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.

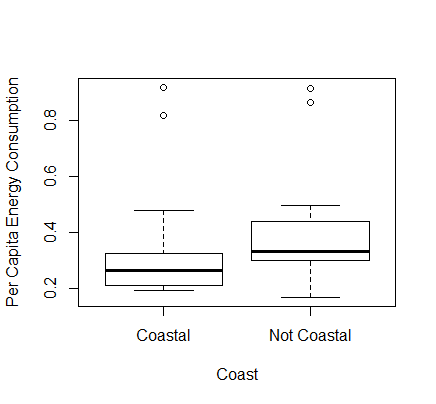
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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: Energy consumption does not differ depending on whether a state is found on a coast or not.

Alternate hypothesis: Energy consumption differs depending on whether a state is found on a coast or not.

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



The plot shows us that the mean energy of coastal and noncoastal states looks different, with the right (no coast) using more electricity per capita. However, the boxes overlap which suggests the difference in means may not be statistically significant.

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

The test to use is a two sample t-test. Testing assumptions:

* Equal Variance: Yes, variance is equal. The null hypothesis would be that the ratio of variance is equal to 1, which means they're equal. The p-value is 0.5098, so we cannot reject the null, so our assumption that variance is equal is met.

var.test(Coast[,"PerCapita"], NoCoast[,"PerCapita"])

F test to compare two variances

data: Coast[, "PerCapita"] and NoCoast[, "PerCapita"]

F = 1.3021, num df = 22, denom df = 27, p-value =

0.5098

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.5861139 2.9930256

sample estimates:

ratio of variances

1.302084

* Normal distribution: I know I must be doing something wrong, but my Coast PerCapita data is not passing the test for normal distribution. P-values are smaller than .05 (5.037e-06), so we must reject the null that the data is normally distributed.

> shapiro.test(Coast[,"PerCapita"])

Shapiro-Wilk normality test

data: Coast[, "PerCapita"]

W = 0.66467, p-value = 5.037e-06

shapiro.test(NoCoast[,"PerCapita"])

Shapiro-Wilk normality test

data: NoCoast[, "PerCapita"]

W = 0.77249, p-value = 3.627e-05

* Observations sampled independently: I am assuming that the observations were sampled independently since they are different states with different energy uses, populations, etc.
  1. Please run the statistical test and interpret the result (1 point).

t.test(Coast[,"PerCapita"], NoCoast[,"PerCapita"], paired=FALSE)

Welch Two Sample t-test

data: Coast[, "PerCapita"] and NoCoast[, "PerCapita"]

t = -1.2319, df = 44.185, p-value = 0.2245

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.16246970 0.03918745

sample estimates:

mean of x mean of y

0.3238174 0.3854585

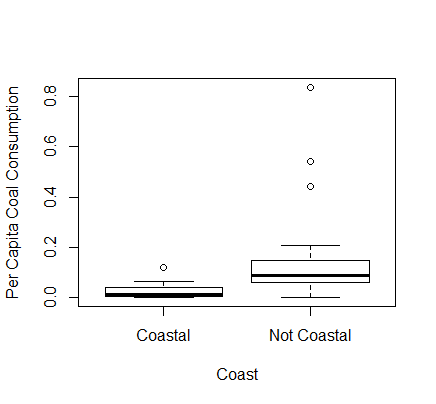
The t-test results show a p-value of 0.2245 which is higher than 0.05 and means we cannot reject the null hypothesis. So, per capita energy use in coastal states is not statistically significantly different from per capita energy use in noncoastal states.

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).

Null hypothesis: Coal consumption does not differ depending on whether a state is found on a coast or not.

Alternate hypothesis: Coal consumption differs depending on whether a state is found on a coast or not.

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



Based on the boxplot, per capita coal consumption means are different between coastal and noncoastal (left and right respectively), with coastal consuming less coal than noncoastal states. The boxes look like they don’t quite overlap, so the difference in means might be statistically significant.

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

Two sample t-test

Assumptions:

* Equal Variance: I must have done something wrong again to fail this test—based on my variance test the p-value is lower than 0.05 and so I must reject the null that variance is equal, meaning this assumption is not met.

