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.

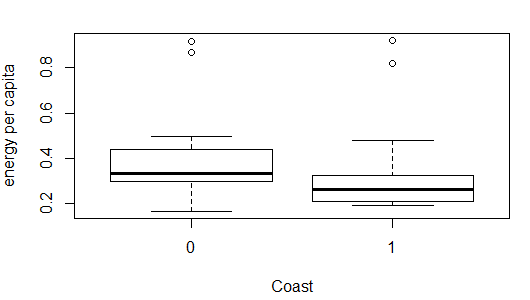
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized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

HA: Per capita energy consumption differs between coastal and inland (non-coastal) states

* 1. Please create a visual plot to answer this question (1 point).



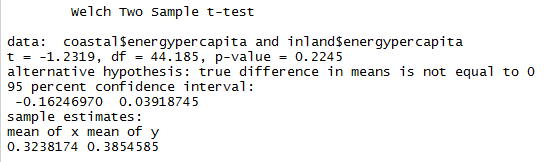
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

Test to run: two-sample two-tailed t-test

Check Assumptions

* Energy consumption per capita is continuous: yes
* Sample is randomly selected: yes (covers all the states in US)
* Independent observation: yes. Energy consumption and population in one state do not affect other ones
* Values are nearly normal/ large enough: Likely violated. 23 coastal states and 28 inland states, close to 30 but probably not big enough for CLT to apply. The data are not normally distributed (histogram, qqplot and failed shapiro test)
* Equal variance: yes (confirmed by var.test)
  1. Please run the statistical test and interpret the result (1 point).

\*The assumption of normal distributed data is probably violated if we apply N>=30 in this case (I’m not sure if the threshold can be lowered. 23 and 28 are close to 30 after all)



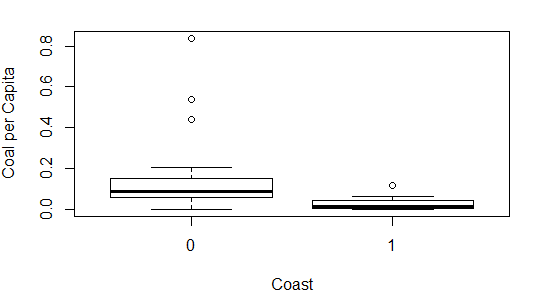
p>0.05. We fail to reject H0. There’s no significant difference in per capita energy consumption between coastal and inland (non-coastal) states.

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 does not differ between coastal and inland (non-coastal) states

HA: Per capita coal consumption differs between coastal and inland (non-coastal) states

* 1. Please create a visual plot to answer this question (1 point).



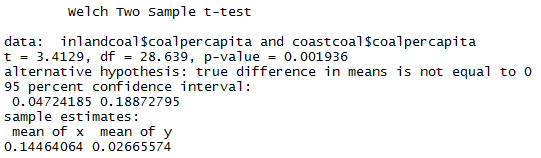
* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

Test to use: t-test (two-sample, two-tailed)

Check Assumptions

* Coal consumption per capita is continuous: yes
* Sample is randomly selected: yes (covers all the states in US)
* Independent observation: yes
* Values are nearly normal/ large enough: Likely violated. 23 coastal states and 28 inland states, close to 30 but probably not big enough for CLT to apply. The data are not normally distributed (histogram, qqplot and failed shapiro test)
* Equal variance: not met. p-value for var.test<0.05 (it’s OK as we are using Welch’s t-test, which does not require equal variance)
  1. Please run the statistical test and interpret the result (1 point).

\* The assumption of normal distributed data is probably violated if we apply N>=30 in this case (I’m not sure if the threshold can be lowered. 23 and 28 are close to 30 after all); the unequal variance is not an issue as we are using Welch’s t-test



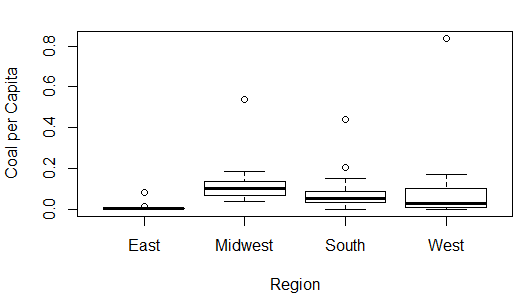
p-value<0.05 so we reject the null hypothesis. There is significant difference in per capita coal use between coastal states and inland states (inland state significantly higher from box-plot)

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 is not different among regions.

