Building a Housing Multiple Linear Regression Model Using R

Didem Bulut Aykurt

MIS500-1 – Foundations of Data Analytics

Colorado State University-Global Campus

Dr. Steve Chung

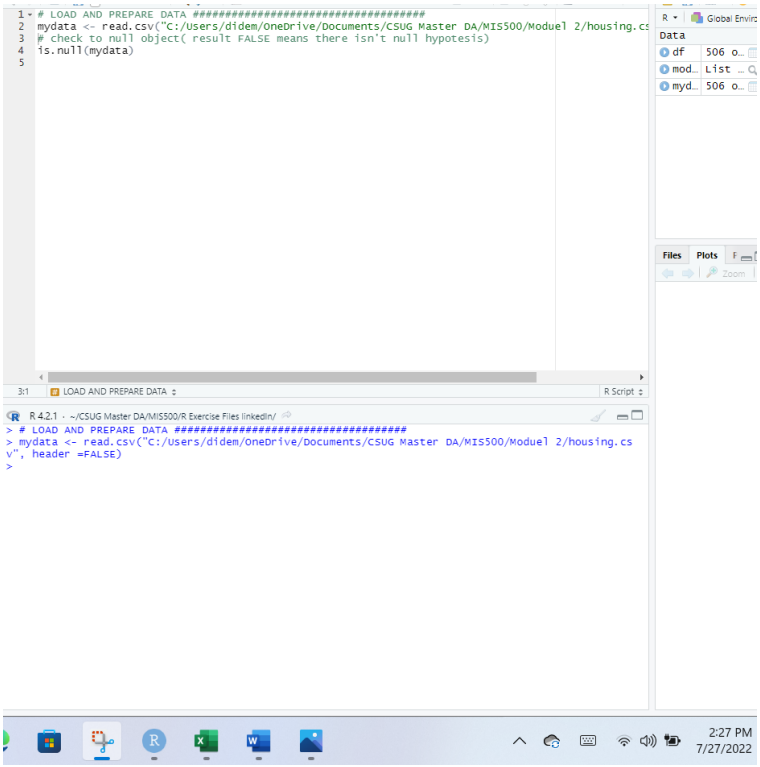
July 30, 2022

Mod 2-Critical Thinking-Option 1

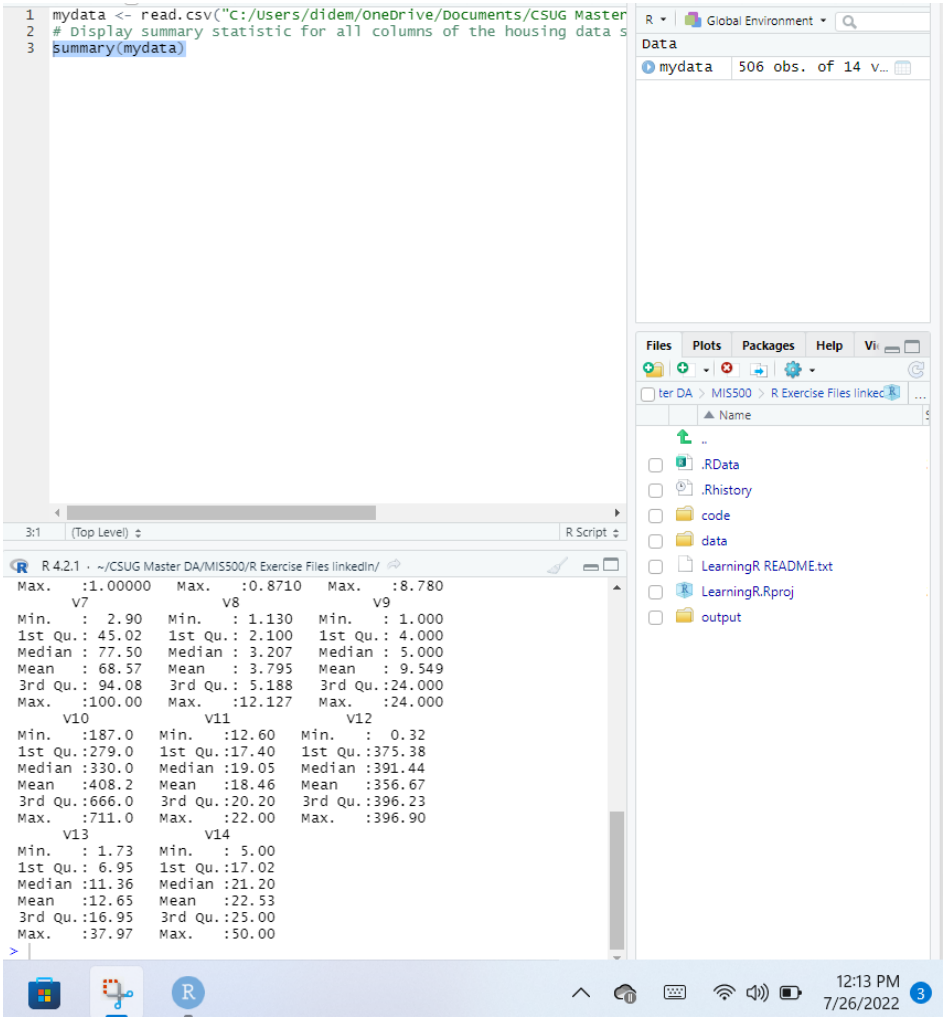
The Boston Housing Dataset; describes concerns housing values in suburbs of Boston. The data sources originate from the StatLib library at Carnegie Mellon University on July 7, 1993, as follows housing data set available at the UCI Machine Learning repository. Five hundred six intakes represent aggregated data and 13 attributes(features) with the dependent variable(price).

