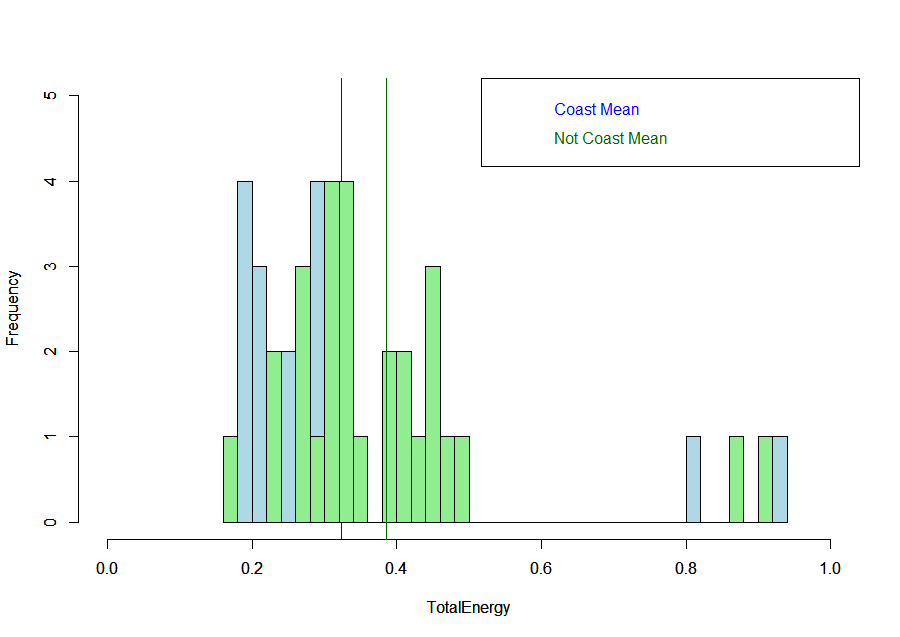
If you run any ANOVAs, you can use the Levene test for equality of variances (leveneTest). If your data violate an assumption about normality, please decide if this is really a problem. In many cases you can still run your parametric test with non-normal data assuming other conditions are met (see lecture notes). If you choose to run a parametric test any way despite the data not being normally distributed, state why you are able to do this. HINT: there is only one analysis in the entire exam (which is clearly marked) where you should run into real problems with normality. For this one analysis, you can get bonus points for transforming your data. If you are unable to transform your data, run the statistical test any way as if your data were normally distributed but make it clear that you violated this assumption in your answer (you won’t lose any points for violating this assumption). I’ve also updated Lecture15.R due to one mistake in the code.

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).
   2. Please create a visual plot to answer this question (1 point).
      1. From this plot, it is difficult to say whether total energy per capita on the coast and not coast areas are significantly different. We can see that their means are different, but it’s unclear if it’s statistically different.



* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. I would use a non-paired, two-sample, two-tailed t-test because we are attempting to determine if the means from two groups are significantly different from each other, and not different from a specific number (which would be one-sample). Additionally, the data are not related to one another (under different time conditions or in the same field plot), so it would be unpaired.
     2. There are five assumptions with a two-sample t-test:
        1. Equal variance between the two populations
           1. Test using the var.test

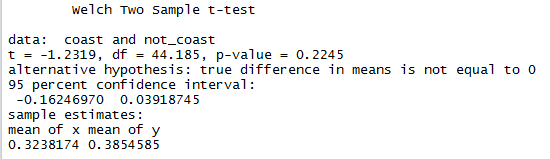
This assumption is met, because the p-value is 0.5098, which is greater than 0.05.

This means that we cannot reject our null hypothesis that the variances in “coast” and “not coast” are equal.

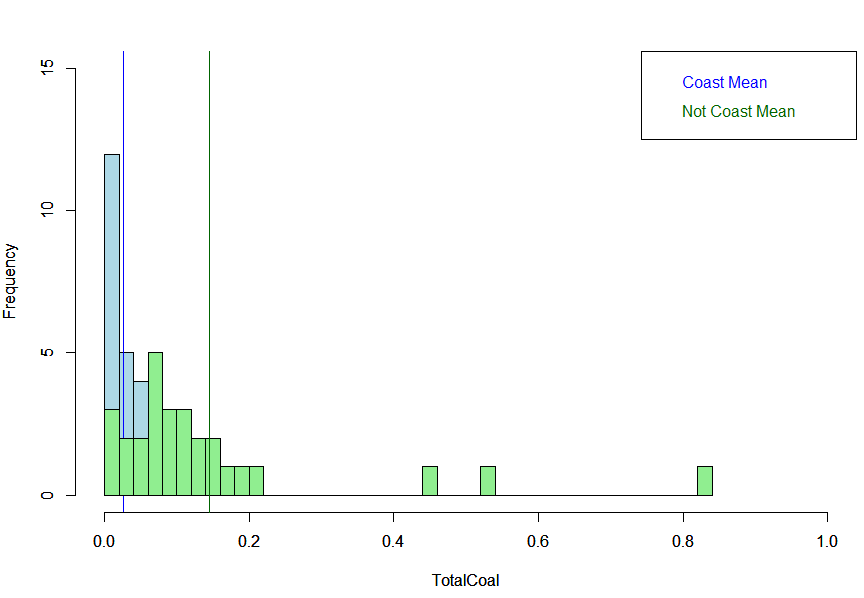
* + - 1. Samples are normally distributed
         1. Test using the Shapiro test (shapiro.test).

