Alabama Population Estimate

Christopher Flowers

College of Information Technology, Western Governors University

C997 Programming in R

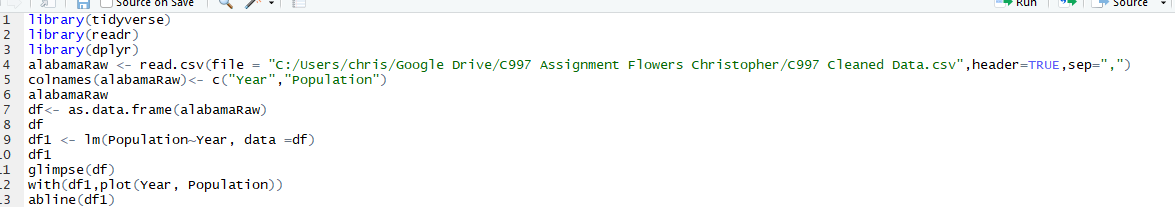
Dr. Sewell

December 27, 2020

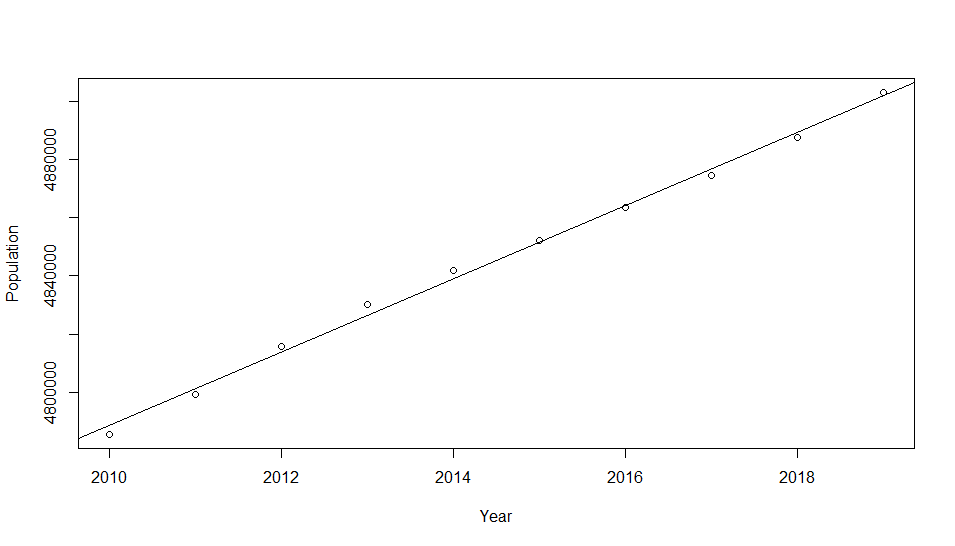
This course is programming in R. With a goal to get you familiar with programming in the language of R. R is used a lot for statistical data and for analysis of data. In this course we are to take the state we live in for me which is the great state of Alabama and look at the population from 2010 to 2019 and test its linear regression. After that we will predict the future population for the state of Alabama.

First, we need to get create a linear regression to show the population growth from 2010 to 2019. R comes lots of preinstalled libraries and just must import those libraries. The libraries used are going to be ‘tidyverse’, ‘readr’, and ‘dplyer’. After the libraries are imported let us read the csv file into a variable called alabamaRaw. Then change the columns names in the data table alabamaRaw to Year and Population. This needs to be done because when reading in the headers from the file the Year header gets missed up and after this just call the alabamaRaw to check the data. We need to take the alabamaRaw variable and turn it into a data frame and the variable will be called df. Now let us use the lm function in R, which is the linear regression function. We will set a variable called df to the lm function with two parameters inside. The first parameter is Population~Year and the second one is the dataset which is data = df1. Let us take glimpse at the data using the glimpse function to make sure everything. To plot the linear regression, we use the ‘with’ statement with the variable df1 and plot(Year, Population) as the parameters this will give the basic it graphs needed but not the linear regression. To get the linear regression we need to add the function abline with the parameter being df1. Now we have a linear regression for the state of Alabama from 2010 to 2019 and the data increases every year. The script and the results will be below.

