Estimating Population Size in R

By: Anthony Marcolongo

Part A: Creating a Linear Regression Analysis

First, I loaded in all the packages needed for this analysis:

```
library(tidyverse)
library(ggplot2)
library(ggthemes)
library(dplyr)
install.packages("rmarkdown")
library(rmarkdown)
library(devtools)
install_github("yihui/tinytex")
library(tinytex)
install.packages("readxl")
library(readxl)
```

Part B: Data Preparation

Next, I imported the data using R Library readxl.

```
df <- read_excel("C:/Users/yankeeh8er/Desktop/Rproject/nst-est2019-alldata.xlsx")
```

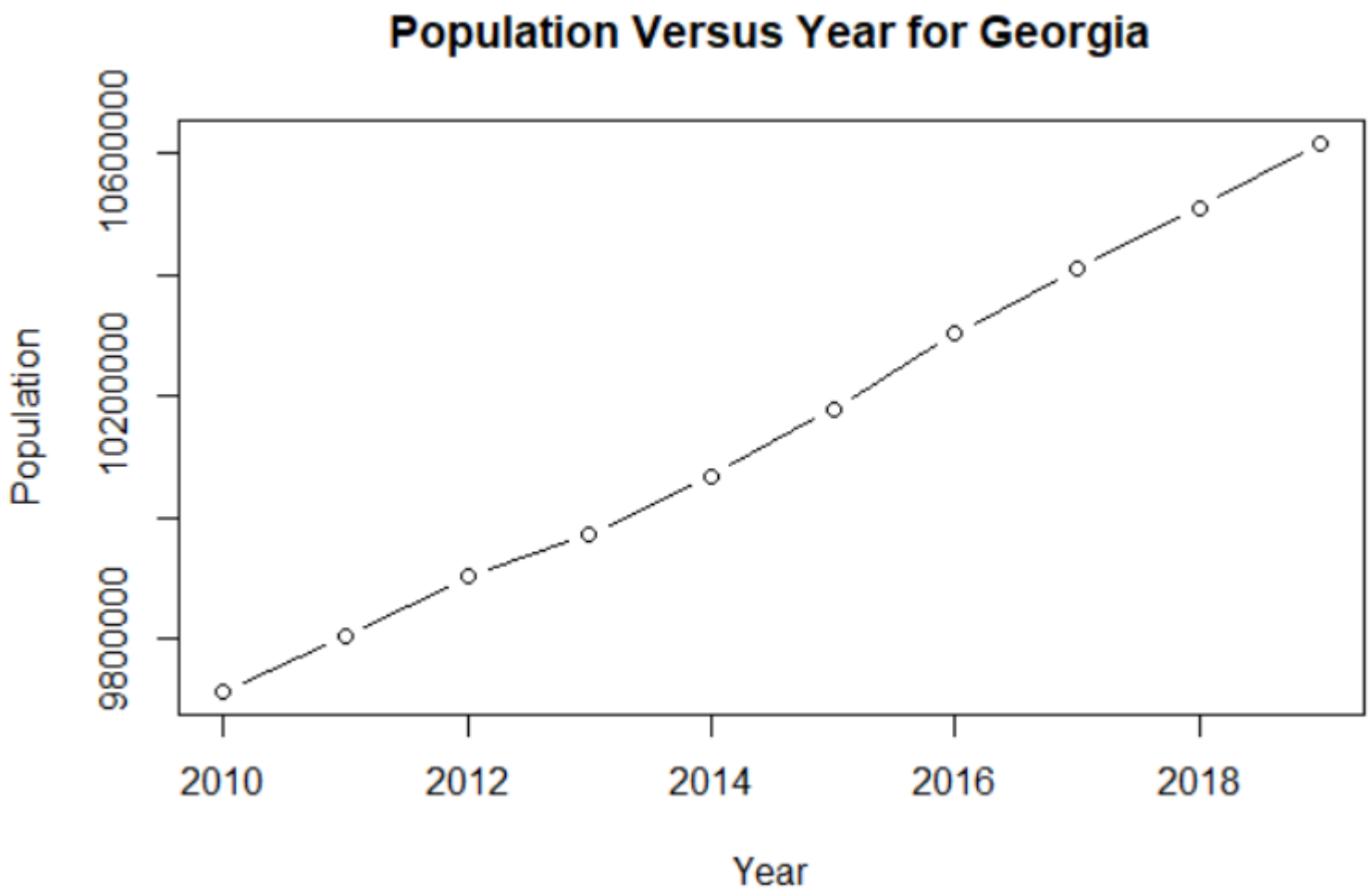
DF has 57 observations of 151 variables, most of which needs to be removed for the analysis.