Unfortunately, this assumption is not met. The p-value for the “TotalCoal” per capita data is 8.808e-8, which means that we should reject our null hypothesis that the samples are normally distributed.

However, since the sample size is greater than 30, we can assume normality based on the central limit theorem (CLT).

* + - 1. Observations are sampled independently
         1. This one is hard to test, so you need to have some background on the data, which we do, and can assume the observations are independent of one another.
      2. Samples are randomly selected from the population
         1. We are assuming these are randomly selected, and that there is no bias. This is just data collected based on real-time energy consumption, so there is no reason to believe there is bias here, and I will assume there isn’t.
      3. Data for the dependent variable are continuous
         1. As we can see from the spreadsheet, the total energy per capita data are, in fact, continuous.
  1. Please run the statistical test and interpret the result (1 point).
     1. After running the t-test, we see that the p-value is 0.2245, which is greater than 0.05 (the somewhat arbitrary number that we use to determine significance). Since the p-value is greater than 0.05, we cannot reject the null hypothesis that the two samples are equal, so we therefore can conclude that the difference between per capita energy consumption for the coast and non-coast states is not significantly different from one another.
     2. 

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).
   2. Please create a visual plot to answer this question (1 point).
      1. From this graph, it is pretty clear that there is a significant difference between the amount of energy from coal consumed on the coast versus the amount of coal energy consumed on the non-coast states. We can use statistics to prove that this difference is significant.



* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
     1. This question can be set up in exactly the same way as the previous question, only using coal energy instead of total energy. As such, we would use the same test and assumptions.
     2. I would use a non-paired, two-sample, two-tailed t-test because we are attempting to determine if the means from two groups are significantly different from each other, and not different from a specific number (which would be one-sample). Additionally, the data are not related to one another (under different time conditions or in the same field plot), so it would be unpaired.
     3. There are five assumptions with a two-sample t-test:
        1. Equal variance between the two populations
           1. Test using the var.test

The p-value is lower than 0.05 (it’s 5.995e-13), so we have to reject our null hypothesis that they are the same, which means our data do not pass this test.

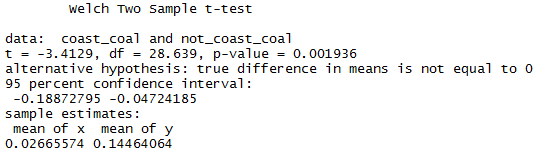
This means that the variances in “coast\_coal” and “not coast\_coal” are equal.

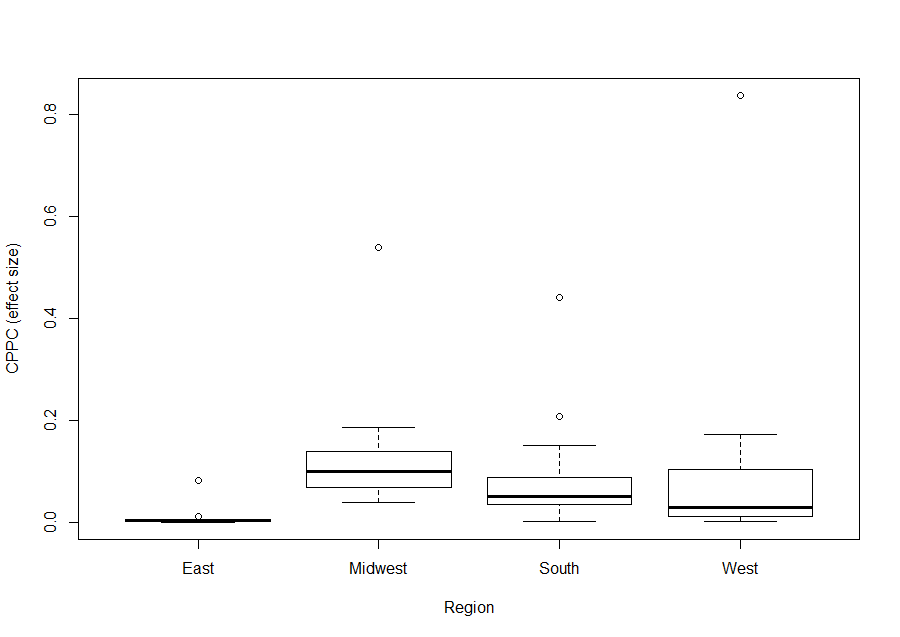
As a result, we will have to use a Welch’s t-test

* + - 1. Samples are normally distributed
         1. Test using the Shapiro test (shapiro.test).

