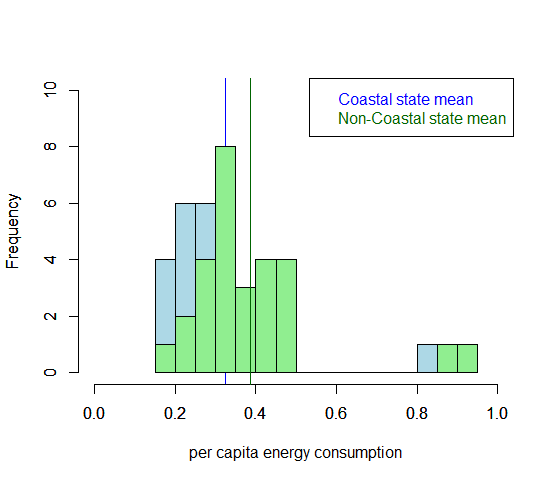
If you run any ANOVAs, you can use the Levene test for equality of variances. If your data violate an assumption about normality and a normal distribution is required for your analyses, you can get bonus points for transforming your data. Otherwise please run the statistical test anyway as if your data 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).
      1. Null: There is no difference in per capita energy consumption between coastal and non-coastal states
      2. Alternate: There is a difference in per capita energy consumption between coastal and non-coastal states
   2. Please create a visual plot to answer this question (1 point).  
      
   3. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
      1. For this question, I should use a two-sample, unpaired t-test because there is one categorical independent variable (coast vs. non-coast) and one continuous dependent variable (per capita energy consumption); need to test assumptions for:
         1. Normality: I tested for normality using the Shapiro-Wilk normality test, and had a resulting p-value of 8.808e-08 (greater than .05), so the distribution is not normal. However, we have a large enough sample size (51 > 30) so the normality assumption is still met.
         2. Equal variances: I used var.test to test equal variance between per capita energy consumptions in coastal and non-coastal states. The results were p-value = 0.5098 (above .05), therefore equal variances can be assumed and the assumption is met.

We also know, but don’t have to directly test:

* + - 1. Continuous dependent variable – have
      2. Independent random sample – have (all states)
  1. Please run the statistical test and interpret the result (1 point).
     1. The p-value = 0.2179 (greater than .05), therefore we cannot reject the null hypothesis that there is a difference in per capita energy consumption between coastal and non-coastal states.
     2. Output:

Two Sample t-test

data: coast[, "pcenergy"] and ncoast[, "pcenergy"]

t = -1.2482, df = 49, p-value = 0.2179

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

95 percent confidence interval:

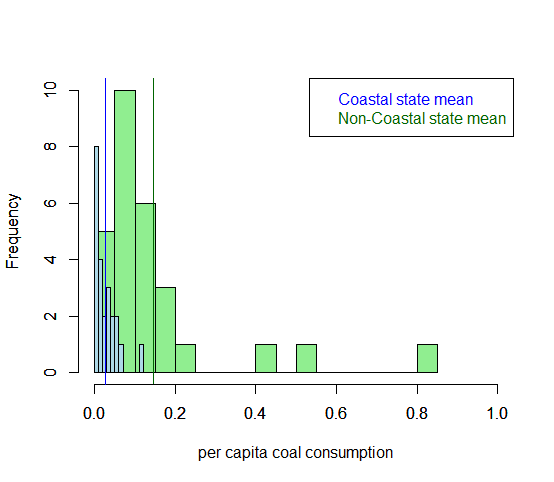
-0.16088003 0.03759778

sample estimates:

mean of x mean of y

0.3238174 0.3854585

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).
      1. Null: There is no difference in per capita coal consumption between coastal and non-coastal states
      2. Alternate: There is a difference in per capita coal consumption between coastal and non-coastal states
   2. 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. For this question, I should use a two sample, unpaired t-test because there is one categorical independent variable (coast vs. non-coast) and one continuous dependent variable (per capita coal consumption);   
        need to test assumptions for:
        1. Normality: I tested for normality using the Shapiro-Wilk normality test, and had a resulting p-value of 6.024e-11 (greater than .05), so the distribution is not normal. However, we have a large enough sample size (51 > 30) so the normality assumption is still met.
        2. Equal Variances: I used var.test to test equal variance between per capita coal consumption in coastal and non-coastal states. The results were p-value = 5.995e-13 (below .05), therefore equal variances cannot be assumed and we should run a Welch’s two sample t-test.

