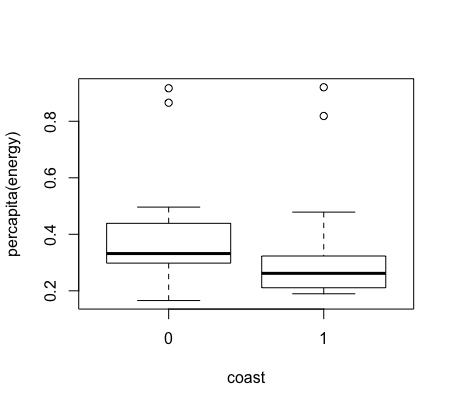
Take home quiz 4

Yihan Wang

March 26th, 2017

1. a. Null Hypothesis: per capita energy consumption does not differ depending on whether a state is found on the coast or not.

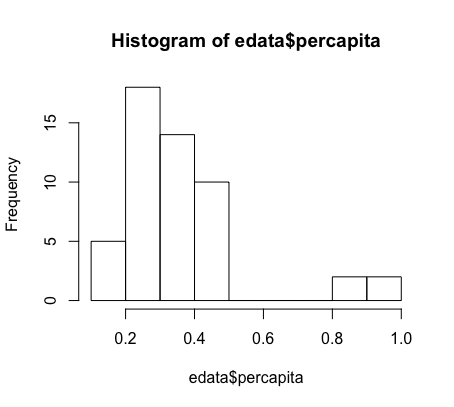
Alternative Hypothesis: per capita energy consumption differs depending on whether a state is found on the coast or not

b. 

c. Two Sample t-test

Check Normality:

Shapiro test shows that the per capita data are not normally distributed. Drawing histogram to confirm the result as follows:



The histogram shows there are outliers in percapita data. To make sure normality, we take out the four outliers located on the far right (per capita >0.8).

Data Transform:

After taking out the four outliers, we run Shapiro test to check normality again. The result shows that p value is greater than 0.05. So we conclude that the per capita data is now normally distributed.

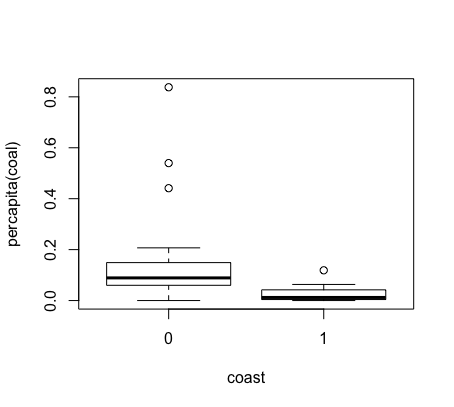
Check equal variance:

Variance test in R shows p value is less than 0.05. It indicates non-equal variances between per capita and Coast.

d. The Welch Two Sample t-test table shows that there are significant differences between energy per capita consumption between states found on the coast and not (p- value is less than 0.05). We reject the null hypothesis. a.

2. a Null Hypothesis: per capita coal consumption does not differ depending on whether a state is found on the coast or not.

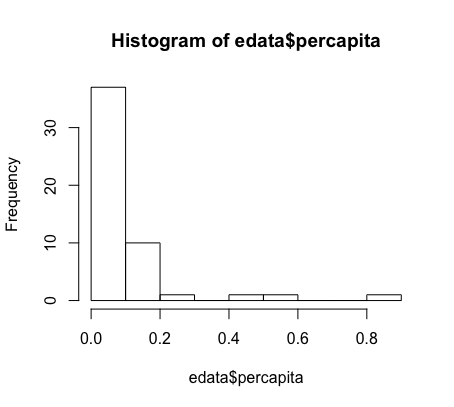
Alternative hypothesis: per capita coal consumption differ depending on whether a state is found on the coast or not.

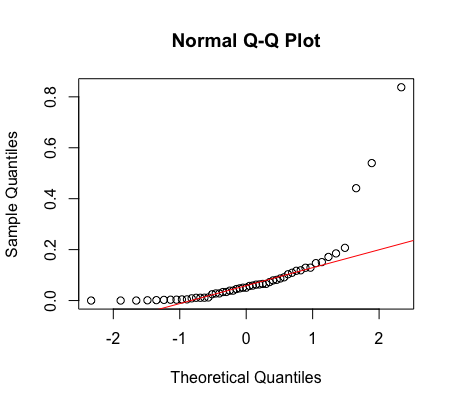
b. 

c. Welch’s Two Sample t-test

Check Normality and data transform:

Shapiro test shows the data is not normally distributed (p value is less than 0.05). Through histogram, we can tell that there is three outstanding outliers in the dataset that falls fall right.

After removing the three outliers that are greater than 0.4, we run the Shapiro test again, but it still violates normality assumption.

