**Student name:** GOUTHAM SELVAKUMAR

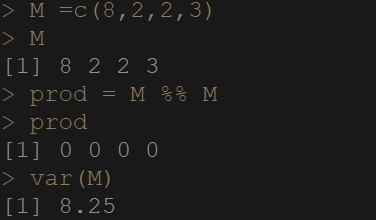
**DSC 324/424 Assignment 1  
Due: Tuesday, April 5th at 11:59pm CST**

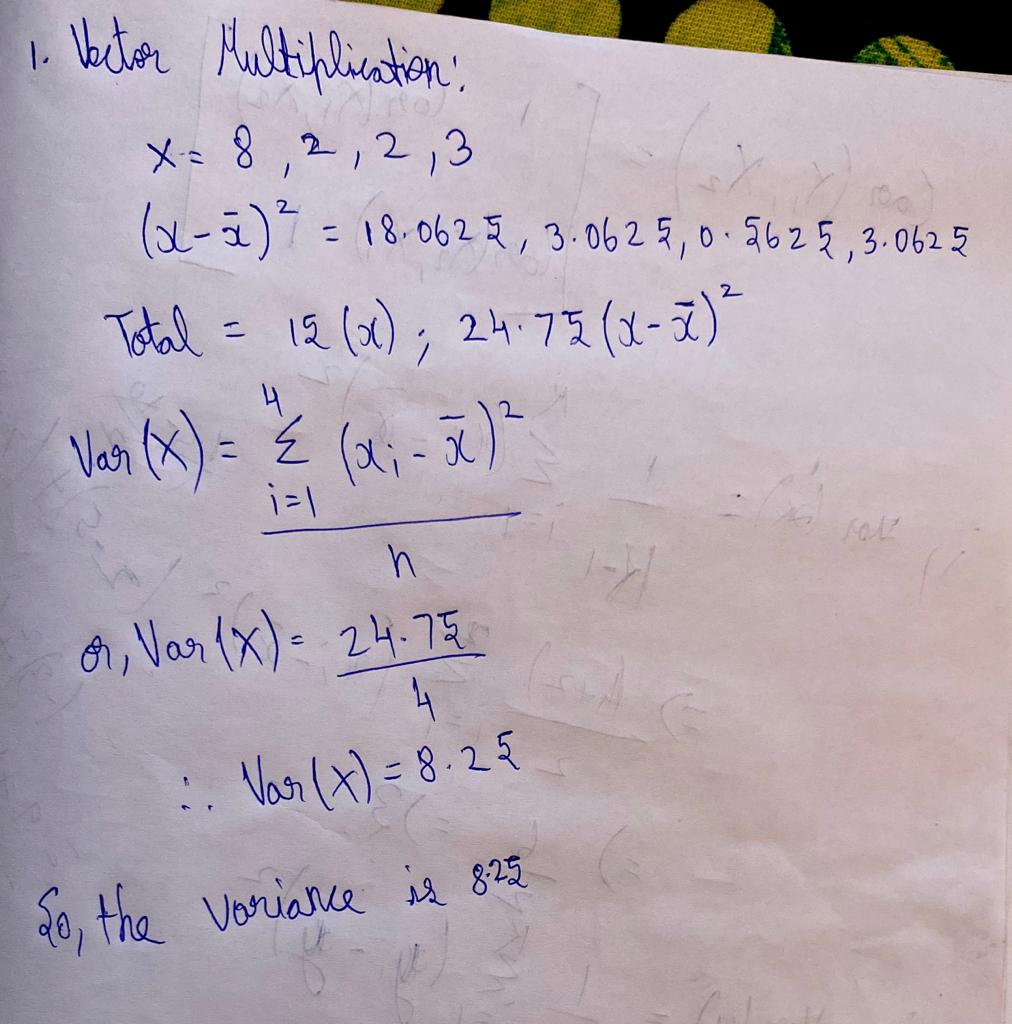
Answer all questions as completely as you can. Submit your answers to the following prompts in a single document to the Assignment 1 submission folder by the due date. **Make sure your name is clearly written at the top of your document.** Along with this document, submit a .R file with the code you used to complete the assignment. You may do any required math on paper, as long as you include *clear* screenshots or scans of your work in your final document.