We also know, but don’t have to directly test:

* + - 1. Continuous dependent variable – have
      2. Independent random samples– have (all states)
  1. Please run the statistical test and interpret the result (1 point).
     1. The p-value = 0.001936 (less than .05), therefore we can reject the null hypothesis that there is no difference in per capital coal consumption between coastal and non-coastal states.
     2. Output:

Welch Two Sample t-test

data: coast[, "pccoal"] and ncoast["pccoal"]

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:

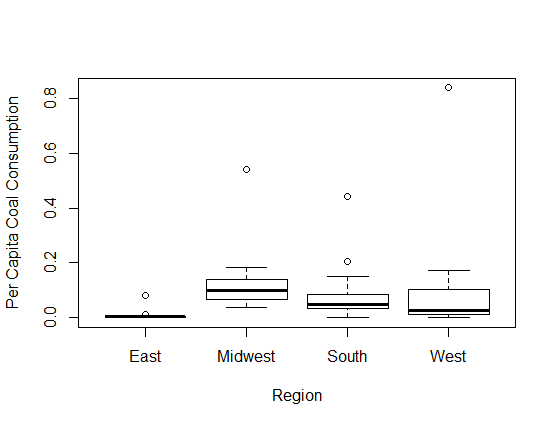
-0.18872795 -0.04724185

sample estimates:

mean of x mean of y

0.02665574 0.14464064

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).
      1. Null: There is no difference in per capita coal consumption between states in different US regions
      2. Alternate: There is a difference in per capita coal consumption between states in different US regions
   2. 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. For this question, I should use a one-way ANOVA, because there is a multi-factor categorical independent variable (state region), and one continuous dependent variable (per capita coal consumption); need to test assumptions for:
        1. Independent random sample - have
        2. Normality of response variable: I tested for response variable normality using the Shapiro-Wilk normality test, and the p value= 6.024e-11 (smaller than .05), so the distribution is not normal; but the number of observations is greater than 30 so we can proceed with the test.
        3. Equal variances: I used Levene’s test to test equal variance between per capita coal consumption between state regions. The results were p-value = 0.5202 (above .05), so equal variances can be assumed and this assumption is met.
  2. Please run the statistical test and interpret the result (1 point).
     1. Output:

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

The p-value= .262 (greater than .05), therefore there is not a significant difference between state region and 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)

The correlation between per capita coal consumption and per capita GDP is 0.03598182. This does not seem like a strong correlation to me, because the closer to 1 or -1, the stronger the correlation is and this correlation is barely above zero. I would consider above .5 to be a strong correlation, therefore per capita coal consumption and per capita GDP do not meet this threshold and are a very weak correlation.

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).

To select three covariates that are not highly correlated, I ran a cor() test with all of the variables besides medv. I knew I wanted to do nitric oxides concentration because I do air pollution research, so based on the correlations, the other two variables I chose were proportion of blacks and pupil to teacher ratio. My threshold for correlations was below .5 (same as question 4).

* 1. Nitric oxides concentration and pupil to teacher ratio correlation value = 0.1034642
  2. Nitric oxides concentration and proportion of blacks correlation value = -0.3584331
  3. Pupil to teacher ratio and proportion of blacks correlation value = -0.08960928
  4. VIF nitric oxides = 1.154209
  5. VIF proportion of blacks = 1.151096
  6. VIF pupil to teacher ratio = 1.014065

From the correlation values and the VIF values, we can see that they these three variables are not highly correlated (correlation value less than .5) and do not have problems with multi-collinearity (VIF less than 10) so I can use them together in the model.

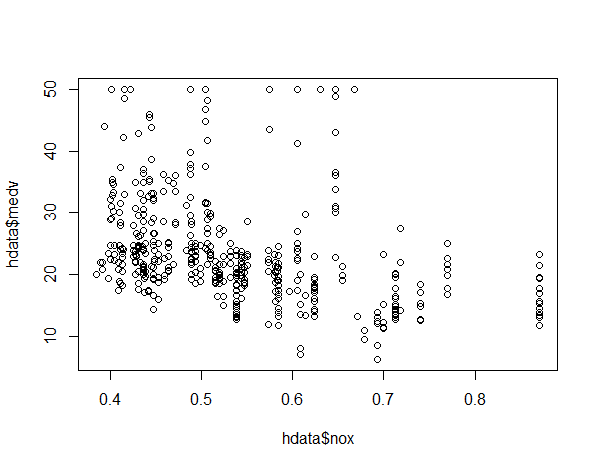
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).

