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

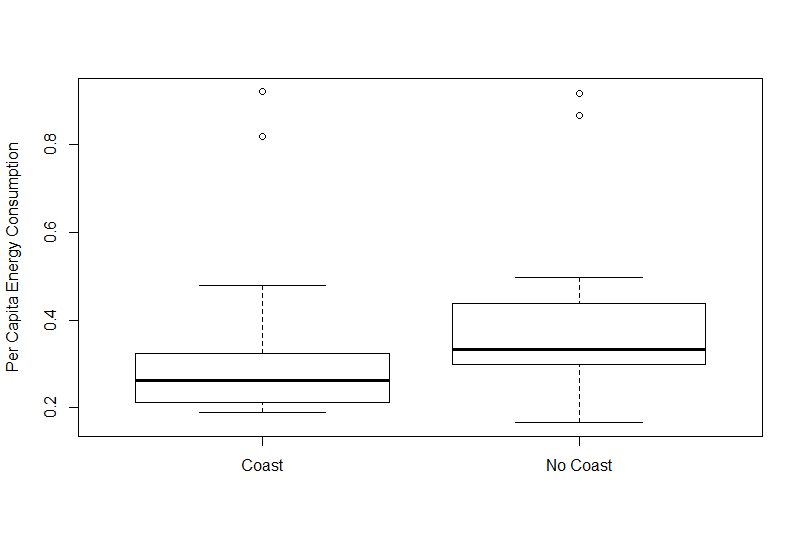
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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 H: There is no significant difference in per capita energy consumption between states found on the coast and states that are not found on the coast.**

**Alternative H: There is a significant difference in per capita energy consumption between states found on the coast and states that are not found on the coast.**

* 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).

**Welch 2 sample, 2-tail t-test. Data are continuous, assume random sampling, assume independent observations, qqplot and shapiro test (p=<0.0001) show data is not normally distributed so we could log-transform, variances are equal as determined by var.test (p=0.5098).**

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

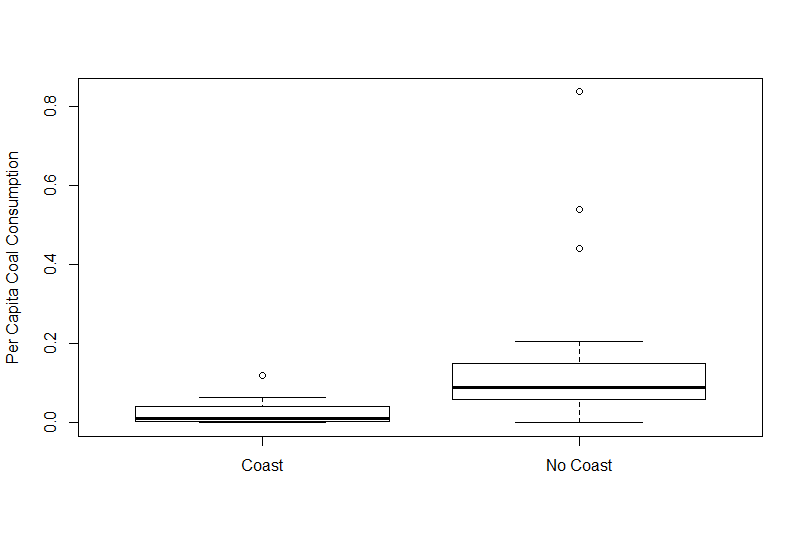
p-value = 0.2245: **differences between the two groups are not statistically different. There is no difference in per capita energy consumption between states on the coast and states not on the coast.**

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 H: There is no significant difference in per capita coal consumption between states found on the coast and states that are not found on the coast.**

**Alternative H: There is a significant difference in per capita coal consumption between states found on the coast and states that are not found on the coast.**

* 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).

**Welch 2 sample, 2-tail t-test. Data are continuous, assume random sampling, assume independent observations, qqplot and shapiro test (p-value = 6.024e-11) show data is not normally distributed so we could log transform our data, variances unequal as well as determined by var.test (p-value = 5.995e-13). Can standardize data to get equal variances.**

