Please use the R script provided to l oad 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.

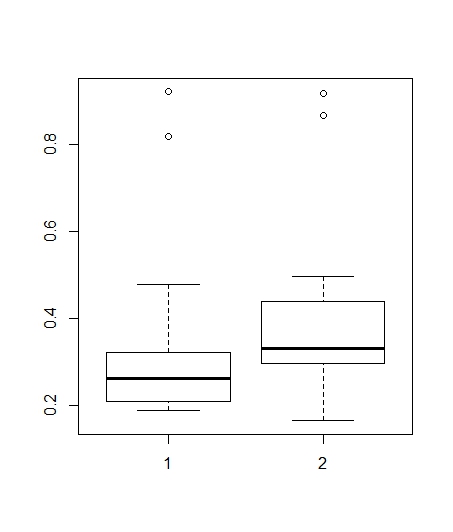
1. Does ***per capita*** energy consumption differ depending on whether a state is found on the coast or not?
   1. Please write the null and alternate hypothesis (1 point).

Null- Energy consumption does not differ between states on the coast and not on the coast

Alternate- Energy consumption does differ between coast and not-coast cities

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

Here is a boxplot- we can see there are outliers. I’m going to try to remove those to run the test… we’ll see how that goes.



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

I think I need to run a t-test. The data looks pretty normal. The variances appear to be unequal.

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

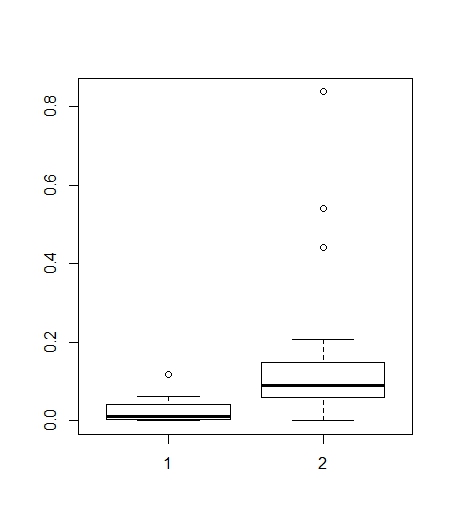
The p-value isn’t significant at the .05 level (it’s 0.2), the means of both the coastal data and the inland data are almost equal (.32, .38). While the null hypothesis is that the difference in means should be zero, and here it’s technically not, the high p value makes me think this test is not significant. I would say there is no statistical difference between energy per capita on the coast vs inland.

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).

Alternative=Per capita coal consumption is higher when a state is not on the coast than when it is on the coast.

Null= There is no difference between per capita coal consumption in coastal and inland 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).

Data seems pretty normal, the variances are equal according to the f test.

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

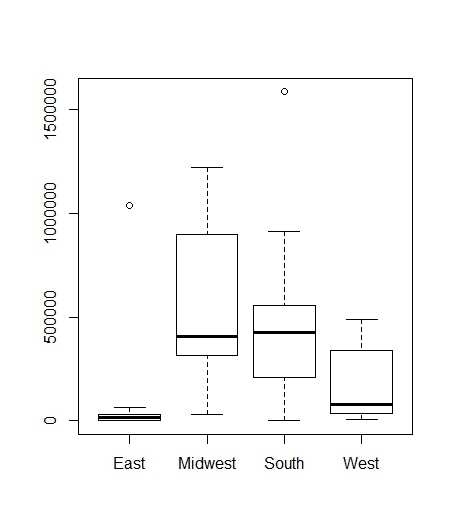
The means are very different (.02 and .14), and the p value is significant (0.001). Looks like they do use more coal in inland, not coastal states.

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= no difference in coal consumption based on region

Alternate= I’m going to say, as a guess, the Midwest uses more coal than any other 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).

One continuous variable and multiple categorical variables probably means I’ll be doing a one way anova, and a tukey post hoc test to find which relationships are significant.