> var.test(Coast3[,"PerCapCoal"], NoCoast3[,"PerCapCoal"])

F test to compare two variances

data: Coast3[, "PerCapCoal"] and NoCoast3[, "PerCapCoal"]

F = 0.025045, num df = 22, denom df = 27, p-value =

5.995e-13

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.01127368 0.05756970

sample estimates:

ratio of variances

0.02504508

* Normal distribution: This assumption is also not met based on my shapiro tests. P values are lower than .05 so I must reject the null that the data are normally distributed.

> shapiro.test(Coast3[,"PerCapCoal"])

Shapiro-Wilk normality test

data: Coast3[, "PerCapCoal"]

W = 0.82584, p-value = 0.001023

> shapiro.test(NoCoast3[,"PerCapCoal"])

Shapiro-Wilk normality test

data: NoCoast3[, "PerCapCoal"]

W = 0.651, p-value = 6.211e-07

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

> t.test(Coast3[,"PerCapCoal"], NoCoast3[,"PerCapCoal"], paired=FALSE)

Welch Two Sample t-test

data: Coast3[, "PerCapCoal"] and NoCoast3[, "PerCapCoal"]

t = -3.4129, df = 28.639, p-value = 0.001936

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.18872795 -0.04724185

sample estimates:

mean of x mean of y

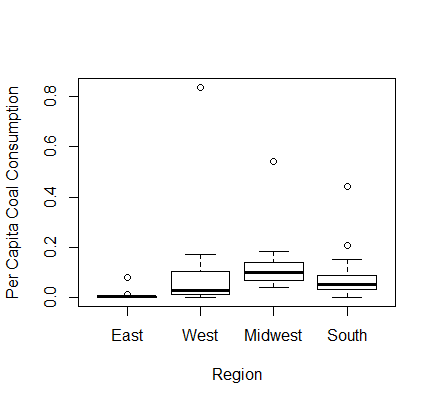
0.02665574 0.14464064

The t-test shows a p-value less than 0.05 (.001) so it suggests that there is a statistically significant difference between per capita coal consumption in coastal and noncoastal states.

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: coal consumption is not different depending on the region in which a state is found.

Alternate hypothesis: coal consumption is different depending on the region in which a state is found.

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

The boxplot shows that there may be some differences between mean per capita coal consumption by region but they may not be statistically significant because the boxes overlap.

* 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

Assumptions:

* Equal Variance: Based on the Levene test, the F value is less than the critical value, so we do not reject the null of equal variance and can assume the variance is equal.

leveneTest(PerCapCoal~Region,data=edata3)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 3 0.7635 0.5202

47

* Normal distribution: Based on my shapiro tests, none of the regions have normally distributed per capita coal consumption data. I don’t think this is a problem because the necessity to meet the assumption of normality can be relaxed.

> shapiro.test(East[,"PerCapCoal"])

Shapiro-Wilk normality test

data: East[, "PerCapCoal"]

W = 0.49468, p-value = 5.535e-06

> shapiro.test(West[,"PerCapCoal"])

Shapiro-Wilk normality test

data: West[, "PerCapCoal"]

W = 0.50477, p-value = 1.103e-05

> shapiro.test(Midwest[,"PerCapCoal"])

Shapiro-Wilk normality test

data: Midwest[, "PerCapCoal"]

W = 0.62528, p-value = 0.0001758

> shapiro.test(South[,"PerCapCoal"])

Shapiro-Wilk normality test

data: South[, "PerCapCoal"]

W = 0.67217, p-value = 5.689e-05

* Independent samples: We can assume the samples were collected independently because they are from different locations around the country.
  1. Please run the statistical test and interpret the result (1 point).

> Region.aov=aov(PerCapCoal~Region, data=edata3)

> Region.aov

Call:

aov(formula = PerCapCoal ~ Region, data = edata3)

Terms:

Region Residuals

Sum of Squares 0.0863652 0.9840767

Deg. of Freedom 3 47

Residual standard error: 0.144699

Estimated effects may be unbalanced

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)

There is a correlation coefficient of 0.036, so there is a small, positive, linear correlation between per capita coal and per capita GDP.

