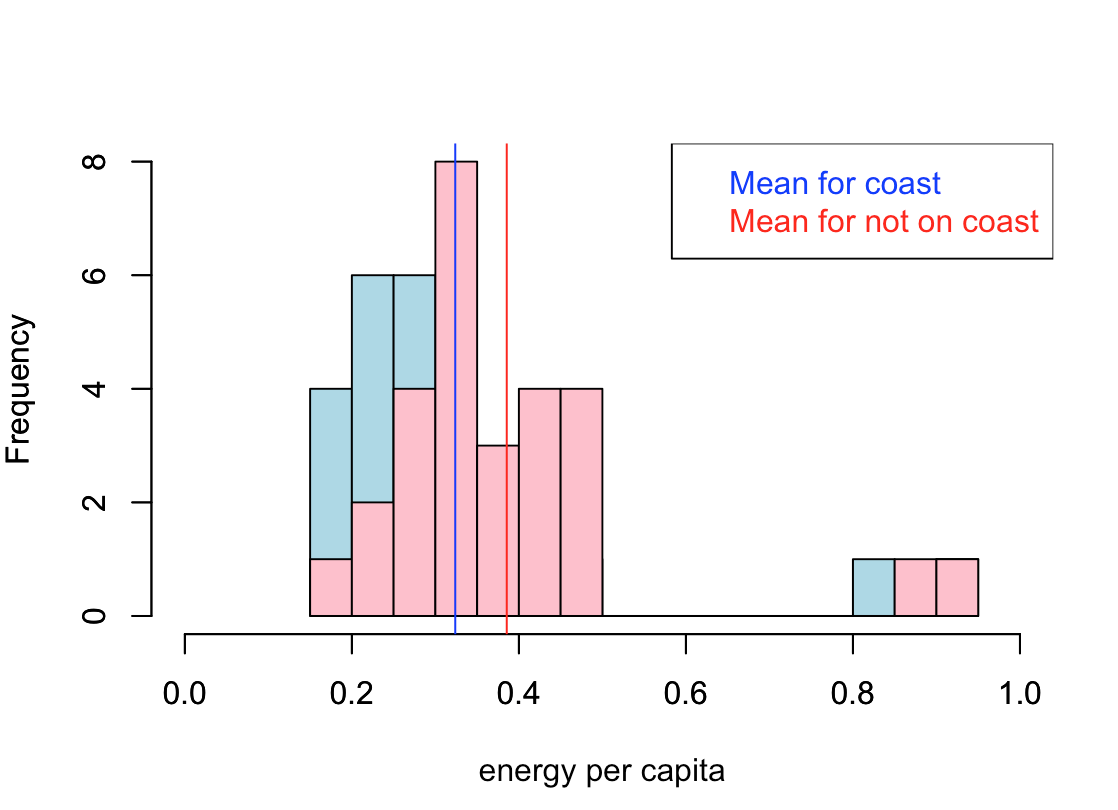
Please use the R script provided to load data and build your script from there.

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).

#NULL : per capita energy consumption for a state on coast is not different than that of a state not on coast

#A : per capita energy consumption for a state on coast is different than that of a state not on coast

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

2 Sample T-Test

Assumptions to be tested:

1. samples have equal variance - TRUE

2. samples are normally distributed -FALSE, this may be because the sample size is less than 30; assumed true for the purpose of the test

3. each observation is sampled independently - ASSUMED TRUE

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

Assuming the assumptions made were all true, the following result holds good

t.test(edata.coast$energy\_pc,edata.notcoast$energy\_pc, paired=FALSE)

data: edata.coast$energy\_pc and edata.notcoast$energy\_pc

t = -1.2319, df = 44.185, p-value = 0.2245

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.16246970 0.03918745

sample estimates:

