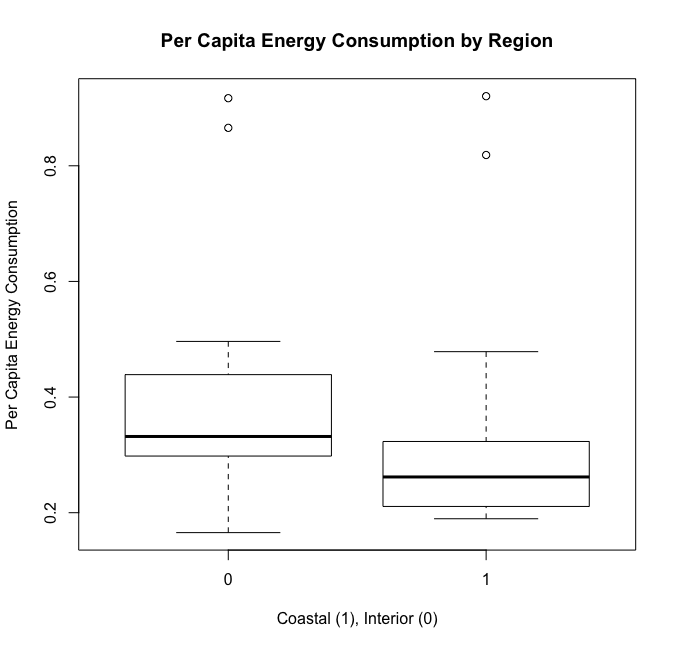
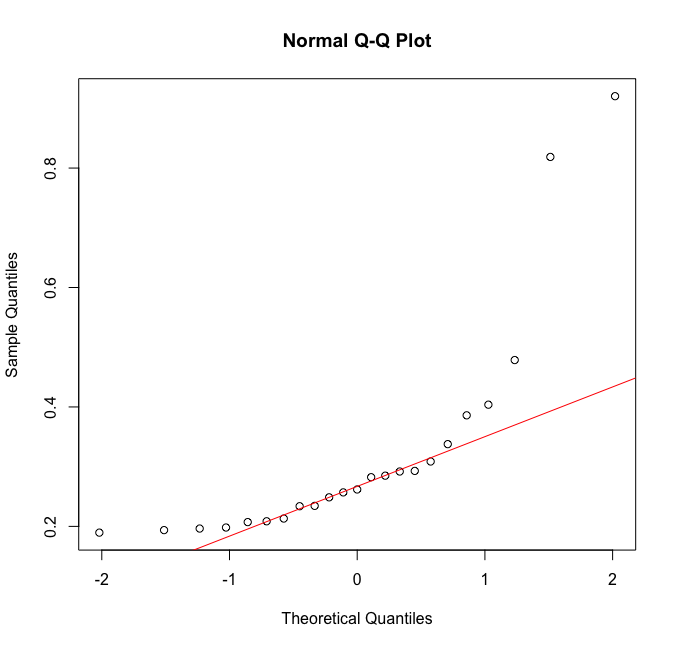
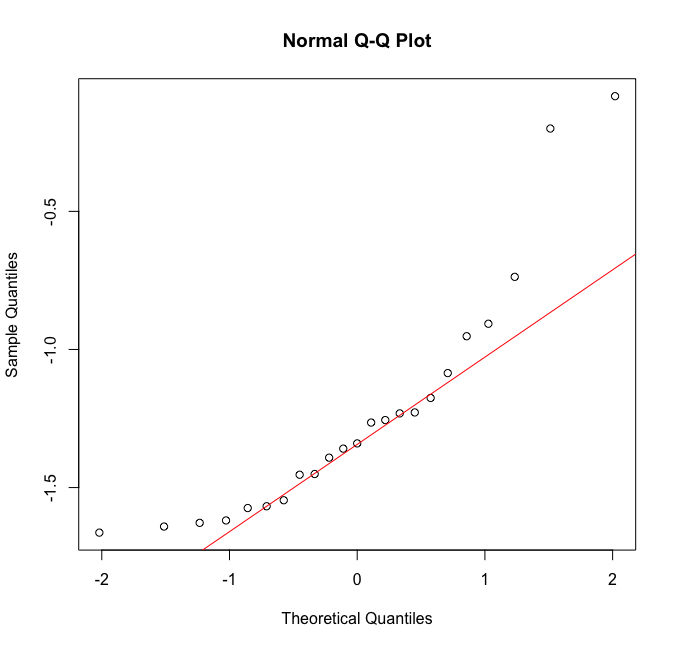
If you run any ANOVAs, you can use the Levene test for equality of variances. If your data violate an assumption about normality and a normal distribution is required for your analyses, you can get bonus points for transforming your data. Otherwise please run the statistical test anyway as if your data 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.