> cor(PerCapGDP, PerCapCoal)

[1] 0.03598182

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: crim (per capita crime rate by town), rm (average number of rooms per dwelling), ptratio (pupil-teacher ratio by town)

The variables do not seem highly correlated with each other based on their relatively low correlation coefficients: crim and rm = -0.142, crim and ptratio = 0.32, rm and ptratio = -0.334.

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

crim rm ptratio

crim 1.0000000 -0.1424577 0.3194701

rm -0.1424577 1.0000000 -0.3341642

ptratio 0.3194701 -0.3341642 1.0000000

The model does not have issues with multicollinearity based on the VIF values of 1.115 for crim, 1.128 for rm, and 1.23 for ptratio, all of which are below the general threshold of 5 to 10.

> vif.crim = 1/(1 - summary(lm(crim~rm+ptratio, data=hdata))$r.squared)

> vif.crim

[1] 1.115444

> vif.rm = 1/(1 - summary(lm(rm~crim+ptratio, data=hdata))$r.squared)

> vif.rm

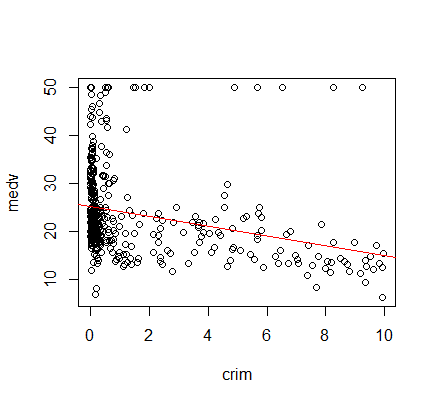
[1] 1.127504

> vif.ptratio = 1/(1 - summary(lm(ptratio~rm+crim, data=hdata))$r.squared)

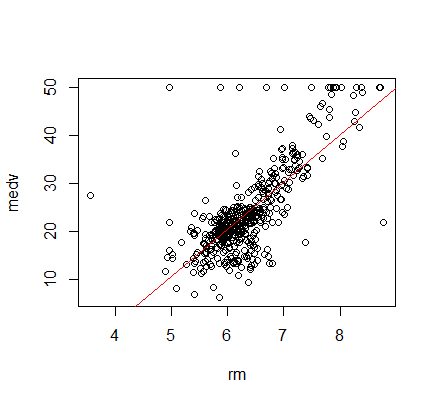
> vif.ptratio

[1] 1.230175

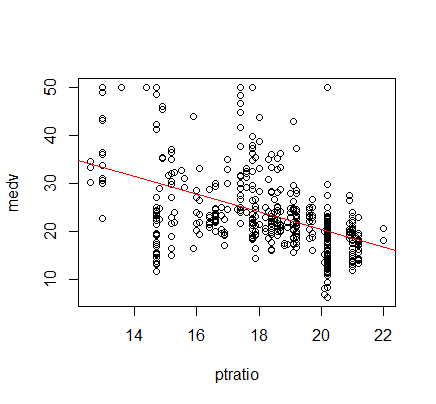
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).



Crime appears to have a negative relationship with medv, but this may not be best described as a linear relationship based on the clustering of many more data points at the lower end of the scale.



Number of rooms appears to have a relatively strong positive linear relationship with medv based on the way the data points appear to fit the line.



Pupil to teacher ratio appears to have a moderate negative relationship with medv, but this also may not be a truly linear relationship based on the pattern in the data.