Unfortunately, this assumption is not met. The p-value for the “TotalCoal” per capita data is 6.024e-11, which means that we should reject our null hypothesis that the samples are normally distributed.

However, since the sample size is greater than 30, we can assume normality based on the central limit theorem (CLT).

* + - 1. Observations are sampled independently
         1. This one is hard to test, so you need to have some background on the data, which we do, and can assume the observations are independent of one another.
      2. Samples are randomly selected from the population
         1. We are assuming these are randomly selected, and that there is no bias. This is just data collected based on real-time energy consumption, so there is no reason to believe there is bias here, and I will assume there isn’t.
      3. Data for the dependent variable are continuous
         1. As we can see from the spreadsheet, the total coal energy per capita data are, in fact, continuous.
    1. Based on the fact that the variances between the two samples are unequal, we will need to conduct a Welch’s test instead of a traditional t-test. The t-test in R uses a Welch’s t-test by default, but we will have to specifically say that the variances are not equal.
  1. Please run the statistical test and interpret the result (1 point).
     1. After running the Welch’s t-test (and specifying that the variances are not equal), we see that the p-value is 0.00194, which is less than 0.05 (the somewhat arbitrary number that we use to determine significance. Since the p-value is less than 0.05, we can reject the null hypothesis that the two samples are equal, so we therefore can conclude that the difference between per capita energy consumption from coal for the coast and non-coast states are significantly different from one another.
     2. 

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. All per capita coal consumption in each region is the same.
        2. I’m not sure how to right this in a formula, but basically, is at least one of these different from the others—they don’t all have to be different.
   2. Please create a visual plot to answer this question (1 point).
      1. Based on this visual, it looks as though Midwest has the highest mean, followed by the South, West, and finally, the East. These data seem to make sense given what I know about coal consumption in the country, and I have a feeling the “outlier” for the West is California. Now we can see if any of these are statistically different from one another.
      2. 
   3. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).
      1. I would choose to run a one-way ANOVA for this test, which is basically like a two-sample t-test, only with more groups (3 or more), which is exactly what we want given our null hypothesis. We are testing to see if there is a statistical difference in the means of four different sample groups.
      2. There are four main assumptions that we need to meet in order to do this test.
         1. Samples are normally distributed
            1. Test using the Shapiro test (shapiro.test).

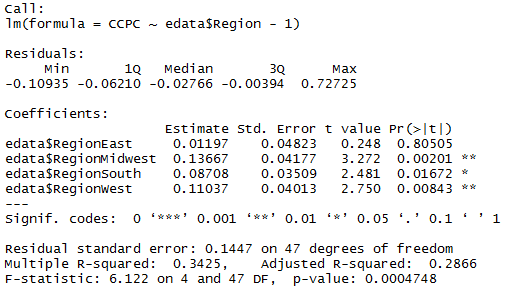
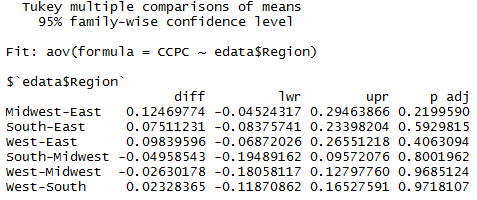
Unfortunately, this assumption is not met. The p-value for the “TotalCoal” per capita data is 6.024e-11 (same as in answer 3), which means that we should reject our null hypothesis that the samples are normally distributed.

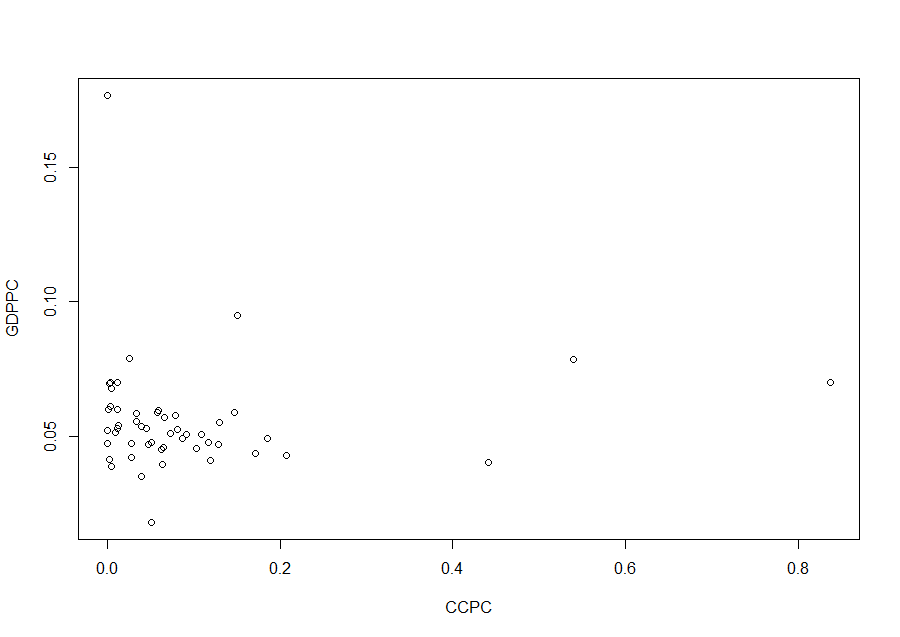
However, since the sample size is greater than 30, we can assume normality based on the central limit theorem (CLT).

