If you run any ANOVAs, you can use the Levene test for equality of variances. If your data violate an assumption about normality and a normal distribution is required for your analyses, you can get bonus points for transforming your data. Otherwise please run the statistical test anyway as if your data were normally distributed but make it clear that you violated this assumption in your answer.

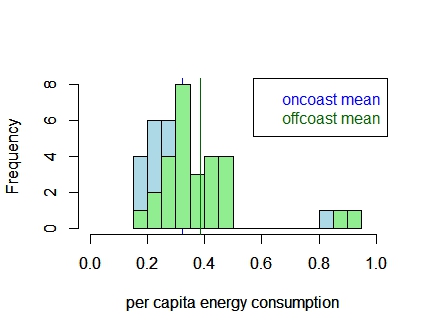
Please use the R script provided to load data and build your script from there.

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: Per capita energy consumption has no difference on whether a state is found on the coast or not.

Alternate hypothesis: Per capita energy consumption has difference 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).

Two samples two tailed t-test is used.

Assumptions checking:

Equal variances

var.test(oncoast[,'percapenergy'],offcoast[,'percapenergy'])

As p-value = 0.5098 which means the result is not significant

We failed to reject null hypothesis and the variances are equal, that follows the assumption

Normality test

shapiro.test(edata$percapenergy)

Since the p-values is 8.808e-08 which is significant, the null hypothesis is rejected

The samples are not normally distributed, that violates the assumption

**Bonus treatment of violation**

from the histogram in Q1b, we found that outliers exit after 0.8 per capita energy consumption.

edata$percapenergy[edata$percapenergy>0.8]=NA

shapiro.test(edata$percapenergy)

Now the p-value is 0.07215 which becomes insignificant, so we failed to reject null hypothesis

The samples are normally distributed now which follows the assumption

Independent sampling

Independent sampling cannot be tested here, we assume each observation is sampled independently

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

t.test(oncoast[,'percapenergy'],offcoast[,'percapenergy'],paired = FALSE)

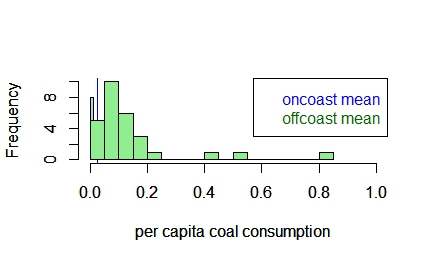
From the t-test, p-value=0.2245 which is larger than 0.05. So the result is not significant and we failed to reject the null hypothesis.

Thus, per capita energy consumption has no significant difference 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).

Null hypothesis: Per capita coal consumption has no difference on whether a state is found on the coast or not.

Alternate hypothesis: Per capita coal consumption has difference 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).

Two samples two tailed t-test is used.

Assumptions checking:

Equal variances

var.test(oncoast[,'percapcoal'],offcoast[,'percapcoal'])

As p-value = 5.995e-13 which means the result is signficant

We reject the null hypothesis and the variances are not equal, that violates the assumption

Normality test

shapiro.test(edata$percapcoal)

Since the p-values is 6.024e-11 which is significant, the null hypothesis is rejected

The samples are not normally distributed, that violates the assumption

Independent sampling

Independent sampling cannot be tested here, we assume each observation is sampled independently

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

Since the variances are not equal, we do the welch t-test

t.test(oncoast[,'percapcoal'],offcoast[,'percapcoal'],paired = FALSE)

\* violated assumptions as the variances are not equal, and the data is not normally distributed.

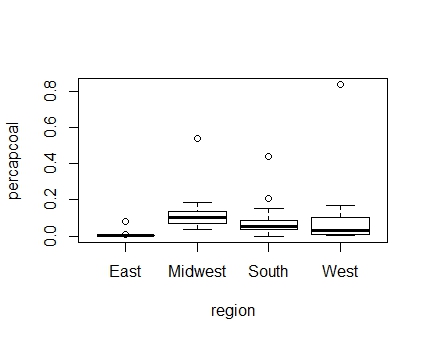
From the t-test, p-value=0.001936 which is less than 0.05. So the result is significant and we reject the null hypothesis.

Thus, per capita coal consumption has significant difference 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).

Null hypothesis: Per capita coal consumption has no difference on the region in which a state is found.