PLOT 1

Independent: Nitric oxides concentration;

Dependent: Median value of owner-occupied homes in $1000’s

Interpretation: There does not seem to be a strong linear relationship between nitric oxide concentration and median value of homes; however the general trend is as nitric oxide concentration decreases, median value of homes decreases.

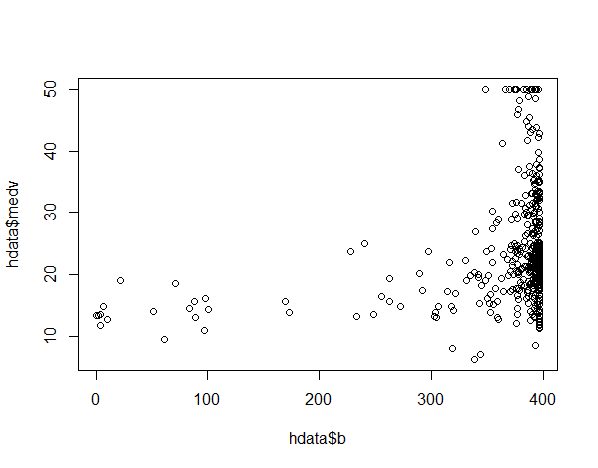


PLOT 2

Independent: Proportion of blacks;

Dependent: Median value of owner-occupied homes in $1000’s

Interpretation: There doesn’t seem to be a strong linear relationship between proportion of blacks by town and median value of homes – it looks like there are a lot more towns with very high proportion of blacks than not, so that may be affecting the distribution. The general trend is that as the proportion of blacks increases, median home value increases.

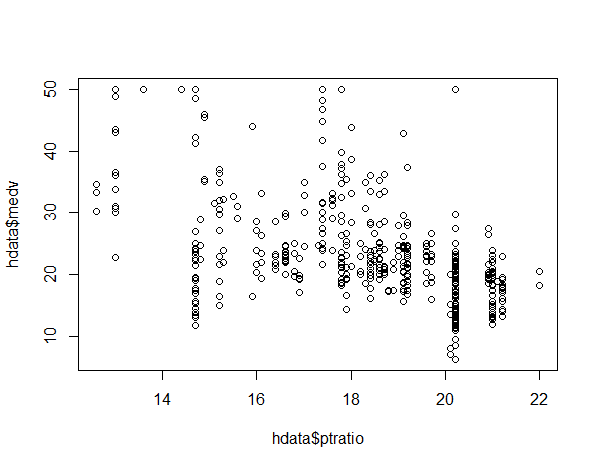


PLOT 3

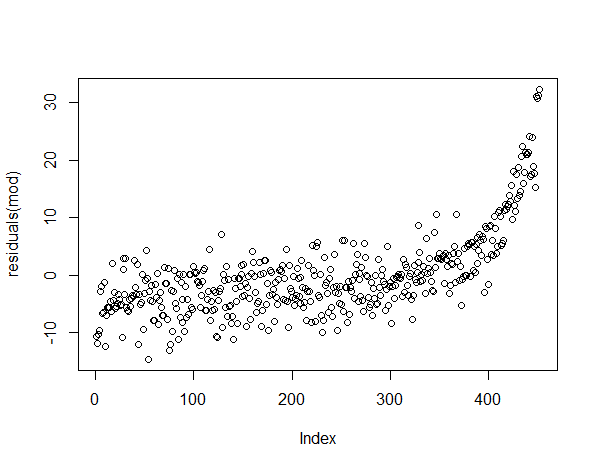
Independent: Pupil to teacher ratio;

Dependent: Median value of owner-occupied homes in $1000’s

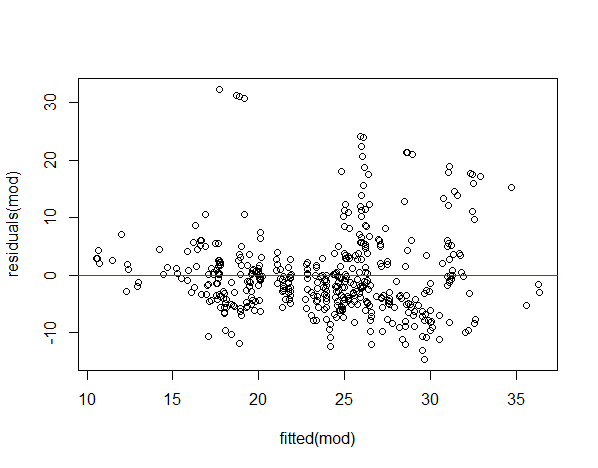
Interpretation: There doesn’t seem to be a strong linear relationship between pupil to teacher ratio and median value of homes, however the general trend is as pupil to teacher ratio increases, median value of home decreases.

