BS803 Final Project

Reghan Lanning rkl1@bu.edu

7 November 2024

**Introduction:**

The R-file *BS803FINALR.r* contains a function *my\_lm* and test cases for a multiple linear regression model. The function uses matrix operations (built into R) and the package *glue*, which can be installedvia *Install.packages(‘glue’)*.

When calling the function *my\_lm*, it requires **three parameters:**

* *Data*: the name of the dataframe in R with the variables of interest for the linear regression. Data must be imported into R before it can be entered as a parameter into the model.
* *Outcome*: the column name in the *data* dataframe that will be the outcome variable. When inputting as a parameter, should be entered as a list:
  + If the outcome is called *Y* in the dataframe, it should be entered into the function as “Y”
* *Predictors*: the column name(s) in the *data* that will be the predictors. When inputting as a parameter, should be entered as a list:
  + If the predictors are called *X1, X2,* and *X3* in the *data*, they should be entered into the function as c('x1', 'x2', 'x3').
  + Recommended: give variable name to list (my\_predictors <- c('x1', 'x2', 'x3'))
* Example using test-case of data entry:
  + my\_lm(airquality, 'Ozone', c('Solar.R','Wind','Temp'))

**Logic**:

To initialize the function, data, outcome, and predictor(s) are inputted.

**Building working dataframe in function:**

The first step is to omit missing values from the data, by writing over variable *data* with a version that has no missing values. Then, create a vector of 1s for the intercept. The intercept is added into an empty data frame *x*. To build the data frame *x*, the list of parameters is iterated through using a *for* loop. If the predictor is either a numeric or integer column vector, it is added into the data frame *x.* At the end of the loop, a message will print stating if any predictors were ignored for being an incorrect data type. Missing values are removed from *x*. *Y* is assigned as a numeric column vector of the predictor.

**Matrix operations:**

R can perform matrix operations using transpose (t(x)), matrix multiplication (%\*%) and inverse (solve(x)). Matrix operations can compute inputs of an ANOVA table, which can compute further estimates:

’Y

**Glue package in R:**

Glue in R allows a user to print output in a manner like f-strings in Python by using brackets to indicate a variable. Additionally, it allows for character output to be printed without quotations as well. I use it within the function to print out data to resemble a table.

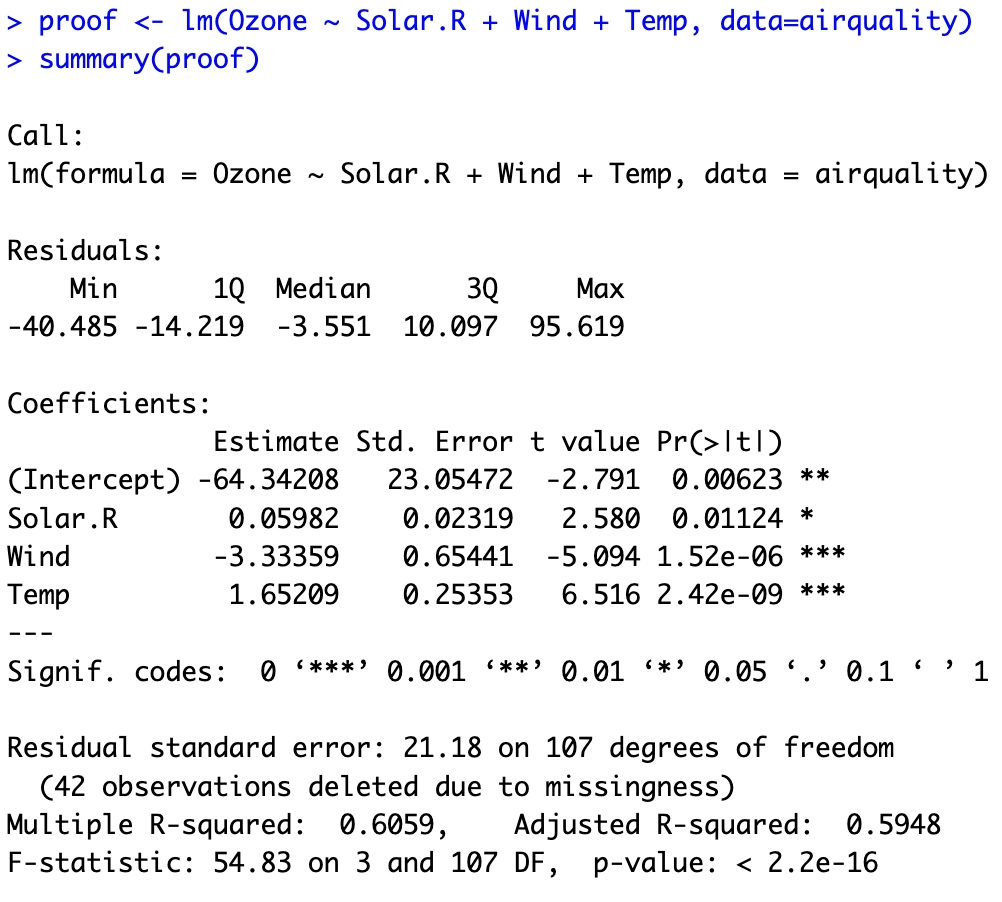
**Test Cases:**

I used the built-in dataset *airquality* in R to test this function. The first test is all numeric predictors. Values were confirmed by using the build-in *lm* and *summary* functions. The second test was randomly generating character-type values, appending that column vector into the *airquality* data set, and using the name of that column (*char\_column*) to test if the implemented warning worked.

Test case 1: all numeric predictors

A white sheet with black text

Description automatically generated



Test case 2: a character vector as one of the predictors

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

Although the two outputs are not identical, they indicate the same ideas of significance for variables. These differences could be due to different missing variables, and approximations using matrix operations.