Alternate hypothesis: Per capita coal consumption has difference 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).

One-way anova is used.

Assumptions checking:

Equal variances

library(car)

leveneTest(percapcoal~Region,data=edata)

p-value is 0.5202 which is not significant which means variances are equal, that follows the assumption

Normality test

qqnorm(edata$percapcoal);qqline(edata$percapcoal,col="Red")

From the QQ-plot test, the sample deviates from the qqline and we cannot assume the samples are normally distributed, that violates the assumption

Independent sampling

Independent sampling cannot be tested here, we assume each observation is sampled independently

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

regioncoal=aov(percapcoal~Region, data=edata)

summary(regioncoal)

\*violated assumptions as the data is not normally distributed.

From the one way anova result, the p-value is 0.262 which is not significant.

We failed to reject the null hypothesis, the means of per capita coal have no significant difference in the 4 regions.

Thus, per capita coal consumption has no difference on the region in which a state is found.

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)

edata$percapgdp<-edata$TotalGDP/edata$Population

cor(edata[,c("percapcoal","percapgdp")])

Per capita coal use and per capita GDP is positively correlated. From the test above, the correlation coefficient is 0.0398 which means that when per capita coal use increase by 1unit, the probability for per capita GDP to increase 1unit is 3.98%. Thus, this is not a strong correlation due to the low correlation coefficient.

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).

Correlation values:

medv crim indus age

medv 1.0000000 -0.2862450 -0.4119145 -0.2998932

crim -0.2862450 1.0000000 0.5738223 0.4476638

indus -0.4119145 0.5738223 1.0000000 0.6059816

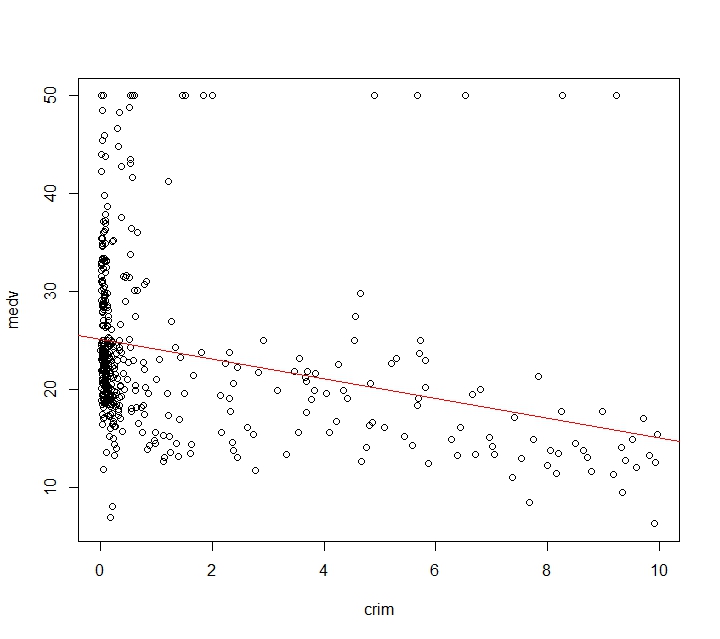
age -0.2998932 0.4476638 0.6059816 1.0000000

VIF= 1.214 which is small and it means the degree of collinearity among these

chosen covariates is small. And thus, the model is applicable for evaluation.

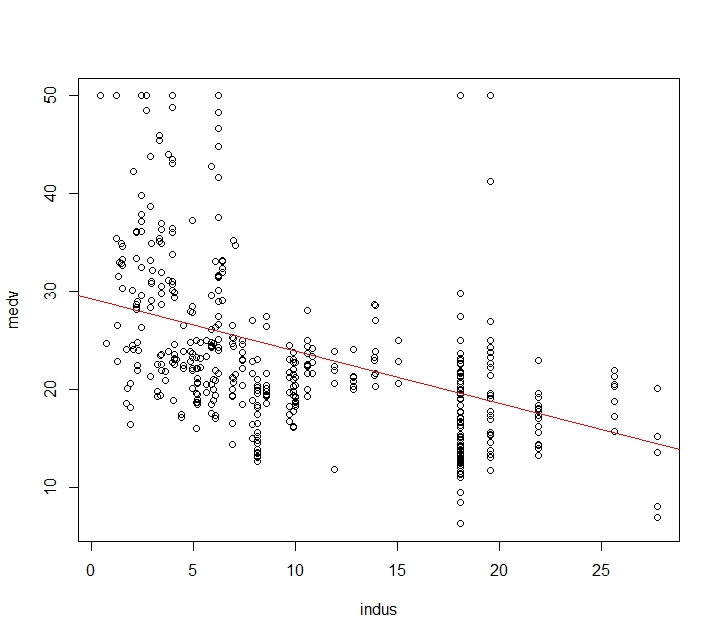
1. Plot the relationship between each of your three independent variables and the dependent variable (medv). **Include each plot in this answer and state whether and how you think each variable is related to median housing prices** (medv; 3 points).