* + - 1. Equal variance between the populations
         1. Test using the leveneTest from the Rcmdr package

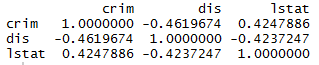
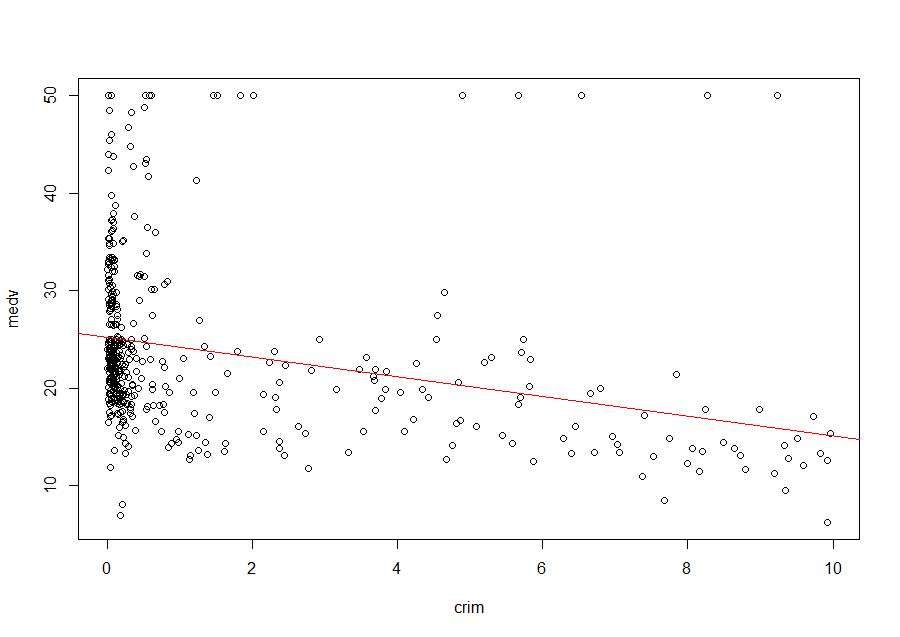
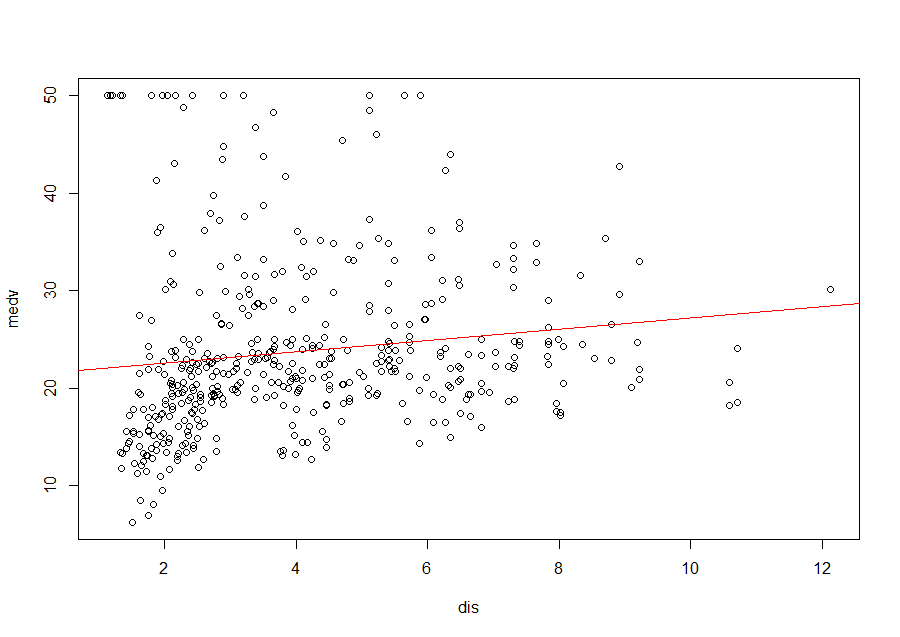