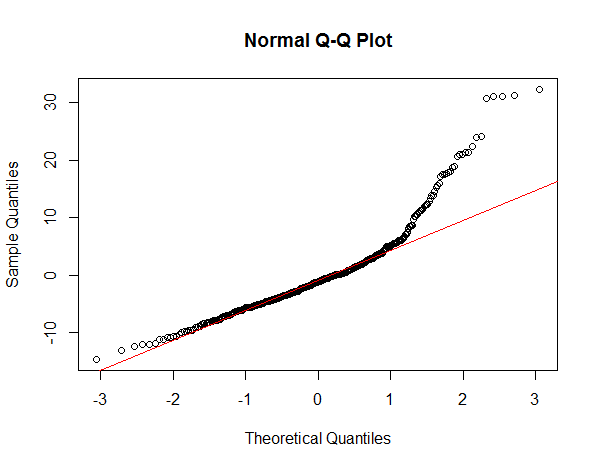


1. Run your multiple linear regression model. Check whether any assumptions are violated. Please state **which assumptions** you checked, **whether they were violated**, and **how you know** whether or not they were violated. If any assumptions are violated (e.g. normality), we will give you bonus points if you are able to identify a way to overcome this problem (3 points, plus additional 1 point bonus).
   1. Assumptions:
      1. Residual independency
         1. [Plot: definitely looks like there is a trend] I ran the Durbin-Watson statistic test to determine autocorrelation. The p value < 2.2e-16, this means that residuals show a clear temporal autocorrelation (are not independent) and I should manipulate the data before doing the linear regression. ASSUMPTION VIOLATED.



* + 1. Residual homoscedasticity
       1. [Plot: looks like the residuals are not homoscedastic, more above line than below]. I ran the studentized Breusch-Pagan test to test for residual heteroscedasticity. The p-value = 0.004039, which means the residuals ARE heteroscedastic, therefore NOT homoscedastic. ASSUMPTION VIOLATED.



* + 1. Residual normality
       1. [Plot: also looks like the residuals are not normal give the many dots above the line near the high end of the theoretical quantiles]. I ran the Shapiro-Wilk test to test for normality of residuals. The p-value < 2.2e-16, which means that residuals are not normally distributed. ASSUMPTION VIOLATED.

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).

(Answers assume that assumptions were met, even though they were not)

* 1. Intercept
     1. Coefficient: 57.970000, p-value: < 2e-16

The intercept is significant at the .05 level, therefore we can say that if all other factors are 0, the median value of owner-occupied homes would be 57.97 $1000’s of dollars ($57,970).

* 1. Nitric oxide concentration
     1. Coefficient: -18.437433, p-value: 2.71e-08

The coefficient is significant at the .05 level, therefore we can say that nitric oxide concentration affects median value of owner-occupied homes, and that for every $1000 increase in median value of owner-occupied homes, there is a 18.44 parts per 10 million decrease in nitric oxide concentration.

* 1. Pupil to teacher ratio
     1. Coefficient: -1.697178, p-value: < 2e-16

The coefficient is significant at the .05 level, therefore we can say that pupil to teacher ratio affects median value of owner-occupied homes, and that for every $1000 increase in median value of owner-occupied homes, there is a 1.70 decrease in pupil to teacher ratio.

* 1. Proportion of blacks in town
     1. Coefficient: 0.018172, p-value: 0.000834

The coefficient is significant at the .05 level, therefore we can say that the proportion of blacks affects median value of owner-occupied homes, and that for every $1000 increase in median value of owner-occupied homes, there is a 0.02 increase in proportion of blacks.

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

The Adjusted R-squared value is 0.3076, which tells us how much variance can be explained by the model. I would say this model is a bad fit because the Adjusted R-squared tells me the model explains less than half of the variance (30.76%) and because there are a lot of other factors that likely impact median value of homes that were not included.