HA: per capita coal consumption is different among regions.

* 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 to use: one-way ANOVA (and Tukey HSD if there is difference)

Check assumptions:

* Independent observation: yes
* Normally distributed dependent variable: No (failed to pass shapiro test)
* Equal variance (as sample size is not equal: 9 East, 12 mid-west, 17 south, 13 west): yes (confirmed by leveneTest)
  1. Please run the statistical test and interpret the result (1 point).

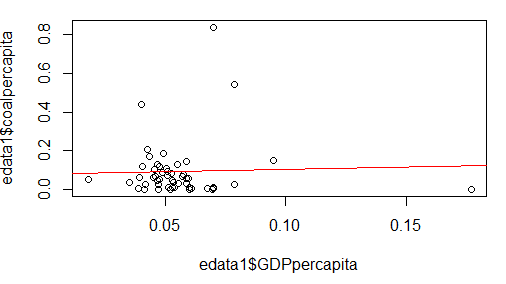
\*The dependent variable is not normally distributed



p-value>0.05 and we fail to reject H­0. There is no significant difference in per capita coal consumption across regions

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)

Per capita coal use and per capita GDP have a weak positive (close to zero) correlation, and it does not seem strong to me. The correlation coefficient is only 0.036, and when plotted against each other the line’s almost horizontal.



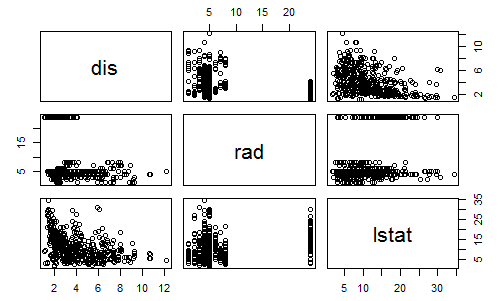
For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple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

* Covariates: DIS (weighted distance to employment centers), RAD (accessibility to radial highways), LSTAT (% lower status of population)
* Correlation value



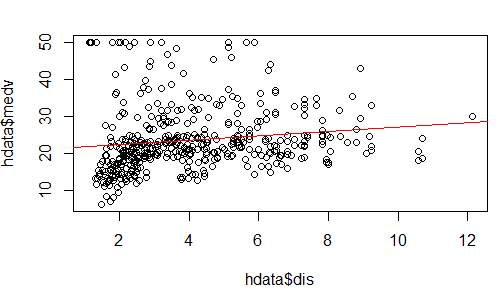
* VIFdis =1.34, VIF­rad =1.21, VIFlstat =1.26
* Plot



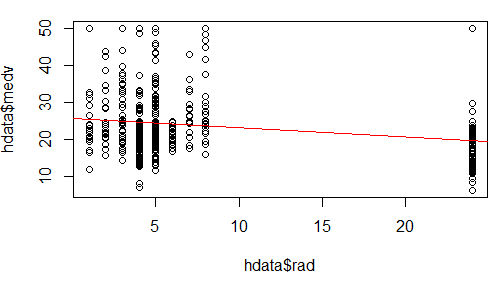
* There does not seem to be multi-collinearity issues (probably a bit between lstat and dis, but does not seem too bad)

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).

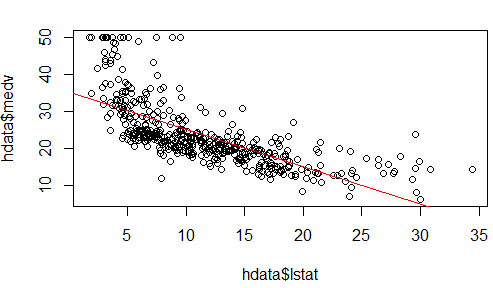
* dis and medv: a slightly positive relationship



