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?

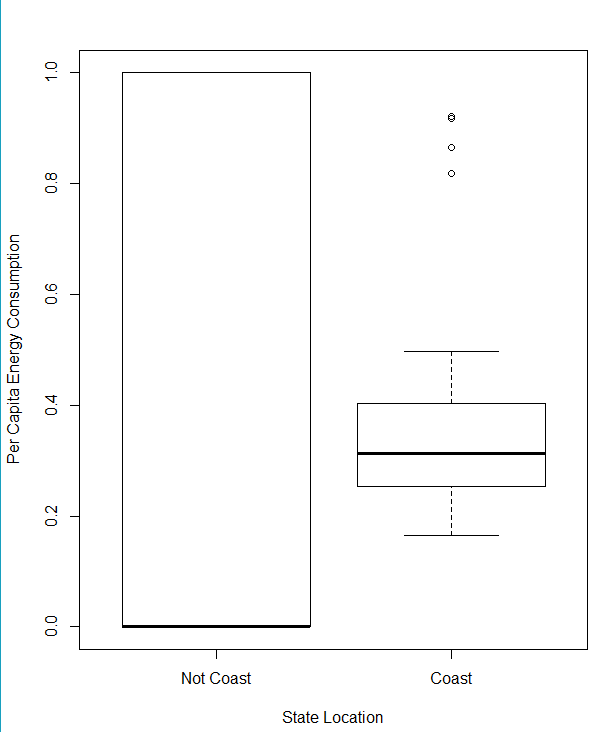
TotEnergyCap=edata$TotalEnergy/edata$Population

* 1. Please write the null and alternate hypothesis (1 point).

Ho: The energy consumption per capita does not differ whether the state is on the coast or not

Ha: The energy consumption per capita does differ if the state is on the coast or not

* 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 because dep. is continuous, independent is categorical, and we are comparing two populations.

Assumptions: data are continuous; sample is randomly selected; observations are independent; values are nearly normal or sample size is large enough; equal variance between 2 populations. The data meet the assumptions of being continuous, randomly selected, and large enough. Variance is tested with var.test:

F test to compare two variances

data: edata$Coast and TotEnergyCap

F = 8.1106, num df = 50, denom df = 50, p-value = 8.978e-12

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

4.629474 14.209430

sample estimates:

ratio of variances

8.110621

Variances are unequal; use Welch’s t-test.

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

Welch Two Sample t-test

data: edata$Coast and TotEnergyCap

t = 1.2512, df = 62.145, p-value = 0.2155

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

95 percent confidence interval:

-0.05575982 0.24240139

sample estimates:

mean of x mean of y

0.4509804 0.3576596

The output of the t-test tells us that the per capita energy consumption for coast/not coast is not significantly different. However, the boxplot and the difference of the means show us that there might be a relationship between higher consumption and coast, but it is not strong.

1. Does ***per capita*** coal consumption differ depending on whether a state is found on the coast or not?

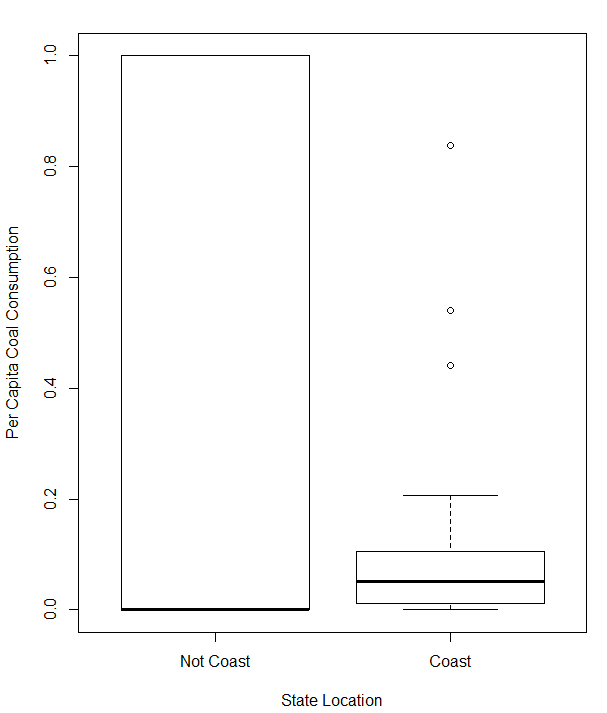
TotCoalCap=edata$TotalCoal/edata$Population

* 1. Please write the null and alternate hypothesis (1 point).