The qq plot confirms again the data for per capita coal use is not normally distributed.

Check Equal Variance:

The var. test shows that the data violates equal variance assumption (p value is less than 0.05).

d. Since the data is not normally distributed, we choose to use Kruskal-Wallis rank sum test

data: percapita by Coast

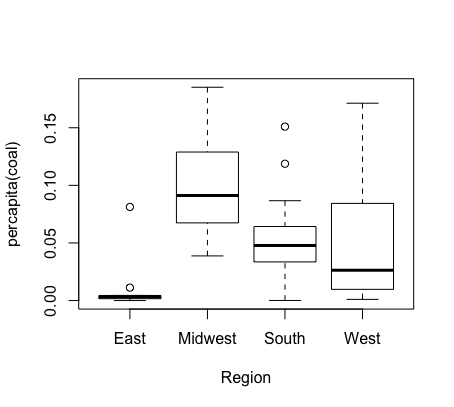
Kruskal-Wallis chi-squared = 18.222, df = 1, p-value =

1.966e-05

The table shows there are significant differences in per capita coal consumption between coast and non-coast states (p value is less than 0.05). We reject the null hypothesis.

1. a. Null hypothesis: per capita coal consumption does not differ in region in which a state is found.

Alternative hypothesis: per capita coal consumption differs in region in which a state is found.

b. 

c. I decide to use ANOVA. The assumptions are normality, equal variance, and independent observations.

Check Normality:

Shapiro test shows p value equals to 0.001. We reject the null hypothesis and the data is not normally distributed

Check equal variance:

Levene test shows p value is 0.2. We fail to reject the null hypothesis and so the data meets equal variance assumption.

d. Call:

lm(formula = percapita ~ Region, data = edata)

Residuals:

Min 1Q Median 3Q Max

-0.06127 -0.02623 -0.00882 0.01274 0.12155

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.01197 0.01429 0.838 0.4067

RegionMidwest 0.08805 0.01927 4.570 4.08e-05 \*\*\*

RegionSouth 0.04351 0.01807 2.407 0.0204 \*

RegionWest 0.03779 0.01890 1.999 0.0519 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04286 on 43 degrees of freedom

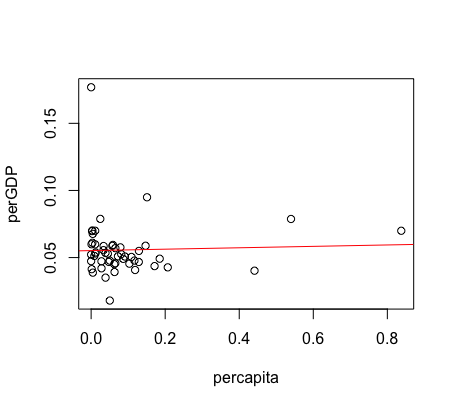
(4 observations deleted due to missingness)

Multiple R-squared: 0.3318, Adjusted R-squared: 0.2852

F-statistic: 7.118 on 3 and 43 DF, p-value: 0.0005497

The ANOVA summary table show that per capita coal consumption differ in states found in the Midwest, South, but not in the East or West. So per capita coal consumption differ depending on northern and southern region of the states.

1. The correlation between per capita coal use and per capita GDP is 0.04. This statistic indicates this is not a strong correlation. By plotting x and y, we do not see significant correlation between the two variables. Both statistic and plot indicate that there is not a strong or clear correlation between the two variables.



1. I choose crim, rm, and age as three covariates. The correlation and VIF are as follows:

cor(hdata$age, hdata$crim)

[1] 0.4476638

cor(hdata$age, hdata$rm)

[1] -0.1878709

cor(hdata$rm, hdata$crim)

[1] -0.1424577

The correlation coefficient values among three covariates are small (correlation coefficient range from -1 to +1).

vif(lm(age~crim+rm, data=hdata))

crim rm

1.020715 1.020715

vif(lm(crim~rm+age, data=hdata))

rm age

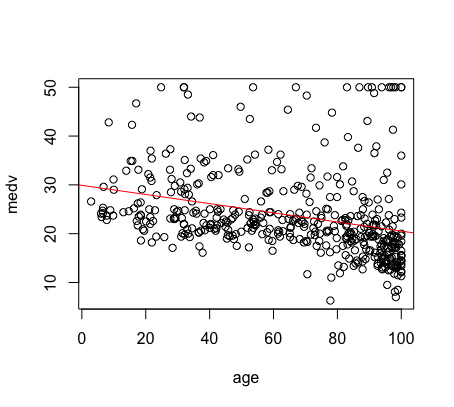
1.036587 1.036587

vif(lm(rm~age+crim, data=hdata))