The file matrix\_snippets.R on d2l has some helpful code for problems 1, 2, and 3.

**1. Vector multiplication (dot product)**: Using a dot product, calculate the variance of the following variable values by hand. Show your work and include your final answer. Use the matrix operators in R to calculate the same and confirm your result. Include a screenshot of your code and output.

[8 2 2 3]

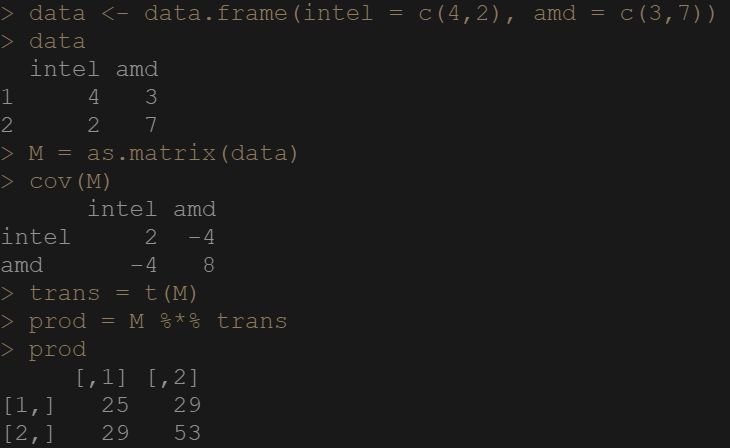
****

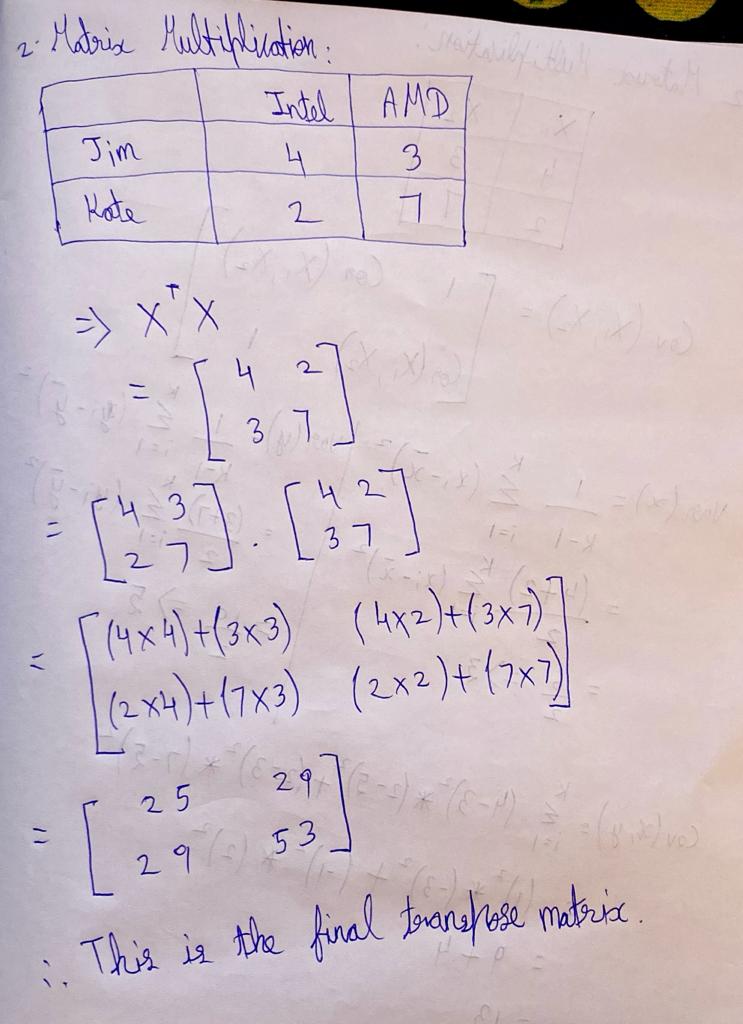
****

**2. Matrix Multiplication:** Calculate the covariance matrix of the following stock portfolios by hand. Show your work and include your final matrix:   
Covariance of all variables in a dataset = XTX

|  |  |  |
| --- | --- | --- |
|  | Intel | AMD |
| Jim | 4 | 3 |
| Kate | 2 | 7 |

Use the matrix operators in R to calculate the same and confirm your result. Include a screenshot of your code and output. Is there anything notable about this matrix?

