If you run any ANOVAs, you can use the Levene test for equality of variances. If your data violate an assumption about normality and a normal distribution is required for your analyses, you can get bonus points for transforming your data. Otherwise please run the statistical test anyway as if your data were normally distributed but make it clear that you violated this assumption in your answer.

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

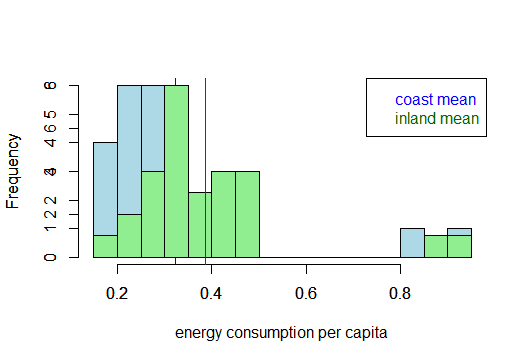
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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: Per capita energy consumption in coastal states is not different from inland states.

H1: Per capita energy consumption in coastal states is different than in inland 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).
     + - * Test: One-tailed, Two sample t-test.
         * Assumptions:

1. Normality: The data failed the Shapiro test and the sample size of each group is less than 30, therefore you cannot assume normality. The data is not normally distributed.

2. Variance: The data has unequal variance

3. Independence: Because I do not know details about research design, I assume that each data point has been sampled independently

* + - * + Because the data does not meet the assumptions of the t-test, I decided to run a non-parametric test: the Mann-Whitney-Wilcoxon test.
  1. Please run the statistical test and interpret the result (1 point).

**R Output Wilcoxon test:**

Wilcoxon rank sum test with continuity correction

data: edata$TEp.cap and edata$Coast

W = 1428, p-value = 0.3876

alternative hypothesis: true location shift is not equal to 0

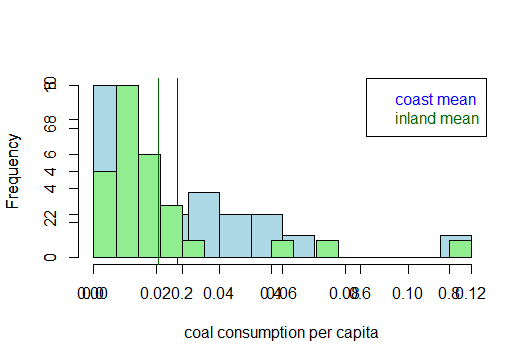
Interpretation: The result of Mann-Whitney-Wilcoxon test indicates that per capita energy consumption does not vary between coastal and inland states (the differences are not significant as shown by the p-value which is greater than 0.05), therefore the null hypothesis is rejected.

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: Per capita coal consumption in coastal states is not different from inland states.

H1: Per capita coal consumption in coastal states is different than in inland 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).

1. Normality: The data failed the Shapiro test and the sample size of each group is less than 30, therefore you cannot assume normality. The data is not normally distributed.

2. Variance: The data has unequal variance

3. Independence: Because I do not know details about research design, I assume that each data point has been sampled independently

* + - * + Because the data does not meet the assumptions of the t-test, I decided to run a non-parametric test: the Mann-Whitney-Wilcoxon test.
  1. Please run the statistical test and interpret the result (1 point).

**R Output:**

Wilcoxon signed rank test with continuity correction

data: edata$Cp.cap

V = 1225, p-value = 1.145e-09

alternative hypothesis: true location is not equal to 0

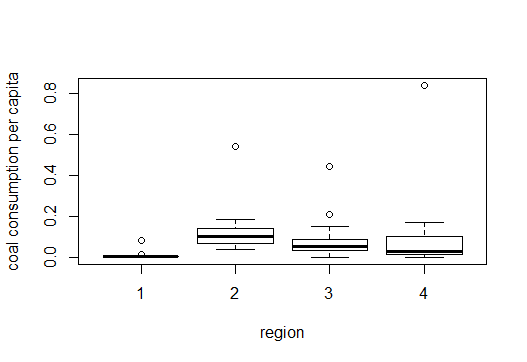
Interpretation: The result of Mann-Whitney-Wilcoxon test indicates that per capita coal consumption does not vary between coastal and inland states (the differences are not significant as shown by the p-value which is greater than 0.05), therefore the null hypothesis is rejected.

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: Per capita coal consumption does not vary across regions.

H1: Per capita coal consumption does vary across regioins

* 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).

1. Normality: The data failed the Shapiro test and the sample size of each group is less than 30, therefore you cannot assume normality. The data is not normally distributed.

