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

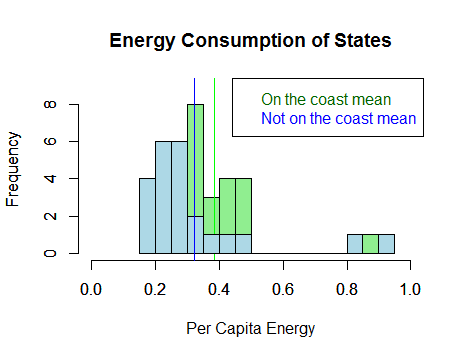
For Questions 1 – 4, please use the energy dataset ‘energy\_data.csv’. It is a dataset that includes the amount of energy consumed (TotalEnergy), the amount of coal consumed (TotalCoal), the GDP (TotalGDP), and the population (Population) of each state in the US in 2014. The states also are categorized 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:** The per capita energy consumption is not different between the states on the coast and the states not on the coast.

**Alternate hypothesis**: The per capita energy consumption is different between the states on the coast and the states not on the coast.

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



Mean of per capita energy consumption of both samples seem close on the plot.

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

Since the two independent variables are categorical (states on the coast / states not on the coast) and the dependent variable is continuous (per capita energy), t-test will be used to analyze the difference.

**Assumption test:**

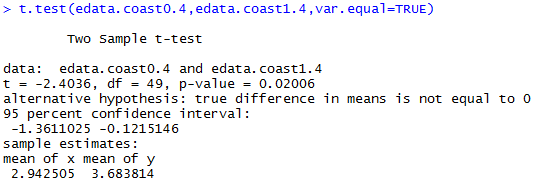
1. Each observation is sampled independently: This is unlikely to test; however, samples are collected by total energy consumption of states separately, so we can assume samples are independently collected.

2. Samples are normally distributed:

* Both samples do not pass shapiro test (pass: p-value>0.05 🡪 normally distributed), so we transform them by *log*.
* Log transformed samples of states on the coast still do not pass shapiro test, so we use box cox power transformation.
* Lambda values are -0.4859927 and -1.574563, so we transform samples by (1/Y) based on box cox transformation table.
* Both (1/Y) transformed samples pass shapiro test, so we meet the normality assumption.

3. Samples have equal variance: P-value of variance test between both (1/Y) transformed samples = 0.3544 > 0.05, so we meet the equal variance assumption.

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



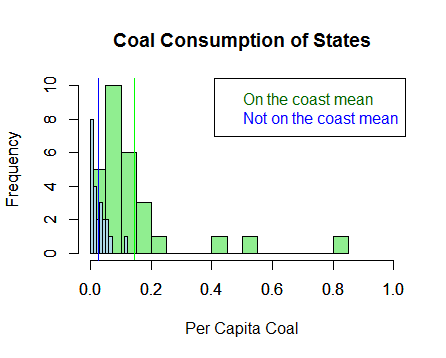
P-value = 0.02006 > 0.05, so we fail to reject the null hypothesis. Therefore, there is no difference of per capita energy consumption between the states on the coast and the states not on the coast.

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 hypothesis:** The per capita coal consumption is not different between the states on the coast and the states not on the coast.

**Alternate hypothesis**: The per capita coal consumption is different between the states on the coast and the states not on the coast.

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



Mean of per capita coal consumption of both samples seem a bit far from each other on the plot.

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

Since the two independent variables are categorical (states on the coast / states not on the coast) and the dependent variable is continuous (per capita coal), t-test will be used to analyze the difference.

**Assumption test:**

1. Each observation is sampled independently: This is unlikely to test; however, samples are collected by total coal consumption of states separately, so we can assume samples are independently collected.