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

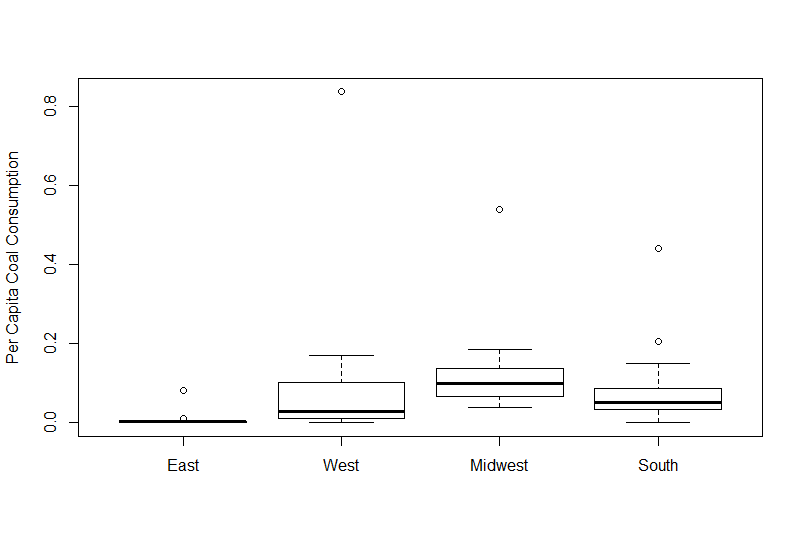
p-value = 0.001936**: per capita coal consumption is statistically significantly higher in states not on the coast than states on the coast.**

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 H: There is no significant difference in per capita coal consumption between states found in the four different regions of study.**

**Alternative H: There is a significant difference in per capita coal consumption between states found in the four different regions of study.**

* 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 way ANOVA. We assume independent samples, the population size does not have to be the same, data does not meet assumptions of normality from qqplots and shapiro tests (p-value = 6.024e-11). We could log transform this data. Unequal variances for each region (across most groups) as seen from var.tests.**

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

**Df Sum Sq Mean Sq F value Pr(>F)**

**Region: 3 0.0864 0.02879 1.375 0.262: High p-value suggests that there is no statistically significant different between the 4 regions of study in terms of per capita coal consumption. No need to run Tukey HSD test.**

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 between per capita coal use and per capita GDP is 0.03598. This value is very small and close to zero which means that the correlation between the two variables is very low to almost non-existent. This means that there is no real change in one variable as the other changes, or if there is, it’s a very small positive change.**

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).

**My model: mod = lm(medv ~ ptratio + crim + lstat, data = hdata). Model has high correlation between crim and lstat (0.42). VIF of ~2.3 suggests this is not a problem.**

ptratio crim lstat

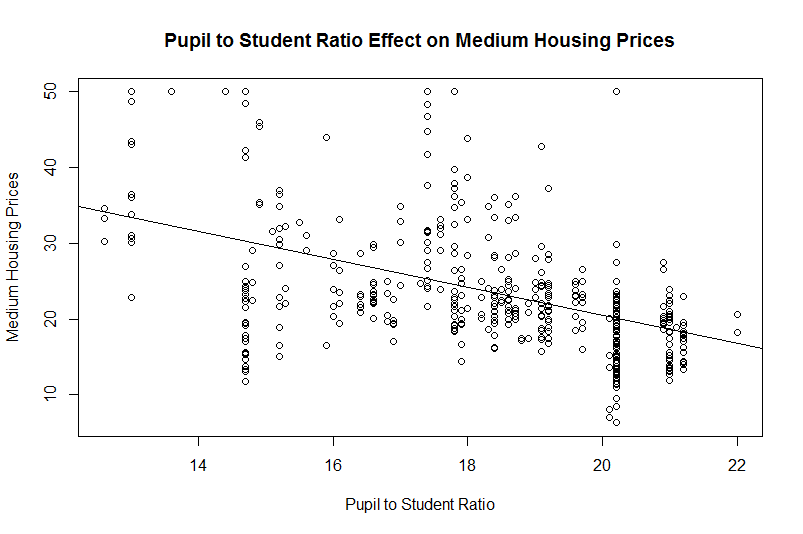
ptratio 1.0000000 0.3194701 0.3030431

crim 0.3194701 1.0000000 0.4247886

lstat 0.3030431 0.4247886 1.0000000

VIF:[1] 2.334327

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).