****

****

**4. Variable selection and regression:** Use the *housing.csv* dataset to answer the following prompts concerning automatic variable selection techniques.

This Housing dataset contains housing values in the suburbs of Boston. A brief description of the variables can be found below. A more detailed explanation of the dataset can be found at the UCI machine learning repository http://archive.ics.uci.edu/ml/datasets/Housing

* CRIM: per capita crime rate by town
* ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
* INDUS: 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: proportion of owner-occupied units built prior to 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
* LSTAT: % lower status of the population
* MEDV: Median value of owner-occupied homes in $1000's (output variable)

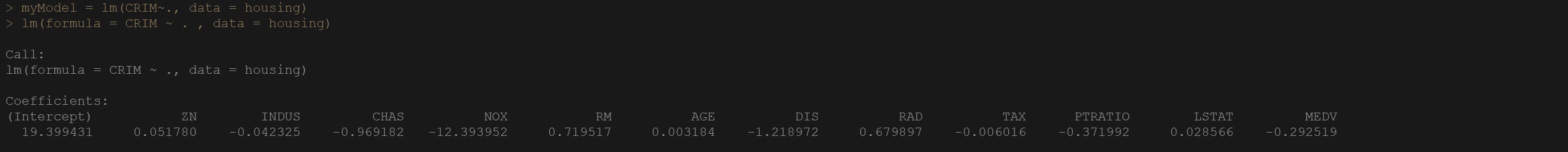
1. Fit a linear regression model of CRIM based on all the other variables and include a screenshot of the statistics generated by the lm () function.

Housing = read.csv (“housing.csv”)

Data(“housing”)

MyModel = lm(CRIM ~ . , data = housing)

lm(formula = CRIM ~ . , data = housing)