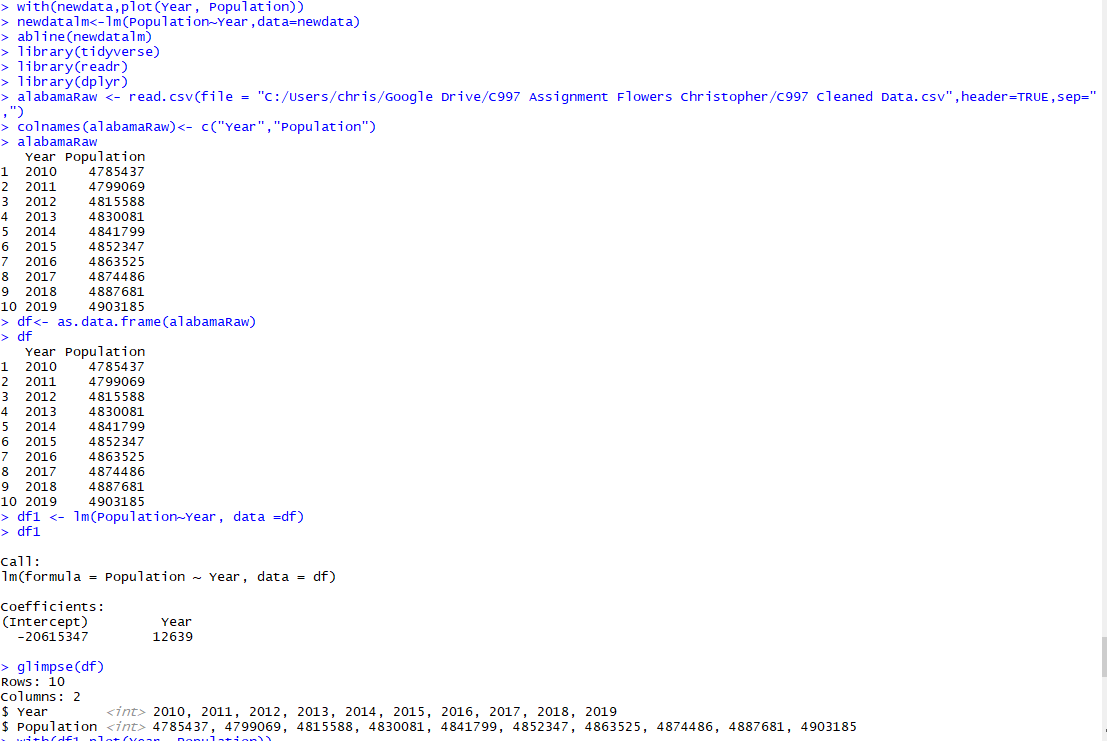
Script



Linear Regression

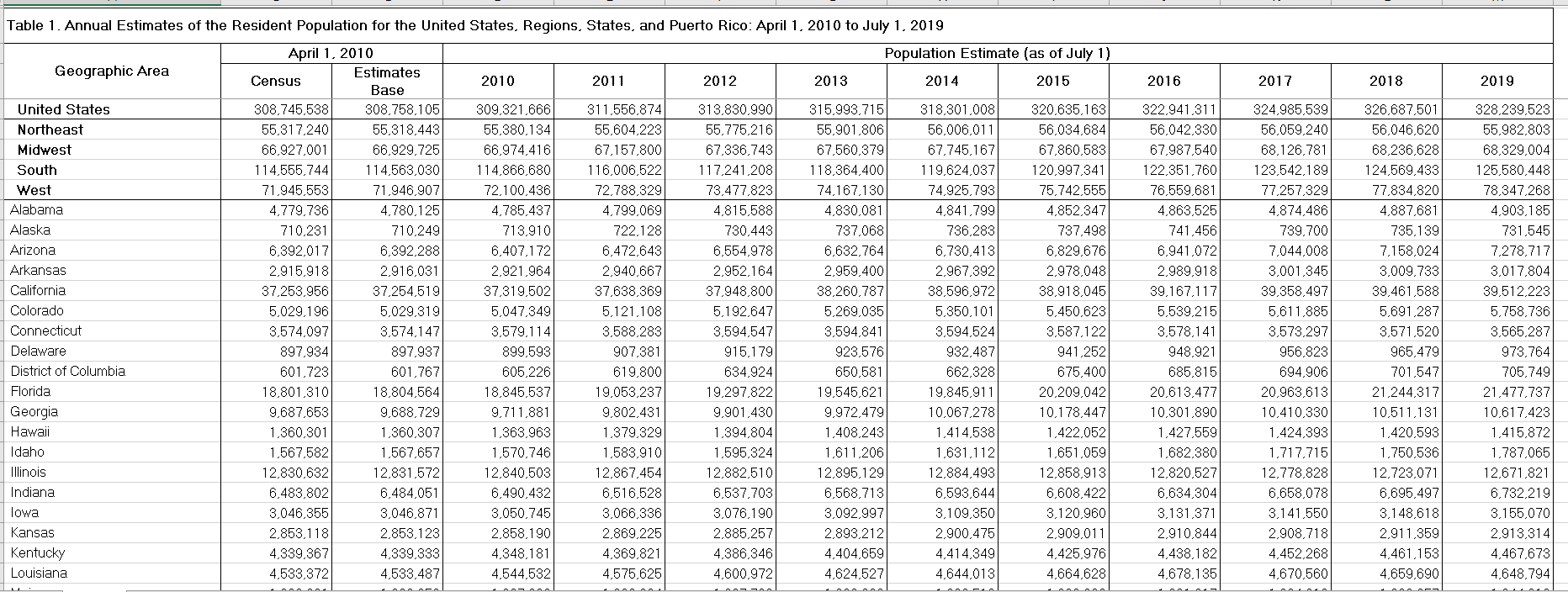


Console results

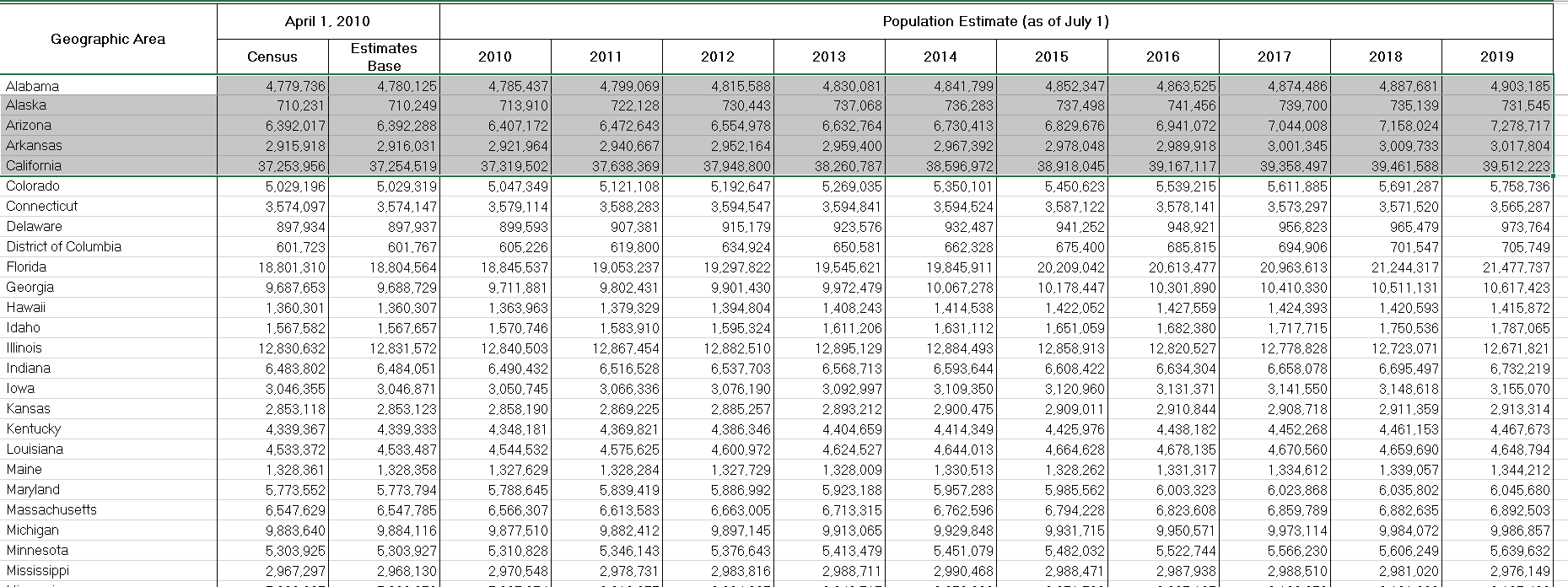


Now we need to transform the data we will do this manually in the excel spreadsheet because the data is not formatted to be imported into R. First, we need to get rid of the Table 1 Row then the United States and directional data as well. The project only wants us to