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?

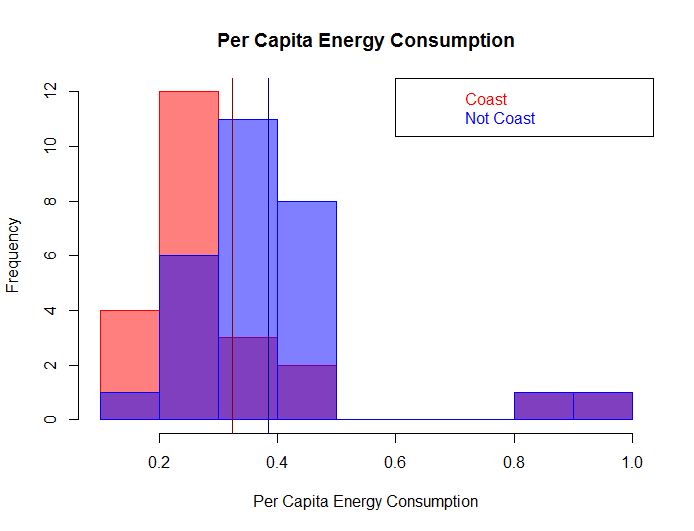
edata = mutate(edata,PerCapEnergy=TotalEnergy/Population)

* 1. Please write the null and alternate hypothesis (1 point).

H0: Per capita energy consumption is the same for states on the coast as for states that are not on the coast.

Ha: Per capita energy consumption is different for states on the coast than for states not on the coast.

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

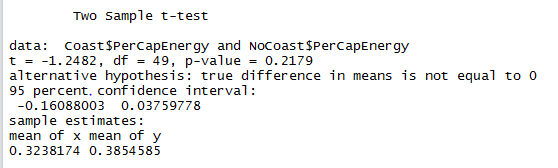


* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

2-sample two-tailed un-paired t-test

Assumptions:

* Data are continuous
  + Yes (the y values are numeric.)
* Randomly selected population
  + Yes! We have 51/51 states, so we actually have selected the WHOLE population, but this seems ok.
* Independent observations
  + Yes! Each state is independent/different, so observations do not affect each other.
* Values are normal
  + Sort of. Shapiro tests indicates that we break this, but since we have a large sample: 51>30, the CLT says it is “good enough”
* Variance is equal
  + Yes. The F test to control 2 variances gives a high p value (p=0.5) so we cannot reject the null hypothesis that variances are the same.
  1. Please run the statistical test and interpret the result (1 point).



p=0.2, so we cannot reject the null hypothesis that there is no difference between the sets. It seems like states on the coast have about the same per capita energy consumption as states that are not on the coast.

1. Does ***per capita*** coal consumption differ depending on whether a state is found on the coast or not?

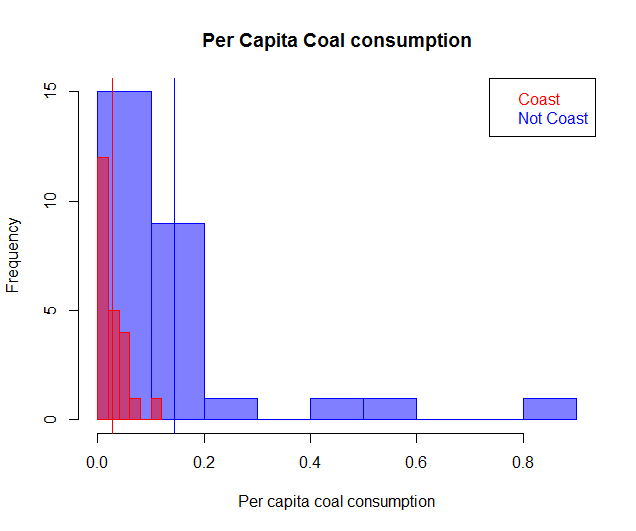
edata = mutate(edata,PerCapCoal=TotalCoal/Population)

* 1. Please write the null and alternate hypothesis (1 point).

H0: Per capita coal consumption is the same for states on the coast as for states that are not on the coast.

Ha: Per capita coal consumption is different for states on the coast than for states not on the coast.

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

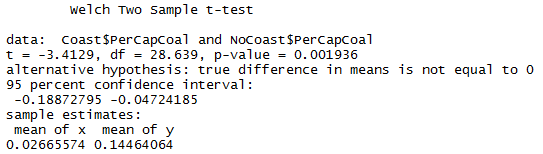


* 1. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

2-sample two-tailed un-paired ~~t-test~~ Welch’s t-test

Assumptions:

* Data are continuous
  + Yes (the y values are numeric.)
* Randomly selected population
  + Yes! We have 51/51 states, so we actually have selected the WHOLE population, but this seems ok.
* Independent observations
  + Yes! Each state is independent/different, so observations do not affect each other.
* Values are normal
  + Sort of. Shapiro tests indicates that we break this, but since we have a large sample: 51>30, the CLT says it is “good enough”
* Variance is equal
  + No. the F test we use to control variance yields p<0.05 so we reject the null hypothesis that variances are the same and know that variances are unequal. This is a problem – but when variances are unequal, we can run the Welch’s t-test.
  1. Please run the statistical test and interpret the result (1 point).