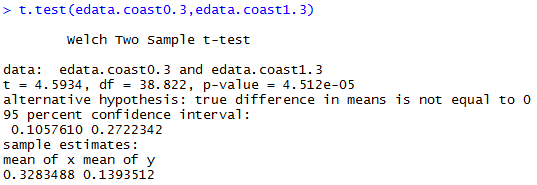
2. Samples are normally distributed:

* Both samples do not pass shapiro test (pass: p-value>0.05 🡪 normally distributed), so we transform them by *sqrt*.
* Sqrt transformed samples of states on the coast becomes normally distributed. Although sqrt transformed samples of states not on the coast still do not pass shapiro test, they are far better than no transformed ones. (p-value = 6.211e-07 🡪 0.008021)

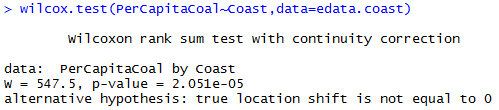
3. Each sample has equal variance: P-value of variance test between both sqrt transformed samples = 0.000254 < 0.05, so we do not meet the equal variance assumption. Therefore, we will use Welch’s t-test.

We will use Welch’s t-test to analyze; however, since sqrt transformed samples of states not on the coast still do not pass shapiro test, and both sample sizes are smaller than 30 (On the coast 23; Not on the coast 28), we will also use non-parametric test (Wilcoxon-Mann-Whitney U test) to check.

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



P-value = 4.512e-05 < 0.05, so we reject the null hypothesis.



P-value = 2.051e-05 < 0.05, so we reject the null hypothesis.

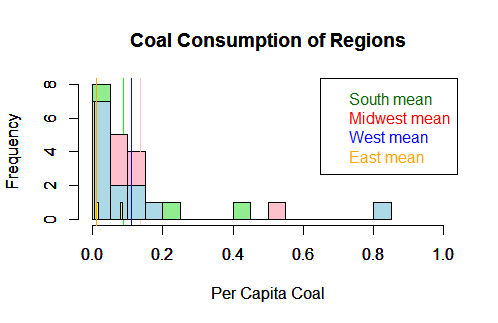
P-value of both tests < 0.05, so we reject the null hypothesis. Therefore, there is significant difference of per capita coal consumption between the states on the coast and the states not on the coast.

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 hypothesis:** The per capita coal consumption is not different between regions where states are found.

**Alternate hypothesis**: The per capita coal consumption is different between regions where states are found.

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



On the plot, mean of per capita coal consumption of south, midwest, and west regions seem close to one another, but mean of per capita coal consumption of east regions seem far from the group of the other three regions.

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

Since we have four independent variables that are categorical (South, West, Midwest, or East) and the dependent variable is continuous (per capita coal), ANOVA will be used to analyze the difference.

**Assumption test:**

1. Each observation is sampled independently: This is unlikely to test; however, samples are collected by total coal consumption of states separately, so we can assume samples are independently collected.

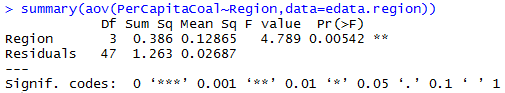
2. Samples are normally distributed:

* All samples do not pass shapiro test (pass: p-value>0.05 🡪 normally distributed), so we transform them by *sqrt*.
* Sqrt transformed samples of the south region become normally distributed. Although samples of the other three regions still do no pass shapiro test, they are far better than no transformed ones. (p-value: West 1.103e-05 🡪 0.003132; Midwest 0.0001758 🡪 0.006807; East 5.535e-06 🡪 0.003827)

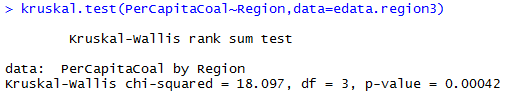
3. Samples have equal variance: P-value of variance test between sqrt transformed samples (South & West = 0.08855 > 0.05; West & Midwest = 0.1044 > 0.05; Midwest & East = 0.1754 > 0.05), and p-value of Levene test = 0.3803 > 0.05, so we meet the equal variance assumption.