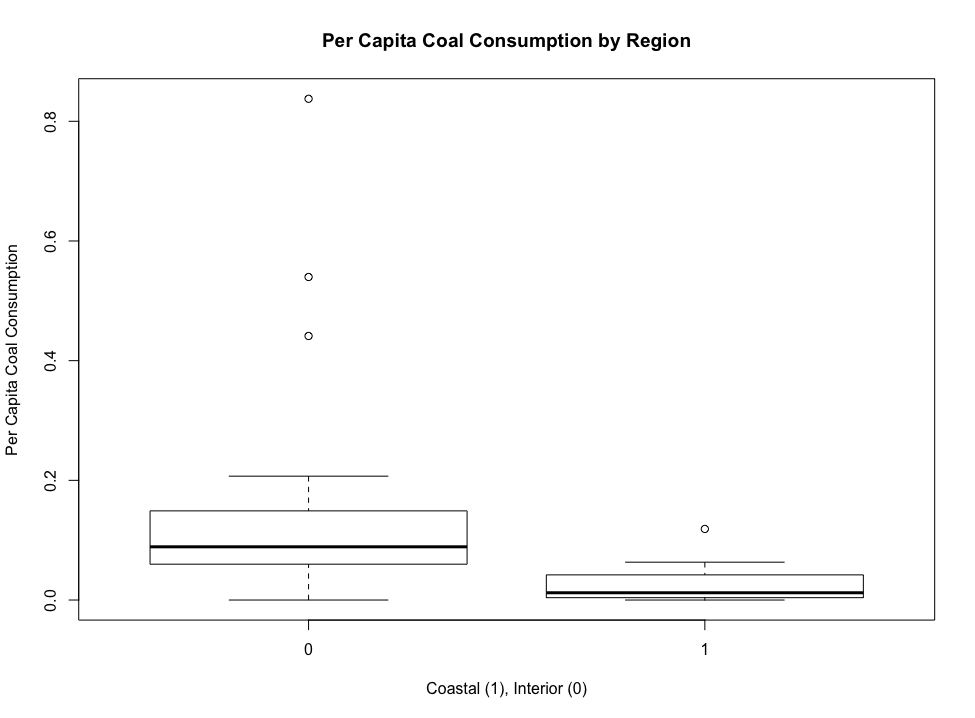
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized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