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

The one way anova was significant, with an F value of 5 and a p value of 0.004. TukeyHSD showed significant relationships between: well, only the West and the Midwest (0.01). A close second is Midwest and East, with a p value of 0.02. Not significant, but it’s very close. So, it looks like we can reject the Null, and say that the Midwest does use more coal than other regions of the United States.

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 cor.test spit out the number 0.035. It also gave me a low t stat, a very high p value (0.8!), and I think the confidence interval falling close to zero on both ends also indicates less significant results. Knowing that, we can say any correlation probably isn’t accurate, but 0.035 is a small number in the first place.

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).

Here is the output for correlation, I apologize in advance for how large and not attractive this is (it continues for several pages):

crim zn indus chas

crim 1.0000000 -0.28123893 0.5738223 0.050065901

zn -0.2812389 1.00000000 -0.5142733 -0.059696546

indus 0.5738223 -0.51427332 1.0000000 0.103016306

chas 0.0500659 -0.05969655 0.1030163 1.000000000

nox 0.6369411 -0.50072854 0.7385173 0.134064175

rm -0.1424577 0.30662341 -0.3648947 0.076579589

age 0.4476638 -0.55567886 0.6059816 0.123140665

dis -0.4619674 0.65615208 -0.6693615 -0.140662124

rad 0.8979883 -0.26660593 0.5133060 0.057336696

tax 0.8256677 -0.26944592 0.6733123 0.016894085

ptratio 0.3194701 -0.36412219 0.3173361 -0.099622729

b -0.4130144 0.15037968 -0.3167522 0.013415202

lstat 0.4247886 -0.41116521 0.5654025 -0.009429928

medv -0.2862450 0.33156988 -0.4119145 0.154408725

nox rm age dis

crim 0.6369411 -0.14245773 0.4476638 -0.4619674

zn -0.5007285 0.30662341 -0.5556789 0.6561521

indus 0.7385173 -0.36489467 0.6059816 -0.6693615

chas 0.1340642 0.07657959 0.1231407 -0.1406621

nox 1.0000000 -0.26459441 0.7071478 -0.7458124

rm -0.2645944 1.00000000 -0.1878709 0.1387741

age 0.7071478 -0.18787087 1.0000000 -0.7203343

dis -0.7458124 0.13877413 -0.7203343 1.0000000

rad 0.5424992 -0.09593148 0.3593263 -0.3883852

tax 0.6151198 -0.21494783 0.4270947 -0.4441307

ptratio 0.1034642 -0.33416415 0.1929558 -0.1522533

b -0.3584331 0.10835237 -0.2237652 0.2344494

lstat 0.5368242 -0.60728892 0.5732663 -0.4237247

medv -0.3327782 0.74018080 -0.2998932 0.1387984

rad tax ptratio b

crim 0.89798831 0.82566775 0.31947010 -0.41301437

zn -0.26660593 -0.26944592 -0.36412219 0.15037968

indus 0.51330603 0.67331226 0.31733613 -0.31675222

chas 0.05733670 0.01689409 -0.09962273 0.01341520

nox 0.54249915 0.61511982 0.10346419 -0.35843313

rm -0.09593148 -0.21494783 -0.33416415 0.10835237

age 0.35932632 0.42709472 0.19295579 -0.22376517

dis -0.38838521 -0.44413069 -0.15225328 0.23444940

rad 1.00000000 0.87287642 0.38748427 -0.35258569

tax 0.87287642 1.00000000 0.38451066 -0.36708374

ptratio 0.38748427 0.38451066 1.00000000 -0.08960928

b -0.35258569 -0.36708374 -0.08960928 1.00000000

lstat 0.30978792 0.41092675 0.30304309 -0.29109383

medv -0.21790210 -0.34589757 -0.46121356 0.26479723v

lstat medv

crim 0.424788624 -0.2862450

zn -0.411165213 0.3315699

indus 0.565402489 -0.4119145

chas -0.009429928 0.1544087

nox 0.536824184 -0.3327782

rm -0.607288917 0.7401808

age 0.573266278 -0.2998932

dis -0.423724689 0.1387984

rad 0.309787921 -0.2179021

tax 0.410926755 -0.3458976

ptratio 0.303043086 -0.4612136

b -0.291093833 0.2647972

lstat 1.000000000 -0.7062551

medv -0.706255059 1.0000000

What I’ve highlighted are the variables I intend to throw out, because they’re highly correlated. I think I’ll pick the first three, “crim”, “nox”, and “rm” as my variables.