plot(medv~crim,data=hdata)

abline(lm(medv~crim,data=hdata),col="red")

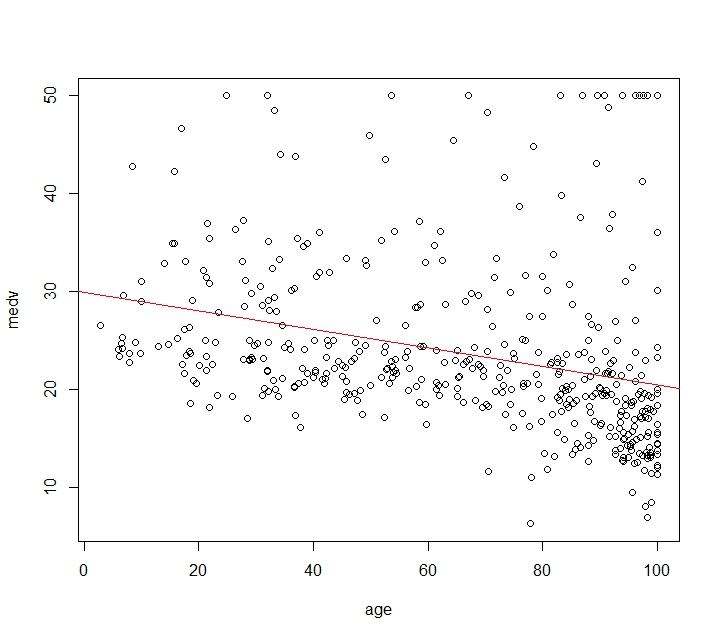
From just eyeballing the plot, per capita crime rate is negatively correlated to the housing value. The effect is quite significant as being represented by the slope of the trendline above.

plot(medv~indus,data=hdata)

abline(lm(medv~indus,data=hdata),col="red")

From just eyeballing the plot, proportion of non-retail business acres per town is negatively correlated to the housing value. The effect is significant as being represented by the steep slope of the trendline above.

plot(medv~age,data=hdata)

abline(lm(medv~age,data=hdata),col="red")

From just eyeballing the plot, proportion of owner-occupied units built prior to 1940 is negatively correlated to the housing value. The effect is not so significant as being represented by the comparatively flat slope of the trendline above.

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).

Run the multiple linear regression model:

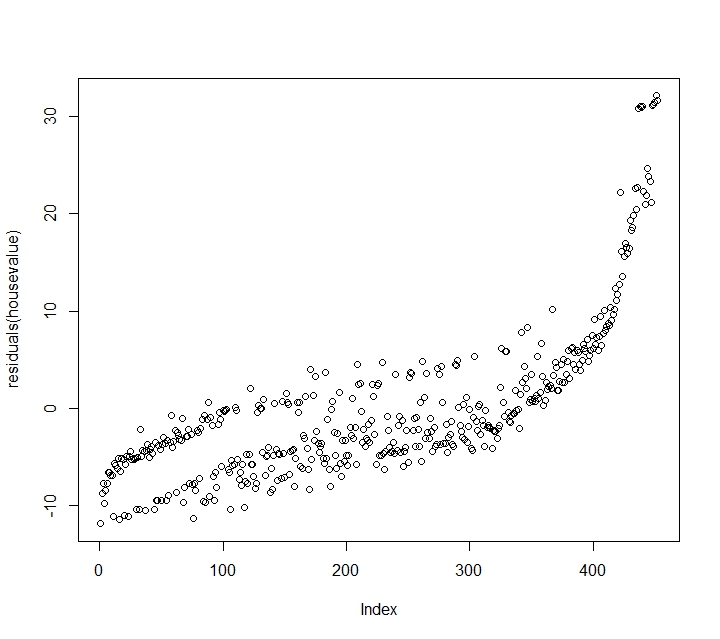
housevalue=(lm(medv~crim+indus+age, data=hdata))

summary(housevalue)