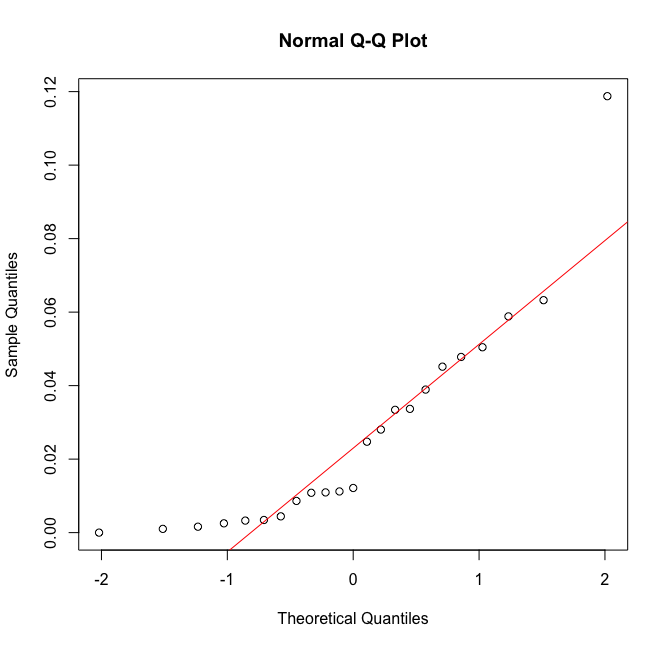
1. Does ***per capita*** energy consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).
      1. H0: Whether a state is coastal or not has no effect on per capita energy consumption.
      2. HA: Coastal states have different per capita energy consumptions than interior states.
   2. Please create a visual plot to answer this question (1 point).
      1. boxplot(percapitaE~Coast, data=edata, main="Per Capita Energy Consumption by Region", xlab="Coastal (1), Interior (0)",ylab="Per Capita Energy Consumption")
      2. 
   3. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
      1. As the data is comparing the difference in per capita energy consumption between two groups the best statistical test to use would be the two-sample t-test.
      2. The assumptions of a t-test are:
         1. Dependent variable data is continuous–The dependent variable, per capita energy consumption, is continuous.
         2. Observations are independent–The observations were independently collected
         3. Variance is equal
            1. var.test(Coastal$percapitaE, Interior$percapitaE)
            2. The variance for each data set, coastal and noncoastal, is equal based on the var.test(), which returned a p-value >.05. This means the null hypothesis, which is that the variance is equal, cannot be rejected.
         4. Values are normal or nearly normal–The data is not normal. For both coastal and interior per capita energy use, the shapiro-wilkes test returns a p-value < .05, meaning the null hypothesis (the data is normal) should be rejected.
            1. Shapiro-Wilk normality test for Coastal per capita energy used: W = 0.66467, p-value = 5.037e-06
            2. 
            3. To adjust for this, a data transformation can be used, such as log or square-root transformations, or the Wilcox-Mann-Whitney test can be used. Although a log transformation normalized the Interior per capita energy use data, it did not normalize the Coastal per capita energy use data, nor the total edata set.

Shapiro-Wilk normality test, data=Coastal$logE : W = 0.83398, p-value = 0.001413

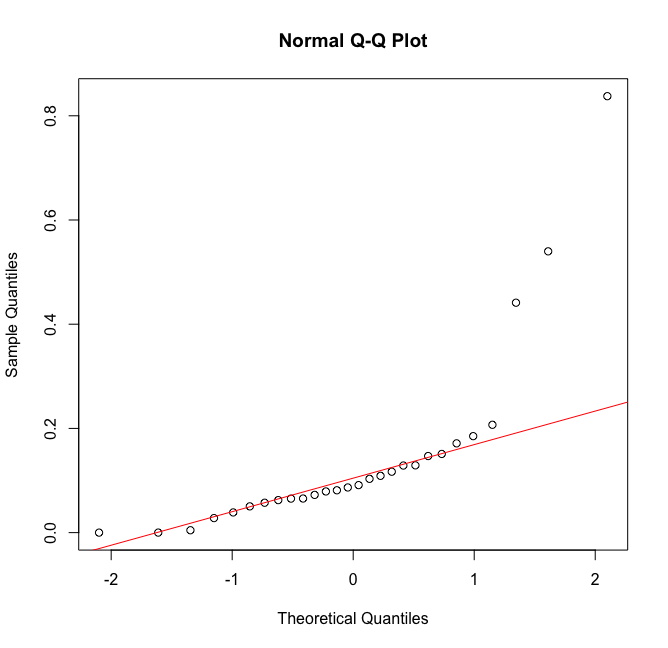


* + - * 1. As neither the log nor the square root transformations yielded a normal distribution of the data for the coastal data set or full edata set, a Wilcox-Mann-Whitney test will have to be used rather than a two-sample t-test.
  1. Please run the statistical test and interpret the result (1 point).
     1. wilcox.test(percapitaE~Coast, data=edata)
     2. This test returns a p-value <.05, indicating that the null hypothesis is not supported. Thus, the alternative hypothesis, that there is a difference between the per capita energy consumption of coastal and interior states is supported.

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. H0: Per capita coal consumption is not affected by a state’s location on the coast or the interior.
      2. HA: States found on the coast have a different per capita coal consumption than states found in the Interior.
   2. Please create a visual plot to answer this question (1 point).
      1. 
   3. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
      1. As the data is comparing the difference in per capita coal consumption between two groups the best statistical test to use would be the two-sample t-test.
      2. The assumptions of a t-test are:
         1. Dependent variable data is continuous–The dependent variable, per capita energy consumption, is continuous.
         2. Observations are independent–The observations were independently collected
         3. Variance is equal
            1. var.test(Interior$percapitaCoal, Coastal$percapitaCoal)
            2. The variance for each data set, coastal and noncoastal, is equal based on the var.test(), which returned a p-value >.05. This means the null hypothesis, which is that the variance is equal, cannot be rejected.
         4. Values are normal or nearly normal
            1. The data is not normal. For both coastal and interior per capita coal use, the shapiro-wilkes test returns a p-value < .05, meaning the null hypothesis (the data is normal) should be rejected.

