Alabama Population Estimate

Christopher Flowers

College of Information Technology, Western Governors University

C997 Programming in R

Dr. Sewell

December 27, 2020

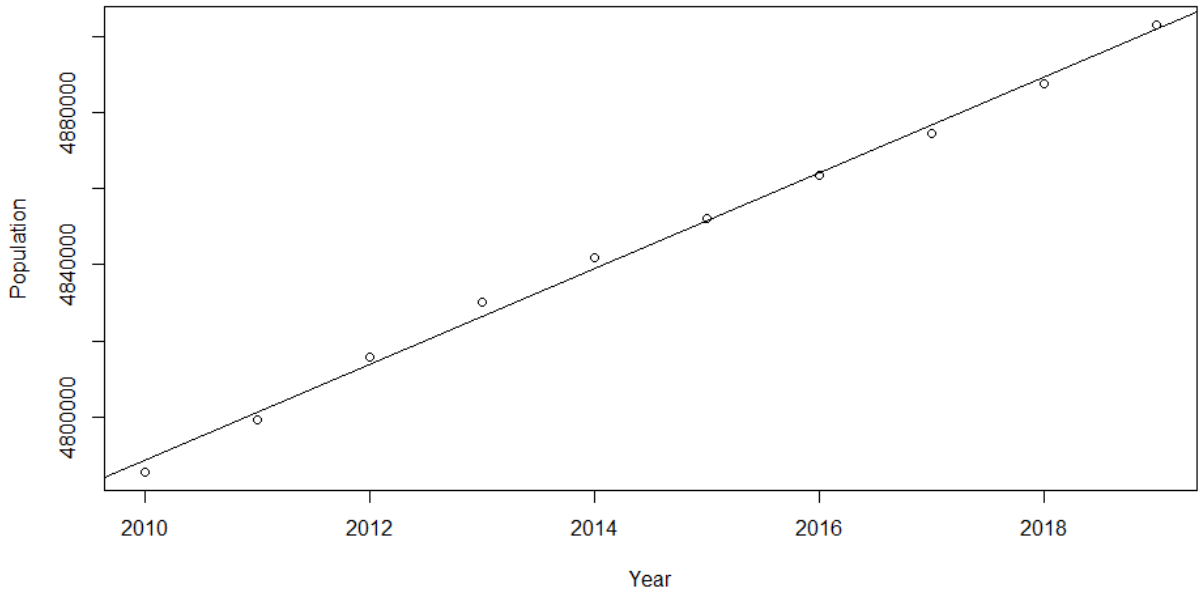
This course is programming in R. With a goal to get you familiar with programming in the language of R. R is used a lot for statistical data and for analysis of data. In this course we are to take the state we live in for me which is the great state of Alabama and look at the population from 2010 to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

```r
library(tidyverse)
library(readr)
library(dplyr)
alabamaRaw <- read.csv(file = "C:/Users/chris/Google Drive/C997 Assignment Flowers Christopher/C997 Cleaned Data.csv",header=TRUE,sep=",")
colnames(alabamaRaw)<- c("Year","Population")
alabamaRaw
df<- as.data.frame(alabamaRaw)
df
df1 <- lm(Population~Year, data =df)
df1
glimpse(df)
with(df1,plot(Year, Population))
abline(df1)
```

Linear Regression



Console results

```
> with(newdata,plot(Year, Population))
> newdatalm<-lm(Population~Year,data=newdata)
> abline(newdatalm)
> library(tidyverse)
> library(readr)
> library(dplyr)
> alabamaRaw <- read.csv(file = "C:/Users/chris/Google Drive/C997 Assignment Flowers Christopher/C997 Cleaned Data.csv",header=TRUE,sep="
,")
> colnames(alabamaRaw)<- c("Year","Population")
> alabamaRaw
   Year Population
1  2010    4785437
2  2011    4799069
3  2012    4815588
4  2013    4830081
5  2014    4841799
6  2015    4852347
7  2016    4863525
8  2017    4874486
9  2018    4887681
10 2019    4903185
> df<- as.data.frame(alabamaRaw)
> df
   Year Population
1  2010    4785437
2  2011    4799069
3  2012    4815588
4  2013    4830081
5  2014    4841799
6  2015    4852347
7  2016    4863525
8  2017    4874486
9  2018    4887681
10 2019    4903185
> df1 <- lm(Population~Year, data =df)
> df1

Call:
lm(formula = Population ~ Year, data = df)

Coefficients:
(Intercept)         Year
  -20615347        12639

> glimpse(df)
Rows: 10
Columns: 2
$ Year       <int> 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019
$ Population <int> 4785437, 4799069, 4815588, 4830081, 4841799, 4852347, 4863525, 4874486, 4887681, 4903185
```