Assumptions checking:

Residual indpendency

plot(residuals(housevalue))



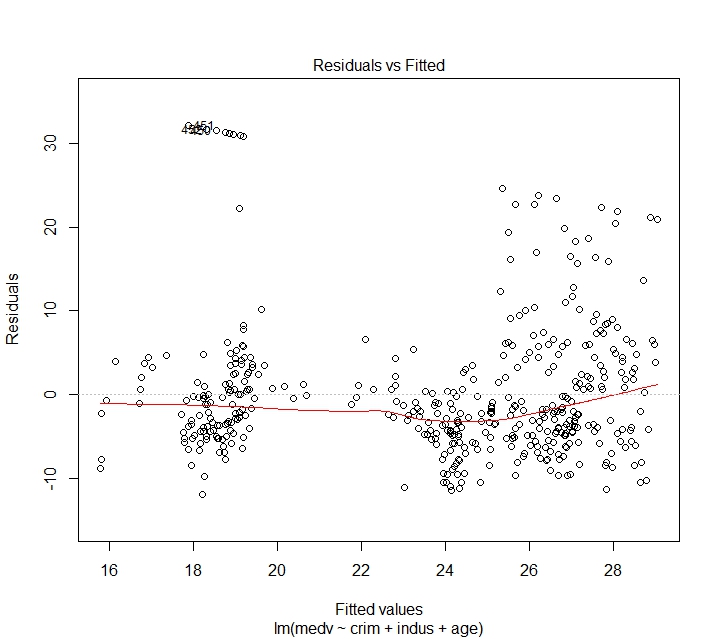
From the residuals plot, there is a clear pattern existed

library(lmtest)

dwtest(housevalue, alternative=c("two.sided"))

From the dwtest, the p-value is <2.2e-16 which is significant, it means autocorrelation exist in the model and violates the assumption

Residual homoscedasticity

plot(housevalue, which=c(1)) 

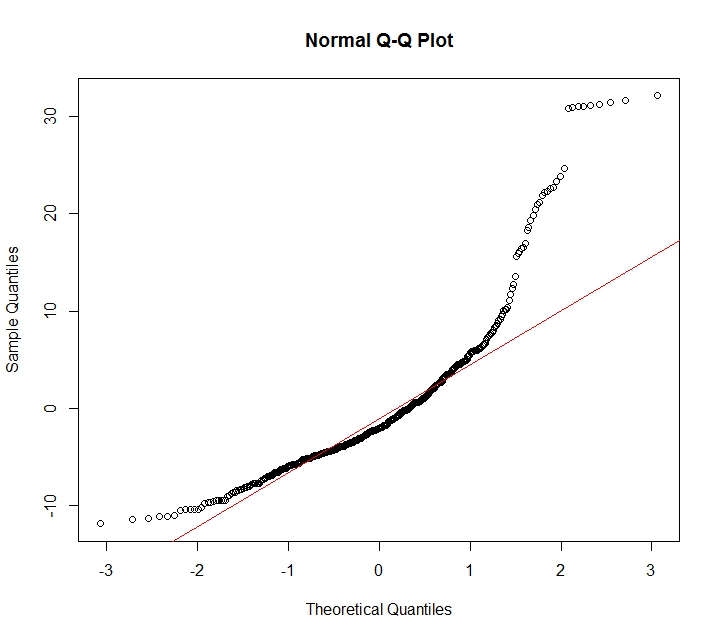
From just eyeballing the plot, the result seems to be homoscedastic

bptest(housevalue)

From the bptest, the p-value is 0.008422 which is not significant, so the residuals is homoscedastic and follows the assumption

Residual normality

qqnorm(residuals(housevalue))

qqline(residuals(housevalue),col="red")

From just eyeballing the qq plot, the residual seems to be not normally distributed

shapiro.test(residuals(housevalue))

From the shapiro test, the p-value is <2.2e-16 which is significant, so the residuals are not normally distributed and it violates the assumption

**Bonus part:**

From all the results above, the assumptions of residual independency and residual normality are being violated.