Ho: The coal consumption per capita does not differ whether the state is on the coast or not.

Ha: The coal consumption per capita does differ if the state is on the coast or not.

* 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 because dep. is continuous, independent is categorical, and we are comparing two populations.

Assumptions: data are continuous; sample is randomly selected; observations are independent; values are nearly normal or sample size is large enough; equal variance between 2 populations. The data meet the assumptions of being continuous, randomly selected, and large enough (approximately 25 states in each category). Variance is tested with var.test:

F test to compare two variances

data: edata$Coast and TotCoalCap

F = 11.796, num df = 50, denom df = 50, p-value = 3.997e-15

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

6.733333 20.666889

sample estimates:

ratio of variances

11.79648

Variance is not equal; use welch’s t-test.

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

Welch Two Sample t-test

data: edata$Coast and TotEnergyCap1

t = 17.419, df = 95.269, p-value < 2.2e-16

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

95 percent confidence interval:

1.388973 1.746285

sample estimates:

mean of x mean of y

0.4509804 -1.1166488

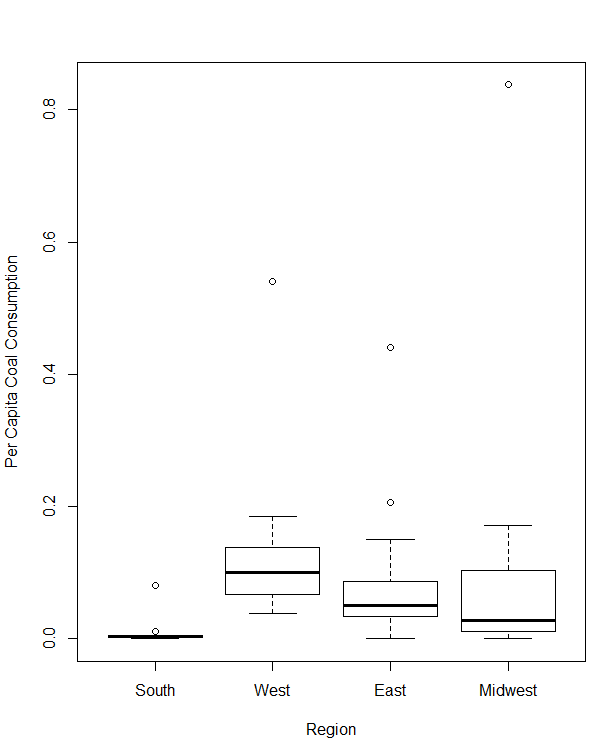
The p-value in this output tells us that the results are significant. Also, it returns the alternative hypothesis, meaning that the coal consumption is related to whether or not the state is on the coast.

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).

Ho: The coal consumption per capita does not differ depending on the region of the state

Ha: The coal consumption per capita does differ depending on the 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).

Anova is a good test because the dependent variable is continuous and the independent variable is categorical but we are testing more than two groups.

Assumptions: populations of interest must be normally distributed; samples must be independent of one another; each population must have the same variance; does not require sample sizes to be equal. Samples are approximately normally distributed and are independent of each other. Check variance with Levene’s test:

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 3 0.7635 0.5202

47

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

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

edata$Region 3 0.0864 0.02879 1.375 0.262

Residuals 47 0.9841 0.02094

From this output, the region is not significant in predicting the per capita 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)

cor(TotCoalCap, TotGDPCap)