mean of x mean of y

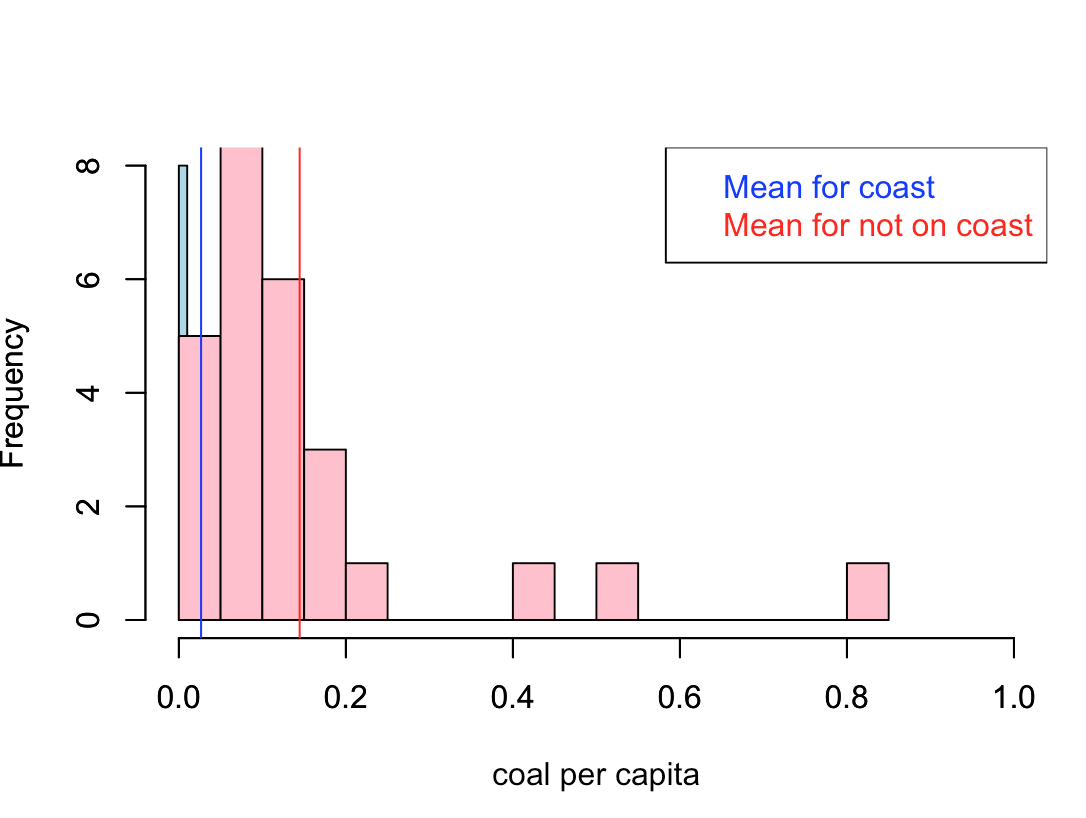
0.3238174 0.3854585

#p-value: 0.2245. Since p-value is greater than 0.05 we accept the null hypothesis that per capita energy consumption for a state on coast is not different than that of a state not on coast.

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).

#NULL : per capita coal consumption for a state on coast is not different than that of a state not on coast

#A : per capita coal consumption for a state on coast is different than that of a state not on coast

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

2 Sample T-Test

Assumptions to be tested:

1. samples have equal variance - FALSE

2. samples are normally distributed -FALSE, this may be because the sample size is less than 30; assumed true for the purpose of the test

3. each observation is sampled independently - ASSUMED TRUE

t.test(edatac.coast$coal\_pc,edatac.notcoast$coal\_pc, paired=FALSE, var.equal = FALSE)

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

Assuming the assumptions made were all true, the following result holds good

Welch Two Sample t-test

data: edatac.coast$coal\_pc and edatac.notcoast$coal\_pc

t = -3.4129, df = 28.639, p-value = 0.001936

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.18872795 -0.04724185

sample estimates:

