

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

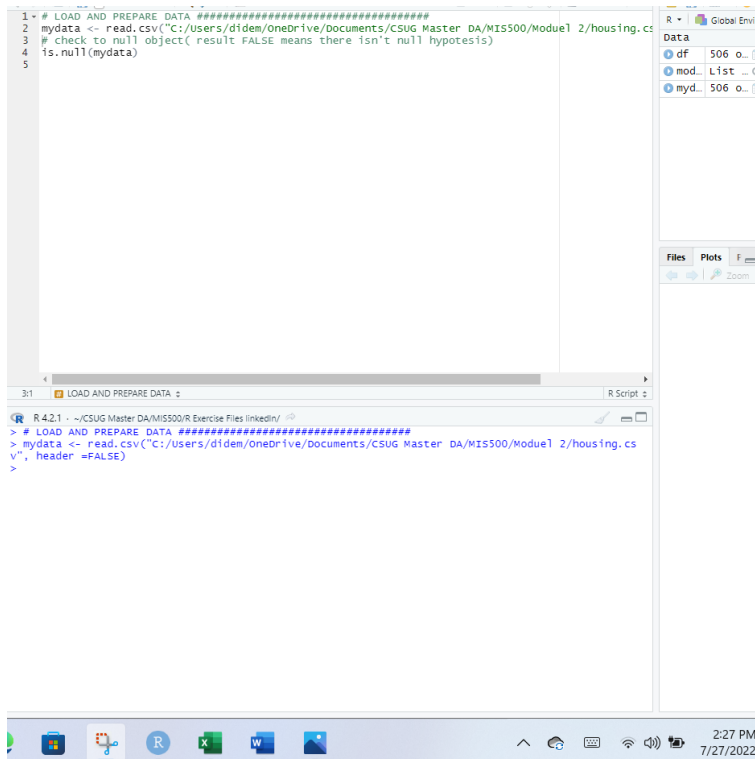
July 30, 2022

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](https://www.statlib.org/datasets/boston.html). 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 - $1000(B_k - 0.63)^2$ where B_k 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 screenshot shows the RStudio interface with the following components:

- Script Editor:** Contains R code to load and prepare the data.

```
1 # LOAD AND PREPARE DATA #####
2 mydata <- read.csv("C:/Users/didem/OneDrive/Documents/CSUG Master DA/MIS500/Module 2/housing.csv")
3 # check to null object( result FALSE means there isn't null hypothesis)
4 is.null(mydata)
5
```
- Environment:** Shows the loaded object 'mydata' with 506 observations and 13 variables.
- Files:** Shows the location of the data file: 'C:/Users/didem/OneDrive/Documents/CSUG Master DA/MIS500/Module 2/housing.csv'.
- Console:** Shows the execution of the code, confirming the file path and the number of observations (506).

Mod 2-Critical Thinking-Option 1

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.

The screenshot shows the R Studio environment. The script editor on the left contains the following code:

```
1 mydata <- read.csv("C:/Users/didem/OneDrive/Documents/CSUG Master  
2 # Display summary statistic for all columns of the housing data s  
3 summary(mydata)
```

The console on the right displays the output of the `summary(mydata)` command, showing summary statistics for 14 variables (V7 through V20). The statistics include Maximum, Minimum, 1st Quartile, Median, Mean, 3rd Quartile, and Maximum for each variable.

Variable	Max.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
V7	1.00000	2.90	45.02	77.50	68.57	94.08	100.00
V8	0.8710	1.130	2.100	3.207	3.795	5.188	12.127
V9	8.780	1.000	4.000	5.000	9.549	24.000	24.000
V10	187.0	12.60	17.40	19.05	18.46	20.20	22.00
V11	375.38	0.32	375.38	391.44	356.67	396.23	396.90
V12	711.0	1.73	6.95	11.36	12.65	16.95	37.97
V13	5.00	17.02	21.20	22.53	22.53	25.00	50.00
V14	5.00	17.02	21.20	22.53	22.53	25.00	50.00

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.