2. Variance: The results of the Levene test for Homogeneity of varian (see below) indicate unequal variance

**R Output:**

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

Df F value Pr(>F)

group 3 0.7635 0.5202

47

3. Independence: Because I do not know details about research design, I assume that each data point has been sampled independently

4. Sample size is uneven across groups

* + - * + Because the data does not meet the assumptions of the t-test, I decided to run a non-parametric test: the Kruskal wallis test

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

**R Output:**

Kruskal-Wallis rank sum test

data: Cp.cap by region.i

Kruskal-Wallis chi-squared = 18.097, df = 3, p-value = 0.00042

Interpretation: The result of the Kruskal-Wallis test indicates that there is a significant difference in per capita coal consumption in states across regions. The adhoc test used for the Kruskal Wallis test is typically the Dunn-test; however, I was unsuccessful in running the test in R.

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)

**ROutput:**

Cp.cap gdp.cap

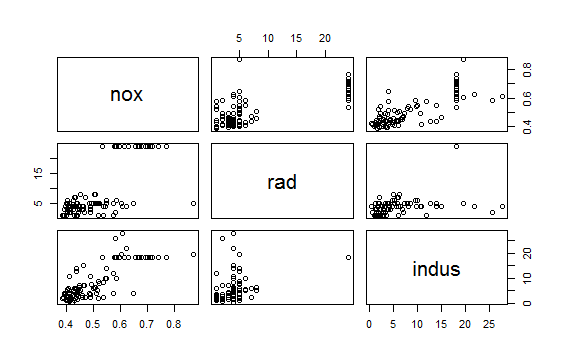
Cp.cap 1.00000000 0.03598182

gdp.cap 0.03598182 1.00000000

Interpretation: The correlation coefficient between per capita coal use and per capita GDP indicates that strength of the relationship between the variables is low (0.036).

**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).
   * + - * Variables selected: NOx ([NOx] in pp10m), RAD (accessibility to radial highways), INDUS (industry)



* + - * + Correlation: All three variables show some degree of correlation with each other (shown below); however the variables “nox” and “indus” have the highest degree of correlation, with an r value of 0.74.

**R Output:**

nox rad indus

nox 1.0000000 0.5424992 0.7385173

rad 0.5424992 1.0000000 0.5133060

indus 0.7385173 0.5133060 1.0000000

* + - * + Multicollinearity: The calculated VIF values (shown below) are all low and less that five, indicating that collinearity between independent variables in the model is low.

**R Output:**

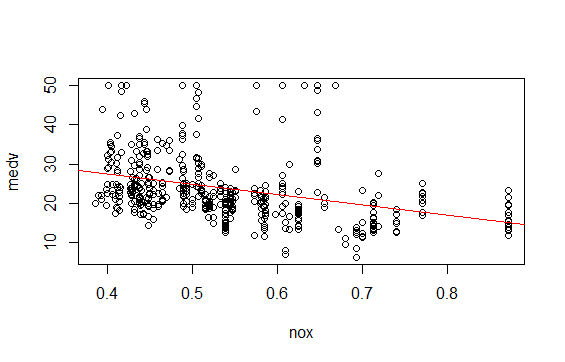
> vif(MR.Bhousing)

nox rad indus

2.390429 1.475418 2.290393

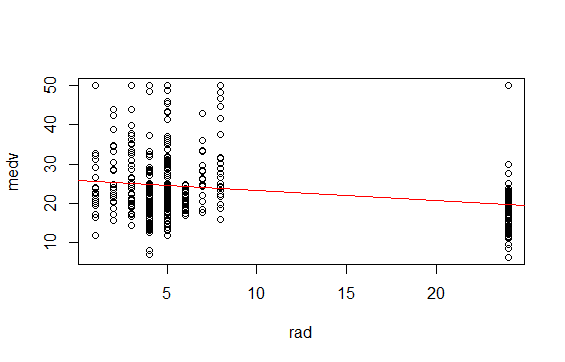
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).

a. Relationship between Nitrogen Oxide levels and median housing prices

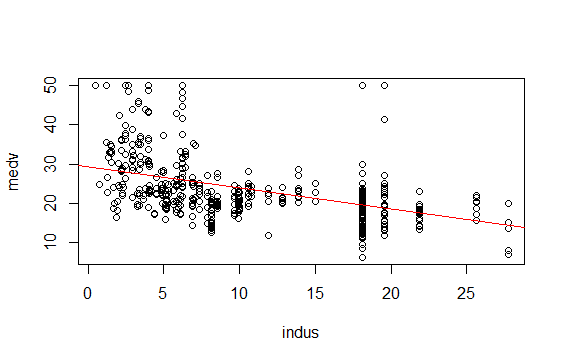


Interpretation: This plot line indicates that median housing prices decrease as Nitrogen Oxide levels increase; however there is a lot of variation in the data (many outliers) that lessen the strength of the relationship between the two variables.