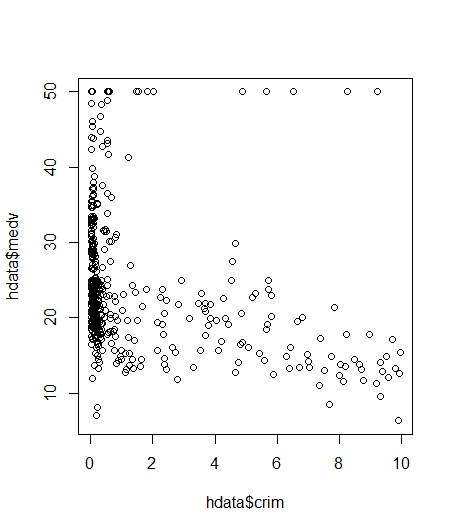
VIF values:

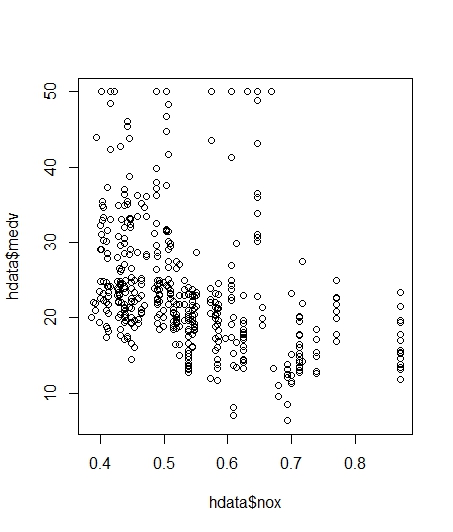
crim nox rm

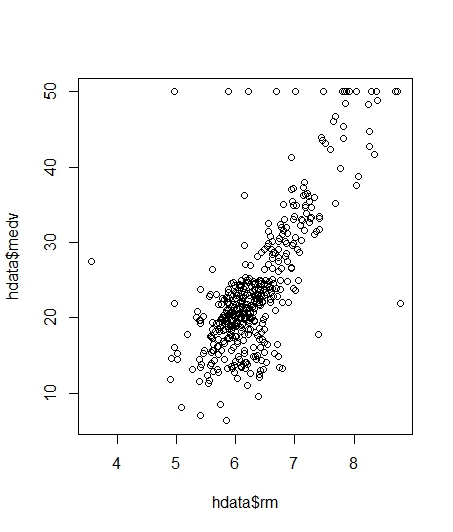
1.684707 1.774769 1.076605

These look good to me

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).

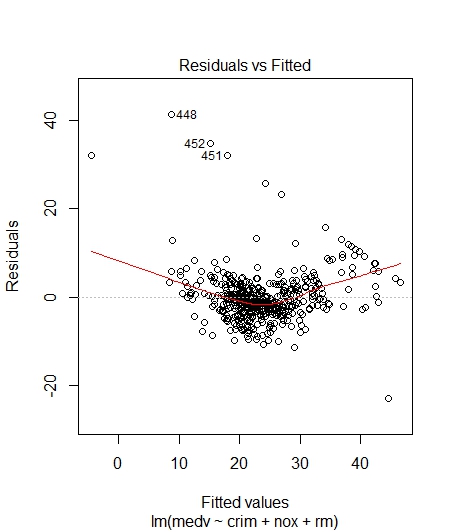
I would imagine housing values go down as crime increases. You can see this a little in this graph in the dark, black blob of points on the left side of this graph.

