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.

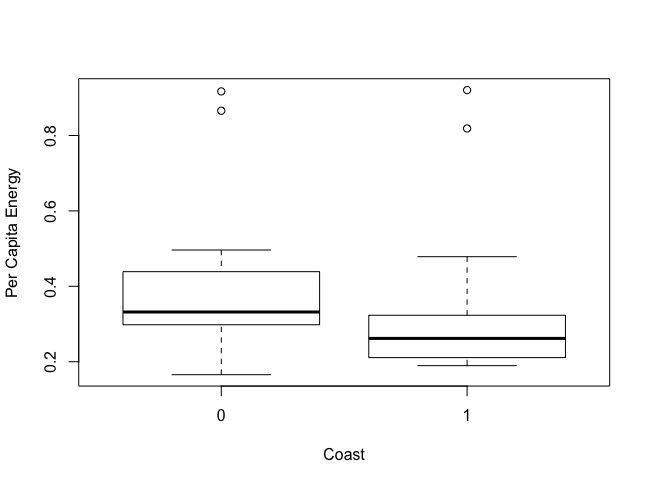
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized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

Null hypothesis: Per capita energy consumption has no difference in costal and no costal states.

Alternate hypothesis: per capita energy consumption differs depending on whether a state is found 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).

I wanted to use two-sample t-test. I used shapiro test to know if the sample is normal distribution. Since both p value of per capita energy consumption in coast group and p value of per capita energy consumption in inland group are smaller than 0.05, data of these two groups are not normally distributed. Sample size of two groups are smaller than 30, so we cannot use two-sample t-test. I also checked if variance of these two group is same by using F-test. Their variance is same. I select to use **Wilcoxon-Mann-Whitney U test.**

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

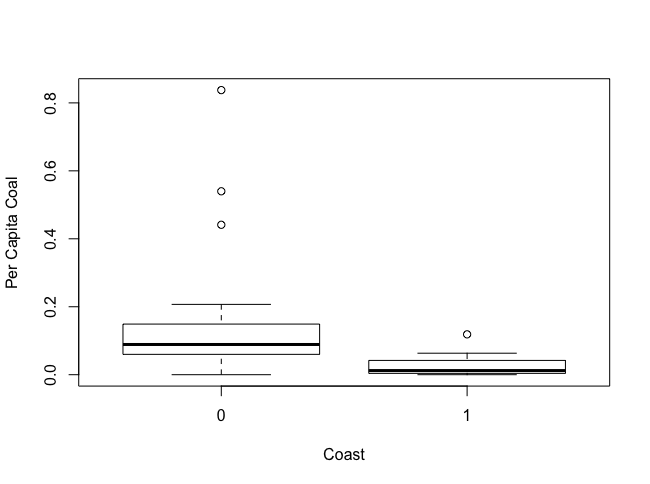
The p-value of Wilcoxon-Mann-Whitney U test is 0.008417 (smaller than 0.05). So the per capita energy consumption in coast and inland states is significantly different.

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).

Null hypothesis: per capita coal consumption is not significantly different depending on whether a state is found on the coast or not.

Alternate hypothesis: per capita coal consumption is significantly different between coast and inland group.

* 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).

I wanted to use two-sample t-test. I used shapiro test to know if the sample is normal distribution. Since both p value of per capita coal consumption in coast group and p value of per capita energy consumption in inland group in shapiro test are smaller than 0.05, data of these two groups are not normally distributed. Sample size of two groups are smaller than 30, so we cannot use two-sample t-test. I also used F-test to check if the variance is different. I got a p-value that is smaller than 0.05. So the variance is different. Based on condition I meet, I **resample** the per capita coal consumption in inland, and per capita coal consumption in coast area, then use **Welch two sample t-test**.

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

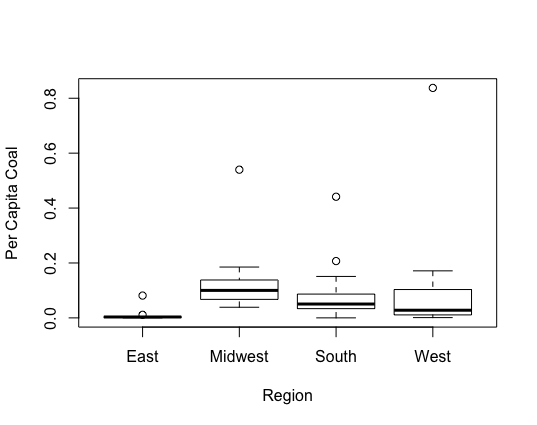
The p value of Welch two sample t-test is smaller than 2.2e-16, so the per capita coal consumption in inland and coast area is significantly different from each other.