We will use ANOVA to analyze; however, since most of the transformed samples still do not pass shapiro test, and their sample sizes are smaller than 30 (South 17; West 13; Midwest 12; East 9), we will also use non-parametric test (Kruskal Wallis test) to check.

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



P-value = 0.00542 < 0.05, so we reject the null hypothesis.



P-value = 0.00042 < 0.05, so we reject the null hypothesis.

P-value of both tests < 0.05, so we reject the null hypothesis. Therefore, per capita coal consumption of at least one region is significantly different from one another.

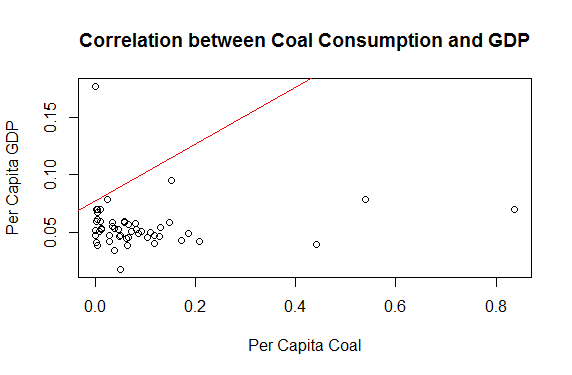
1. What is the correlation between ***per capita*** coal use and ***per capita*** GDP? Does this seem like a strong correlation to you? Why or why not? (2 points)



r = 0.03598182, which is not a strong correlation to me. Since positive correlation coefficient can range between 0 < r ≤ 1 (strongest correlation), r = 0.03598182 is really small within this range.

In addition, from the plot below, a data, with extremely high per capita GDP when per capita coal is low (top-left), largely skews the best fit line, which also decreases the strength of the relationship between per capita coal and per capita GDP.

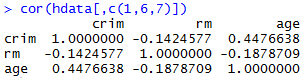




For questions 5-9, please use the ‘housedata.csv’ dataset that shows housing information for the Boston area. Information on what each of the variables are can be found here: <http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>. In this exercise, the goal is to create a multiple 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.

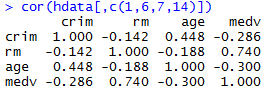
1. Please select three covariates that you will include in your model as independent variables. Please check if these variables are highly correlated with one another to make sure you do not run into problems of multi-collinearity. Check if this model has issues with multi-collinearity using the variance inflation factor. **Report correlation values and VIF values in your answer** (3 points).

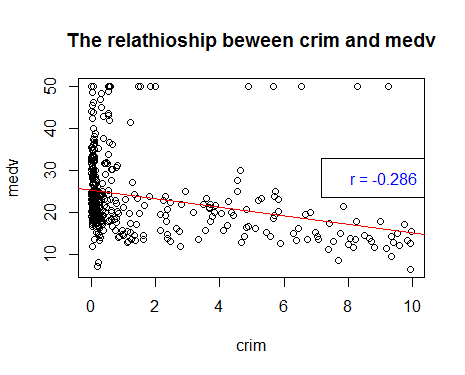
Three covariates that will be included in the model as independent variables are *crim*, *rm*, and *age*. (*crim*: per capita crime rate by town; *rm*: average number of rooms per dwelling; *age*: proportion of owner-occupied units built prior to 1940)



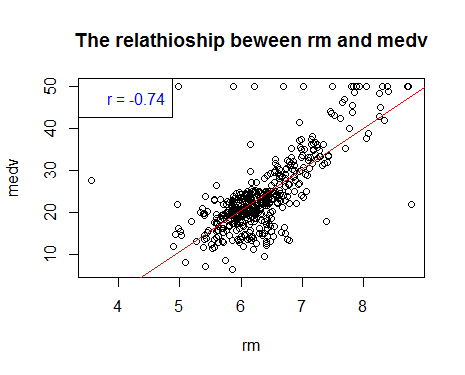
All correlation coefficients between each independent variable are < 0.5, and VIF = 2.439356 < 10, so there is a low possibility to run into the problem of multi-collinearity between independent variables.