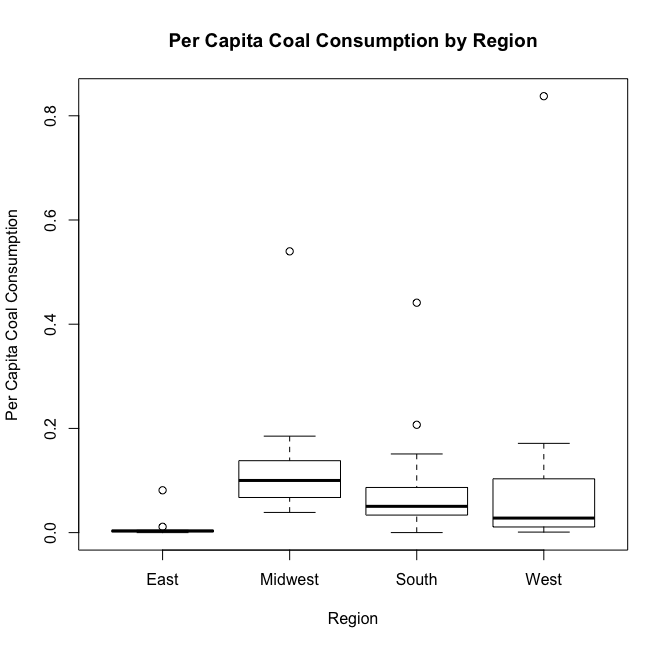
Coastal

Shapiro-Wilk normality test, data= Coastal$percapitaCoal: W = 0.82584, p-value = 0.001023

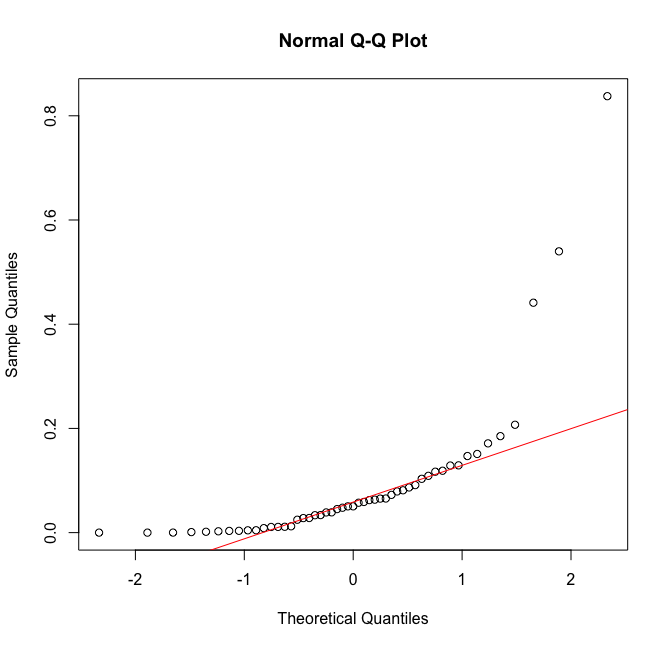
Interior

Shapiro-Wilk normality test, data= Interior$percapitaCoal: W = 0.651, p-value = 6.211e-07