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

Table 1. Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019

Geographic Area	April 1, 2010		Population Estimate (as of July 1)									
	Census	Estimates Base	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
United States	308,745,538	308,758,105	309,321,666	311,556,874	313,830,990	315,993,715	318,301,008	320,635,163	322,941,311	324,985,539	326,687,501	328,239,523
Northeast	55,317,240	55,318,443	55,380,134	55,604,223	55,775,216	55,901,806	56,006,011	56,034,684	56,042,330	56,059,240	56,046,620	55,982,803
Midwest	66,927,001	66,929,725	66,974,416	67,157,800	67,336,743	67,560,379	67,745,167	67,860,583	67,987,540	68,126,781	68,236,628	68,329,004
South	114,555,744	114,563,030	114,866,680	116,006,522	117,241,208	118,364,400	119,624,037	120,997,341	122,351,760	123,542,189	124,569,433	125,580,448
West	71,945,553	71,946,907	72,100,436	72,788,329	73,477,823	74,167,130	74,925,793	75,742,555	76,559,681	77,257,329	77,834,820	78,347,268
Alabama	4,779,736	4,780,125	4,785,437	4,799,069	4,815,588	4,830,081	4,841,799	4,852,347	4,863,525	4,874,486	4,887,681	4,903,185
Alaska	710,231	710,249	713,910	722,128	730,443	737,068	736,283	737,498	741,456	739,700	735,139	731,545
Arizona	6,392,017	6,392,288	6,407,172	6,472,643	6,554,978	6,632,764	6,730,413	6,829,676	6,941,072	7,044,008	7,158,024	7,278,717
Arkansas	2,915,918	2,916,031	2,921,964	2,940,667	2,952,164	2,959,400	2,967,392	2,978,048	2,989,918	3,001,345	3,009,733	3,017,804
California	37,253,956	37,254,519	37,319,502	37,638,369	37,948,800	38,260,787	38,596,972	38,918,045	39,167,117	39,358,497	39,461,588	39,512,223
Colorado	5,029,196	5,029,319	5,047,349	5,121,108	5,192,647	5,269,035	5,350,101	5,450,623	5,539,215	5,611,885	5,691,287	5,758,736
Connecticut	3,574,097	3,574,147	3,579,114	3,588,283	3,594,547	3,594,841	3,594,524	3,587,122	3,578,141	3,573,297	3,571,520	3,565,287
Delaware	897,934	897,937	899,593	907,381	915,179	923,576	932,487	941,252	948,921	956,823	965,479	973,764
District of Columbia	601,723	601,767	605,226	619,800	634,924	650,581	662,328	675,400	685,815	694,906	701,547	705,749
Florida	18,801,310	18,804,564	18,845,537	19,053,237	19,297,822	19,545,621	19,845,911	20,209,042	20,613,477	20,963,613	21,244,317	21,477,737
Georgia	9,687,653	9,688,729	9,711,881	9,802,431	9,901,430	9,972,479	10,067,278	10,178,447	10,301,890	10,410,330	10,511,131	10,617,423
Hawaii	1,360,301	1,360,307	1,363,963	1,379,329	1,394,804	1,408,243	1,414,538	1,422,052	1,427,559	1,424,393	1,420,593	1,415,872
Idaho	1,567,582	1,567,657	1,570,746	1,583,910	1,595,324	1,611,206	1,631,112	1,651,059	1,682,380	1,717,715	1,750,536	1,787,065
Illinois	12,830,632	12,831,572	12,840,503	12,867,454	12,882,510	12,895,129	12,884,493	12,858,913	12,820,527	12,778,828	12,723,071	12,671,821
Indiana	6,483,802	6,484,051	6,490,432	6,516,528	6,537,703	6,568,713	6,593,644	6,608,422	6,634,304	6,658,078	6,695,497	6,732,219
Iowa	3,046,355	3,046,871	3,050,745	3,066,336	3,076,190	3,092,997	3,109,350	3,120,960	3,131,371	3,141,550	3,148,618	3,155,070
Kansas	2,853,118	2,853,123	2,858,190	2,869,225	2,885,257	2,893,212	2,900,475	2,909,011	2,910,844	2,908,718	2,911,359	2,913,314
Kentucky	4,339,367	4,339,333	4,348,181	4,369,821	4,386,346	4,404,659	4,414,349	4,425,976	4,438,182	4,452,268	4,461,153	4,467,673
Louisiana	4,533,372	4,533,487	4,544,532	4,575,625	4,600,972	4,624,527	4,644,013	4,664,628	4,678,135	4,670,560	4,659,690	4,648,794