I used the following code to remove the unnecessary rows and columns:

```
df2 <- df[16,]
df2
df2[,5]
df3 <- df2[,5:151]
df3
df4 <- df3[,1:13]
my_data <- df4[-(1:3)]
```

This leaves me with a dataframe, my_data, containing 1 observation and 10 variables.

Next, I created a plot of population versus year for my_data.


Population Versus Year for Georgia
Population
9800000
10200000
10600000
2010
2012
2014
2016
2018
Year

Part C: Executing the Script

I used R's built-in summary function on my_data and got the following results:

2010 2011 2012 2013
Min. :9711881 Min. :9802431 Min. :9901430 Min. :9972479
1st Qu.:9711881 1st Qu.:9802431 1st Qu.:9901430 1st Qu.:9972479
Median :9711881 Median :9802431 Median :9901430 Median :9972479
Mean :9711881 Mean :9802431 Mean :9901430 Mean :9972479
3rd Qu.:9711881 3rd Qu.:9802431 3rd Qu.:9901430 3rd Qu.:9972479
Max. :9711881 Max. :9802431 Max. :9901430 Max. :9972479
2014 2015 2016
Min. :10067278 Min. :10178447 Min. :10301890
1st Qu.:10067278 1st Qu.:10178447 1st Qu.:10301890
Median :10067278 Median :10178447 Median :10301890
Mean :10067278 Mean :10178447 Mean :10301890
3rd Qu.:10067278 3rd Qu.:10178447 3rd Qu.:10301890
Max. :10067278 Max. :10178447 Max. :10301890
2017 2018 2019
Min. :10410330 Min. :10511131 Min. :10617423
1st Qu.:10410330 1st Qu.:10511131 1st Qu.:10617423
Median :10410330 Median :10511131 Median :10617423
Mean :10410330 Mean :10511131 Mean :10617423
3rd Qu.:10410330 3rd Qu.:10511131 3rd Qu.:10617423
Max. :10410330 Max. :10511131 Max. :10617423

Part D: Executing the Script

Next, I created a model to predict the popluation of Georgia for the next 5 years.
Linear regression in R | An easy step-by-step guide. (2020, March 2). Scribbr.

I defined the following two vectors: years and pop.

pop <- c(9711881, 9802431, 9901430,9972479,10067278,10178447,10301890,10410330,10511131,10617423)
years <- c(2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019)