Mod 2-Critical Thinking-Option 1

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

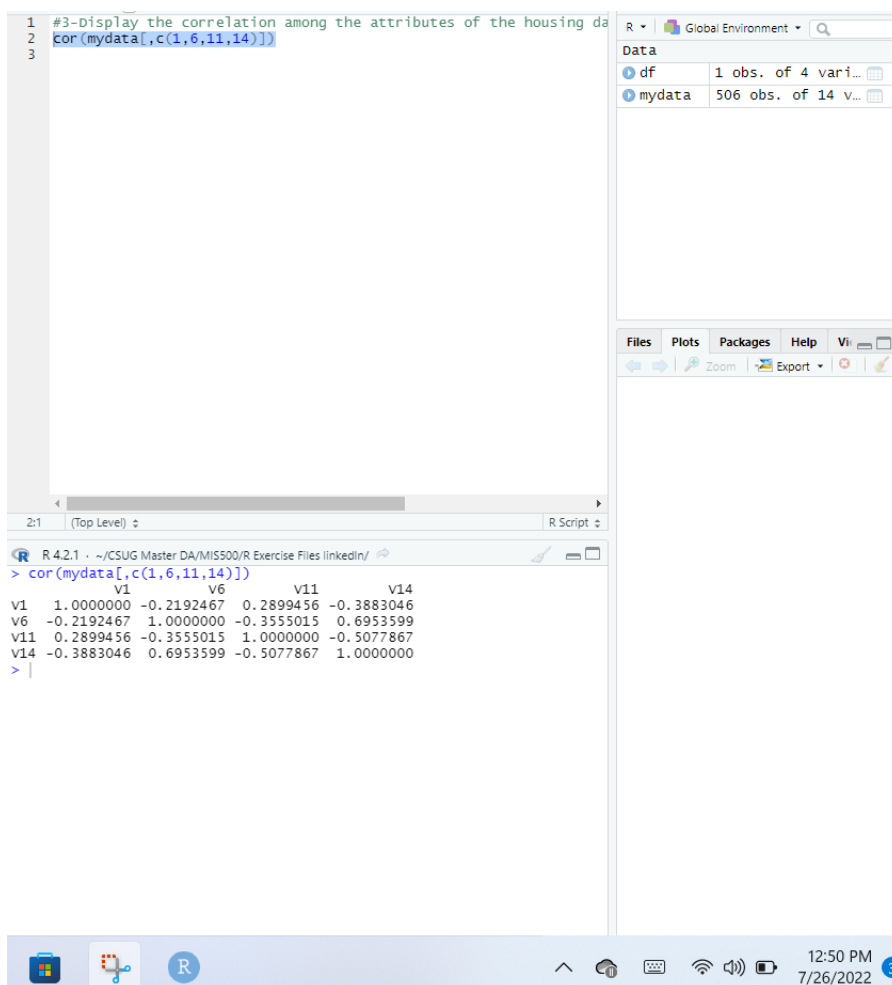
Min-Max

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 screenshot shows the R Studio environment. The script editor on the left contains the following code:

```
1 #3-Display the correlation among the attributes of the housing da
2 cor(mydata[,c(1,6,11,14)])
3
```

The console on the right displays the output of the `cor()` function:

```
> cor(mydata[,c(1,6,11,14)])
      v1      v6     v11     v14
v1  1.000000 -0.2192467 0.2899456 -0.3883046
v6  -0.2192467 1.0000000 -0.3555015 0.6953599
v11 0.2899456 -0.3555015 1.0000000 -0.5077867
v14 -0.3883046 0.6953599 -0.5077867 1.0000000
```