After the first part of the data transformation.

Geographic Area	April 1, 2010		Population Estimate (as of July 1)									
	Census	Estimates Base	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Alabama	4,779,736	4,780,125	4,785,437	4,799,069	4,815,588	4,830,081	4,841,799	4,852,347	4,863,525	4,874,486	4,887,681	4,903,185
Alaska	710,231	710,249	713,910	722,128	730,443	737,068	736,283	737,498	741,456	739,700	735,139	731,545
Arizona	6,392,017	6,392,288	6,407,172	6,472,643	6,554,978	6,632,764	6,730,413	6,829,676	6,941,072	7,044,008	7,158,024	7,278,717
Arkansas	2,915,918	2,916,031	2,921,964	2,940,667	2,952,164	2,959,400	2,967,392	2,978,048	2,989,918	3,001,345	3,009,733	3,017,804
California	37,253,956	37,254,519	37,319,502	37,638,369	37,948,800	38,260,787	38,596,972	38,918,045	39,167,117	39,358,497	39,461,588	39,512,223
Colorado	5,029,196	5,029,319	5,047,349	5,121,108	5,192,647	5,269,035	5,350,101	5,450,623	5,539,215	5,611,885	5,691,287	5,758,736
Connecticut	3,574,097	3,574,147	3,579,114	3,588,283	3,594,547	3,594,841	3,594,524	3,587,122	3,578,141	3,573,297	3,571,520	3,565,287
Delaware	897,934	897,937	899,593	907,381	915,179	923,576	932,487	941,252	948,921	956,823	965,479	973,764
District of Columbia	601,723	601,767	605,226	619,800	634,924	650,581	662,328	675,400	685,815	694,906	701,547	705,749
Florida	18,801,310	18,804,564	18,845,537	19,053,237	19,297,822	19,545,621	19,845,911	20,209,042	20,613,477	20,963,613	21,244,317	21,477,737
Georgia	9,687,653	9,688,729	9,711,881	9,802,431	9,901,430	9,972,479	10,067,278	10,178,447	10,301,890	10,410,330	10,511,131	10,617,423
Hawaii	1,360,301	1,360,307	1,363,963	1,379,329	1,394,804	1,408,243	1,414,538	1,422,052	1,427,559	1,424,393	1,420,593	1,415,872
Idaho	1,567,582	1,567,657	1,570,746	1,583,910	1,595,324	1,611,206	1,631,112	1,651,059	1,682,380	1,717,715	1,750,536	1,787,065
Illinois	12,830,632	12,831,572	12,840,503	12,867,454	12,882,510	12,895,129	12,884,493	12,858,913	12,820,527	12,778,828	12,723,071	12,671,821
Indiana	6,483,802	6,484,051	6,490,432	6,516,528	6,537,703	6,568,713	6,593,644	6,608,422	6,634,304	6,658,078	6,695,497	6,732,219
Iowa	3,046,355	3,046,871	3,050,745	3,066,336	3,076,190	3,092,997	3,109,350	3,120,960	3,131,371	3,141,550	3,148,618	3,155,070
Kansas	2,853,118	2,853,123	2,858,190	2,869,225	2,885,257	2,893,212	2,900,475	2,909,011	2,910,844	2,908,718	2,911,359	2,913,314
Kentucky	4,339,367	4,339,333	4,348,181	4,369,821	4,386,346	4,404,659	4,414,349	4,425,976	4,438,182	4,452,268	4,461,153	4,467,673
Louisiana	4,533,372	4,533,487	4,544,532	4,575,625	4,600,972	4,624,527	4,644,013	4,664,628	4,678,135	4,670,560	4,659,690	4,648,794
Maine	1,328,361	1,328,358	1,327,629	1,328,284	1,327,729	1,328,009	1,330,513	1,328,262	1,331,317	1,334,612	1,339,057	1,344,212
Maryland	5,773,552	5,773,794	5,788,645	5,839,419	5,886,992	5,923,188	5,957,283	5,985,562	6,003,323	6,023,868	6,035,802	6,045,680
Massachusetts	6,547,629	6,547,785	6,566,307	6,613,583	6,663,005	6,713,315	6,762,596	6,794,228	6,823,608	6,859,789	6,882,635	6,892,503
Michigan	9,883,640	9,884,116	9,877,510	9,882,412	9,897,145	9,913,065	9,929,848	9,931,715	9,950,571	9,973,114	9,984,072	9,986,857
Minnesota	5,303,925	5,303,927	5,310,828	5,346,143	5,376,643	5,413,479	5,451,079	5,482,032	5,522,744	5,566,230	5,606,249	5,639,632
Mississippi	2,967,297	2,968,130	2,970,548	2,978,731	2,983,816	2,988,711	2,990,468	2,988,471	2,987,938	2,988,510	2,981,020	2,976,149