b. Relationship between accessibility to radial highways and median housing prices



Interpretation: The plot indicates that median housing prices are concentrated at the extremes of the x-axis, where rad is less than 10 or greater than 20, with a high degree of variability between the two variables. The plot line indicates that as the index of accessibility to radial highways increases, median house prices decrease.



Interpretation: The plotline indicates that median housing prices decrease as the proportion of non-retail business acres per town increases. Again, there is high variability in the data which lessens the strength of the relationship between the variables.

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).

**R Output:**

> summary(MR.Bhousing)

Call:

lm(formula = medv ~ nox + rad + indus, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-13.086 -4.998 -1.735 2.992 31.000

Coefficients:

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

(Intercept) 31.313974 2.206574 14.191 < 2e-16 \*\*\*

nox -5.020630 5.145789 -0.976 0.330

rad 0.006559 0.060996 0.108 0.914

indus -0.475464 0.084343 -5.637 3.06e-08 \*\*\*

---

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

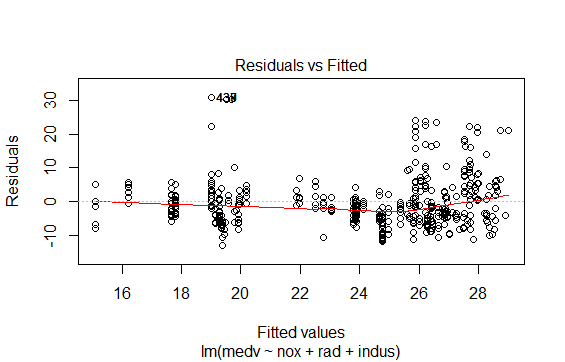
Residual standard error: 8.045 on 448 degrees of freedom

Multiple R-squared: 0.1715, Adjusted R-squared: 0.1659

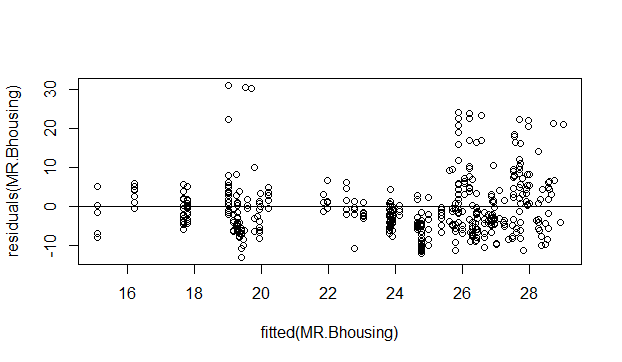
F-statistic: 30.91 on 3 and 448 DF, p-value: < 2.2e-16

Assumptions:

1. Linearity: The graph below indicates that the variation of the residuals against the fitted values is not constant. Additionally, the red plot line is not flat (there is some curvature at the far right end) which indicates that the data may not be linear.



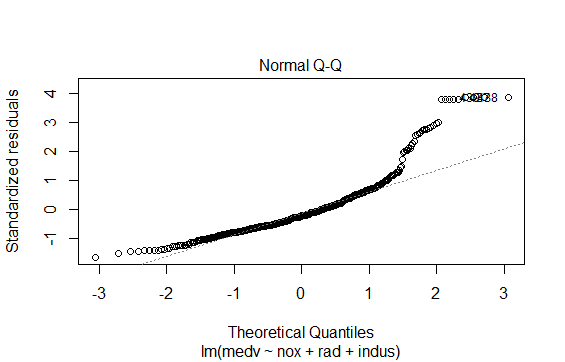
2. Homoscedasticity: The residuals (plotted below) are homoscedastic; however, there is a greater concentration of residual points on top of the line, indicating bias. The results of the Breusch-Pagan test confirm that the residuals are homoscedastic.



studentized Breusch-Pagan test

data: MR.Bhousing

BP = 2.4256, df = 3, p-value = 0.4889

3. Normality: The QQ-plot (shown below) indicates that the data is not normal. 

4. Independence of errors (correlation): The results of the Durbin-Watsin test (results below) indicate that there may be some autocorrelation between the residuals of the three independent variables.