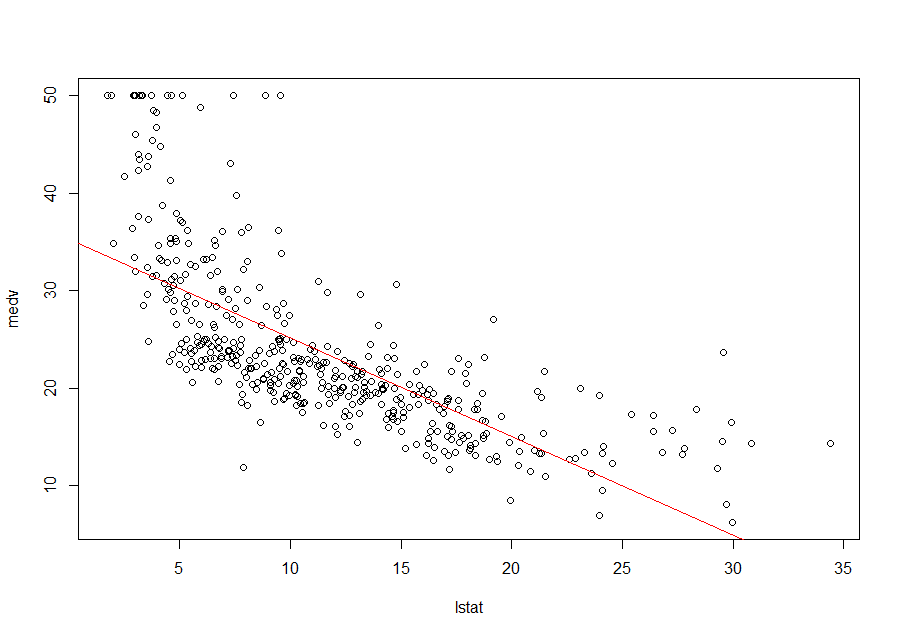
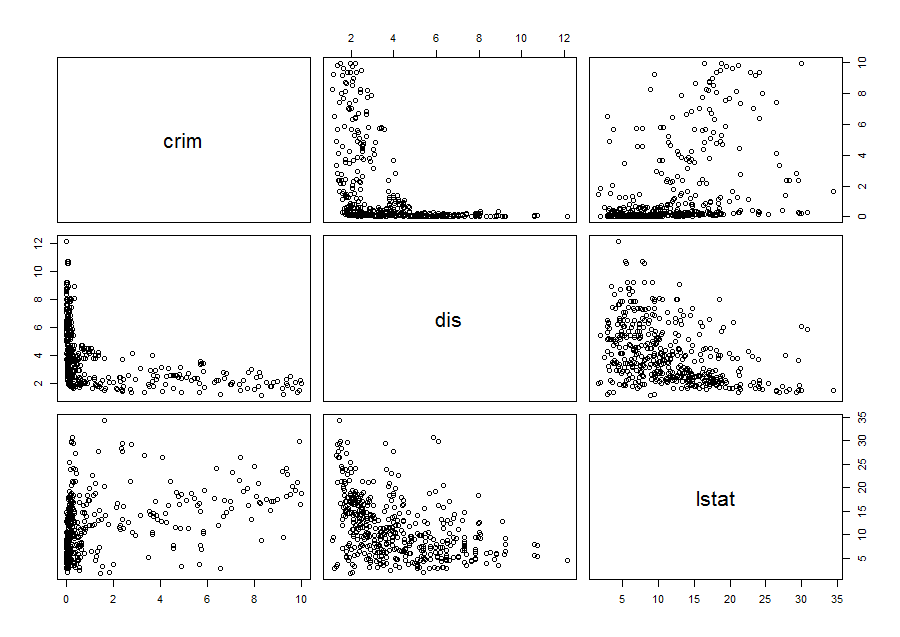
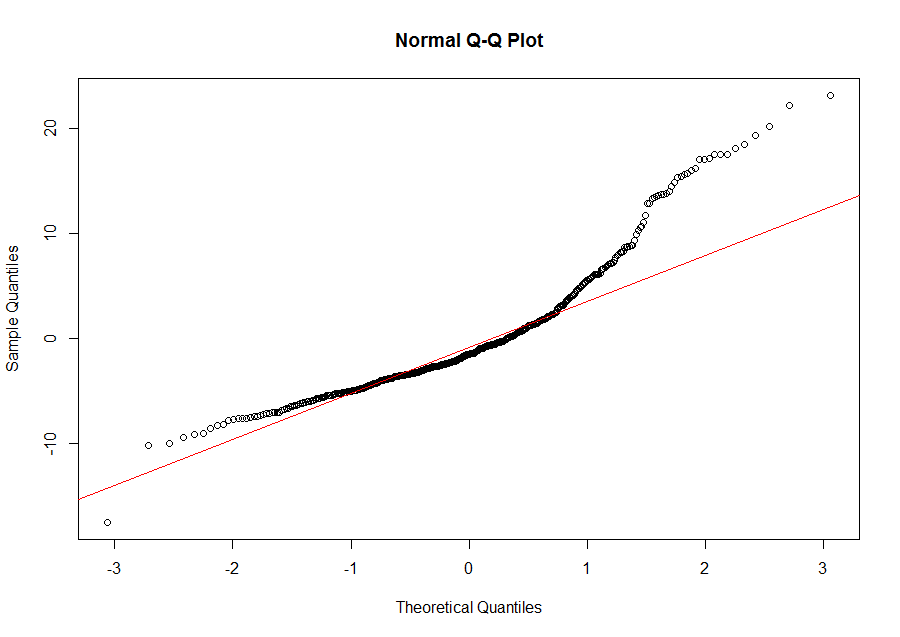
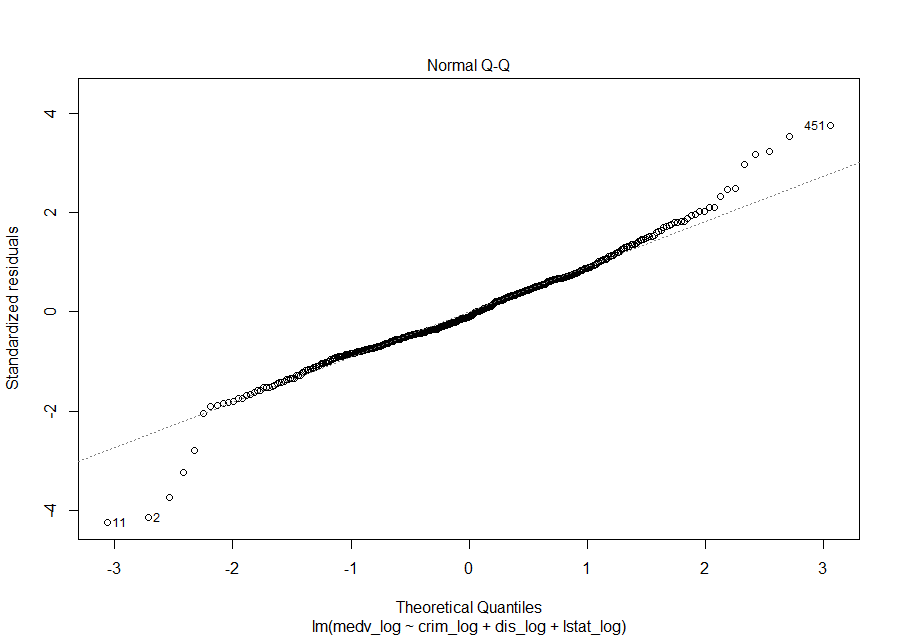
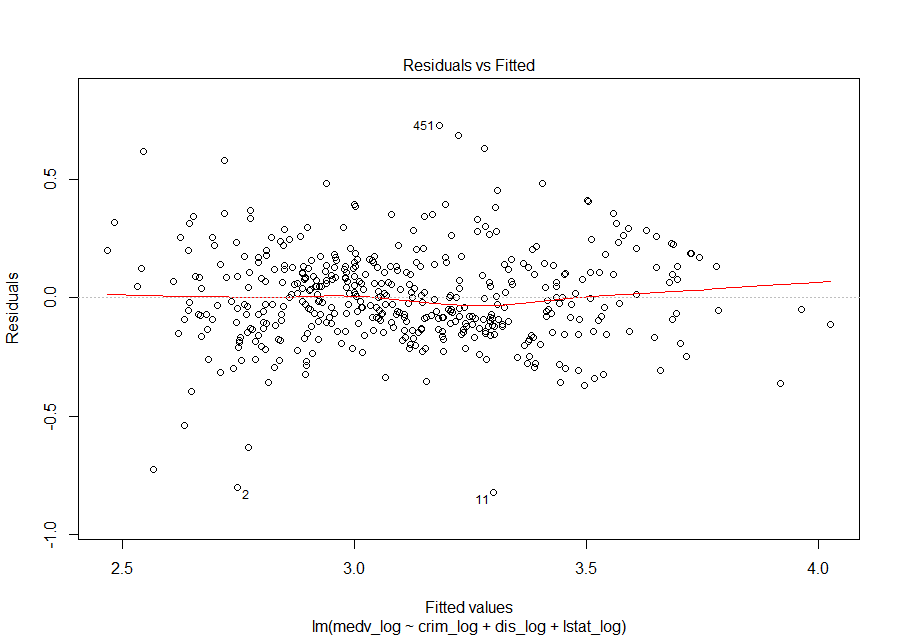