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).

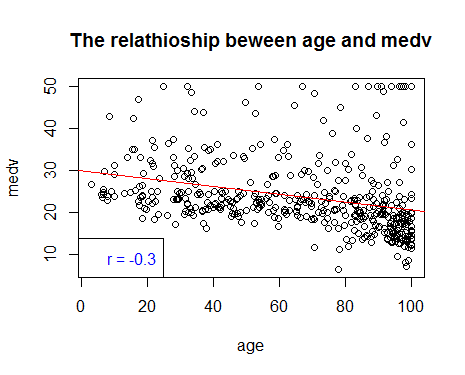




From the plot, the per capita crime rate will slightly decrease the median housing prices. (r = -0.286)



From the plot, the average number of rooms per dwelling will largely increase the median housing prices. (r = 0.74)



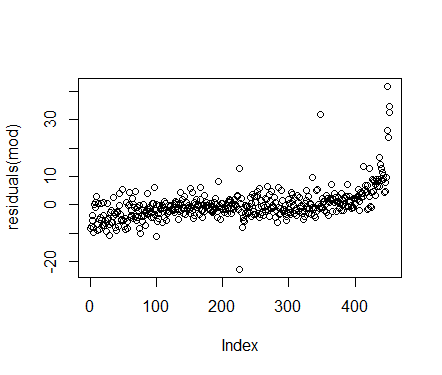
From the plot, the proportion of owner-occupied units built prior to 1940 will slightly decrease the median housing prices. (r = -0.3)

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).



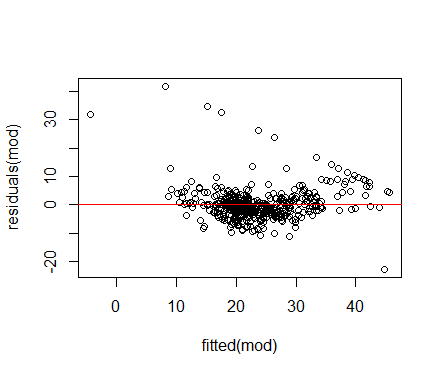
**Assumption test:**

1. There is a linear relationship between variables: From the plots in question 6, we can assume the independent variables and the dependent variable have linear relationships.
2. Residual independency:



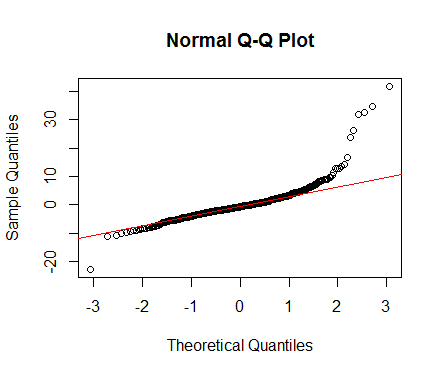
* From the plot, there is a clear pattern that residuals are not randomly scattered, which indicates that the residuals are autocorrelated.
* P-value of Durbin-Watson test = 2.2e-16 < 0.05, which shows the residuals have autocorrelation.

1. Residual homoscedasticity:



* From the plot, the residuals do not seem to distribute evenly along the red line of fitted values because points on the top area show a decreasing pattern with the increase of fitted values, which indicates the variance of errors is not constant.
* P-value of Breusch-Pagan test = 8.811e-07 < 0.05, which shows the residuals are not homoscedastic.

