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

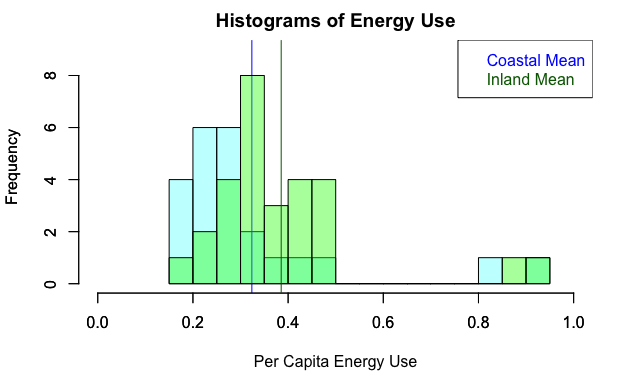
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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).

H0: Energy consumption does not differ depending on if the state is coastal or not.

Ha: Energy consumption is different depending on whether or not a state is on the coast.

* 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).

Two sample T-test.

Normality: Shapiro-Wilk: Not normally distributed.

data: energy\_coast$PerCapitaEnergy

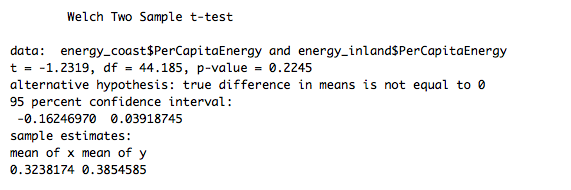
W = 0.66467, p-value = 5.037e-06

data: energy\_inland$PerCapitaEnergy

W = 0.77249, p-value = 3.627e-05

Similar Variance: F-test: p-value = 0.5098. Variance are similar.

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



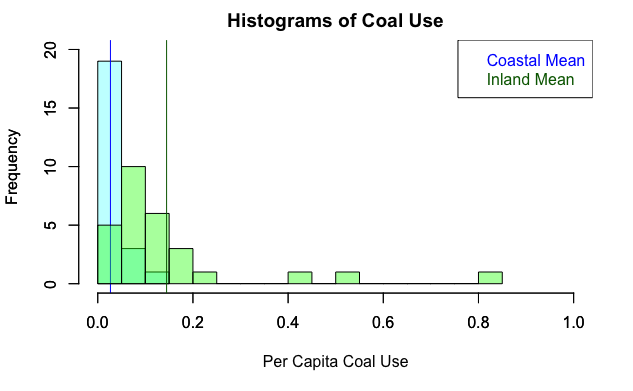
With a p-value of 0.2245 the null hypothesis would not be rejected. Per capita energy consumption is not statistically different between coastal and inland states.

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).

H0: Coal consumption does not differ depending on if the state is coastal or not.

Ha: Coal consumption is different depending on whether or not a state is on the coast.

* 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).

Two sample T-test.

Normality: Shapiro-Wilk: Not normally distributed.

data:

coal\_coast$PerCapitaCoal

W = 0.82584, p-value = 0.001023

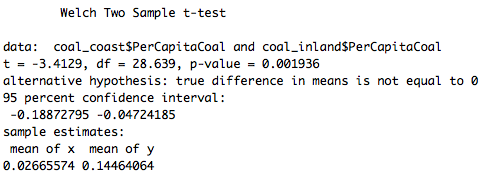
data:

coal\_inland$PerCapitaCoal

W = 0.651, p-value = 6.211e-07

Similar Variance: F-test: p-value = 5.995e-13. Variance are not similar.