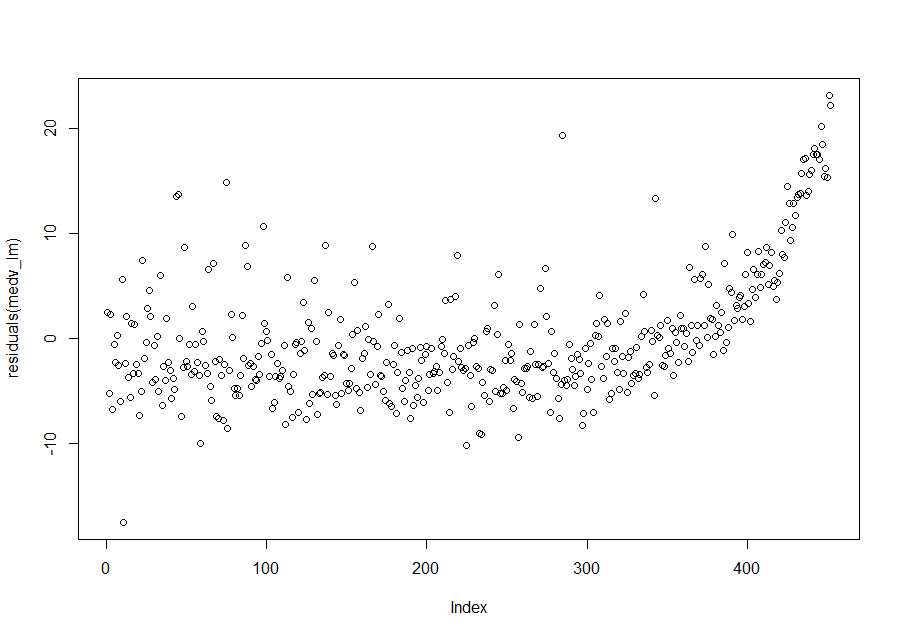
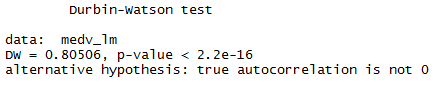
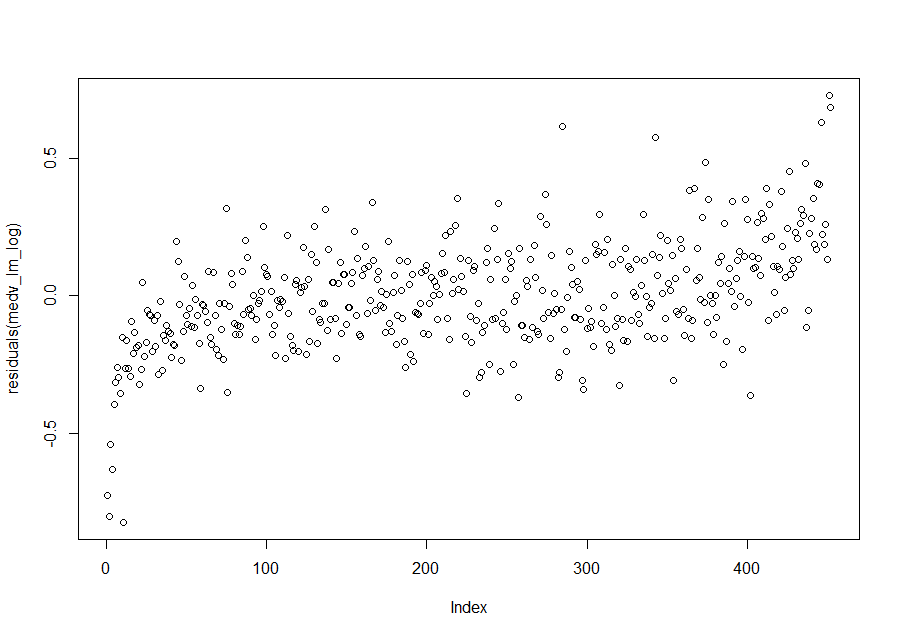
This assumption is met, because the p-value is 0.5202, which is greater than 0.05.