worry about the state we reside so since I live in Alabama, we will only keep the Alabama row. The next thing to do is to make the columns unmerged and simpler. Remove all the state, census, estimate columns and delete row 1. Now we will transpose the data into two columns instead of two rows. Now copy the two rows with data then click on a cell in column A and paste using past special with transpose option checked. The data should be transposed starting in the cell pasted. Now to do is clean the old data and move up the new transposed data. We need to add column headers on to the columns to have headers when it is read into R and then change the first column from ‘Population 2010’ to just ‘2010’and do this for all the years listed. Lastly save the file.

Before cleaning starts



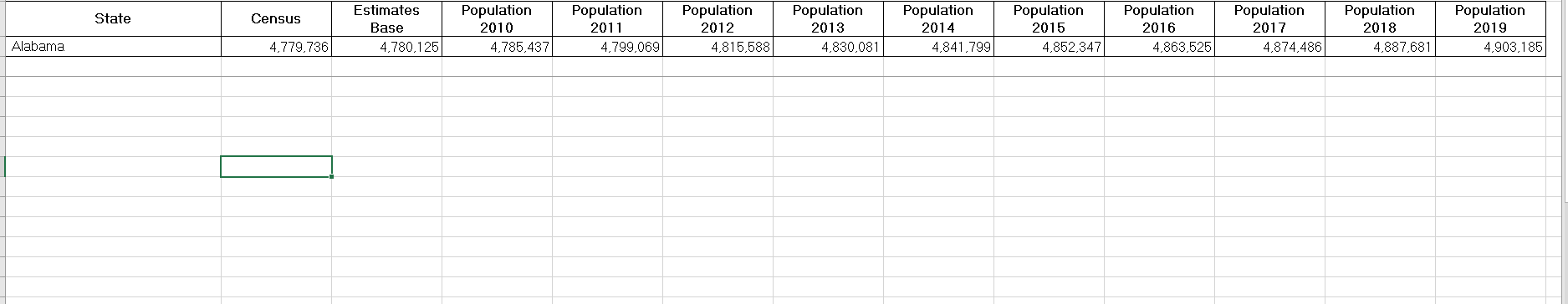
After the first part of the data transformation.