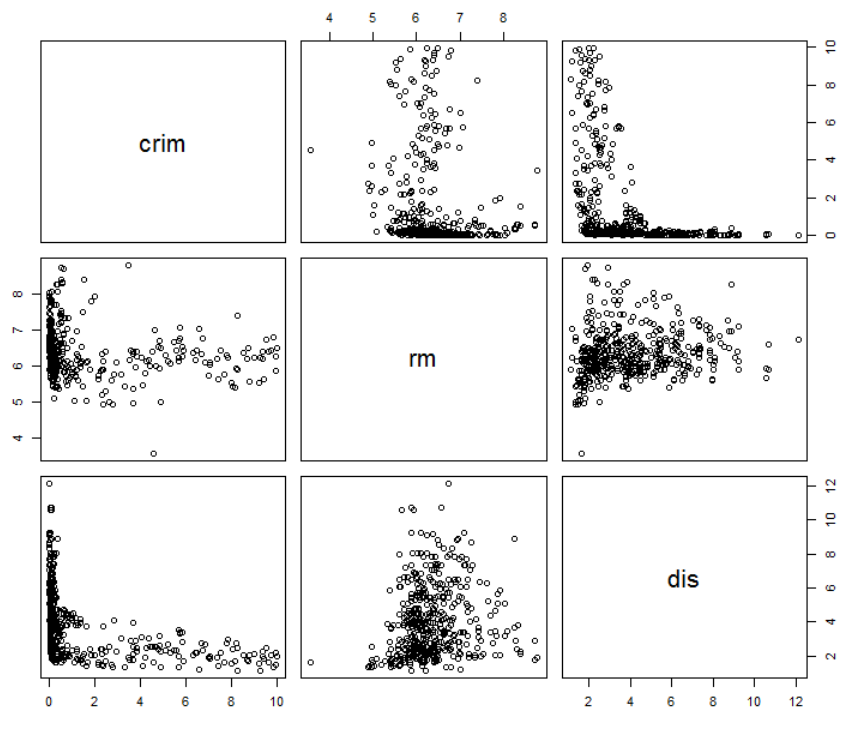
#correlation = 0.0359

This correlation is not very strong. Correlation coefficient can range from -1 to 1; values close to 0 are not 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).

I picked crime rate, number of rooms, and distance to city center.



pairs(hdata[,c(1,6,8)])