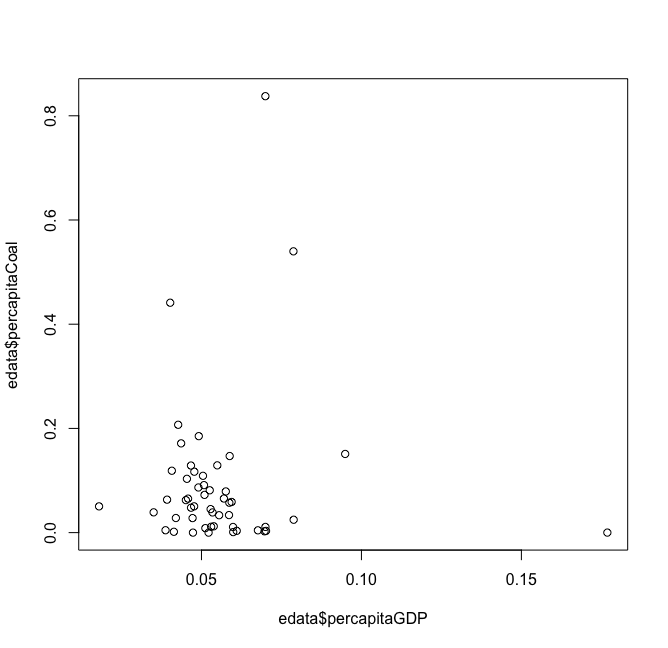
* 1. Please run the statistical test and interpret the result (1 point).
     1. t.test(Coastal$percapitaCoal,Interior$percapitaCoal,paired=FALSE)
     2. Running the Welch’s Two Sample t-test (although the data violates the normality assumptions of the t-test) results in a p-value<.05, indicating that the null hypothesis is not supported. Thus, the alternative hypothesis, that there is a difference between the per capita coal consumption of coastal and interior states is supported.

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. H0: Per capita coal consumption is not affected by a state’s region.
      2. HA: States in different regions have different per capita coal consumptions.
   2. Please create a visual plot to answer this question (1 point).
      1. 
      2. boxplot(percapitaCoal~Region, data=edata, main="Per Capita Coal Consumption by Region", xlab="Region",ylab="Per Capita Coal Consumption")
   3. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
      1. As the data has a continuous dependent variable and more than two categorical independent variables, the best test to use would be an ANOVA.
      2. The assumptions of an ANOVA are
         1. Normal distribution of population
            1. The data is not normal as the shapiro-wilkes test returns a p-value<.05 and the qqnorm plot shows that the data does not map onto a straight line.

Shapiro-Wilk normality test, edata$percapitaCoal : W = 0.57291, p-value = 6.024e-11



* + - 1. Independent samples
         1. The data was collected independently as far as I know.
      2. Equal variance of population
         1. Using leveneTest(percapitaCoal~Region, data=edata) returns a p-value>.05, indicating that the population is of equal variance.
  1. Please run the statistical test and interpret the result (1 point).
     1. mod <- lm(percapitaCoal~Region, data=edata)
     2. summary(mod)
     3. The results of this model indicate that only the Midwest region is significantly different from the East in terms of percapita coal consumption, and this difference is only marginally significant. The low R-squared suggests that the model does not explain a large amount of the data variance.

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)
   1. Based on cor(edata$percapitaGDP, edata$percapitaCoal), the correlation between per capita coal use and per capita GDP is 0.03598182. This does not seem like a strong correlation to me as the correlation coefficient is between 1 and -1. This result is close to zero. Creating a plot of the data also shows a low correlation.
   2. 

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).
   1. The covariates I selected are rm, nox, and ptratio.
   2. The correlation values for these are:

cor(hdata[, c("rm", "nox", "ptratio")], use="na.or.complete")