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:

* Residual independency: based on the DW test with a p value of less than .05, there is autocorrelation in the model. Also the plot shows that there is a pattern present in the residuals, which suggests data are autocorrelated

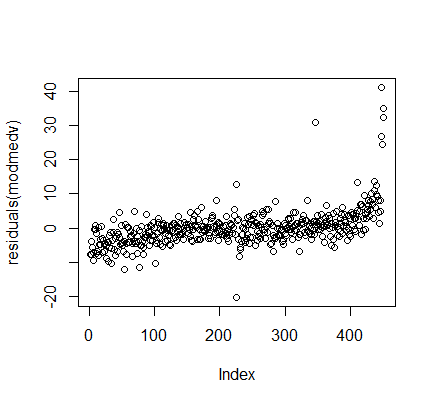
> dwtest(modmedv, alternative=c("two.sided"))

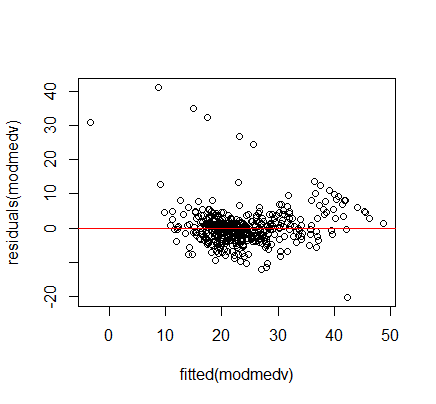
Durbin-Watson test

data: modmedv

DW = 0.90527, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is not 0



* Residual homoscedasticity: Based on the plot, the residuals are not distributed evenly and a sort of pattern does exist. The BP test tells us that, with a p-value of 1.396e-06 (less than .05), we can reject the null that the model is homoscedastic and therefore the model does not meet the assumption of homoscedasticity.

bptest(modmedv)

studentized Breusch-Pagan test

data: modmedv

BP = 29.976, df = 3, p-value = 1.396e-06

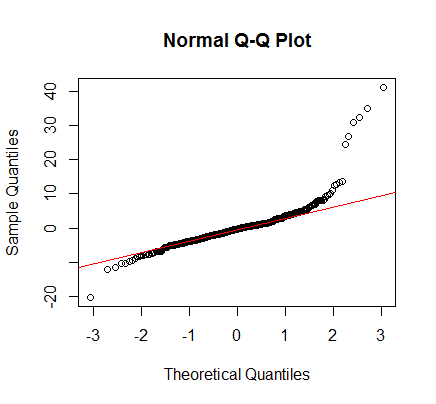
* Residual normality: Based on the qqplot, the data looks like it could be normally distributed. But the shapiro test shows us that the model does not follow normal distribution, because the p-value is 2.2e-16 and so we must reject the null of normality and say the model fails the assumption of normal distribution.

> shapiro.test(residuals(modmedv))

Shapiro-Wilk normality test

data: residuals(modmedv)

W = 0.78982, p-value < 2.2e-16



RUN THE MODEL:

> modmedv=lm(formula=medv~crim+rm+ptratio,data=hdata)

> summary(modmedv)

Call:

lm(formula = medv ~ crim + rm + ptratio, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-20.411 -2.774 -0.372 1.679 41.280

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -15.7432 4.0252 -3.911 0.000106 \*\*\*

crim -0.4541 0.1094 -4.151 3.96e-05 \*\*\*

rm 8.6468 0.4116 21.006 < 2e-16 \*\*\*

ptratio -0.8063 0.1303 -6.187 1.38e-09 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.49 on 448 degrees of freedom

Multiple R-squared: 0.6142, Adjusted R-squared: 0.6116

F-statistic: 237.7 on 3 and 448 DF, p-value: < 2.2e-16

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).

The regression coefficient for the intercept, -15.7432, is significant with its p-value of 0.0001. This means that when all the variables are set to zero, the housing value is -15.7432. This suggests there have to be a combination of variables involved in order to have an actual, positive housing price. The coefficient for crime is -0.4541, and is significant with its p-value of 3.96e-05. This means that with each unit increase in crime, housing prices decrease by 0.454. The coefficient for room number is 8.6468 and is significant with its p-value smaller than 2e-16. This means that with each unit increase in room number, housing prices increase by 8.65. The coefficient for pupil to teacher ratio is -0.806, and is significant with its p-value of 1.38e-09. This means that with each unit increase in pupil to teacher ratio, housing prices decrease by 0.806.

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

I don’t think the model is the best fit because the tests used to check the assumptions suggest that a linear model is not the best fit for the data. Plots for those tests as well as the plots of each variable in question 5 show that there are other patterns present and other types of relationship that likely explain the data better than a linear model.