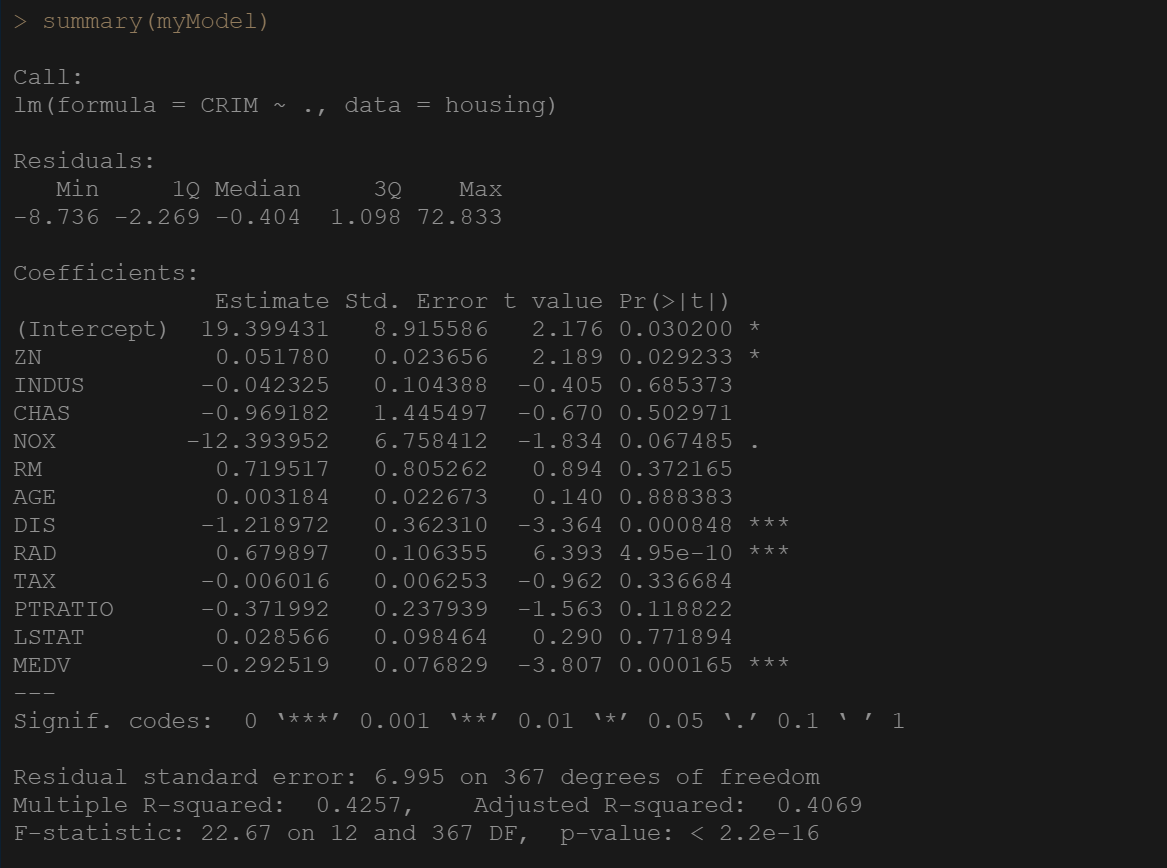


1. Evaluate this model's performance using the evaluation metrics you've learned for multiple linear regression. In your evaluation, make a case for whether or not this model is performing sufficiently well in the context of this dataset.

Housing: The predicted CRIM equation for given variables is,

**CRIM = 19.399431 + ZN\*0.051780 +INDUS\*-0.042325+CHAS\*-0.969182+NOX\*-12.393952+RM\*0.719517+AGE\*0.003184+DIS\*-1.218972+RAD\*0.679897+TAX\*-0.006016+PTRATIO\*-0.371992+LSTAT\*0.028566+MEDV\*-0.292519**

From given analysis we see that from all variables the variables which are significant to predict the per capita crime rate by town (CRIM) are ZN, DIS, RAD, and MEDV. In this Model all variables except DIS, RAD, MEDV have p-value > 0.05 which tells that these variables have no relation with CRIM in this model, thus they don’t contribute to produce fitted values for this model. Also, this model has R-Squared values of 0.4257 which tells that 42.5% of variations are there within CRIM (Response Variable) of this model with respect to variations in other independent variables, which is quite moderate and thus model is acceptable. Overall we see p-value of the model = 2.2e-16 < 0.05 so, this model accepts the alternative hypothesis and tells us that this model produces a relation with at least 1 independent variable and CRIM. So, Model is acceptable.



c. Fit two more linear regressions with the same variables, one using the forward selection and one using the backwards elimination variable selection techniques. Include screenshot of both models' statistics as you did in part a.