- CRIM - per capita crime rate by town
- ZN – the proportion of residential land zoned for over 25,000 sq. ft.
- INDUS – the proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE – the proportion of owner-occupied units built before 1940
- DIS - weighted distances to five Boston employment centers
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town
- B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's

Let's deep dive into the housing dataset with RStudio. First, load and prepare data.



The data set doesn't have a null hypothesis. We don't use a header column, so we don't need to change the useless column name—the next step is to view the statistic summary.



Mean

The first thing to check the mean value is to see an average data point for each variable. V12 is the median value of owner-occupied homes with a $22.53 price. The price is reasonable and very low because the average home price in the USA was $138 in 1993.

Age the average age of a home was 68, older than in New York, with the median age at 63 in 1993. The room average is six, meaning big houses in 1993. And the house wasn't so far from the employment center as the mean of DIS is 3.7 miles.

Mean-Median

The median is preferably measured at a distance from the mean if there are outliers or high variability in the dataset. The data is usually distributed if the mean and median length is small. If the median is much bigger than the mean as distribution is left-skewed. Suppose the median is much smaller than the mean, the right-skewed distribution.

The average of room numbers would have a normal distribution. Age has a left-skewed distribution. And TAX has a right-skewed distribution.

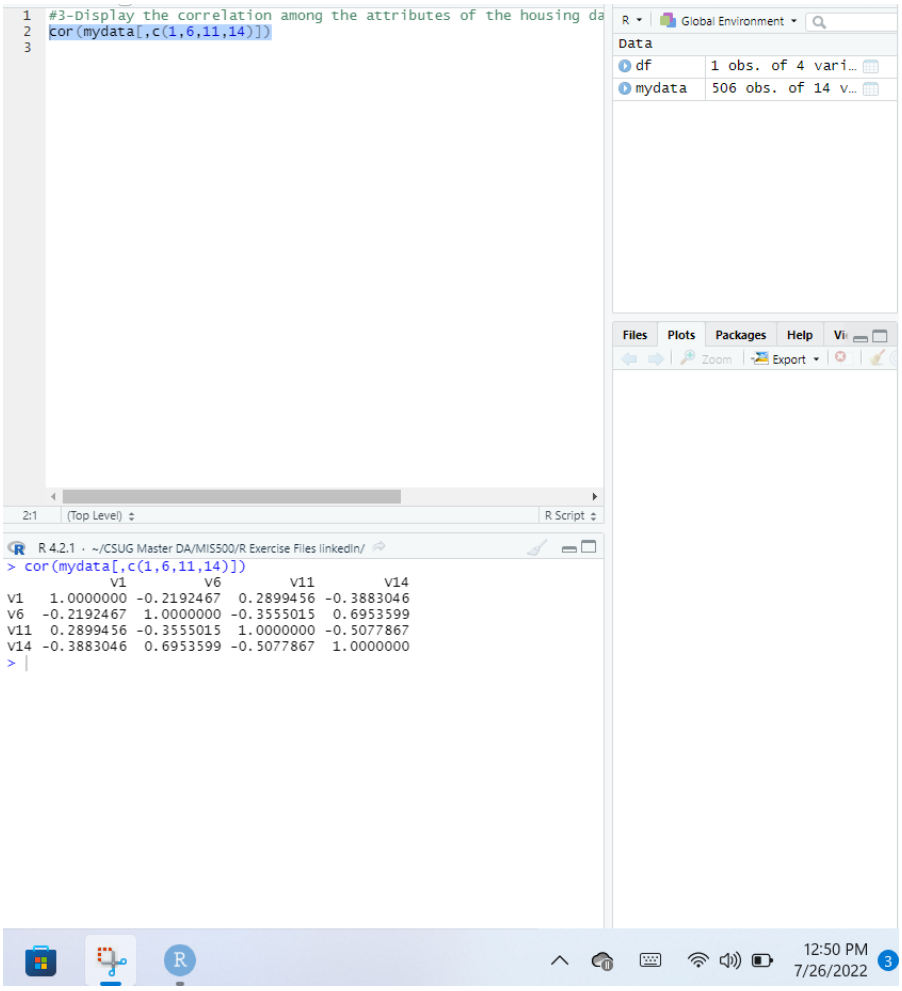
Min-Max

Price of house dataset spans between $5-$50. The max of the price was reasonable in 1993 but min price of house look bit lower, mainly the average price is $22.53.

Outliers

The 3rd quartile and max difference big that difference include potential more outliers. Same as 1st quartile and min value difference. The price of the home has outliers, mainly for the upper side.

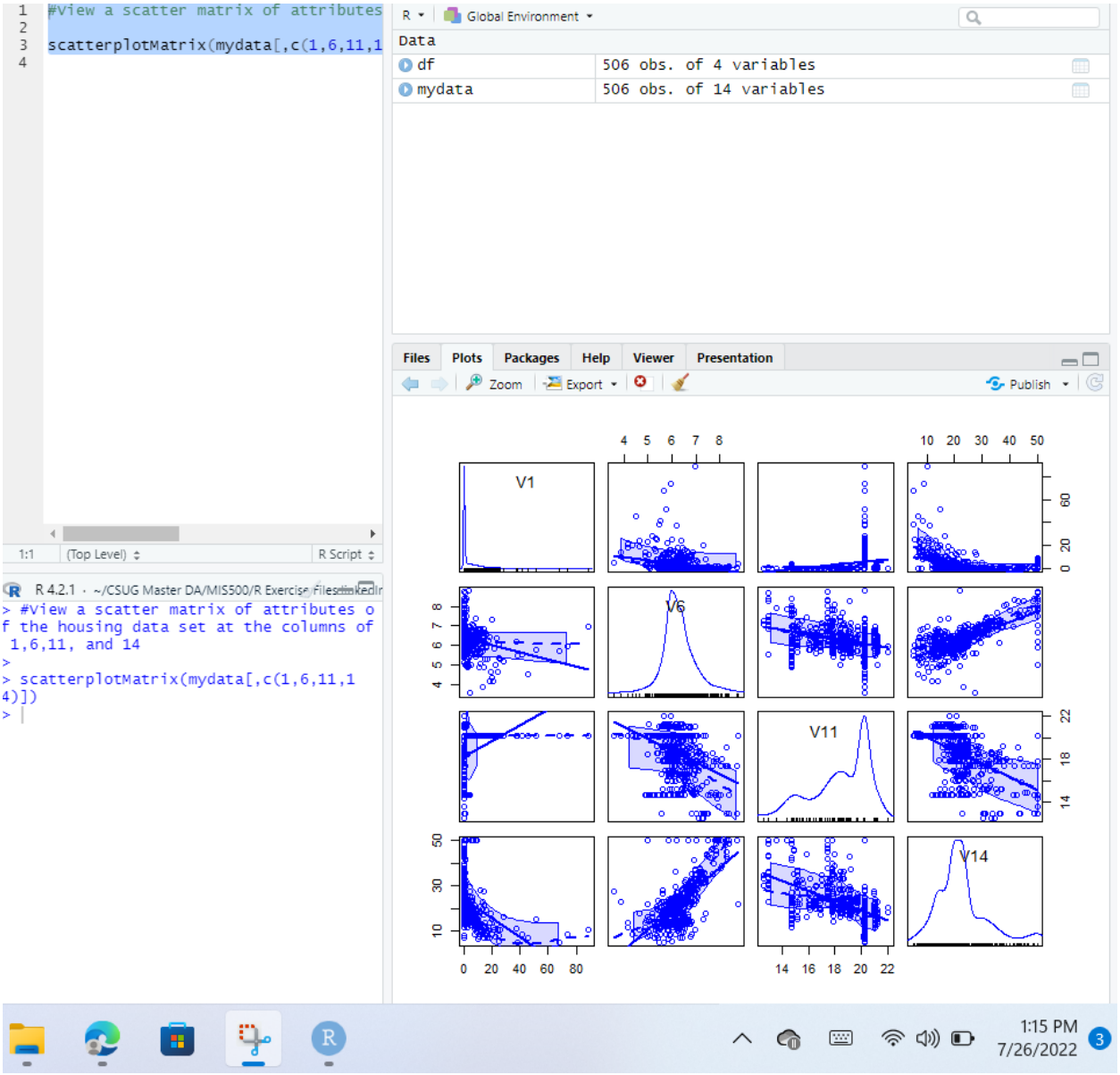
Move the next step to examine the correlation with CRIM, RM, PTRATIO, and MEDV.