mean of x mean of y

0.02665574 0.14464064

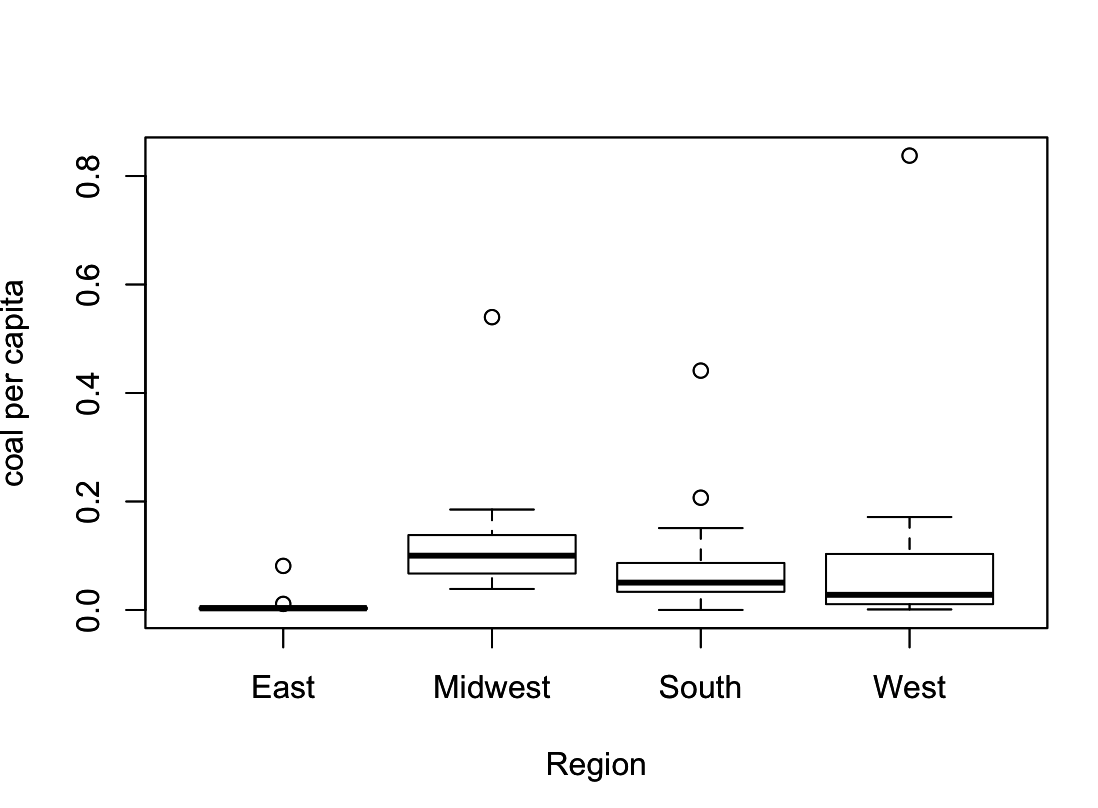
#p-value: 0.001936 Since p-value is lesser than 0.05 we reject the null hypothesis that per capita coal consumption for a state on coast is not different than that of a state not on coast.

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).

#NULL : per capita coal consumption doesn't differ depending on the region in which a state is found

#A : per capita coal consumption differs depending on the region in which a state is found

* 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).

ANOVA

modQ1\_3=aov(coal\_pc~Region, data=edata)

summary(modQ1\_3)

Assumptions

1. samples have equal variance - TRUE

2. samples are normally distributed -FALSE, may be because of the small sample size, assumed true for the purpose of the test (the qqplot in the r script shows that they are almost normal)

3. each observation is sampled independently - ASSUMED TRUE

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

modQ1\_3=aov(coal\_pc~Region, data=edata)

summary(modQ1\_3)

Df Sum Sq Mean Sq F value Pr(>F)

Region 3 0.0864 0.02879 1.375 0.262

Residuals 47 0.9841 0.02094

#Since the p-value is greater than 0.05, the NULL hypothesis is true that per capita coal consumption doesn't differ depending on the region in which a state is found

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)

cor(edata$coal\_pc,edata$gdp\_pc)

#0.03598182.

As the correlation coefficient is very small, i.e. 0.03598, per capita coal use and per capita GDP are not strongly correlated.

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. **Report correlation values and VIF values in your answer** (3 points).

The three covariates are :

1. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
2. AGE - proportion of owner-occupied units built prior to 1940
3. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

ZN and AGE

-0.5556789, the coefficient value is high thus these 2 variables are negatively correlated.