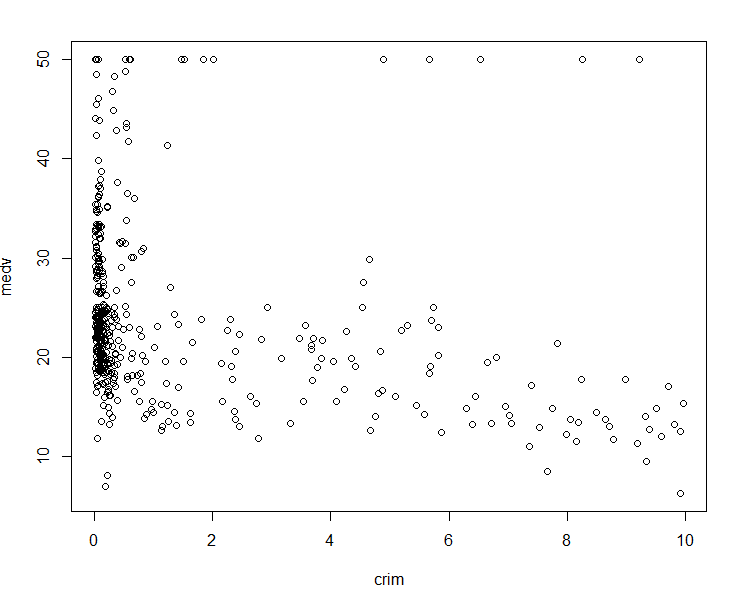
cor(hdata$crim, hdata$rm) #-0.1424577

cor(hdata$rm, hdata$dis) #0.1387741

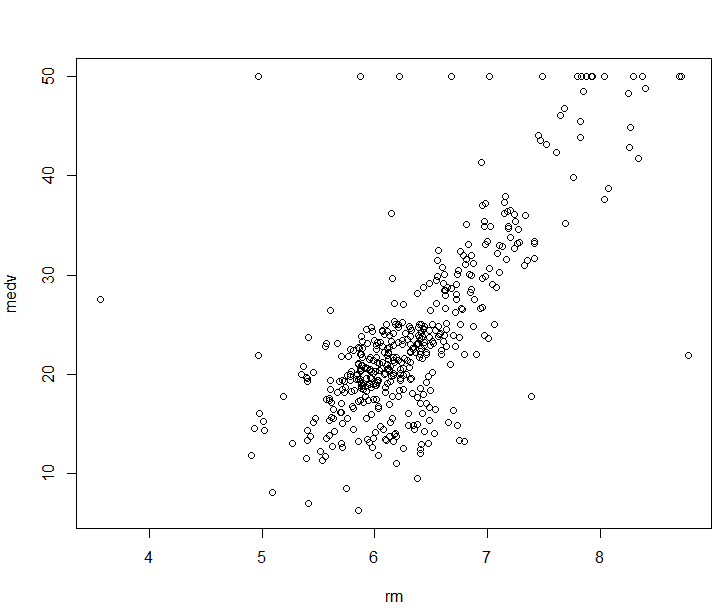
cor(hdata$crim, hdata$dis) #-0.4619674

None are too correlated. Crime rate and distance are getting close, but are usable.

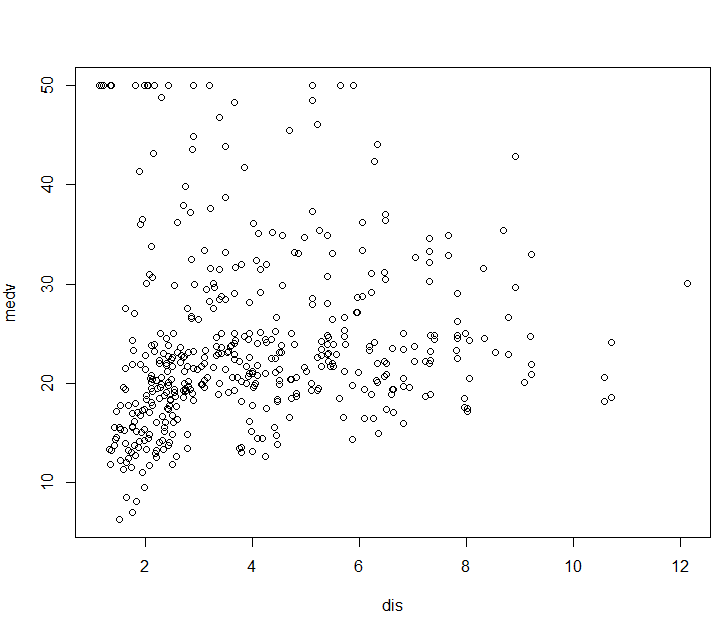
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).



Inverse relationship; higher value at low crime rates



Direct relationship; greater number of rooms means greater value



This is harder to interpret; in general, the value increases farther from the city.

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:

– Linear relationship: Number of rooms seems to be clearly linear, but distance and crime rates could be non-linear relationships

– Independent errors: There appears to be some autocorrelation since the dw test returns the alternative hypothesis

Durbin-Watson test

data: housemodel

DW = 0.95719, p-value < 2.2e-16

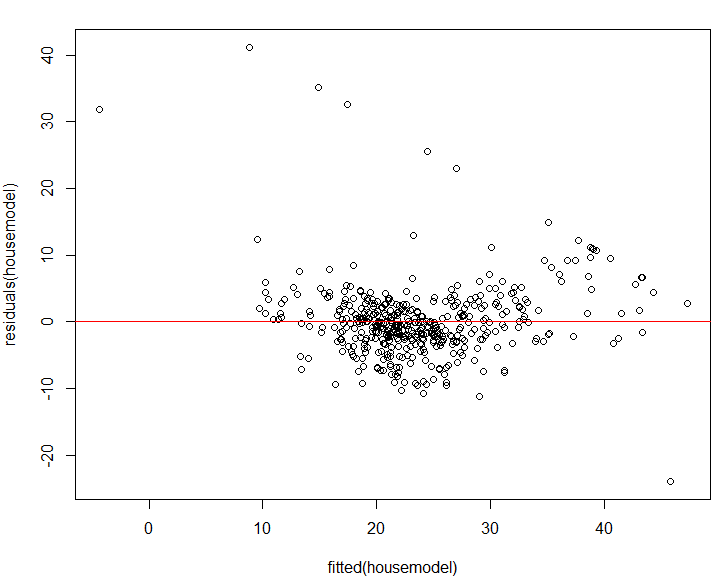
alternative hypothesis: true autocorrelation is not 0

– Homoscedascity: The p-value is small, but there looks like some sort of relationship with the variances.