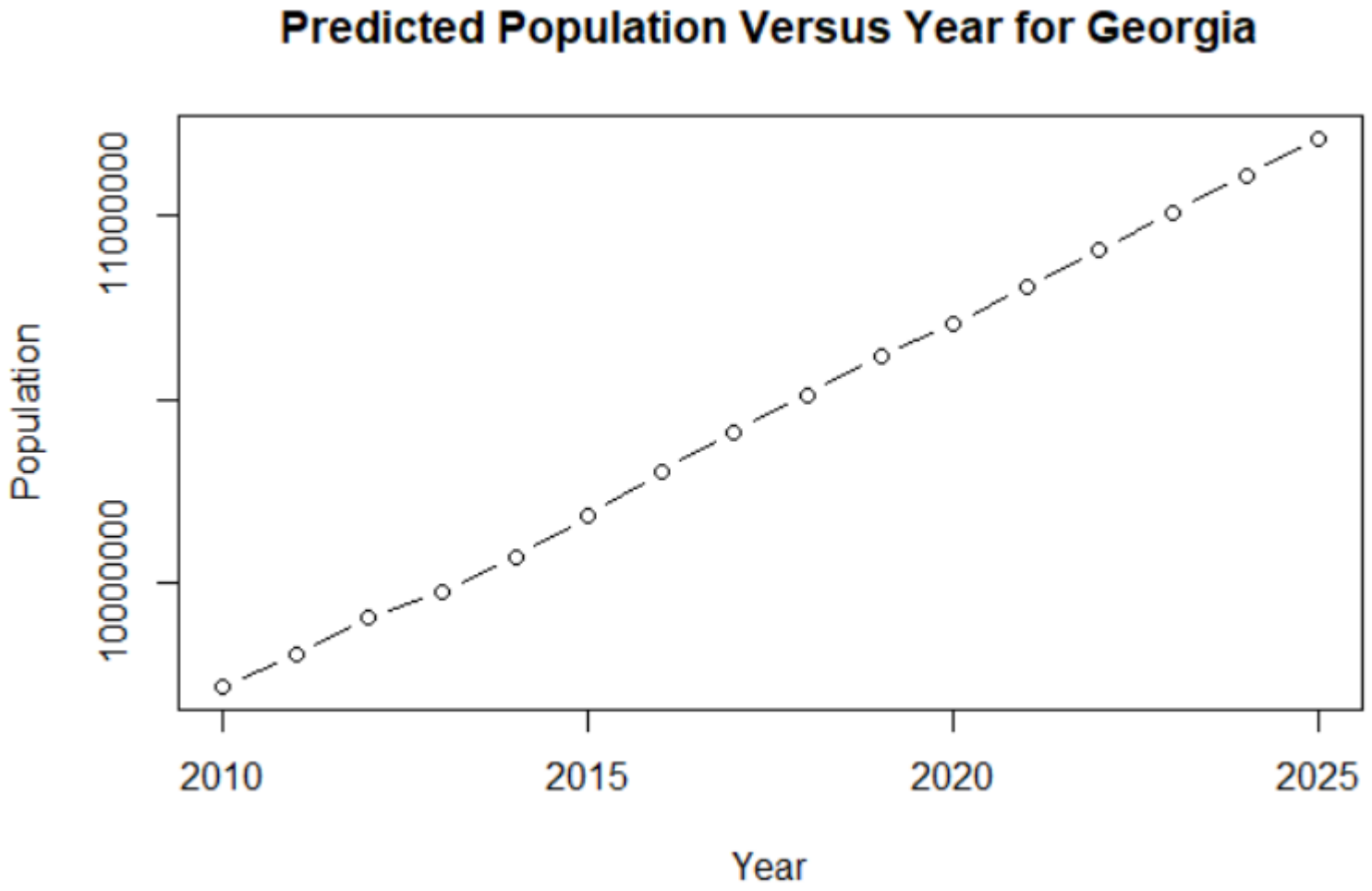
I then used the following code:

mod=lm(pop~years, data=my_data)
f3<-data.frame(years=c(2020,2021,2022,2023,2024,2025))
prediction<-predict(mod,f3)
How to combine vectors in R. (2016, March 26). dummies.

This gave me the following predictions:

10705961 10807505 10909048 11010592 11112135 11213679

Finally, I made a graph of the population data versus year, including the predicted population data.


Predicted Population Versus Year for Georgia
Population
10000000
11000000
2010
2015
2020
2025
Year

References

How To Combine Vectors in R

Andrie Vries-Andrie Vries- Revolution Analytics - https://www.dummies.com/programming/r/how-to-combine-vectors-in-r/ (https://www.dummies.com/programming/r/how-to-combine-vectors-in-r/)

https://www.scribbr.com/statistics/linear-regression-in-r/ (https://www.scribbr.com/statistics/linear-regression-in-r/)