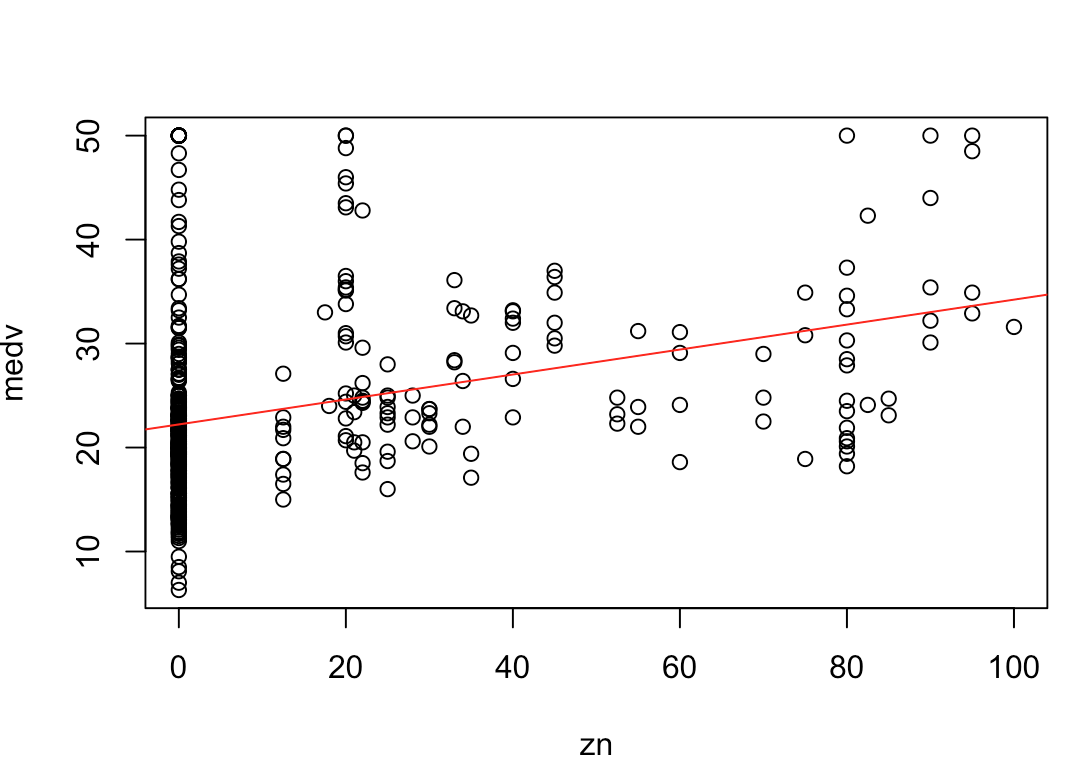
ZN and B

0.1503797, the coefficient value is very low thus these 2 variable are not that correlated.

AGE abd B

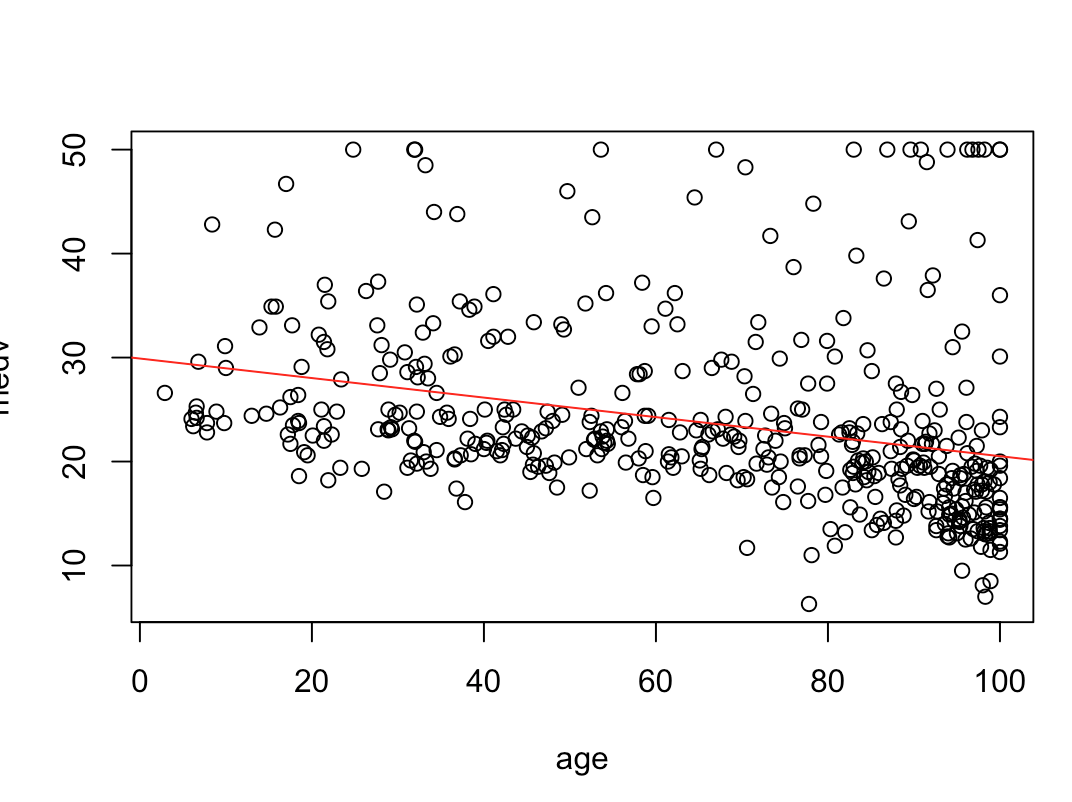
-0.2237652, the coefficient value is very low thus these 2 variable are not that correlated.