The correlation coefficient evaluates the strong relationship between two variables. There are three different types of relationship negative relation means one increases, the other is a decrease, positive relation is both a boost, the last one on the relation doesn't have linear is zero. Our data set has a positive relationship between the average room and the home price of 69%, with the average of pupil-teacher and number of rooms of 29%. Additional highest negative correlation between the standard of

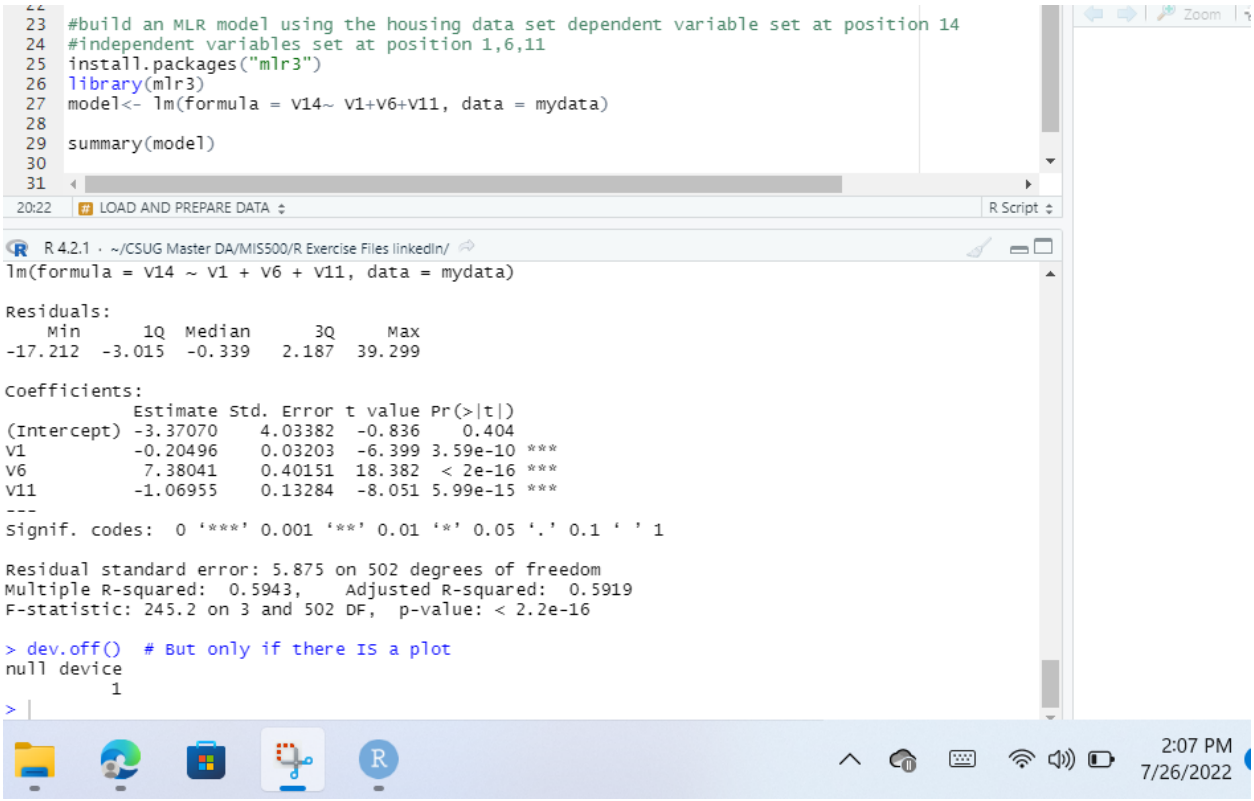
pupil-teacher and cost of a home -50% than per capita crime rate and price of a house -38% and others all negative correlations with each other except themselves.