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



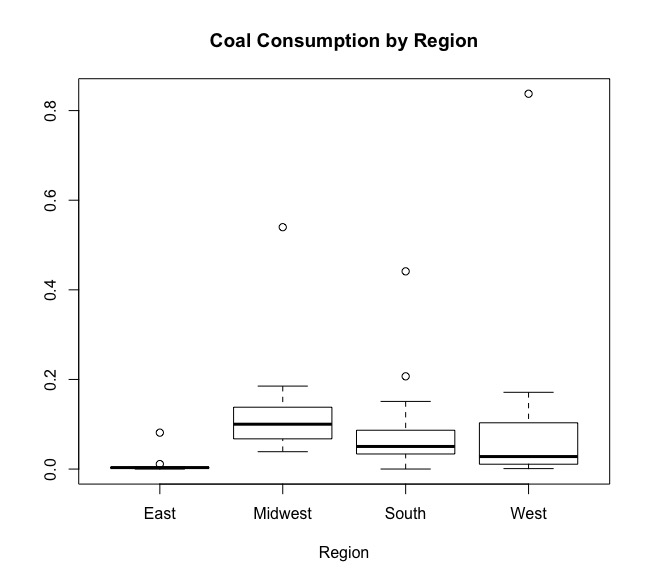
With a p-value of 0.001936 the null hypothesis is rejected. There is a statistically significant difference between per capita coal consumption between coastal and inland states.

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).

H0: Coal consumption does not differ by region.

Ha: Coal consumption is different depending on what part of the US a state is.

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



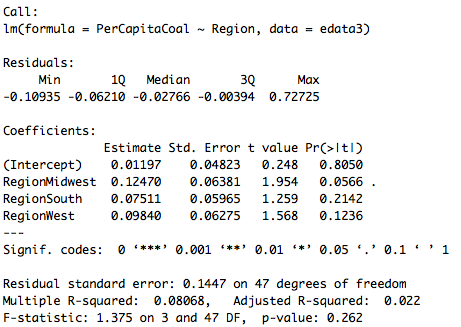
1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

ANOVA.

Normality: Shapiro-Wilk: p-value = 5.037e-06. Not normally distributed.

Levene Test: p-value: 0.5202 ; Variance is equal.

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



There are no p-values less than 0.05, so the null hypothesis is not rejected. There is not a statistically significant effect of region on coal consumption.

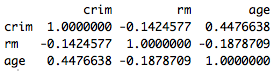
1. What is the correlation between ***per capita*** coal use and ***per capita*** GDP? Does this seem like a strong correlation to you? Why or why not? (2 points)

The correlation between the two is -0.2085636, which is not a strong correlation. Correlations above 0.5 are considered to be highly correlated.

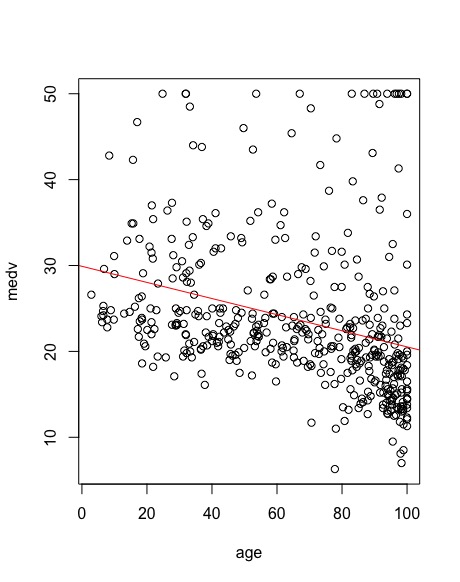
For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple 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).

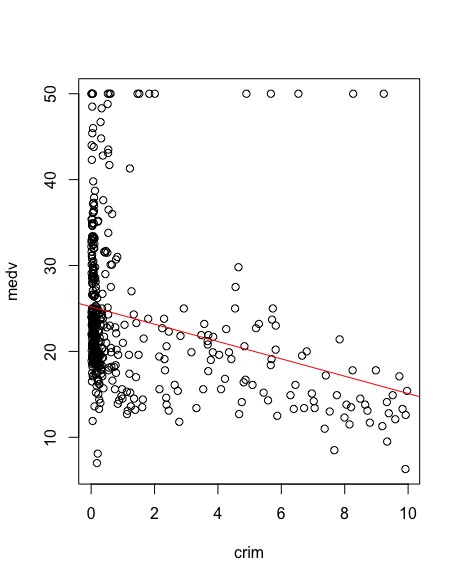
Independent variables: crim, rm, age Collinearity is low.