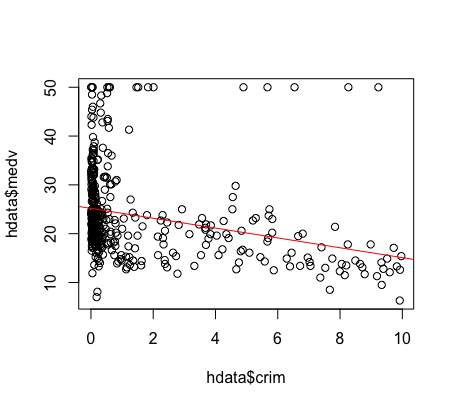
age crim

1.25063 1.25063 plot (age and medv)

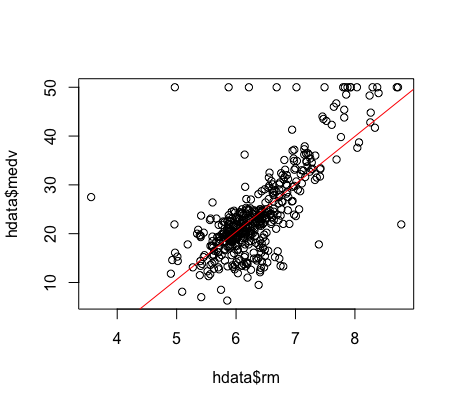
The VIF values among three covariates are less than 5. It shows that the three models do not have multi-collinearity.

6. 

This plot shows that age and the median housing price is negatively correlated but the correlation is not strong



This plot shows that crime rate and median housing prices are negatively correlated but the correlation is not strong.



This plot shows that rm rate and housing price are strongly and positively correlated. The correlation is a clear linear correlation.

1. The multiple linear regressions show that there are autocorrelation between three independent variables (dwtest p value is less than 0.05); residuals are heteroscedastic (bptest p value is less than 0.05); and residuals are not normally distributed (Shapiro test p value is less than 0.05). The model violates are three assumptions.
2. Call:

lm(formula = medv ~ age + rm + crim, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-22.812 -2.727 -0.709 1.863 41.809

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -32.07178 2.78319 -11.523 < 2e-16 \*\*\*

age -0.03322 0.01070 -3.105 0.00202 \*\*

rm 9.25306 0.40775 22.693 < 2e-16 \*\*\*

crim -0.49045 0.11965 -4.099 4.93e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.659 on 448 degrees of freedom

Multiple R-squared: 0.5901, Adjusted R-squared: 0.5873

F-statistic: 214.9 on 3 and 448 DF, p-value: < 2.2e-16

The summary table shows the coefficient for intercept is -32.07. It means when age, rm, and crim are zero, the medv is -32.07. The p value is less than 0.05, which means this result is significant.

The coefficient for age is -0.03. It indicates that for every additional unit increase of age, you can expect the median housing price decrease by 0.03 unit. The P value for this coefficient is less than 0.05, which suggests that this result is significant.

The coefficient for rm indicates that for every additional unit of average number of rooms per dwelling, you can expect the median housing price increase by 9.25 unit. The p value shows this result is significant.

The coefficient for crim indicates that for every additional unit of per capita crime rate by town, you can expect the median housing price decrease by 0.4 unit. The p value shows this result is significant.

Call:

lm(formula = medv ~ age + rm \* crim, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-13.523 -2.920 -0.588 1.828 34.832

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -45.10726 3.07206 -14.683 < 2e-16 \*\*\*

age -0.02662 0.01005 -2.648 0.00838 \*\*

rm 11.23193 0.45470 24.702 < 2e-16 \*\*\*

crim 9.30912 1.22855 7.577 2.05e-13 \*\*\*

rm:crim -1.58084 0.19736 -8.010 1.01e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.298 on 447 degrees of freedom

Multiple R-squared: 0.6415, Adjusted R-squared: 0.6383

F-statistic: 200 on 4 an d 447 DF, p-value: < 2.2e-16

In this model with interaction term, the coefficient for the interaction term (rm:crim) means that the effects of rm on housing prices depend on crime rate. When crime rate by town is zero, for every addition unit in rm, the median housing price increase by 11.23. However, with crim rate taking into calculation, for every additional unit of rm, the median housing price decrease by 11.23-1.58=9.65 unit. The intercept means when age, crim, and rm are zero, the median housing price is -45.11. The coefficient estimate of crim ( 9.31 ) means when rm is not in place, for every additional increase in crime rate, the housing price increase by 9.31 unit. All of these results are significant because all of their p value is less than 0.05.

9. The R squre is 0.64 in this model. It is a fairly good fit. However, because the model violates normality, homoscadesticity, and non autocorrelation assumptions, so linear model may not be the best fit.