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).

Null hypothesis: the per capita coal consumption in different region is same.

Alternative hypothesis: the per capita coal consumption is different in different 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).

I use one-way Anova. So I will use shapiro test to check if the data is normally distributed, and use F-test to check if variance of these two samples is different. I also check the sample size of this distribution. Variance of these four groups are not same, and the data is not normally distributed. The sample size of each group is not larger than 3 times. So I run the **kruskal test**.

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

P value of kruskal test is 0.00042 (smaller than 0.05), this means the per capita coal consumption is different in at least one region.

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)

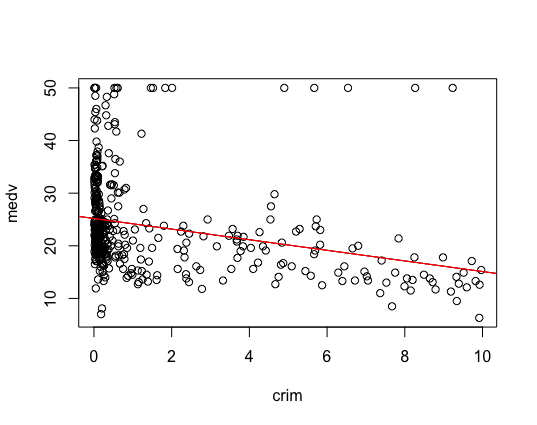
The correlation value between per capita coal use and per capita coal use and per capita GDP is 0.03598182. This is a positive correlation but not very strong, because this correlation value is not high.

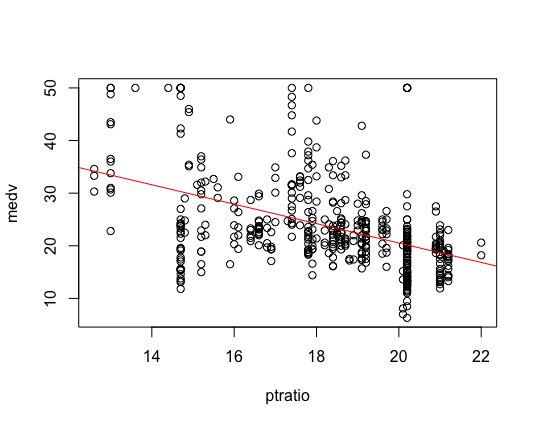
For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple 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.

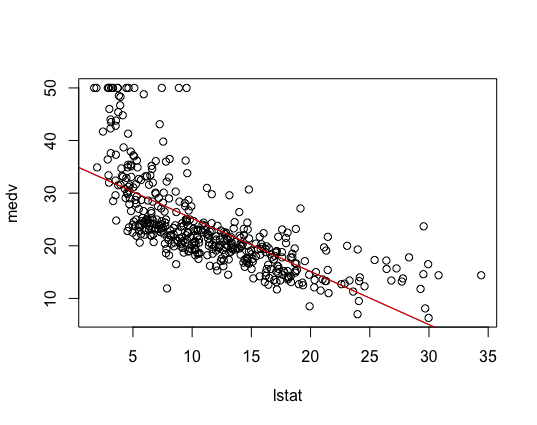
1. Please select three covariates that you will include in your model as independent variables. Please check if these variables are highly correlated with one another to make sure you do not run into problems of multi-collinearity. Check if this model has issues with multi-collinearity using the variance inflation factor. **Report correlation values and VIF values in your answer** (3 points).

I select crim (per capita crime rate by town), ptratio(pupil-teacher ratio by town) and lstat (%lower status of the population) in the dataset. The correlation value between these three independent variables are: cor1(crim, ptratio) = 0.3194701, cor2(crim, lstat) = 0.4247886, cor3(ptratio, lstat) = 0.3030431. Then I use VIF to check their multi-collinearity. VIF1(crim~ptration+lstat) = 1.282884, VIF2 (ptratio~lstat+crim) = 1.157712, VIF3(lstat~crim+ptratio) = 1.268439. All VIF values are lower than 5, therefore, these three independent variables are not high correlated with each other.