**Mdev~ptratio: As the pupil-student ratio decreases by 1 unit, the medv value decreases by 1.84 units. This variable explains around 21% of the variation in this variable and the effect is significantly different from 0 (as seen from summary function).**

Call:

lm(formula = medv ~ ptratio, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-18.5006 -4.9561 -0.9568 2.5721 29.8557

Coefficients:

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

(Intercept) 57.4457 3.0779 18.66 <2e-16 \*\*\*

ptratio -1.8466 0.1675 -11.03 <2e-16 \*\*\*

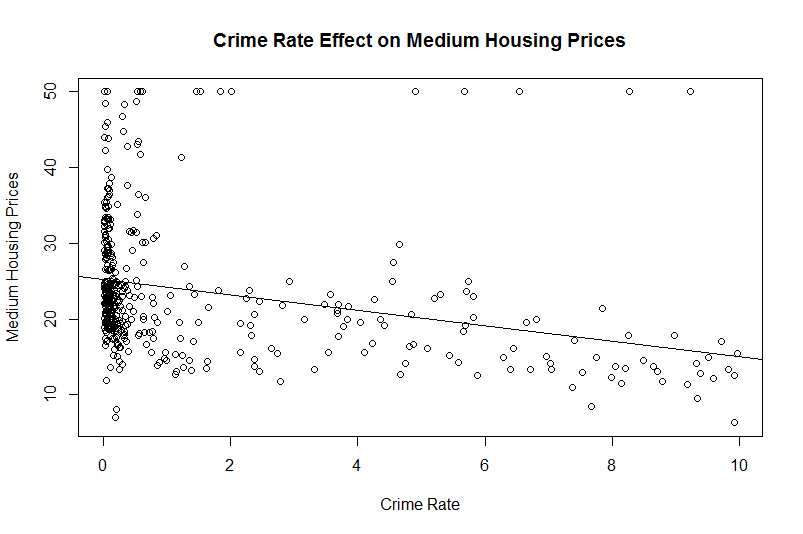
---

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

Residual standard error: 7.824 on 450 degrees of freedom

Multiple R-squared: 0.2127, Adjusted R-squared: 0.211

F-statistic: 121.6 on 1 and 450 DF, p-value: < 2.2e-16



**Mdev~crim: As the crime rate decreases by 1 unit, the medv value decreases by 1.01 units. This variable explains around 8.1% of the variation in this variable and the effect is significantly different from 0.**

Call:

lm(formula = medv ~ crim, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-18.001 -5.100 -2.190 2.791 34.141

Coefficients:

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

(Intercept) 25.1858 0.4574 55.059 < 2e-16 \*\*\*

crim -1.0102 0.1594 -6.337 5.68e-10 \*\*\*

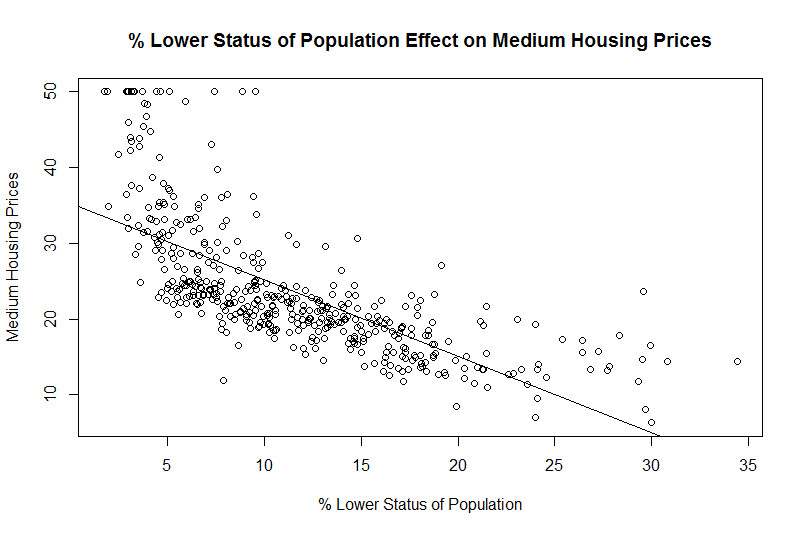
---

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

Residual standard error: 8.449 on 450 degrees of freedom

Multiple R-squared: 0.08194, Adjusted R-squared: 0.0799

F-statistic: 40.16 on 1 and 450 DF, p-value: 5.683e-10



**Mdev~lstat: As the % lower status of the population decreases by 1 unit, the medv value decreases by 1.01 units. This variable explains around 50% of the variation in this variable and the effect is significantly different from 0.**

Call:

lm(formula = medv ~ lstat, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-15.450 -4.136 -1.493 1.879 24.318

Coefficients:

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

(Intercept) 35.31254 0.62027 56.93 <2e-16 \*\*\*

lstat -1.01051 0.04775 -21.16 <2e-16 \*\*\*

---

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

Residual standard error: 6.243 on 450 degrees of freedom

Multiple R-squared: 0.4988, Adjusted R-squared: 0.4977

F-statistic: 447.8 on 1 and 450 DF, p-value: < 2.2e-16

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 independence (no autocorrelation): dwtest: p-value < 2.2e-16**