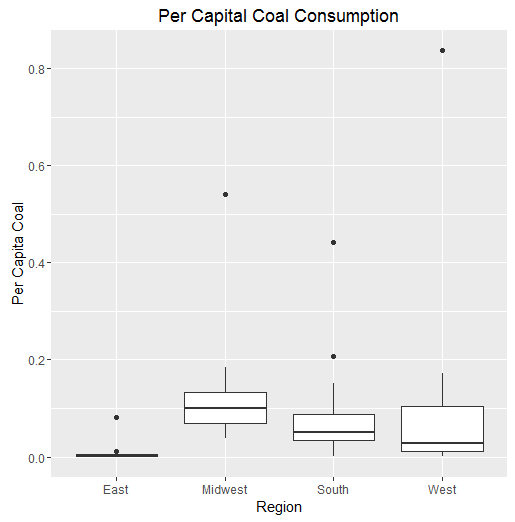
P<0.05. We reject the null hypothesis that the difference in means = 0; in other words, we can say that states on the coast have a significantly different per capita coal consumption than states not on the coast. Our sample estimates show that states not on the coast have higher per capita coal consumption.

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 the same for states in all regions.

Ha: Per capita coal consumption is different for states in different 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).

One-way ANOVA (because > two categories within only one factor of interest)

Assumptions:

* Normal distribution within population
  + Sort of. Shapiro tests indicates that we break this, but since we have a large sample: 51>30, the CLT says it is “good enough”
* Independent samples
  + Yes! Each state is independent/different, so observations do not affect each other.
* Variance is equal
  + The Levene test shows that I fail to reject my null hypothesis that the variances are equal. This lets me assume that variances are equal among my samples.
  1. Please run the statistical test and interpret the result (1 point).



There is not a significant difference between the per capita coal consumption in different regions. p>0.05 so we fail to reject the null hypothesis that the regions are all the same. All regions consume about the same amount of coal per capita.

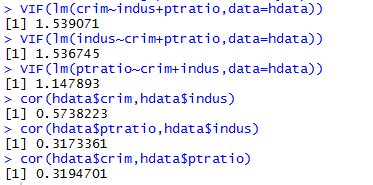
1. What is the correlation between ***per capita*** coal use and ***per capita*** GDP? Does this seem like a strong correlation to you? Why or why not? (2 points)



The correlation between per capita coal use and per capita GDP is 0.036. This indicates that as per capita coal use increases, per capita GDP is expected to increase very slightly. This does not seem like a strong correlation, because a change in x yields only a very tiny change in y. I would consider this to be an **extremely weak positive correlation** between the two variables.

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).

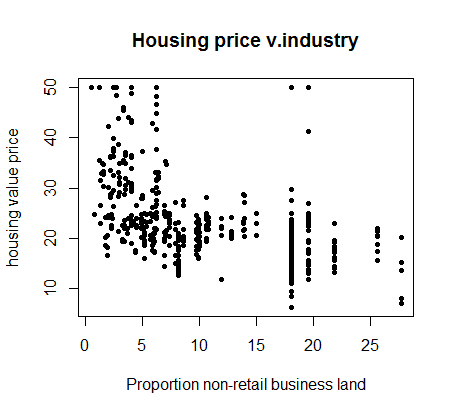


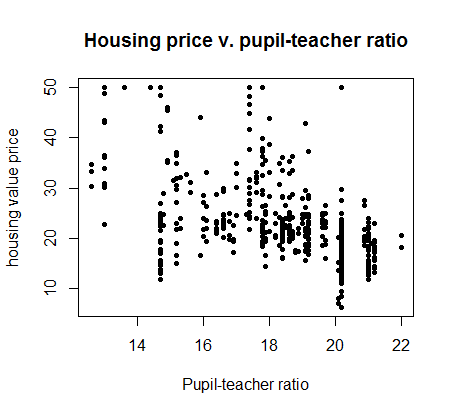
I expect that housing value prices (medv) can be predicted by the following covariates:

* Per capita crime rate (crim)
* Pupil-teacher ratio (ptratio)
* Proportion of non-retail business land (indus)

My variance inflation factors (VIF) are less than 10, so I do not think I have multi-collinearity issues. My correlation coefficients are all around 0.3 so there is significant correlation, but not full correlation.

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).