For the first problem, we can take a look back at the correlation value table:

Correlation values:

medv crim indus age

medv 1.0000000 -0.2862450 -0.4119145 -0.2998932

crim -0.2862450 1.0000000 0.5738223 0.4476638

indus -0.4119145 0.5738223 1.0000000 0.6059816

age -0.2998932 0.4476638 0.6059816 1.0000000

From the table, we find that the correlation coefficient between crim and indus, age and indus are both >0.5. That means in general, they are a bit ‘too’ correlated. We can solve the problem by replacing another variable with indus to make the correlation values in the table all below 0.5. Then run the dwtest again to look for a insignificant p-value (>0.05)

For the residual normality problem, we can solve it by transforming the data into suitable format (log, square root, or power) in order to get normally distributed residuals. The interpretation will depend on the form of transformation.

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).

summary(housevalue)

From the regression result, medv = 29.9-0.226(crim)-0.432(indus)-0.0217(age)

For the first line of the coefficients table, the intercept coefficient means when all other variables hold zero, the housing value = 29.9

The significant p-value (<2e-16) indicates that the coefficient is significantly different from zero

For the second line, the coefficient means that when indus and age hold zero, a unit increase in crim will lead to 0.226unit decrease in housing value

The insignificant p-value (0.227) indicates that the effect of crim towards housing value is negligible

For the third line, the coefficient means that when crim and age hold zero, a unit increase in indus will lead to 0.432unit decrease in housing value

The significant p-value (3.87e-08) indicates that the effect of indus towards housing value is significant

For the forth line, the coefficient means that when crim and indus hold zero, a unit increase in age will lead to 0.0217unit decrease in housing value

The insignificant p-value (0.204) indicates that the effect of age towards housing value is negligible

**Bonus part:**

housevalue1=(lm(medv~crim+indus+age\*dis, data=hdata))

summary(housevalue1)

From the regression result, medv = 36.7-0.411(crim)-0.706(indus)+0.0401(age)-0.330(dis)-0.0288(age)(dis)

For the first line of the coefficients table, the intercept coefficient means when all other variables hold zero, the housing value = 36.7

The significant p-value (<2e-16) indicates that the coefficient is significantly different from zero

For the second line, the coefficient means that when indus,age, and dis hold zero, a unit increase in crim will lead to 0.411unit decrease in housing value

The significant p-value (0.0231) indicates that the effect of crim towards housing value is significant

For the third line, the coefficient means that when crim,age, and dis hold zero, a unit increase in indus will lead to 0.706unit decrease in housing value

The significant p-value (<2e-16) indicates that the effect of indus towards housing value is significant

For the forth line, the coefficient means that when crim,indus, and dis hold zero, a unit increase in age will lead to 0.0401unit increase in housing value

The insignificant p-value (0.303) indicates that the effect of age towards housing value is negligible

For the fifth line, the coefficient means that when crim,indus, and age hold zero, a unit increase in dis will lead to 0.330unit decrease in housing value

The insignificant p-value (0.484) indicates that the effect of dis towards housing value is negligible

For the sixth line, the coefficient means that when dis increase by 1unit, the effect of age towards housing value will decrease by 0.0288

The significant p-value (0.000246) indicates that the interaction effect between age and dis is significant

However, in this case since both the individual effect of age and dis are negligible, even though the interaction is significant, the interaction effect still cannot be expressed

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

One way to tell whether the model is good or bad is to examine R-squared

which tells how much of the variance is being explained by the independent variables we put in the model.

summary(housevalue)

From the summary table, R-squared is 0.1764, which means that only 17.64% of the variability of the response data round its mean.

So this is not really a good model

Meanwhile, the p-value of the F-statistic is significant, 2.2e-16, so that means the R-squared is reliable even though it is small.

As i created a second model with interaction term, we can use model comparison to tell whether it is a good model.

anova(housevalue,housevalue1)

From the result, the p-value is 5.397e-12 which is significant, that means the probability that model 2 can explain the same variance with model 1 is very small.

So model 2 is explaining different variances.

Also, model 2 has a lower residual sum of squares, which means it has more predictable sum of squares than model 1. So model 2 is a better model.