**R Output:**

Durbin-Watson test

data: MR.Bhousing

DW = 0.23836, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is not 0

Non normal dist

Considering that the data set violates the assumptions needed to run the regression, I transformed the data with the boxcox function (see R Output below). The data transformation greatly normalized the residuals (see QQ plot below) and flattened the lined in the residuals vs. fitted plot, indicating greater linearity.

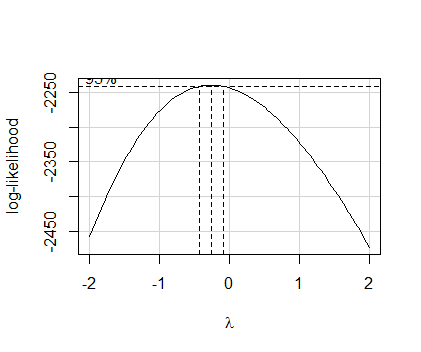
**R Output:**

Data transformation with boxcox:

# bc = boxCox(hdata$medv~hdata$nox+hdata$rad+hdata$indus)

#(trans = bc$x[which.max(bc$y)])

#[1] -0.2626263



**R Output:**

Call:

lm(formula = hdata$medv^trans ~ hdata$nox + hdata$rad + hdata$indus)

Residuals:

Min 1Q Median 3Q Max

-0.112953 -0.019560 0.002466 0.020152 0.144664

Coefficients:

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

(Intercept) 0.3945076 0.0095533 41.295 < 2e-16 \*\*\*

hdata$nox 0.0457842 0.0222785 2.055 0.0405 \*

hdata$rad 0.0002117 0.0002641 0.802 0.4232

hdata$indus 0.0022495 0.0003652 6.160 1.62e-09 \*\*\*

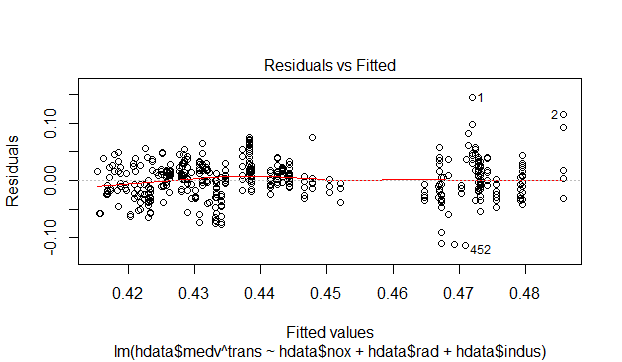
---

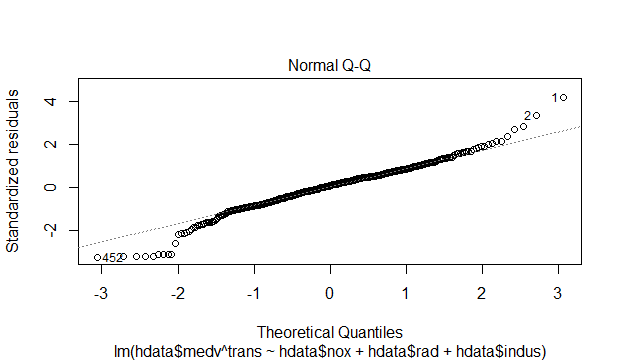
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03483 on 448 degrees of freedom

Multiple R-squared: 0.2563, Adjusted R-squared: 0.2513

F-statistic: 51.47 on 3 and 448 DF, p-value: < 2.2e-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).

Results of the *transformed* data: The fit of the model is low (as indicated by the low Adjusted R-squared value) – only 25% of the variation of median housing prices is accounted for by the three variables. The regression coefficient for “nox” estimates the effect of nox on on median housing prices while controlling for the effect of radial highway access and the proportion of non-retail business acres per town on median housing prices. The same is true for the regression coefficients of “rad” and “indus”.

The regression coefficient for “nox” is 0.046 (and is statistically significant) which means that after controlling for the other two variables, for every unit increase in Nitrogen Oxide levels, median housing prices increase 0.046 units.

The regression coefficient for “rad” is 0.00021; however, it is not statistically significant and its effect on median housing prices is not considered.

The regression coefficient for “indus” is 0.0022 (and is statistically significant) which means that after controlling for the other two variables, for every unit increase in the proportion of non-retail business acres per town, median housing prices increase 0.0022 units.

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

The model is not a great fit for the data: The model tells us that the effect is statistically significant, but it only accounts for about 25% of the variation in the data.