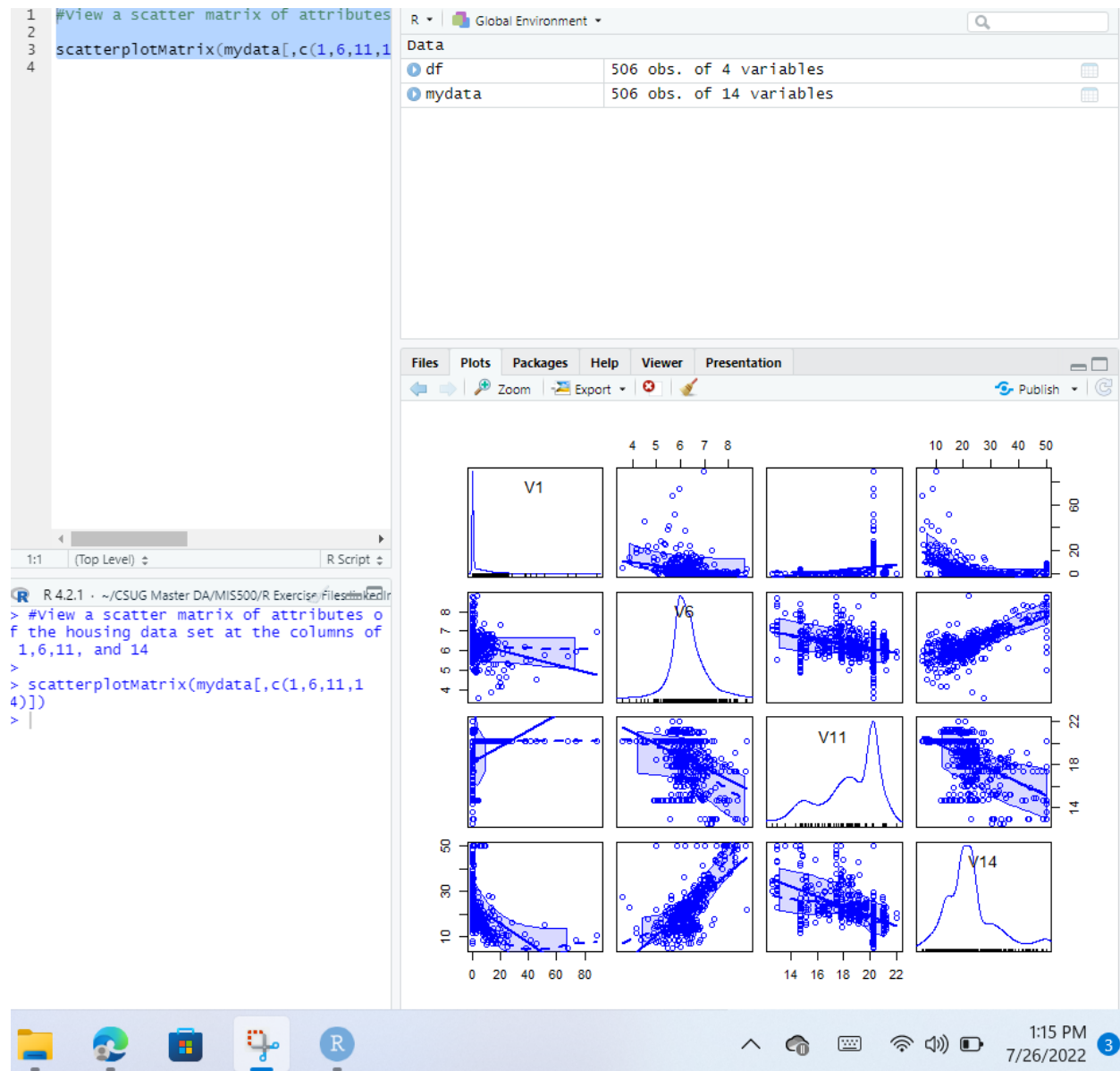
The R Studio interface also shows the 'Data' pane on the right, indicating that the data frame 'mydata' has 506 observations and 14 variables. The bottom status bar shows the date and time as 12:50 PM on 7/26/2022.

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

Mod 2-Critical Thinking-Option 1

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 $y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$

Mod 2-Critical Thinking-Option 1

```
23 #build an MLR model using the housing data set dependent variable set at position 14
24 #independent variables set at position 1,6,11
25 install.packages("mlr3")
26 library(mlr3)
27 model<- lm(formula = v14~ v1+v6+v11, data = mydata)
28
29 summary(model)
30
31
```

20:22 LOAD AND PREPARE DATA R Script

R 4.2.1 ~ /CSUG Master DA/MIS500/R Exercise Files linkedin/

```
lm(formula = v14 ~ v1 + v6 + v11, data = mydata)

Residuals:
    Min       1Q   Median       3Q      Max
-17.212  -3.015  -0.339   2.187  39.299

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.37070    4.03382  -0.836   0.404
v1          -0.20496    0.03203  -6.399 3.59e-10 ***
v6           7.38041    0.40151  18.382 < 2e-16 ***
v11         -1.06955    0.13284  -8.051 5.99e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.875 on 502 degrees of freedom
Multiple R-squared:  0.5943,    Adjusted R-squared:  0.5919
F-statistic: 245.2 on 3 and 502 DF,  p-value: < 2.2e-16

> dev.off() # But only if there is a plot
null device
1
>
```

2:07 PM 7/26/2022

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

```
25 install.packages("mlr3")
26 library(mlr3)
27 model<- lm(formula = v14~ v6+v11, data = mydata)
28
29 summary(model)
30
31
```

R 4.2.1 - ~/CSUG Master DA/MIS500/R Exercise Files/linkedlin/ | R Script

```
> model<- lm(formula = v14~ v6+v11, data = mydata)
> summary(model)
```

Call:
lm(formula = v14 ~ v6 + v11, data = mydata)

Residuals:

	Min	1Q	Median	3Q	Max
	-17.672	-2.821	0.102	2.770	39.819

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.5612	4.1889	-0.611	0.541
v6	7.7141	0.4136	18.650	<2e-16 ***
v11	-1.2672	0.1342	-9.440	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.104 on 503 degrees of freedom
Multiple R-squared: 0.5613, Adjusted R-squared: 0.5595
F-statistic: 321.7 on 2 and 503 DF, p-value: < 2.2e-16

> |

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 \[WITH EXAMPLES\] \(r-coder.com\)](#)