**Homoscedasticity: bptest: p-value = 3.703e-09**

**Normality: shapiro test: p-value = 9.946e-16**

**There model residuals are not independent, display no homoscedasticity, and are not normally distributed (all assumptions violated, all tests below p=0.05). Normality can be restored with a transformation (like a log transformation). To deal with heteroscedasticity, we can use more robust standard errors, or log transform data. To deal with residual autocorrelation, we may want to use a model other than a linear 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 ~ ptratio + crim + lstat, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-12.1728 -3.5742 -0.9775 2.0661 24.4036

Coefficients:

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

(Intercept) 55.21590 2.35297 23.466 <2e-16 \*\*\*

ptratio -1.16062 0.13321 -8.712 <2e-16 \*\*\*

crim 0.29738 0.12361 2.406 0.0165 \*

lstat -0.93603 0.04983 -18.784 <2e-16 \*\*\*

---

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

Residual standard error: 5.785 on 448 degrees of freedom

Multiple R-squared: 0.5716, Adjusted R-squared: 0.5687

F-statistic: 199.3 on 3 and 448 DF, p-value: < 2.2e-16

**The intercept estimate is where the line passes through the y-axis (at 55.216 units) when the x variables are held at 0 and this is significantly different from 0. The ptratio estimate is the regression coefficient or slope of the relationship between ptratio and medv (effect size), and means that as ptratio increases by 1 unit, then medv decreases by 1.16 units. This effect size is significantly different from 0 as seen by the low p-value. The crim estimate is the regression coefficient or slope of the relationship between crim and medv (effect size), and means that as crim increases by 1 unit, then medv increases by 0.297 units. This effect size is significantly different from 0 as seen by the low p-value. The lstat estimate is the regression coefficient or slope of the relationship between lstat and medv (effect size), and means that as lstat increases by 1 unit, then medv decreases by 0.936 units. This effect size is significantly different from 0 as seen by the low p-value. The adjusted R^2 value of 0.5687 means that about 57% of the variation in the data can be explained by this model.**

Call:

lm(formula = medv ~ ptratio + crim \* lstat, data = hdata)

Residuals:

Min 1Q Median 3Q Max

-12.517 -3.439 -1.026 2.175 20.340

Coefficients:

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

(Intercept) 55.17887 2.31249 23.861 < 2e-16 \*\*\*

ptratio -1.23167 0.13206 -9.326 < 2e-16 \*\*\*

crim 1.62261 0.34512 4.702 3.45e-06 \*\*\*

lstat -0.82381 0.05609 -14.686 < 2e-16 \*\*\*

crim:lstat -0.08052 0.01963 -4.102 4.86e-05 \*\*\*

---

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

Residual standard error: 5.685 on 447 degrees of freedom

Multiple R-squared: 0.5872, Adjusted R-squared: 0.5835

F-statistic: 158.9 on 4 and 447 DF, p-value: < 2.2e-16

**The intercept estimate is where the line passes through the y-axis (at 55.179 units) when the x variables are held at 0 and this is significantly different from 0. The ptratio estimate is the regression coefficient or slope of the relationship between ptratio and medv (effect size), and means that as ptratio increases by 1 unit, then medv decreases by 1.23 units. This effect size is significantly different from 0 as seen by the low p-value. The crim estimate is the regression coefficient or slope of the relationship between crim and medv (effect size), and means that as crim increases by 1 unit, then medv increases by 1.62 units. This effect size is significantly different from 0 as seen by the low p-value. The lstat estimate is the regression coefficient or slope of the relationship between lstat and medv (effect size), and means that as lstat increases by 1 unit, then medv decreases by 0.82 units. This effect size is significantly different from 0 as seen by the low p-value. The crim:lstat value is the interaction term and the estimate gives an understanding of how dependent the crim variable is on the lstat variable (based on slope). This interaction is significantly different from zero as seen from the low p-value. The adjusted R^2 value of 0.5835 means that about 58% of the variation in the data can be explained by this model.**

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

**If we had log transformed the data to correct the violated assumptions then maybe this would have been a good fit. The autocorrelation between the residuals means that a linear model may not be the best fit in any case. Despite this, the significant variable effects on medv and the high R2 value (0.56) means that is the assumption were not violated this model would fit very well. Unfortunately, the violated assumptions mean that this linear model is not the best fit to the data.**