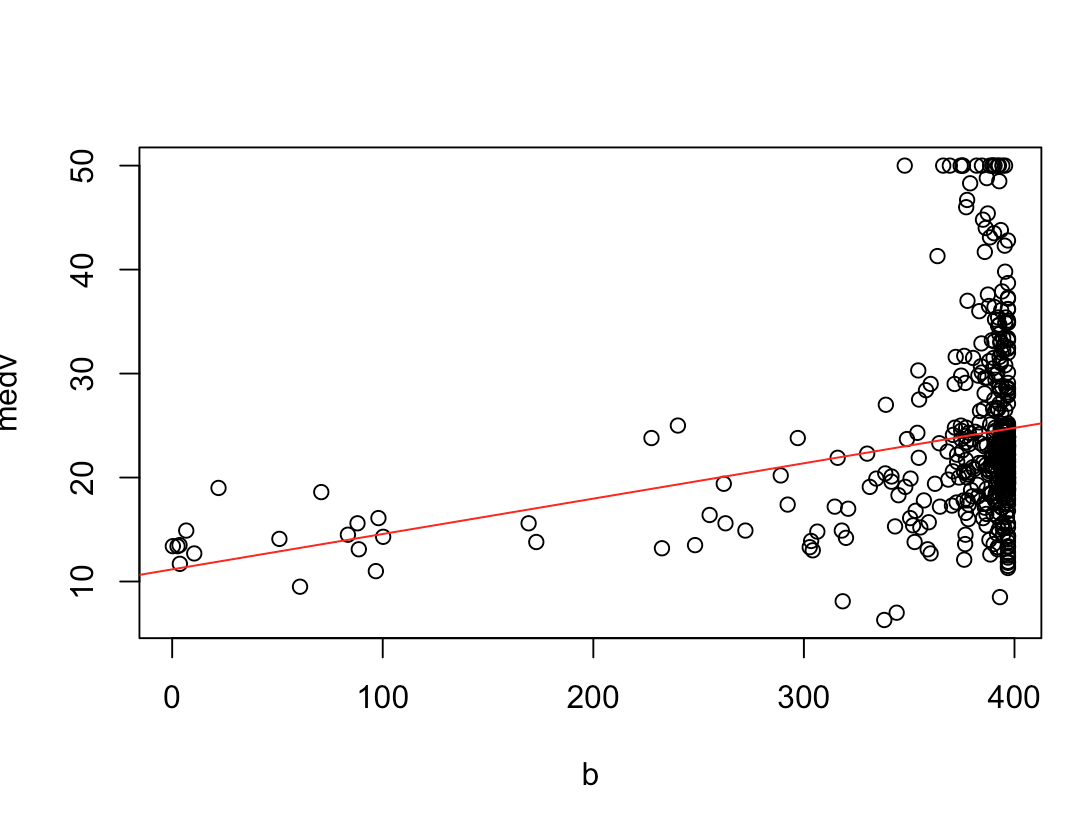
VIF = 1.201813, since VIF value is less than 5 the collinearity of the variable does not affect the model.