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).





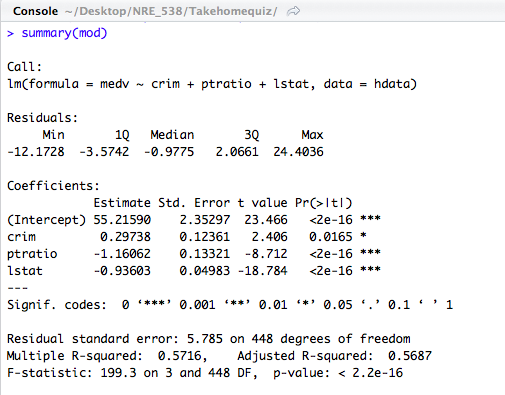


From three plots, we can see that all three independent variables are negatively correlated with median housing prize, because when these three independent variables increase, the medv decreases. However, the correlation between lower status percent or crime rate and housing prize may not be a linear relationship.

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. Check the residual homoscedasticity. I use studentized Breusch-Pagan test to check the residual homoscedasticity. The p-value is equal to 3.703e-09 (smaller than 0.05), so the residual is not homoscedastic.
   2. Check the residual independency. I use Durbin-Watson test to check the residual independency. The p-value is smaller than 2.2e-16. That means the residual has an autocorrelation, so the residual is not independent.
   3. Check the residual normality. I use shapiro-test to check the independency of the residual. The p-value is 9.946e-16 (smaller than 0.05), so the residual is not normally distributed.
   4. When looked at plots from last question, I guess the data is not linear regressed.

To solve problems, I tried to transform my data. First I tried to log transform my data, but the residual is still not normally distributed. I then square root the medv, but the residual is not normally distributed. I build a power model but the residual is still not normally distributed.

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).



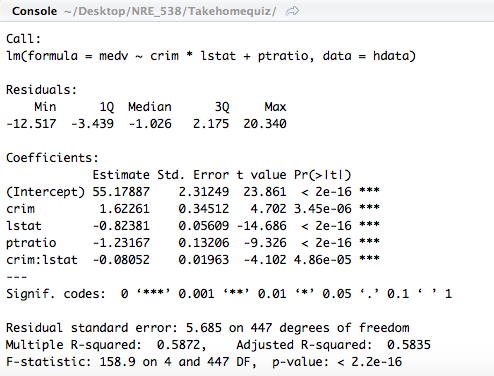
In the first row, the intercept means when crime rate, pupil-teacher ratio by town, and the low status rate are all zero, the estimate prize of house is 55.2159, and this number is significantly different from zero, since p-value is smaller than 0.05.

In the second row, the estimate is 0.29728 means when control pupil-teacher ratio and low status rate, one unit increasing of crime rate will cause the median price of house increase 0.29738 units, and this number is significant since p-value is 0.0165.

In the third row, the estimate is -1.16062, that means when we control crime rate and low status rate, one unit increasing of pupil-teacher ratio will cause median house price decrease 1.16062 units. This changing number is significant since p value is smaller than 0.05 in this row.

The fourth row shows, when we control crime rate and pupil-teacher ratio, one unit increasing of low status rate will cause the median house price decrease 0.93603 units. This changing number is significant, since p value is smaller than 0.05.

(In this question, I still run the linear regression but I don’t know how to make these variables meet assumptions of linear regression.)



The figure shows, when all independent variables (crime rate, low status rate, and pupil-teacher ratio) are zero, the median housing price should be 55.17887 units. One unit increasing of crime rate will cause median housing price increase 1.62261 units, when we control other two variables. One unit increasing of low status rate will cause median housing price decrease 0.82381 units, when control other two independent variables. One unit increasing of pupil-teacher ratio will cause the median housing price decrease 1.23167 units, when we control other two factors. And the fourth row shows, the existence of low status rate can affect the effect of crime rate on median housing price. The effect from existence of low status rate is significant. When the low status rate increases 1 unit, the slope between crime rate and median housing price decreases 0.08052 units.

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

Based on the Adjusted R square in figures in last question, we can say these two models both fit our data good, because the adjusted R square is 0.5687 in mod and 0.5835 in the interact model. These two values are both higher than 0.5, and adjusted R square of interact model is higher, so this model fits better.

But our data does not meet requirements of linear regression, so both linear regressions cannot describe data well.