 TakeHomeQuiz/Q5.VIF.png

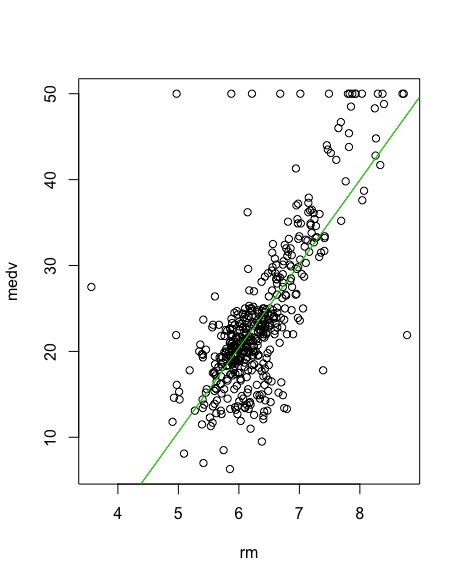
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).

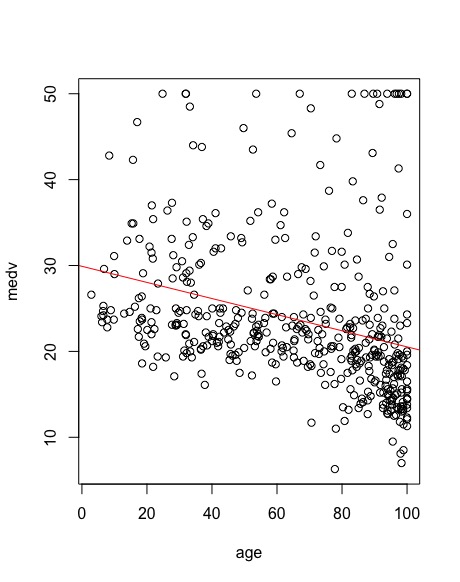


There is a positive relationship between the number of rooms in a house and median housing prices.



There is a negative relationship between the crime rate and median housing prices.





There is a negative relationship between the age of a home and the median housing prices.

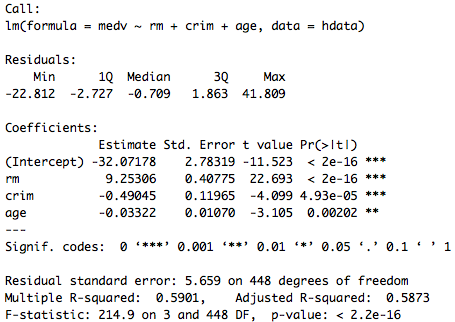
1. Run your multiple linear regression model. Check whether any assumptions are violated. Please state **which assumptions** you checked, **whether they were violated**, and **how you know** whether or not they were violated. If any assumptions are violated (e.g. normality), we will give you bonus points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

Normality of Residuals: shapiro.test: p-value < 2.2e-16 ; not normally distributed

Correlation of Residuals: dwtest: p-value < 2.2e-16 ; autocorrelated

Homoscedasticity of Residuals: bptest: p-value: 8.811e-07 ; errors are heteroscedastic

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 of -32.1 is the median home value when the other coefficients are set to 0. Rm 9.3 means that for an increase in one room lead to an increase in $9,300. There is a negative effect for crim, with a coefficient of -0.5. For every increase in one unit of per capita crime rate, median home values decrease by $490. There is a negative effect of age, with a coefficient of -0.03. For each increase in the proportion of homes built before 1940, the median home value decreases by $30. All beta coefficients are significant.

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

This model has an ok R2 at 0.5873, but all of the assumptions were violated. When looking at the qqplot for the model it looked like there is likely a much better model (maybe exponential?) that would fit the data.