Move on the scatter matrix to compare the relation between CRIM, RM, PTRATIO, and MEDV.



This is the proof of picture what we talk about statistic summary and correlation. The average room number has a left skew (mean<median<mode), and the price of a home has a right skew (Mode<median<mean).

The last step is applying the MLR model to find significant futures and ensuring the data fit my hypothesis. Linear regression equation is y= b0+b1*x1+b2*x2

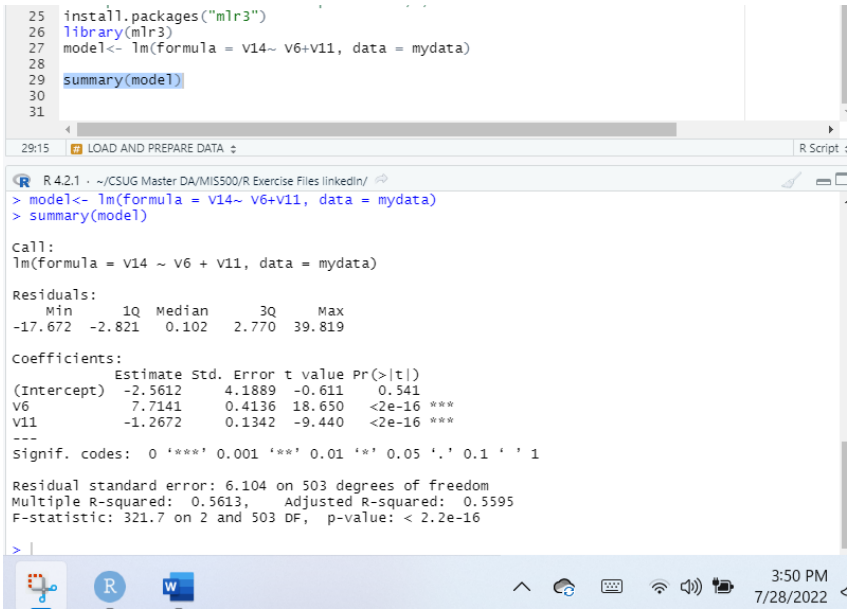


P-Value and Estimate

The multiple regression model has a few main results. One of them is p-<value that is <2.2e-16, which means one of the independent variables is seriously related to the dependent variable. Let's look at the estimate of regression in which independent variables strongly relate to the lowest p-value. The estimated shows the average increase in the dependent variable with a one-section increase independent variable or opposite decrease. The average room number strongly refers to changes in house price as V6's estimate is 7.39 and p-value<2e-16.

The CRIM is not strong enough for MLR, which means CRIM did not significantly change the house price. We can remove it from the list to see how it affects all other variable results.

Mod 2-Critical Thinking-Option 1



In the result, V6 and V11 have the p-values <2e-16e, later removed V1. MLR result formula is price of house = -2.5 + 7.71*V6 -1.26*V11. As we discussed in the correlation summary, V6 has a positive correlation, and V11 has a negative correlation; we can see the estimated result is the same.

R-Squared

R2 is one of the other results in MLR that shows how to fit independent variables with dependent variables to measure the relation between dependent and independent variables. R range from zero to a hundred is one of the significant issues it increases when adding more independent variables. Thus, the adjusted R2 result has a correction for the number of independent variables in the future model. For Boston housing data with V6 and V11 predictor variables adjusted R2=0.56, a 56% difference in the price of a house can be predicted by V6 and V11.

Reference

Linear regression using RStudio. 6 simple steps to design, run and read… | by Santiago Rodrigues Manica | Epidence | Medium

SCATTER PLOT in R programming [WITH EXAMPLES] (r-coder.com)