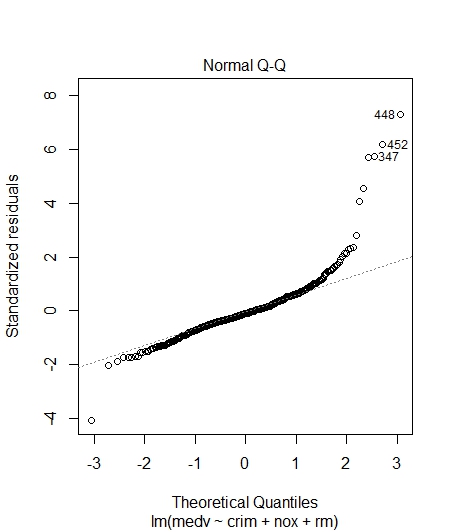
The amount of nitrous oxides compared to housing prices. It looks like there is a slight relationship between them, that the housing prices are slightly higher with less nitrous oxides. I wonder how that’s quantified. Do real estate agents tell their customers this?

This graph shows that as the number of rooms increases, housing prices go up. Again, that seems to make sense.

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).

I made some plots to check my regression and included them in this answer:

 So, the fact that this is a u shape makes me think something is not quite right here. The residuals are drifting pretty far from zero on the trendline, which leads me to think the data may not quite follow a linear pattern. The points themselves seem clustered pretty close to zero evenly across the board. It’s not perfect, but I would guess this is basically homoscedastic. Again, though, with the u shape it may be a little off in reality, but I think that may be ok since this isn’t a textbook, perfect data set.

 On the bright side: this looks pretty normal. The outliers might be a problem but the rest of the data looks good to me.

I’m not sure how to check for independence, so I’m just going to put my faith in the researchers that they collected data properly.

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).

Here’s my output:

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

Residuals:

Min 1Q Median 3Q Max

-22.789 -2.684 -0.550 2.094 41.292

Coefficients:

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

(Intercept) -32.6388 3.4136 -9.561 < 2e-16 \*\*\*

crim -0.5504 0.1398 -3.936 9.62e-05 \*\*\*

nox -3.6182 3.1476 -1.150 0.251

rm 9.3210 0.4184 22.275 < 2e-16 \*\*\*

---

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

Residual standard error: 5.711 on 448 degrees of freedom

Multiple R-squared: 0.5825, Adjusted R-squared: 0.5797

F-statistic: 208.3 on 3 and 448 DF, p-value: < 2.2e-16

So my intercept coefficient is -32.6388, and it’s the y intercept on this regression model. If all of my chosen independent variables were equal to zero, this would be the value of my house. That’s a really cheap house. They would pay YOU for the house. The coefficient for variable 1, criminal activity, is -0.5504, and it is significant. So, for every increase in criminal activity by a value of one the value of the housing goes down by 0.55 thousand dollars. The second coefficient, nitrous oxide, is technically -3.6182. So, for every increase in nitrous oxide by a value of 1 ppm, the value of the house should go down by 3.61, HOWEVER, this relationship is not significant at the 0.05 level, and based on that, nitrous oxide probably doesn’t have an effect on housing value. Finally, the last coefficient explains that every time you add a room to a house (increase rm by a value of 1), the value of the house increases by 9.3210, (thousand dollars!) and this is significant. This is also why I will likely only ever live in a one room cardboard box. I can’t afford an extra $9,000 for more rooms! These changes I am listing, the fact that they increase or decrease by a value of 1, only hold true if the other two variables remain constant.

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

Well, I saw from checking the assumptions that this model may not be linear, or very linear, so right out of the gate the answer to this is this model is a bad fit. Linear models have to have linear relationships in order to work. However, even if I ignore that, the r square and adjusted r square are very low, indicating that this model only accounts for about half the variance in this data. To me, that means this probably wasn’t the best model, even if I did find significant relationships.