* rad and medv: probably a slight negative association

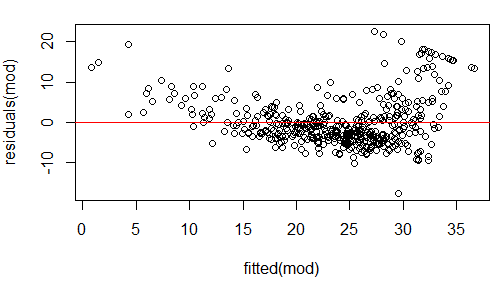


* lstat and medv: a negative relationship between them

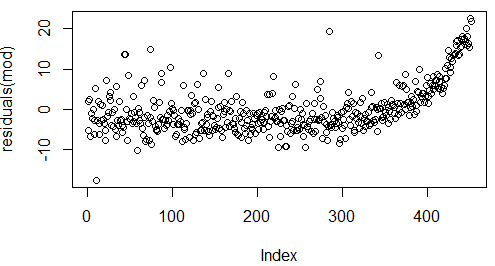


1. Run your multiple linear regression model. Check whether any assumptions are violated. Please state **which assumptions** you checked, **whether they were violated**, and **how you know** whether or not they were violated. If any assumptions are violated (e.g. normality), we will give you bonus points if you are able to identify a way to overcome this problem (3 points, plus additional 1 point bonus).

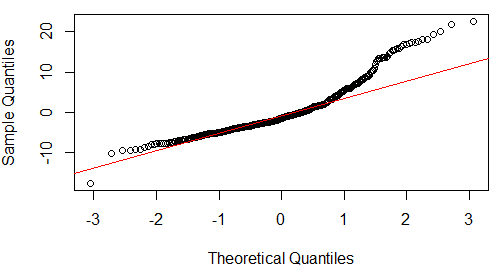
* Linear Relationship: I can’t be very sure with the DIS and RAD. There seem to be a clearer linear relationship between LSTAT and medv
* Homoscedasticity: probably violated (Plot, and did not pass bptest)



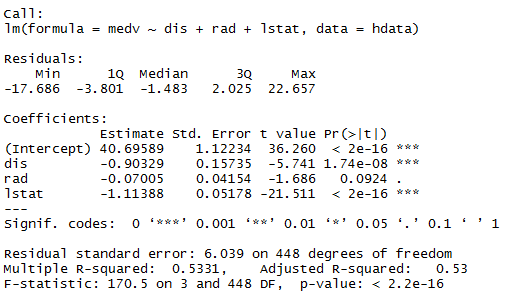
* Independent errors: probably violated. (Plot of residuals, and did not pass dwtest—p value<0.05)



* Normally distributed error: probably violated as well (qqnorm and qqline; also failed shapiro test. How come I chose those independent variables…)

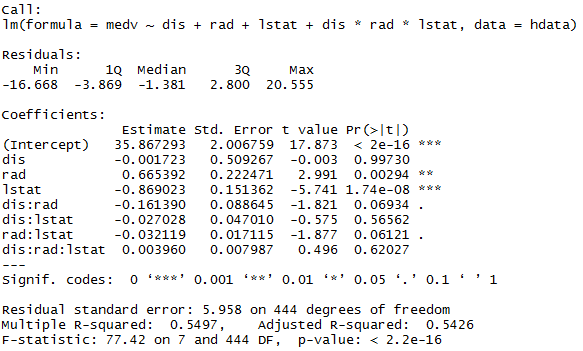


1. Interpret the results of the linear regression model. State **what the coefficient and its significance means** for the intercept and each of your three independent variables. Please explain what each regression coefficient means and do not just state that the coefficient is significant or not significant. For 1 bonus point, add in an interaction term, rerun the model, and interpret the result (3 points plus additional 1 point bonus).



* Coefficient for intercept: the housing price value (medv) when dis, rad and lstat are zero. The intercept is significantly different from 0
* Coefficient for dis (weighted distance to employment centers): with each unit of change of dis, medv decrease by 0.9 unit when controlling for rad and lstat. Medv has a significantly negative relationship with dis
* Coefficient for rad (accessibility to radial highway): there is no significant relationship between rad and medv
* Coefficient for lstat (% lower status of population): with each unit of change in lstat, medv decrease by 1.1 unit when controlling for dis and rad. Medv has a significantly negative relationship with lstat

Bonus



* Coefficient of intercept: same as above
* No significant relationship between dis and medv
* Significant positive relationship between rad and medv. With each unit of rad, medv increase by 0.67 unit when controlling for other variables
* Significant negative relationship between lstat and medv. With each unit of lstat, medv decrease by 0.87 unit when controlling for other variables
* No significant interaction between independent variables (if we go by the 0.05 threshold) although two are barely insignificant

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

From the R2 alone (0.53) I’d say it’s an OK fit, as 53% of the variance is explained by the independent variables.

But the trouble is that the model probably just violated all the assumptions for linear regression, which means I should have done some transformation or try some other model in the first place…