This means that we cannot reject our null hypothesis that the variances in the “region” dataset are equal.

* 1. Please run the statistical test and interpret the result (1 point).
     1. After running the one-way ANOVA, we see that there are some differences among the treatments. From the output below, we see that the p-value for the East region is not significant, which indicates that there are differences between the samples and we may not be able to reject our null hypothesis that we are 95 percent confident that this region is equal to the average coal consumed per person in the other regions. We cannot tell from this output, however, what the differences between each of the samples are—only whether there is difference from each other. We need to do a post-hoc test to determine that.
     2. 
     3. After running a post-hoc test, the Tukey’s honest significant differences (TukeyHSD), we see that none of the regions are significantly different from one another, as indicated by the fact that their adjusted p-values are all higher than 0.05. This means that we would reject our null hypothesis that all groups are the same.
     4. 

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. When looking at the correlation between per capita coal use and per capita GDP, it appears as though there isn’t a strong correlation (the value is 0.03598). Typically, we think of something being highly correlated if it is greater than 0.50.
   2. Just to see what this looks like, I made a plot (below), and it looks as though there are a few outliers, but most of them are centered in the bottom left.
   3. 
   4. In order to examine the correlation between these two variables a bit further, I decided to take out some of these outliers, and ran the test again to see if this made a difference. I took out everything from the CCPC dataset that was above 0.4, and everything from the GDPPC dataset that was above 0.10. The plot is below.