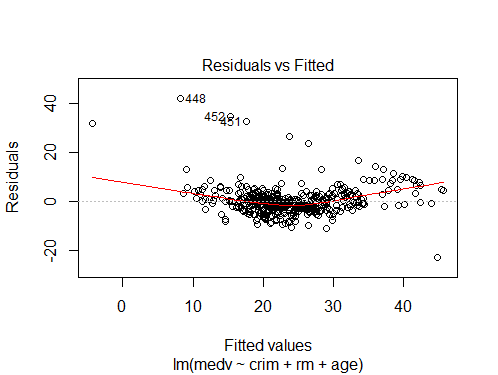
1. Residual normality:

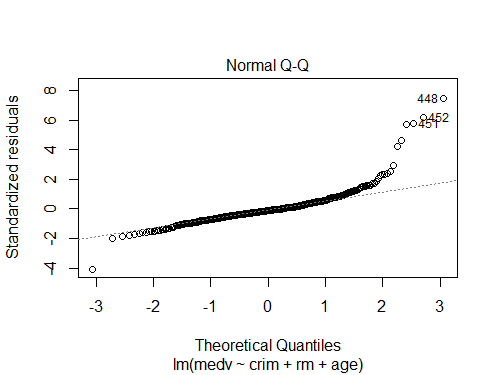


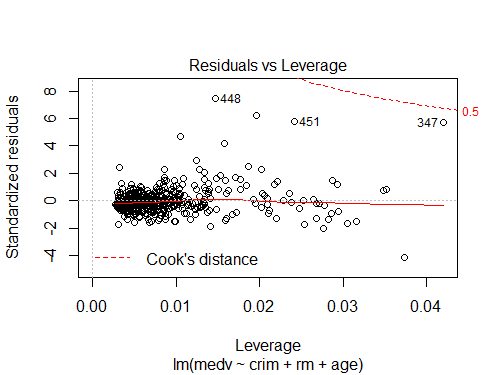
* From the plot, the residuals on the right area do not follow the qqline, which indicates the residuals are not normally distributed.
* P-value of Shapiro-Wilk test = 2.2e-16 < 0.05, which shows the residuals are not normally distributed.

Three assumptions of the residuals are violated. For the violation of residual independency, it suggests that there is room for improvement for the model and linear model may not be the best to fit the data, so other types of model should be considered.

To overcome the violation of residual homoscedasticity and residual normality, one possible way is to remove the outliers based on model plots. The other way to overcome the violation of residual normality is doing data transformation.



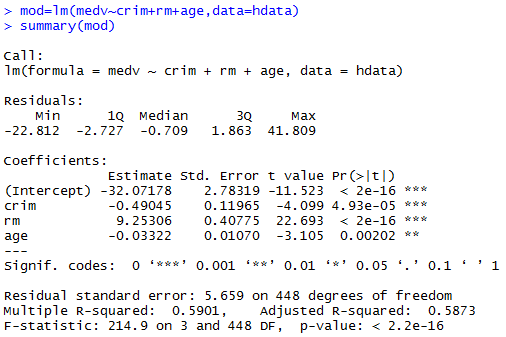




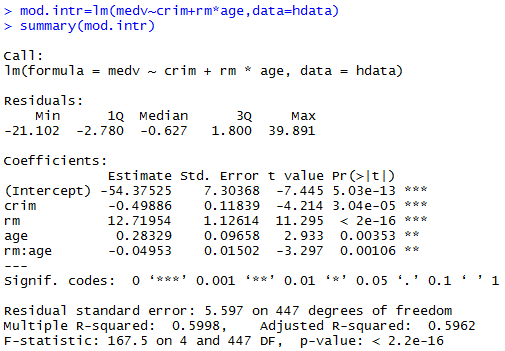
Cook’s distance identifies the possible outliers, so we might take a look at the 347th, 448th, 451st, and 452nd data, and decide whether they are really outliers and remove them.

However, both (1/sqrt) transformed data and outlier removal still cannot resolve the violation of assumptions. (Process in R script)

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).



