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).

The per capita energy consumption in coast state is not significantly different with the state not on the coast. The alternative hypothesis will be *per capita* energy consumption will differ depending on whether a state is found on the coast or not.

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

I will use two sample T-test.

**Assumptions:**

data should be continuous; (YES)

we assume our sample is randomly selected form the population;

observations between coastal and non-coastal region are independent (YES)

samples have equal variance(YES)

samples are normally distributed (According to Shapiro test they are not normal distributed, but according to qqplot, most of the value are on the qqline or close to the line. And my sample size>50 so I will accept the assumption of normality )

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

The p-value = 0.2179, which>0.05, so we cannot reject the null hypothesis, which means ***per capita*** energy consumption DO NOT significantly differ depending on whether a state is found on the coast or not.

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).

The per capita coal consumption in coast state is not significantly different with the state not on the coast. The alternative will be *per capita* coal consumption will differ depending on whether a state is found on the coast or not.

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

I will use two sample Welch T-test.

**Assumptions:**

data should be continuous; (YES)

we assume our sample is randomly selected form the population;

observations between coastal and non-coastal region are independent (YES)

samples have equal variance (NO)

samples are normally distributed (According to Shapiro test they are not normal distributed, but according to qqplot, most of the value are on the qqline or close to the line, and the sample size>30, so I will accept the assumption of normality)

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

The p-value = 0.001936, which<0.05, so we can reject the null hypothesis. Which means ***per capita*** coal consumption is significantly differ depending on whether a state is found on the coast or not.

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).

The per capita coal consumption in coast state is not significantly different among the state in different region. The alternative hypothesis will be *per capita* coal consumption will significantly differ depending on the region in which a state is found.

* 1. Please create a visual plot to answer this question (1 point).
  2. Please decide what statistical test to use and check whether your data meet the assumptions to run this test (1 point).

I wii use non-parametric one way anova test—Kruskal-Wallis test

**Assumptions:**

normally distributed; (NO)

samples need to be independent; (YES)

each population have same variance (YES, according to Levene's Test for Homogeneity of Variance )

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

The p-value = 0.00042, which<0.05, so we can reject the null hypothesis, which means ***per capita*** coal consumption is significantly differ among the state in different region.

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 coefficient=0.03598182 which close to 0, indicate the correlation between ***per capita*** coal use and ***per capita*** GDP is not significant.

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 will use this three in my model:

RM average number of rooms per dwelling

NOX nitric oxides concentration (parts per 10 million)

CRIM per captia clim rates

correlation values———————————————————

### rm nox crim

###rm 1.0000000 -0.2645944 -0.1424577

###nox -0.2645944 1.0000000 0.6369411

###crim -0.1424577 0.6369411 1.0000000

VIF value=2.395009, which <5

According to these result, there is no significant multi-collinearity in this mod.

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).

The relationship between rm and medv indicate a positive relation, when average number of rooms per dwelling increase, the house value will also increase.



The relationship between nox and medv indicate a negative relation, when nitric oxides concentration increase, the house value will decrease.



The relationship between crim and medv indicate a negative relation, when per capita crime rate by town increase, the house value will decrease.

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).

residual independency (NO)

Durbin-Watson statistic to detect the existence of autocorrelation, p-value < 2.2e-16, the residuals show a clear temporal autocorrelation

residual homoscedasticity (NO)

BPtest for homoscedasticity, p-value = 1.969e-06, the results show that the residuals are not homoscedastic.

residual normality (NO)

shapirotest, p-value < 2.2e-16, residuals are not normally distributed.

I tried log and squire transformation for the dependent variables, it still violates my assumption. Then I try to use boxcox to transform my dependent variable, still violate my assumption (mod.trans=update(mod,.~.+boxCoxVariable(medv)) . So before I find out a better solution, I will keep use the original model.

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).

Call:

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

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

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

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

nox -3.6182 3.1476 -1.150 0.251

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

When the number of room in the house increase per unit, the value of house will increase 9.321 unit. Their regression relation is significant;

The coefficient of nox on house value is insignificant;

When the crime rate increase per unit, the house value will decrease 0.5504 unit. Their regression relation is significant.

Bonus:

Call:

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

Coefficients:

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

(Intercept) -34.9754 3.5588 -9.828 <2e-16 \*\*\*

rm 9.3646 0.4171 22.452 <2e-16 \*\*\*

nox 0.2983 3.5985 0.083 0.9340

crim 1.7642 1.0544 1.673 0.0950 .

nox:crim -3.5268 1.5925 -2.215 0.0273 \*

But the interpretation as follows - The interaction between “nox" and “crim" is significant. The association between “nox" and “medv" depends on the effect of “crim”. From the above interaction, we can see that the effect of “nox" on “medv" decreases by 3.5 for every unit increase in crim. There is still significant regression relation between rm and “medv”.

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

Model 1: medv ~ rm

Model 2: medv ~ rm + nox

Model 3: medv ~ rm + nox + crim

Res.Df RSS Df Sum of Sq F Pr(>F)

1 450 15822

2 449 15116 1 705.53 21.633 4.352e-06 \*\*\*

3 448 14611 1 505.17 15.489 9.616e-05 \*\*\*

Model 3 which is my original model have the lowest RSS in F-test, which indicate it perform better than another two model. The Adjusted R-squared value for model is 0.5797 , also indicate nearly 58% variability can be explained by the linear fit model, which is good.