooms, nor can the pupil-teacher ratio be zero.**

**The value for crime rate (-0.45) means that for each unit crime rate increases, housing prices fall by .45 units.**

**The value for rooms (8.65) means that for each average additional room, housing prices fall increase by 8.65 units**

**The value for pupil-teacher ratio (-0.81) means that for each pupil-teacher ratio increases, housing prices fall by .81 units.**

1. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?

**The adjusted R2 value of 0.6116 indicates a pretty good fit, and we know that there is not significant collinearity between variables, so we can be relative confident that the model is a good fit because the variables explain the housing price.**

**R code:**

# Energy data for question 1-4

edata = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/TakeHomeQuiz/energy\_data.csv",

sep=",", fill=TRUE, header=TRUE)

# Question 1

head(edata)

edata$PerCapitaEnergy <- edata$TotalEnergy/edata$Population

head(edata$PerCapitaEnergy)

Coastal.States = subset(edata, edata$Coast=="1")

Interior.States = subset(edata, edata$Coast=="0")

t.test(Coastal.States[,"PerCapitaEnergy"], Interior.States[,"PerCapitaEnergy"], paired=FALSE)

hist(Coastal.States[,"PerCapitaEnergy"], col="light blue", ylim=c(0,12), xlab="Per Capita Energy")

abline(v=mean(Coastal.States[,"PerCapitaEnergy"]), col="blue")

par(new=TRUE)

hist(Interior.States[,"PerCapitaEnergy"], col="green", ylim=c(0,12), xlab="")

abline(v=mean(Interior.States[,"PerCapitaEnergy"]), col="dark green")

legend("topright", legend = c("Coastal mean", "Interior mean"),text.col=c("blue", "dark green"))

# Question 2

edata$PerCapitaCoal <- edata$TotalCoal/edata$Population

head(edata$PerCapitaCoal)

Coastal.States = subset(edata, edata$Coast=="1")

Interior.States = subset(edata, edata$Coast=="0")

t.test(Coastal.States[,"PerCapitaCoal"], Interior.States[,"PerCapitaCoal"], paired=FALSE)

hist(Coastal.States[,"PerCapitaCoal"], col="orange", xlim=c(0, 1), ylim=c(0,12), xlab="Per Capita Energy")

abline(v=mean(Coastal.States[,"PerCapitaCoal"]), col="dark orange")

par(new=TRUE)

hist(Interior.States[,"PerCapitaEnergy"], col="blue", xlim=c(0, 1), ylim=c(0,12), xlab="")

abline(v=mean(Interior.States[,"PerCapitaEnergy"]), col="dark blue")

legend("topright", legend = c("Coastal mean", "Interior mean"),text.col=c("orange", "blue"))

# Question #3

boxplot(PerCapitaCoal~Region, data=edata, xlab="Region", ylab="PerCapitaCoal mean")

ANOVA1 = lm(PerCapitaCoal~Region-1, data=edata)

summary(ANOVA1)

ANOVA3 = aov(PerCapitaCoal~Region, data=edata)

summary(ANOVA3)

TukeyHSD(ANOVA3)

# Question #4

edata$PerCapitaGDP <- edata$TotalGDP/edata$Population

head(edata$PerCapitaGDP)

plot(edata$PerCapitaGDP~edata$PerCapitaCoal)

cor(edata$PerCapitaGDP, edata$PerCapitaCoal)

REG1 = lm(edata$PerCapitaGDP~edata$PerCapitaCoal)

summary(REG1)

####################################################

# Housing data for question 5-9

hdata = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/TakeHomeQuiz/housingdata.csv",

sep=",", fill=TRUE, header=TRUE)

library(RCurl)

cor(hdata[,c(1:13)], use="na.or.complete")

library(car)

VIF(lm(hdata$medv~hdata$crim+hdata$rm+hdata$ptratio,data=hdata))

# if more than 10, multicollinearity

plot(hdata$crim~hdata$medv)

plot(hdata$rm~hdata$medv)

plot(hdata$ptratio~hdata$medv)

MOD1 = lm(hdata$medv~hdata$crim+hdata$rm+hdata$ptratio, data=hdata)

summary(MOD1)