rm nox ptratio

rm 1.0000000 -0.2645944 -0.3341642

nox -0.2645944 1.0000000 0.1034642

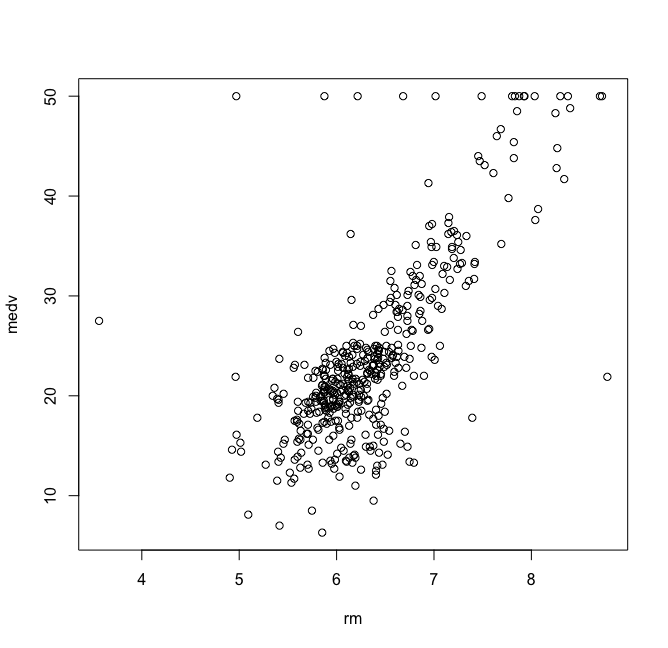
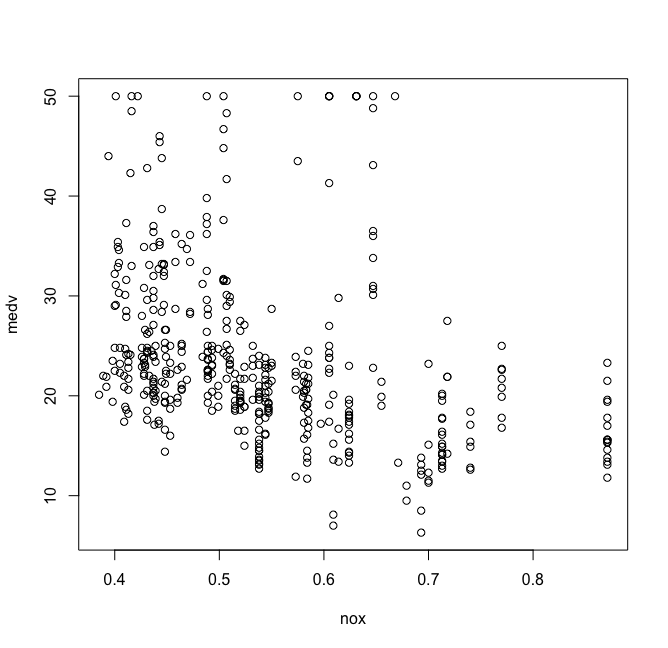
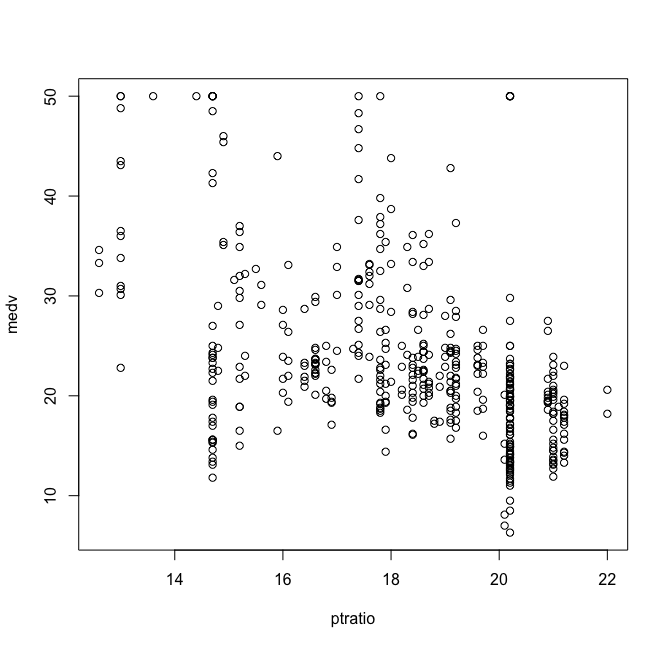
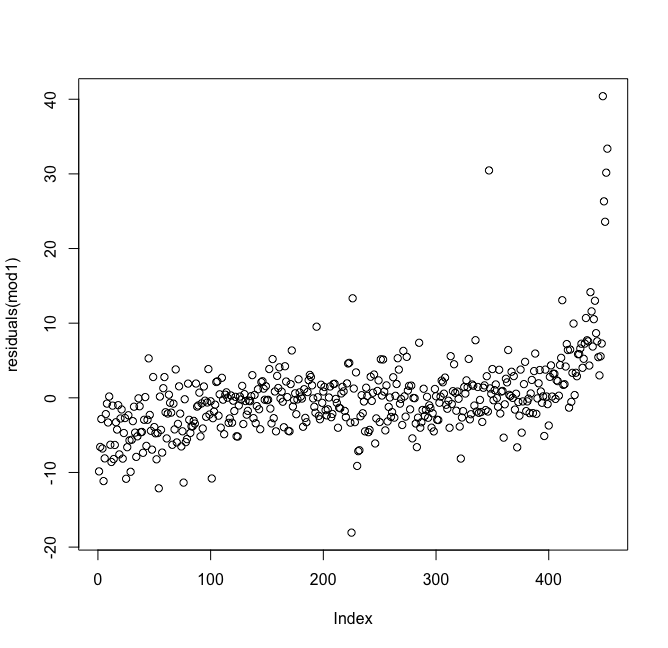
ptratio -0.3341642 0.1034642 1.0000000