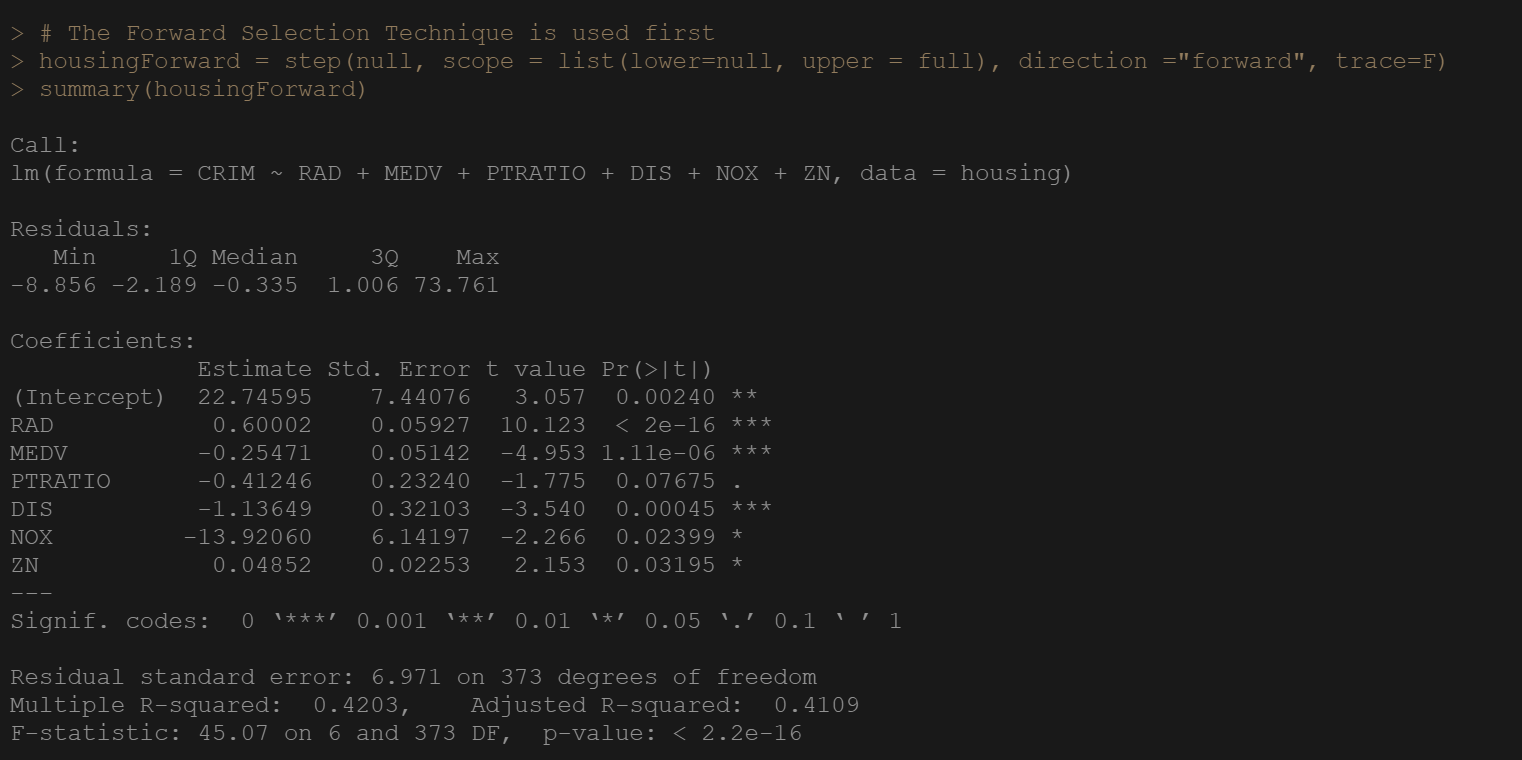
housingForward = step(null, scope = list(lower=null, upper = full), direction ="forward", trace=F)