studentized Breusch-Pagan test

data: housemodel

BP = 32.28, df = 3, p-value = 4.568e-07

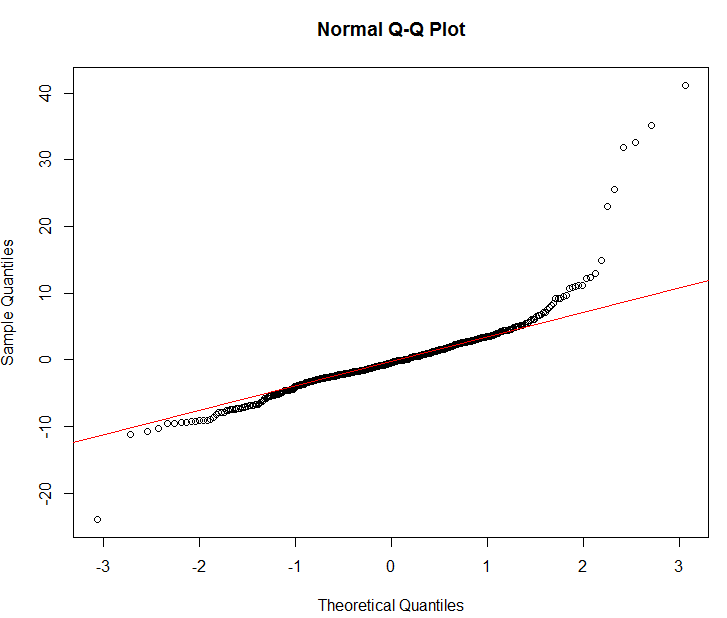


– Non-normal errors: The errors are normal since the p-value is small

Shapiro-Wilk normality test

data: residuals(housemodel)

W = 0.90474, p-value = 3.264e-16



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).

Call:

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

Residuals:

Min 1Q Median 3Q Max

-23.965 -2.683 -0.447 2.271 41.203

Coefficients:

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

(Intercept) -34.3862 2.6589 -12.933 < 2e-16 \*\*\*

crim -0.7441 0.1218 -6.111 2.15e-09 \*\*\*

dis -0.2456 0.1453 -1.690 0.0917 .

rm 9.4879 0.4082 23.246 < 2e-16 \*\*\*

---

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

Residual standard error: 5.701 on 448 degrees of freedom

Multiple R-squared: 0.5839, Adjusted R-squared: 0.5811

F-statistic: 209.5 on 3 and 448 DF, p-value: < 2.2e-16

From the output, the crime rate and number of rooms are significant in the value of the houses. The distance is not significant. The coefficient for the intercept is where the line would cross the y-axis if all other variables were set to zero. The coefficients for the three independent variables give the slopes for each variable. For crime rate, the value of the house decreases by 0.744 units for every increase in unit of crime rate. For every increase in unit of distance, the value of the house would decrease by 0.245 (but is not significant). For every additional room, the value of the house increases by 9.48 units.

Call:

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

Residuals:

Min 1Q Median 3Q Max

-13.811 -2.834 -0.408 1.976 34.371

Coefficients:

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

(Intercept) -47.9083 2.9035 -16.500 < 2e-16 \*\*\*

crim 9.4470 1.2357 7.645 1.28e-13 \*\*\*

rm 11.4269 0.4517 25.297 < 2e-16 \*\*\*

crim:rm -1.6237 0.1980 -8.200 2.55e-15 \*\*\*

---

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

Residual standard error: 5.333 on 448 degrees of freedom

Multiple R-squared: 0.6359, Adjusted R-squared: 0.6334

F-statistic: 260.8 on 3 and 448 DF, p-value: < 2.2e-16

With the interaction term, the intercept point is now -47.9. The slopes of both the crime rate and number of rooms has changed. For one unit increase in crime rate, while keeping rooms constant, the value of the house increases 9.4. For one increase in rooms, while keeping crime rate constant, the value increases 11.4. At higher crime rates, the increase in rooms now decreases the value of the home by 1.62.

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

The first model is not a very good fit, since the R-squared value is 0.58. The second model is better, since the R-squared value is 0.633. Also, the second model is better since it only includes terms that are significant. However, I do not know what a typical real estate model would be getting for an R-squared value, so it is hard to determine if this is a good fit. Since this is coming from real-world data, I would think the second model is a good fit.