* 1. The VIF values for these are:

vif.hdata = 1/(1 - summary(lm(medv~rm+nox+ptratio, dat=hdata))$r.squared)

[1] 2.621043

* 1. These values indicate that the variables do not have issues with mutli-colinearity, as the VIF < 5 and none of the r values are > .5.

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).
   1.  This plot suggests that the average number of homes per dwelling (rm) is positively correlated with housing prices. The relationship appears to be fairly strong, with most data points clustering around a clear linear pattern.
   2. This plot suggests that the nitric oxide concentrations do not correlate significantly with housing prices. There is a slight negative trend, however it is not strong.
   3. This plot suggests that pupil-teacher ratio does not correlate significantly with housing prices. There is no obvious trend in the data, indicating that the relationship is likely not a very significant one.
2. 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. mod1 <- lm(medv~rm+nox+ptratio, data=hdata)
   2. summary(mod1)
   3. Assumptions:
      1. Linear relationship
         1. I checked this assumptions using the plot() function. Medv and rm appear to have a positive linear relationship, however the relationship is not as clear for medv and nox or medv and ptratio. Both of these variables show a possible negative linear trend, but there is a large amount of variance around the trend line.
      2. Independence of errors
         1. I checked this using the plot(residuals(mod1)) function and the Durbin-Watson Test.
            1. 
            2. Durbin-Watson test : DW = 0.93721, p-value < 2.2e-16 – alternative hypothesis: true autocorrelation is not 0
         2. Both of these tests indicate that the errors are not independent, as the plot shows a pattern in the data and the DW test returns a very low p-value.
      3. Homoscedasticity
         1. To test for homoscedasticity, I used

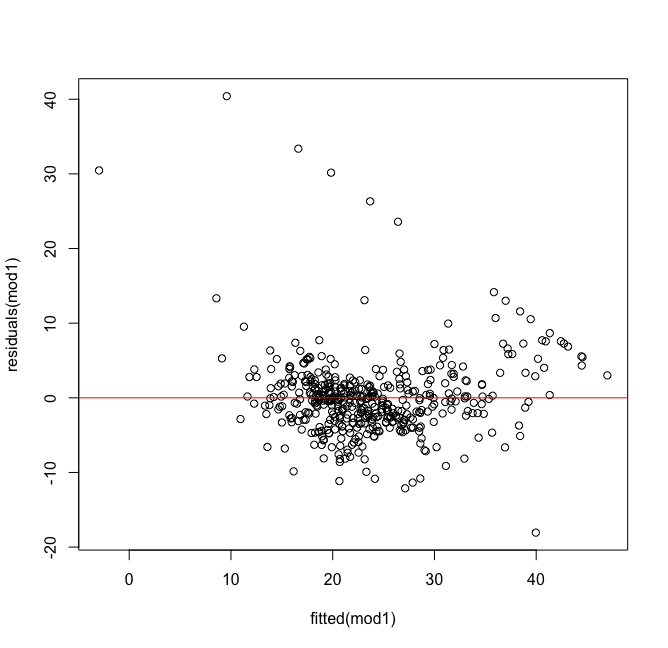
plot(residuals(mod1)~fitted(mod1))

abline(lm(residuals(mod1)~fitted(mod1)), col="red")