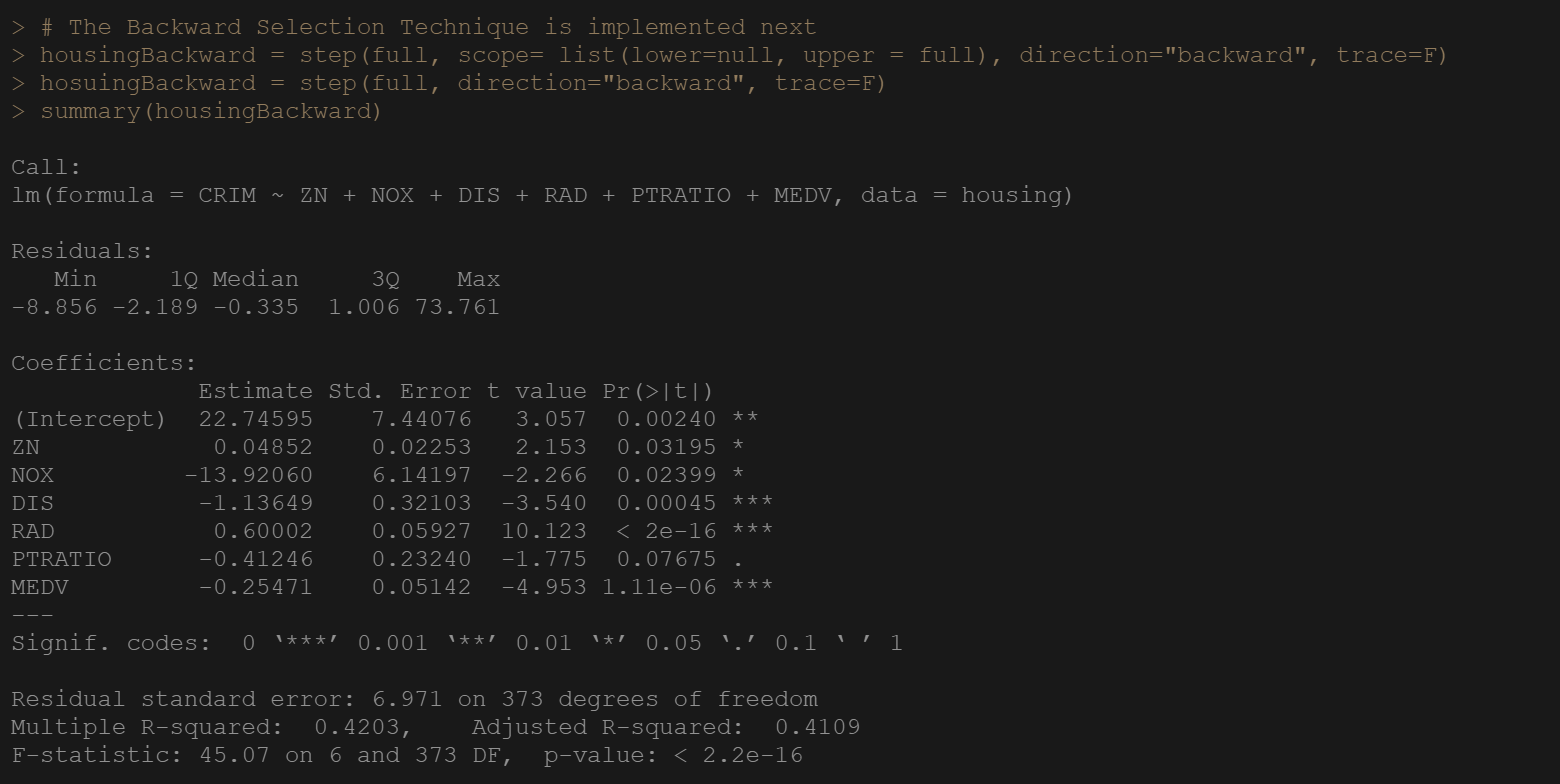
summary(housingForward)