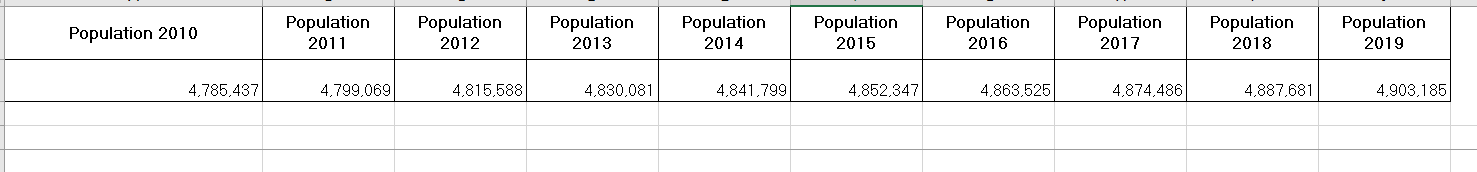
Only Alabama



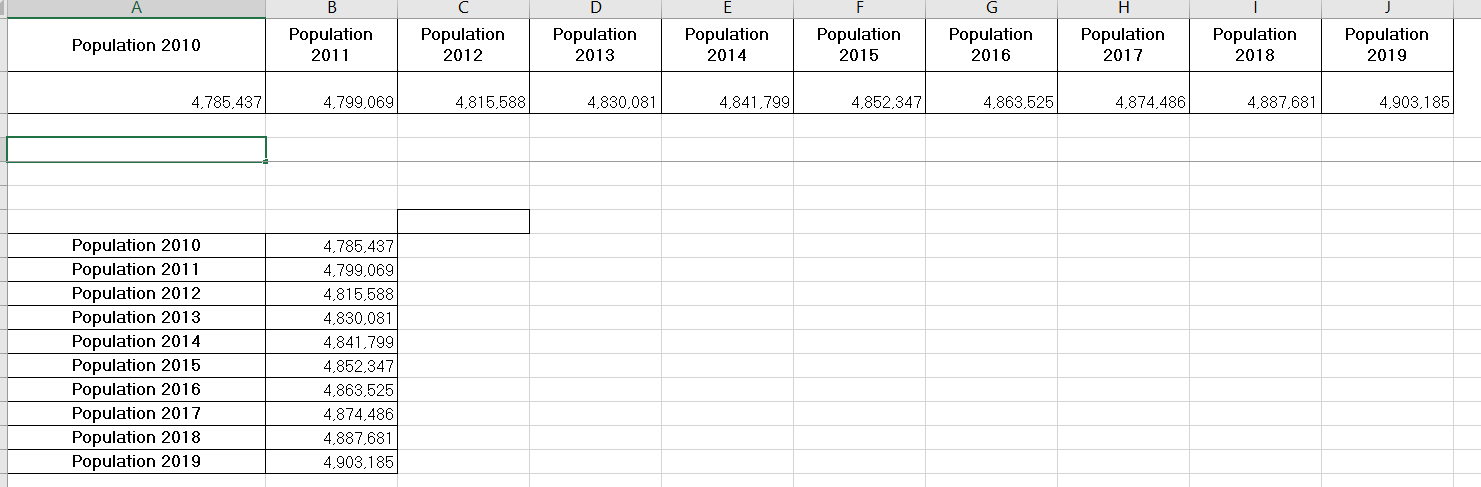
Column headers simpler



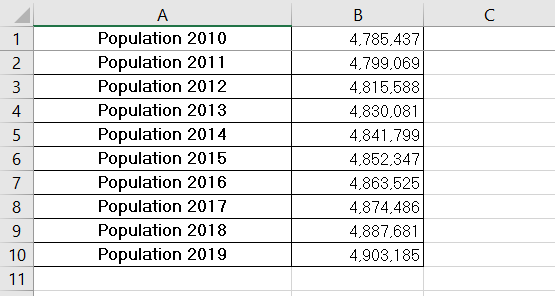
Removed State, census, estimate and delete row 1

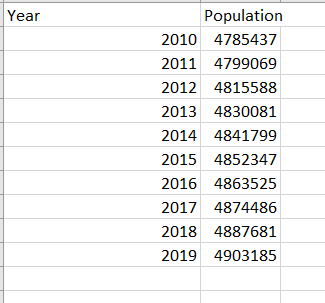


Transpose the data with original data



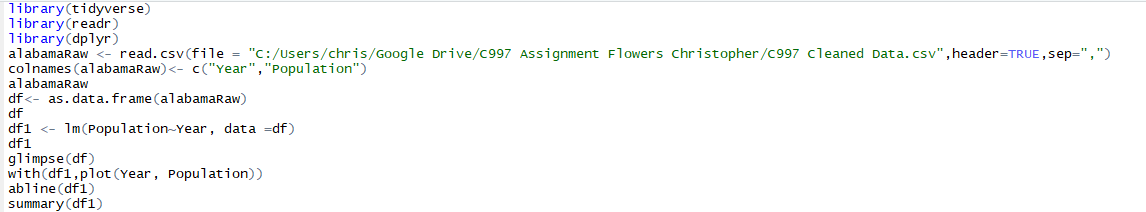
Cleaned and transposed data



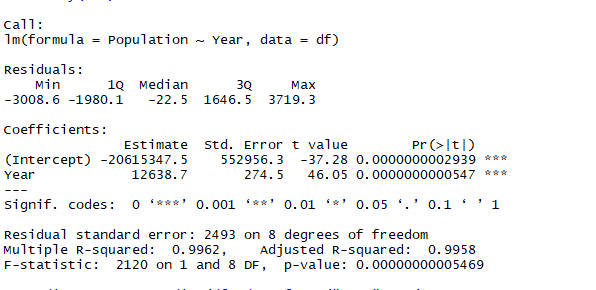


The next thing we are asked to do is provide a summary of the data with the script and results. Well R makes part of this extremely easy because the script to achieve the summary is a single function call with a single parameter. The function is the summary function with the data frame passed in as the parameter, so in this case its summary(df1) is the script. But we will take a glance at the results as well and based on the we have some low p values. Also, we have some significance on the Year and intercept on the coefficients P values.