and

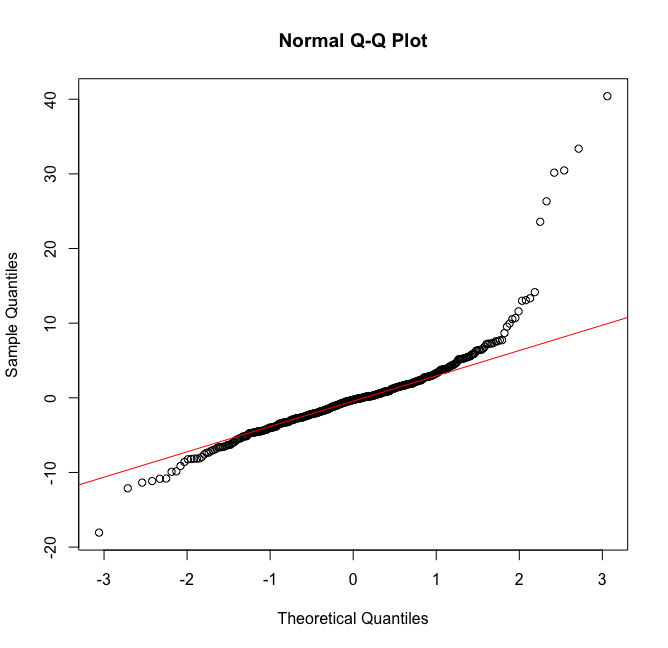
library(lmtest)

bptest(mod1)

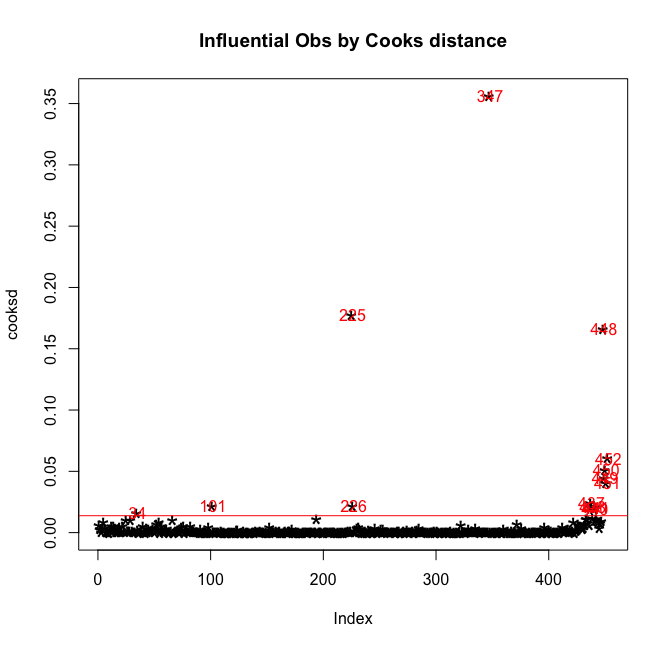


studentized Breusch-Pagan test: BP = 11.333, df = 3, p-value = 0.01006

Both of these tests indicate that the data is not homoscedasticitic (p-value<.05) as it bunches up near the middle. There are also some outlying points that could be skewing the distribution.

* + 1. Normal error distribution
       1. I used qqnorm(residuals(mod1)) and shapiro.test(residuals(mod1)) to test for normality of the residuals.
          1. 
          2. shapiro.test(residuals(mod1)) : W = 0.81107, p-value < 2.2e-16
          3. Both of these tests indicate that the data violates the normality of the residuals assumption.

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).
   1. The intercept is -5.0099, meaning that the y-intercept of the estimated mean housing price is -5.0099.
   2. The coefficient of rm is 8.2246, meaning that for every increase in housing price, average number of rooms per dwelling increases by 8.2246.
   3. The coefficient for nitrous oxide concentrations (nox) is -11.0969, meaning that for the variance that is unexplained by rm decreases 11.0969 for every increase in mean housing price.
   4. The coefficient for prtration is -0.9541, meaning that for the variance unexplained by rm or nox, the pupil-teacher ratio decreases by 0.9541for every increase in mean housing price.
2. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?
   1. The R^2 is 0.6185, which for social science is a decent R^2 and means that 0.6185 of the variance is explained by the model. However, a number of the model assumptions were violated, which suggests that a linear model might not be the best fit for the data. A general linear model might be better, or a non-linear model. Alternatively, using the Cooks Distance function indicates that there are some outliers in the data. Removing these outliers could improve the fit of the model and make a linear model more appropriate.



cooksd <- cooks.distance(mod1)

plot(cooksd, pch="\*", cex=2, main="Influential Obs by Cooks distance")

abline(h = 4\*mean(cooksd, na.rm=T), col="red")

text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4\*mean(cooksd, na.rm=T),names(cooksd),""), col="red") # add labels