1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).

MEDV vs ZN

This plot of medv and zn shows that they are not highly correlated, although some values show linear relationship. Also from this plot it can be concluded that the residuals are not homostedastic and this may be confirmed my bptest.

MEDV vs AGE

This plot of medv and age shows that they are not highly correlated, although values show linear relationship. Also from this plot it can be concluded that the residuals are not homostedastic and this may be confirmed my bptest.

MEDV vs B

This plot of medv and b shows that they are not highly correlated, although values show linear relationship. Also from this plot it can be concluded that the residuals are not homostedastic and this may be confirmed my bptest.

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).

ASSUMPTIONS:

1. Linear relationship: TRUE

From the plots show above for med with zn, age , b independently.

1. No Multicollinearity: TRUE

The VIF test done above shows that the independent variable are not highly correlated with each other.

1. Residuals are independent: FALSE

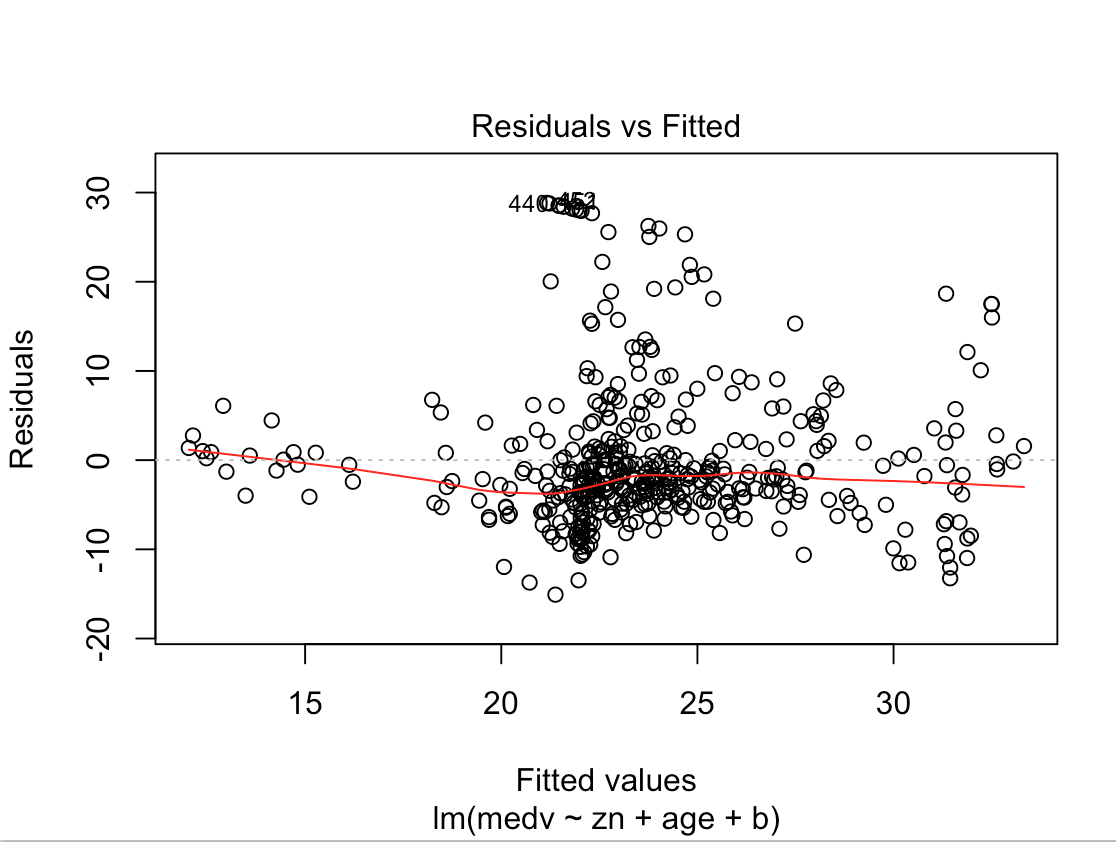
Did the dwtest to check for this. #p-value < 2.2e-16, since p-value is less than 0.05, the alternative hypothesis is true and there exists autocorrelation and the residuals are not independent.