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 three covariates that I initially wanted to include as independent variables in my multiple linear regression model to predict housing value prices (medv) are below, along with VIF values for each answer.
      1. CRIM—per capita crime rate by town
         1. VIP = 1.089 (much smaller than 10, which is what we want)
      2. DIS—weighted distances of five Boston employment centers
         1. VIP = 1.019 (much smaller than 10, which is what we want)
      3. LSTAT—percent lower status of the population
         1. VIP = 1.995 (much smaller than 10, which is what we want)
   2. Fortunately, according to the VIP, none of these factors have issues of multi-collinearity using the variance inflation factor, as shown in the VIP results above, when compare only with the dependent variable. When testing the VIP for the entire proposed model (medv ~ crim + dis + lstat), the results are also very low, at 2.139, which indicates that there is little collinearity among the independent variables in regards to the dependent variable.
   3. In addition, these three variables were also checked with a simple correlation test (see output below) to see if they were correlated with each other in any way, which they aren’t. All of the variables are below 0.5 or -0.5, which indicates that they are mostly not correlated to one another.
      1. 
2. 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. medv ~ crim
      1. The relationship appears to have a decreasing relationship. Although I’ve plotted a line here, the relationship doesn’t really look linear—it appears that it could be someone exponential.
      2. 
   2. medv ~ dis
      1. I would say that this graph has a positive relationship, but perhaps not linear. It look as though the plot are mainly under the “best fit” line, but the relationship is not quite exponential (although it appears to be following that path), and is decelerating rather than accelerating.
      2. 
   3. medv ~ lstat
      1. I think the relationship with the lstat to medv is a decreasing relationship, but again, I don’t think that a linear relationship is the best fit. I think this is a decreasing exponential relationship, given the downward and sharp decline in apparent slope in the beginning of the data points (on the x-axis), and the gradual slope towards the end of the graph (on the x-axis). 
      2. We can also see the relationships of each independent variable in R using the pairs function. As you can see from the graph below, there appears to be a sharply decreasing exponential relationship between crim and dis, and a slightly decreasing exponential relationship between lstat and dis. The crim and lstat relationship is a little less clear, but it appears as though there is a somewhat linear relationship with crim and lstat, but most of the data points are on one axis.
      3. 
3. 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. Based on what we saw from the correlation and relationships between the variables, we already know that a linear model may not be the best fit for some of these—at least without transforming the data. However, it’s good to check and be sure. We should check three different assumptions:
      1. Normal distribution
         1. We need to ensure that the data are normally distributed.
            1. We can start by plotting the data using a QQ plot, which reveals the graph below. From this plot, we can see that the residuals are not normally distributed, because they do not follow the red line. They appear to be somewhat exponential.
            2. 
            3. We can also test for normality based on the Shapiro test. This test confirms that the data are not normally distributed. The p-value is 2.2e-16, which means that we should reject our null hypothesis that the residuals are normally distributed.
            4. In order to correct for this, we can try to manipulate the data. Given that there is a right skew to the graph, we can attempt to log transform the data and see if we can meet the normality assumption. The graphs showing this attempt are below. As we can see, the QQ plot looks a lot more normal than the previous version. However, unfortunately, we still do not meet the normality assumption, because the Shapiro test still says that the p-value is 1.994 e-7 (an improvement), which means we have to reject our hypothesis that the data are normally distributed. This is mostly due to the outliers, which you can see particularly well in the QQ plot below.
            5. 
            6. 
      2. Residual independency
         1. In plotting the residuals prior to transforming the data, we can see that there is a clear pattern, which implies that the residuals are not independent of one another (below).
         2. 
         3. We can also test this using the Durbin-Watson test (dwtest) in the lmtest package. This test confirms what we see in the plot of the residuals. The p-value is 2.2e-16, which means that we should reject our null hypothesis that the residuals are not correlated.
         4. 
         5. After manipulating the data and doing a log transform, we can see that the residuals have less of a pattern, and are approaching homoscedasticity.
         6. 
         7. However, after using the Durbin-Watson test again, we unfortunately still have to reject our null hypothesis that the residuals are not correlated, because the p-value is still below 0.05 (it’s at 1.06e-14).
      3. Homoscedasticity
         1. To determine if a linear model is a good fit, we next need to look at the distribution of the variance of residuals to see if there is any pattern. If there is a pattern, and the variance of the residuals is not constant, a linear model will not be a good fit.
   2. -Checking if we have a good linear model or not, there are three things to do:
      1. Residual independency (possibly with the dwtest (in lmtest package))
         1. Null is that we have a residual that is not correlated to each other
            1. Cannot reject the null hypothesis means that we pass
      2. Homoscedasticity (bp.test)
         1. Testing if the residual of this y depends on this y
         2. Null is the same--that we have no autocorrelation in our residual
         3. If p-value is high, we cannot reject, and pass the test
      3. Normality (shapiro)
4. 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).
5. Discuss the fit of your model and whether you think it is a good or bad fit. Why (2 points)?