Only Alabama

Geographic Area	April 1, 2010		Population Estimate (as of July 1)									
	Census	Estimates Base	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Alabama	4,779,736	4,780,125	4,785,437	4,799,069	4,815,588	4,830,081	4,841,799	4,852,347	4,863,525	4,874,486	4,887,681	4,903,185

Column headers simpler

State	Census	Estimates Base	Population 2010	Population 2011	Population 2012	Population 2013	Population 2014	Population 2015	Population 2016	Population 2017	Population 2018	Population 2019
Alabama	4,779,736	4,780,125	4,785,437	4,799,069	4,815,588	4,830,081	4,841,799	4,852,347	4,863,525	4,874,486	4,887,681	4,903,185

Removed State, census, estimate and delete row 1

Population 2010	Population 2011	Population 2012	Population 2013	Population 2014	Population 2015	Population 2016	Population 2017	Population 2018	Population 2019
4,785,437	4,799,069	4,815,588	4,830,081	4,841,799	4,852,347	4,863,525	4,874,486	4,887,681	4,903,185

Transpose the data with original data

A	B	C	D	E	F	G	H	I	J
Population 2010	Population 2011	Population 2012	Population 2013	Population 2014	Population 2015	Population 2016	Population 2017	Population 2018	Population 2019
4,785,437	4,799,069	4,815,588	4,830,081	4,841,799	4,852,347	4,863,525	4,874,486	4,887,681	4,903,185

Population 2010	4,785,437
Population 2011	4,799,069
Population 2012	4,815,588
Population 2013	4,830,081
Population 2014	4,841,799
Population 2015	4,852,347
Population 2016	4,863,525
Population 2017	4,874,486
Population 2018	4,887,681
Population 2019	4,903,185

Cleaned and transposed data

	A	B	C
1	Population 2010	4,785,437	
2	Population 2011	4,799,069	
3	Population 2012	4,815,588	
4	Population 2013	4,830,081	
5	Population 2014	4,841,799	
6	Population 2015	4,852,347	
7	Population 2016	4,863,525	
8	Population 2017	4,874,486	
9	Population 2018	4,887,681	
10	Population 2019	4,903,185	
11			

Year	Population
2010	4785437
2011	4799069
2012	4815588
2013	4830081
2014	4841799
2015	4852347
2016	4863525
2017	4874486
2018	4887681
2019	4903185

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

```
library(tidyverse)
library(readr)
library(dplyr)
alabamaRaw <- read.csv(file = "C:/Users/chris/Google Drive/C997 Assignment Flowers Christopher/C997 Cleaned Data.csv",header=TRUE,sep=",")
colnames(alabamaRaw)<- c("Year","Population")
alabamaRaw
df<- as.data.frame(alabamaRaw)
df
df1 <- lm(Population~Year, data =df)
df1
glimpse(df)
with(df1,plot(Year, Population))
abline(df1)
summary(df1)
```

