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.

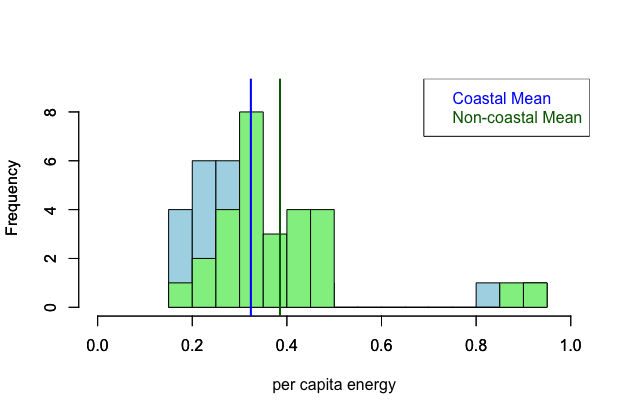
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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: There is no difference in per capita energy consumption between coastal states and non-coastal states.

Alternative: Per capita energy consumption is greater in coastal states than in 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).

T-test. Assumptions: do the samples have equal variances? No. Are the data normally distributed? No.

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

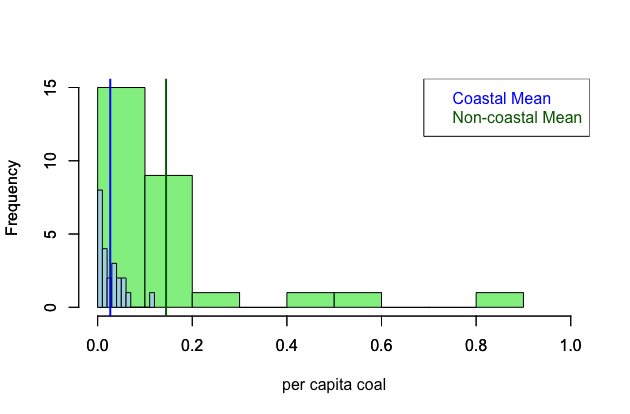
There is no significant difference between the sample means.

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: There is no difference in per capita coal consumption between coastal states and non-coastal states.

Alternative: Per capita coal consumption is greater in non-coastal states than in 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).

T-test. Assumptions: do the samples have equal variances? No. Are the data normally distributed? No.

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

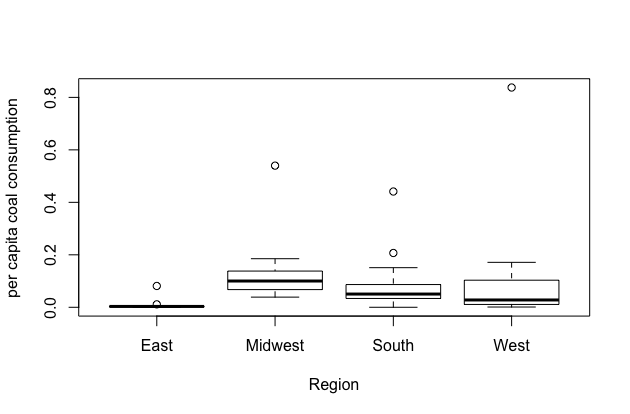
Per capita consumption of coal is significantly greater in non-coastal states than coastal states.

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: Per capita coal consumption does not differ between regions.

Alternative: Per capita

* 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).

ANOVA. Assumptions: are the variances equal between the groups? No.

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

There is no significant difference between the region means.

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 no significant correlation between per capita coal use and per capita GDP according to my Pearson’s test.

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).

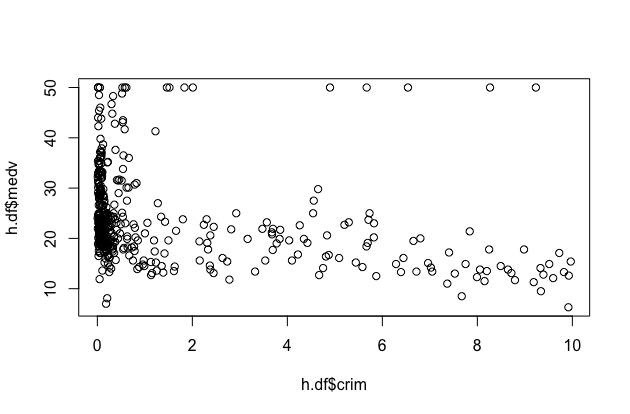
Covariates: rm, crim, b

Correlation values:

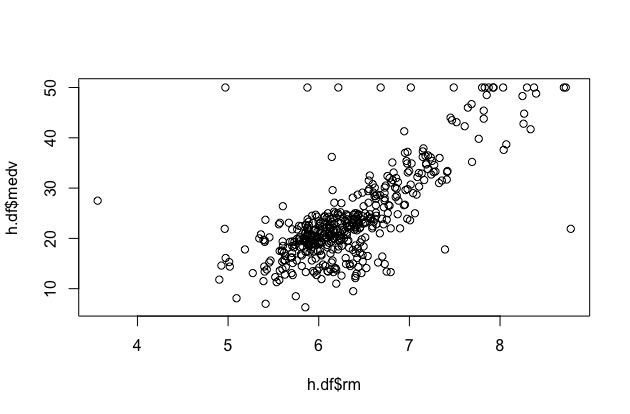
|  |  |  |  |
| --- | --- | --- | --- |
|  | crim | rm | b |
| crim |  | -0.14 | -0.41 |
| rm | -0.14 |  | 0.11 |
| b | -0.41 | 0.74 |  |

rm and b appear to be fairly highly correlated, but VIF < 5 (2.48).

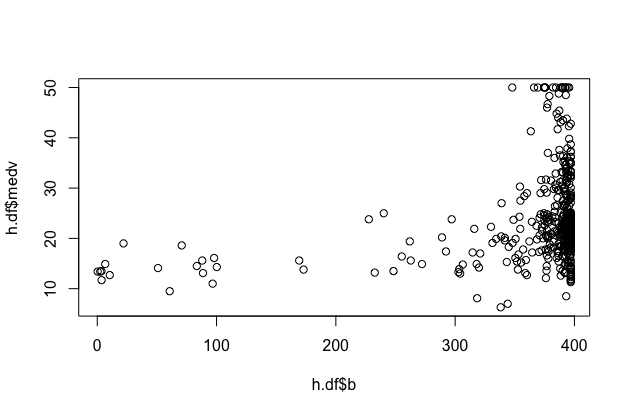
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).



Lower crime rates appear to correlated with higher median values.



Higher numbers of rooms appear to be correlated with higher median values.



Higher proportions of black residents appear to be correlated with higher median values.

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).

Assumption: residuals are homoscedastic and normally distributed. By plotting the residuals, they appear to be heteroscedastic, violating this assumption. From extracting the residuals and performing a Shapiro test, they are not normally distributed. I will perform a Box-Cox transformation on the dependent variable to attempt to correct for these violations.

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 Box-Cox transformation did not appear to help, so I’m using the original model. The intercept itself (-41.24) is difficult to interpret as a real value because it is being fit to a hypothetical house that has no rooms. The rm coefficient (9.34) shows a significant positive effect on home values. The crim coefficient (-0.46) shows a negative effect on home values when the number of rooms is held constant. The b coefficient shows a significant positive effect on home values when the other variables are held constant.

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

The adjusted R squared value is 0.5935. This fit seems to be reasonable for these types of socioeconomic variables, which tend to be quite skewed.