Script



Results

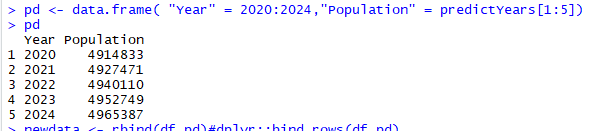


Now we want to predict where Alabama’s population will be in five years. To do this we create a variable called predictYears and set it to the predict function in R with two parameters. This function will predict the population based on our input. The parameters we will pass into the function will be df1 which is the linear regression data frame and data.frame with the parameters of Year equal to new columns “2020,2021,2022,2023,2024” then we call the variable predictYears to see the results. In a new variable called pd we will combine the years and population together, using data.frame with Years equal to 2020:2024 and the second parameter is population equal to [1:5] and call the variable to see the results. Now we need to combine both data frames together by creating a variable called newdata and using the rbind function. The rbind function combines the rows of the two data frames into one data frame with the original data frame df as one parameter and pd as the second parameter, which is the predicted data frame. Call the variable newdata to see the results of the new one. Next thing to do is take this data and create a linear regression. This will be done the same as earlier using with statement and plot with year and population as the parameters but with newdata has the dataframe instead of df1. Also creating a new linear regression as well but with newdata as the data frame here in a variable called newdatalm. The abline function this time will take the variable the newdatalm as its parameter. Below you will the script used and screenshots of the results.

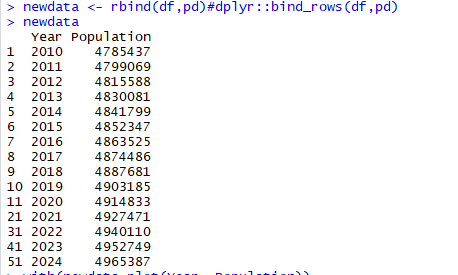
Results of the predict



Results of year and Population



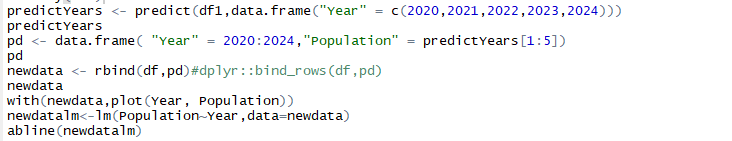
Combined data frame



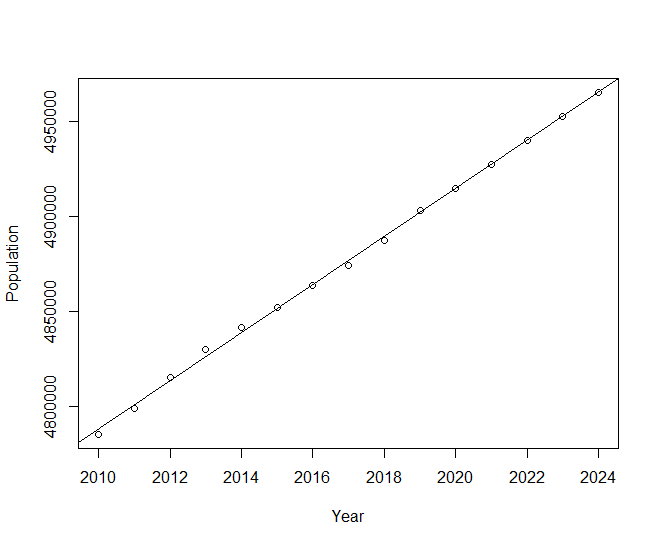
Console of the linear regression



Script



Linear Regression through 2010 through 2024



This course is programming in R, where we are tasked to get familiar with the R language. R is primarily used to for statistical analysis amongst other things. In this course we were to take the population size of the state we reside in, which in case this is Alabama from 2010 to 2019. We are clean the new data spreadsheet the data is on into a new csv file. We read this file into R and change the column names. Then change the table into a data frame and take the linear regression of the data frame then plot the data with a linear regression line, this case Alabama steadily increase and is fitted data. Also, we want a statistical summary of the data to analysis the data. Next, we want to predict the population of Alabama for the next five years and combine this with the original data frame into a single data frame. Then we will take the linear regression of this data and plot with the linear regression line. Now we are familiar programming in the language R.