Results

```
Call:
lm(formula = Population ~ Year, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3008.6 -1980.1   -22.5  1646.5  3719.3

Coefficients:
              Estimate  Std. Error t value        Pr(>|t|)
(Intercept) -20615347.5    552956.3  -37.28 0.0000000002939 ***
Year            12638.7       274.5   46.05 0.0000000000547 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2493 on 8 degrees of freedom
Multiple R-squared:  0.9962,    Adjusted R-squared:  0.9958
F-statistic:  2120 on 1 and 8 DF,  p-value: 0.00000000005469
```

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

```
> predictYears <- predict(df1,data.frame("Year" = c(2020,2021,2022,2023,2024)))
> predictYears
      1       2       3       4       5
4914833 4927471 4940110 4952749 4965387
```

Results of year and Population

```
> pd <- data.frame( "Year" = 2020:2024,"Population" = predictYears[1:5])
> pd
  Year Population
1 2020    4914833
2 2021    4927471
3 2022    4940110
4 2023    4952749
5 2024    4965387
> newdata <- rbind(df,pd)#dplyr::bind_rows(df,pd)
```

Combined data frame

```
> newdata <- rbind(df,pd)#dplyr::bind_rows(df,pd)
> newdata
   Year Population
1  2010    4785437
2  2011    4799069
3  2012    4815588
4  2013    4830081
5  2014    4841799
6  2015    4852347
7  2016    4863525
8  2017    4874486
9  2018    4887681
10 2019    4903185
11 2020    4914833
21 2021    4927471
31 2022    4940110
41 2023    4952749
51 2024    4965387
```

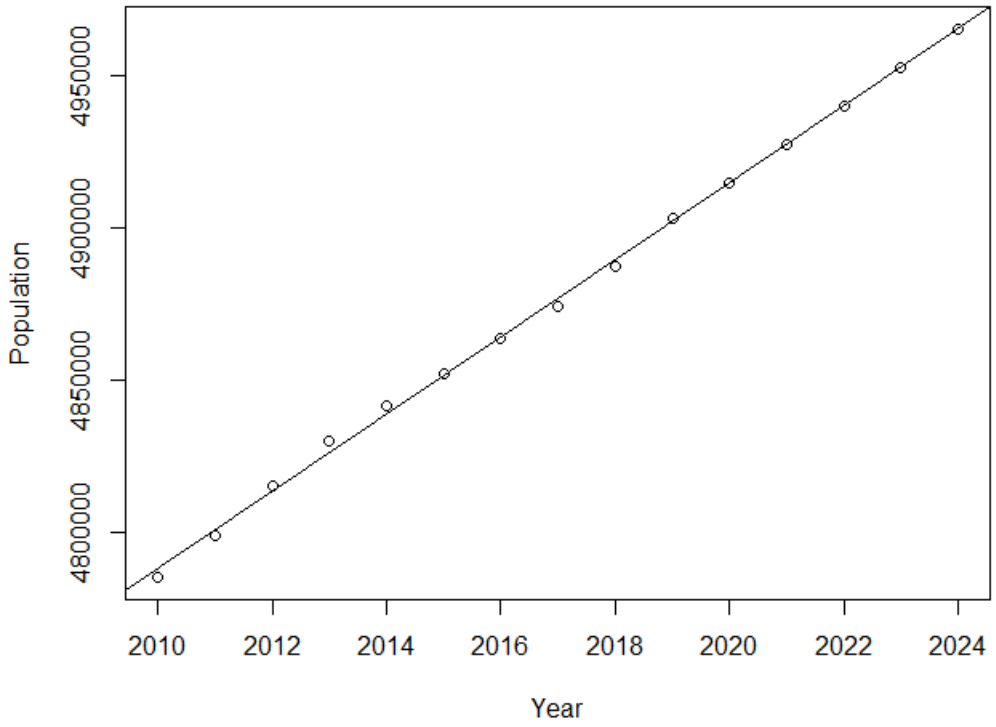
Console of the linear regression

```
51 2024    4965387
> with(newdata,plot(Year, Population))
> newdatalm<-lm(Population~Year,data=newdata)
> abline(newdatalm)
>
```

Script

```
predictYears <- predict(df1,data.frame("Year" = c(2020,2021,2022,2023,2024)))
predictYears
pd <- data.frame( "Year" = 2020:2024,"Population" = predictYears[1:5])
pd
newdata <- rbind(df,pd)#dplyr::bind_rows(df,pd)
newdata
with(newdata,plot(Year, Population))
newdatalm<-lm(Population~Year,data=newdata)
abline(newdatalm)
```

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.