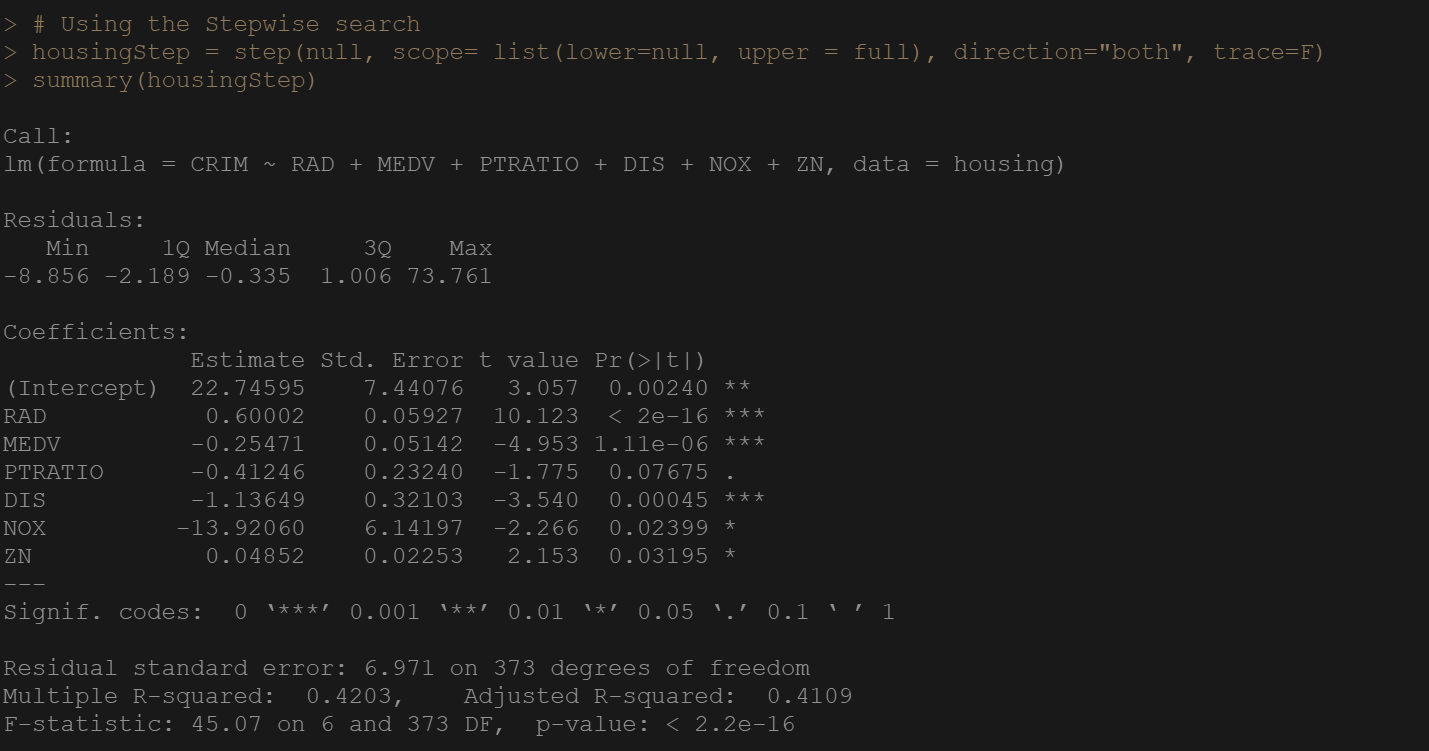
housingBackward = step(full, scope= list(lower=null, upper = full), direction="backward", trace=F)

housingBackward = step(full, direction="backward", trace=F)

summary(housingBackward)







1. Compare the variables the two selection methods decided to keep. Is there any consensus? Were there any variables kept that were not deemed statistically significant by the model in part a?

**For Backward**: The best set parametes to forsee the CRIM are : ZN, NOX, DIS, RAD, MEDV. Except the PTRATIO AND LSTAT all are very much significant to predict CRIM as we see there are p-values < 0.05, and null hypothesis are rejected accepting the alternative ones.

**For Forward:** The best set parameters to predict CRIM are: ZN, DIS, RAD, MEDV. These are much significant to predict the CRIM as we see there p-values < 0.05, and null hypothesis are rejected accepting the alternative ones. Variable black is also significant to determine CRIM but ZN, DIS, RAD, MEDV are the most significant ones to determine the CRIM parameter with proper analysis.

So since the candidate variables < sample size of the variables being used so backward model is considered good. Out of both models the backward model is more predictable since the most of the insignificant variables are removed from this model which are available Forward model which makes the model more clear and understandable, and also the Multiple R-squared value is smaller in cause of backward model. Yes, in case of Forward model all the parameters that are not significant in part a model are included in Forward model, but in backward model those variables are removed, which makes it more precise and acceptable.

1. Compare the performance of all three models and make a case for which is most appropriate to use for predicting crime rates. Consider whether or not the chosen variables are appropriate predictors for this purpose. Also consider whether or not the benefits of having a simpler model justify accepting slightly worse performance than a more complex model.

From all the three models, we see the backward model is morea cceptable ones, to predict the crime rate since as we see it’s summary, we see only 8 variables are required to predict the crime rate, whereas in other 2 models 13 varibles are there to predict the crime rate which makes the model efficient and time saving.

There are no *correct* answers, but *strong* answers will use statistics to backup your decisions.