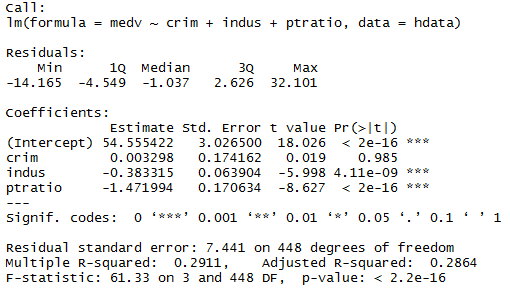
It seems like in general, higher crime rates lead to lower median housing values (although this does not always hold true and there are some outliers). Higher proportion of non-retail business land also seems to lead to lower housing values, although not very significantly. The pupil-teacher ratio does not seem to have a very strong correlation with median housing prices.

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).

Assumptions:

* There exists a linear relationship – not violated.
  + Looking at the plots made in #6, it looks like all 3 of these sets could have linear relationships.
* Data are normally distributed – violated
  + N>>30 (N=452) so CLT allows us to assume normality; BUT crim data is not normal as we can see from the qqplot (or from shapiro tests, which all have p<0.05. The data are significantly right skewed. One way to overcome this is to take the log transformation of crime data. This makes data much more normal – they are still not “normal” by shapiro test, but are better!
* No multi-collinearity – not violated
  + We showed in #5 that we do not have multi-collinearity since we have a low VIF
* Independent values (no auto-correlation) – may be violated
  + I plotted the residuals of each variable from a univariate model and there does not appear to be significant auto-correlation. However, when I ran the Durbin-Watson test on my model, I see that p<0.05 so we reject the null hypothesis that I have no autocorrelation. This is something that is hard to correct for in a linear model…
* Homoscedasticity – may be violated
  + The plots of the residuals I did to check for autocorrelation also do not seem to indicate much heteroscedasticity. There is error throughout the plots – but error does not seem to be much larger in one side than in another side. However, when I use bptest() to check for homoscedasticity, I get p<0.05 for ptratio and log(crim), but not for indus or crim. This causes me to reject the null hypothesis of homoscedasticity, concluding that my data might be heteroscedastic for ptratio. It also makes me think I should \*not\* log transform my crime data even though this made sense to improve normality.

Linear model: I tried using crim and log(crim), but my F-test showed neither was a better predictor than the other. I used crim instead of log(crim) since it is simpler to interpret.

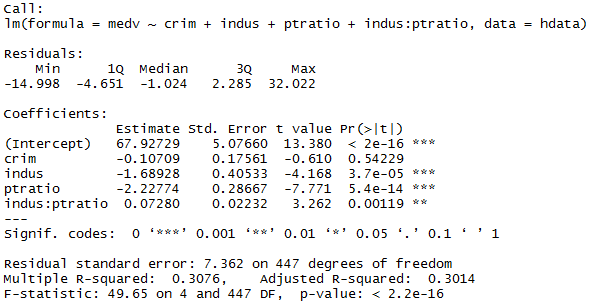


1. Interpret the results of the linear regression model. State **what the coefficient and its significance means** for the intercept and each of your three independent variables. Please explain what each regression coefficient means and do not just state that the coefficient is significant or not significant.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

medv ~ crim + indus + ptratio

The intercept indicates that at the 0 level of all variables (crime rate=0 in the town; 0 non-retail business acres per town; pupil-teacher ratio=0 in the town), we expect the median value of owner-occupied homes to be $54,555. This intercept is highly significant. We expect to see no effect of increased crime rate, since our coefficient is not significant; but the coefficient says that an increase in crime rate of 1 would yield an increase in median home value of $3. An increase in the proportion of non-retail business acres of 1 is expected to decrease median home value by $383, and this relationship is significant. Finally, a highly significant coefficient for ptratio indicates that an increase in pupil:teacher ratio of 1 will decrease home value by $1,472. The significance of coefficients can be interpreted as, for instance, there is a less than 2\*10-16 chance that the intercept is actually 0.

Additional model including interaction term:



This model shows similar trends, but shows that the effect of industry on teacher ratios is also significant. The interpretation of coefficients is the same as above, excluding the interaction term. With all factors=0, we expect the median housing value to be $67,927, etc. We see that the coefficient for industry and pt ratio go down with the addition of the interaction term, and their significance decreases. The interaction term “indus:ptratio” says that with an increase of 1 in the proportion of non-retail business acres, the effect of pupil-teacher ratio will increase by $73 (this is significant at p<0.05). Another way to understand this would be that if I increase both indus and ptratio, my median housing value decreases but not as much as I would expect from simply summing the coefficients.

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

My original model is a fairly poor fit. My R2 value indicates that only 28.64% of variance is explained by my model. However, my coefficients are highly significant, indicating that my explanatory variables are highly correlated with my variable of interest, there is just a lot of additional noise. I would expect this to be unavoidable since houses are just different and people like different styles of houses, for instance, which are variables that we did not measure. Interestingly, when I tried a model that included a log(crim) term replacing the original crim term, it performed nearly exactly the same as the model when I did not log transform my crime variable.