1. The estimated coefficient of intercept = -32.07178, which is the *medv* value when *crim*, *rm*, and *age* are 0; its p-value = 2e-16 < 0.05, which means the intercept is significantly different from 0.
2. The estimated coefficient of *crim* (the effect of *crim* on *medv* / slope) = -0.49045, which means when *rm* and *age* are controlled (hold constant), increasing one unit of *crim* will decrease *medv* by 0.49045. Its p-value = 4.93e-05 < 0.05, which means the estimate is significantly different from slope=0, and thus indicates that the effect of *crim* on *medv* is significant.
3. The estimated coefficient of *rm* (the effect of *rm* on *medv* / slope) = 9.25306, which means when c*rim* and *age* are controlled (hold constant), increasing one unit of *rm* will increase *medv* by 9.25306. Its p-value = 2e-16 < 0.05, which means the estimate is significantly different from slope=0, and thus indicates that the effect of *rm* on *medv* is significant.
4. The estimated coefficient of *age* (the effect of *age* on *medv* / slope) = -0.03322, which means when c*rim* and *rm* are controlled (hold constant), increasing one unit of *age* will decrease *medv* by 0.03322. Its p-value = 0.00202 < 0.05, which means the estimate is significantly different from slope=0, and thus indicates that the effect of *age* on *medv* is significant.



1. The estimated coefficient of intercept = -54.37525, which is the *medv* value when *crim*, *rm*, and *age* are 0; its p-value = 5.03e-13 < 0.05, which means the intercept is significantly different from 0.
2. medv = (-54.37525) + (-0.49886)\*crim + 12.71954\*rm + 0.28329\*age + (-0.04953)\*(rm\*age)
3. The estimated coefficient of *crim* (the effect of *crim* on *medv* / slope) = -0.49886, which means when *rm* and *age* are controlled (hold constant), increasing one unit of *crim* will decrease *medv* by 0.49886. Its p-value = 4.04e-05 < 0.05, which means the estimate is significantly different from slope=0, and thus indicates that the effect of *crim* on *medv* is significant.
4. The estimated coefficient of *rm* (the effect of *rm* on *medv* / slope) = 12.71954, which means when c*rim* and *age* are controlled (hold constant), increasing one unit of *rm* will increase *medv* by 12.71954. Its p-value = 2e-16 < 0.05, which means the estimate is significantly different from slope=0, and thus indicates that the effect of *rm* on *medv* is significant.
5. The estimated coefficient of *age* (the effect of *age* on *medv* / slope) = 0.28329, which means when c*rim* and *rm* are controlled (hold constant), increasing one unit of *age* will increase *medv* by 0.28329. Its p-value = 0.00353 < 0.05, which means the estimate is significantly different from slope=0, and thus indicates that the effect of *age* on *medv* is significant.
6. The estimated coefficient of *rm:age* [the interaction between *rm* and *age* = the influence of *rm* on the effect of *age* on *medv* (the influence of *rm* on *age*’s slope) **OR** the influence of *age* on the effect of *rm* on *medv* (the influence of *age* on *rm*’s slope)] = -0.04953, which means when c*rim* is controlled (hold constant), increasing one unit of *rm* will decrease the effect of *age* on *medv* by 0.04953 **OR** increasing one unit of *age* will decrease the effect of *rm* on *medv* by 0.04953. Its p-value = 0.00106 < 0.05, which means the estimate of interaction term is significantly different from 0, and thus indicates that the influence of *rm* on the effect of *age* on *medv* **OR** the influence of *age* on the effect of *rm* on *medv* (the interaction between *rm* and *age*) is significant.

* When *crim* is controlled, adding *x* units of *rm* will result in the effect of *age* on *medv* (estimated coefficient):

βage’= βrm + βrm:age\* *x* = 0.28329 + (-0.04953)\**x*

* When *crim* is controlled, adding *x* units of *age* will result in the effect of *rm* on *medv* (estimated coefficient):

βrm’= βage+ βrm:age\* *x* = 12.71954 + (-0.04953)\**x*

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

From the table of *summary(mod)*, the adjusted R-squared = 0.5873 indicates the three independent variables we put in this model explain approximate 59% of the variation of the dependent variable (median housing prices), which is a good fit to me. There are many factors (independent variables) might influence the median housing prices. I think only these three independent variables can explain almost 60% of the dependent variables, which is pretty good.