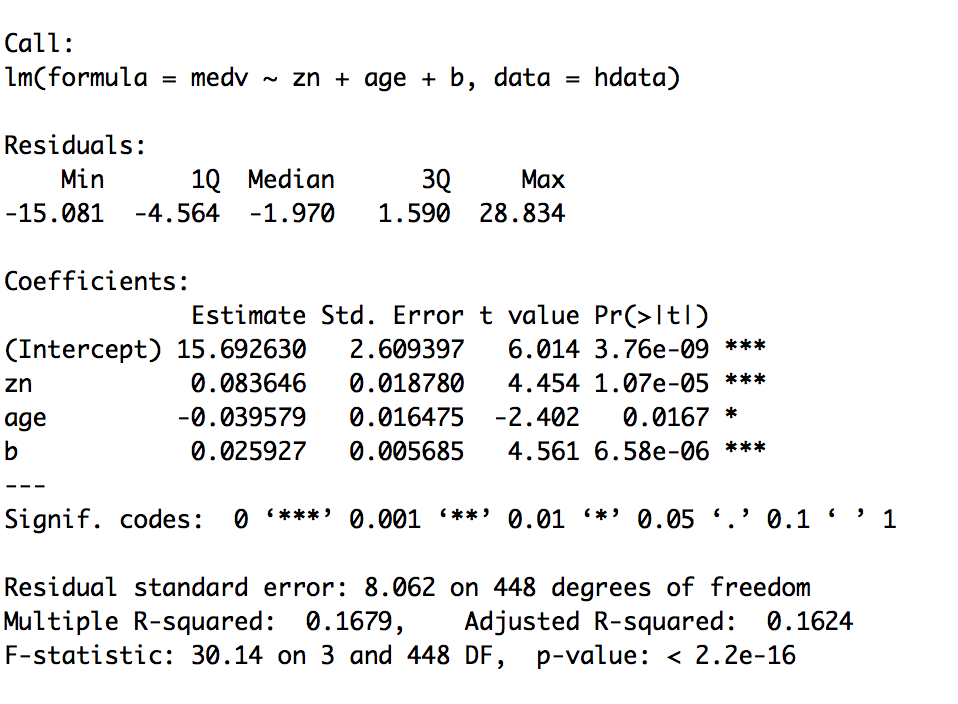
1. Residuals are normally distributed: FALSE

Did the shapiro will test. p:2.2e-16, The residuals are not normally distributed

1. Residual are homostedastic: FALSE

Did the bptest. p:0.001001, the residuals are not homostedastic.



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 value ‘coefficient estimate of the intercept’, shows that the value of medv is 15.6926 when zn=age=b=0. This is a significant value indicated by the p value and significant codes.

The value ‘coefficient estimate of the zn', shows that the value of medv increases by 0.0836 with increase in zn value by 1 unit. This is a significant value indicated by the p value and significant codes.

The value ‘coefficient estimate of the age’, shows that the value of medv decreases by 0.0395 with increase in age value by 1 unit. This is a significant value indicated by the p value and significant codes.

The value ‘coefficient estimate of the b’, shows that the value of medv increases by 0.0259 with increase in b value by 1 unit. This is a significant value indicated by the p value and significant codes.

The coefficient standard error for each variable shows average amount by which the coefficient of estimates vary from the actual average value of dependent variable.

The coefficient t-value indicate how many standard deviations is the coefficient estimate away from 0.

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

The above discussed linear model is not a good fit

1. since the residuals are not homostedastic, not normally distributed and not independent.
2. since the F-value is 30.14 (p < 2.2e-16), indicating that we should reject the null hypothesis that the variables zn, age and b collectively have no effect on medv.
3. also since the R2 value which is 0.1679, which means that the model can explain only 16.79% of the variance exhibited by the data thus the model doesn't fit the data.