C740 Proficiency Assessment

Jonathan DiQuattro

Fundamentals of Data Analytics C997

Emelda Ntinglet, PhD

May 25, 2020

Abstract

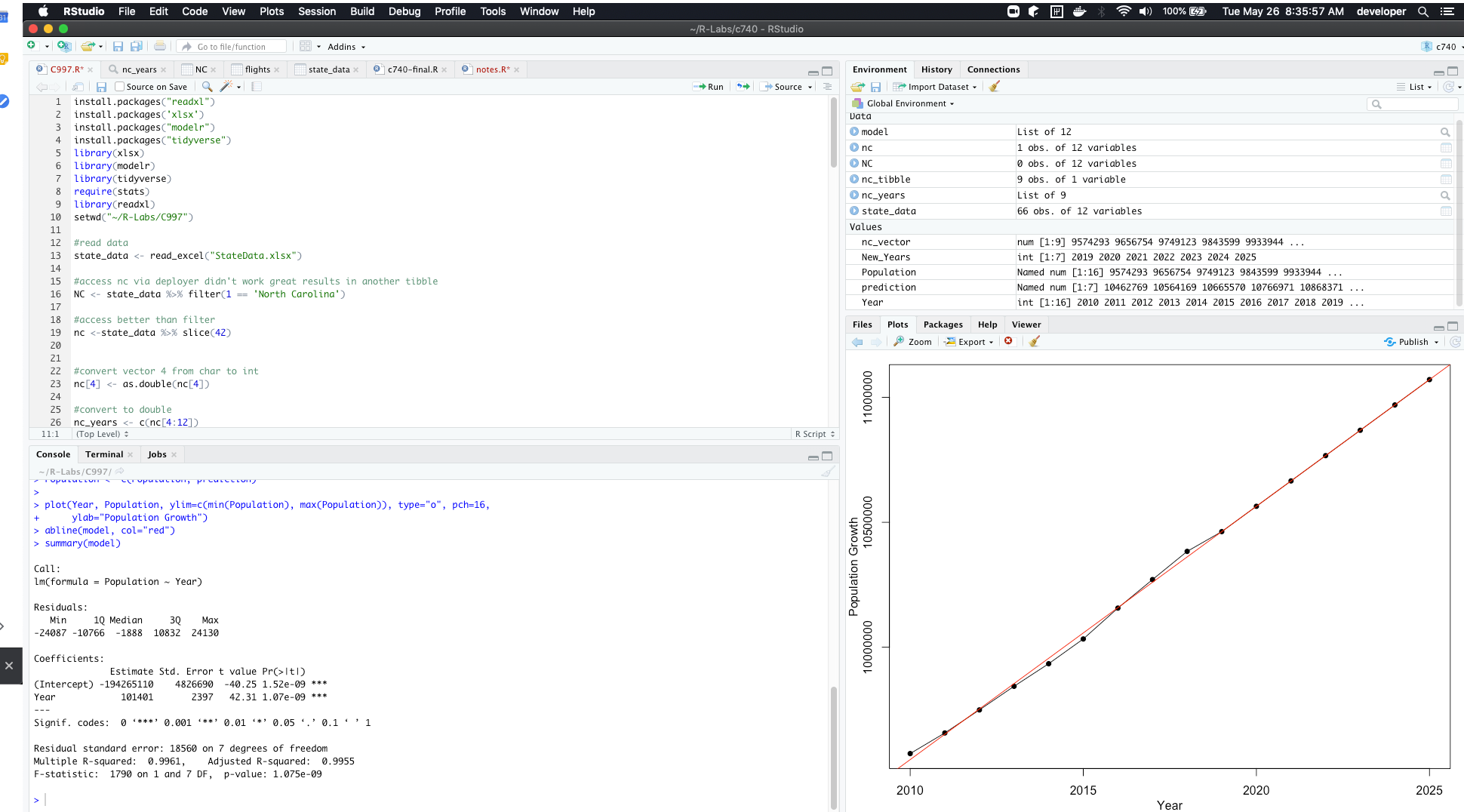
This paper is being submitted as a final assessment for the C997 R for Data Analysts class, and demonstrates how to use regression via the R language to predict population growth for the state of North Carolina. The data for this regression was imported from an Excel sheet located at www2.census.gov/programs-surveys/popest/tables/2010-2018/state/totals/nst-est2018-01.xlsx via the readxl R package.

C740 Proficiency Assessment C740

**For this project, you will use R to create a linear regression model of the population dynamics of your state (North Carolina) and predict the size of its population.**

**A. Create a linear regression analysis with R to predict the size of the population for the state you live in based on the “Current Estimates Data” dataset (see weblink below). Provide a screenshot of your results.**

In the attached screenshots there is an image of the completed regression table that has been taken from R studio, and was based on the population data from 2010-2018. The model is used to predict the population each year until 2025.



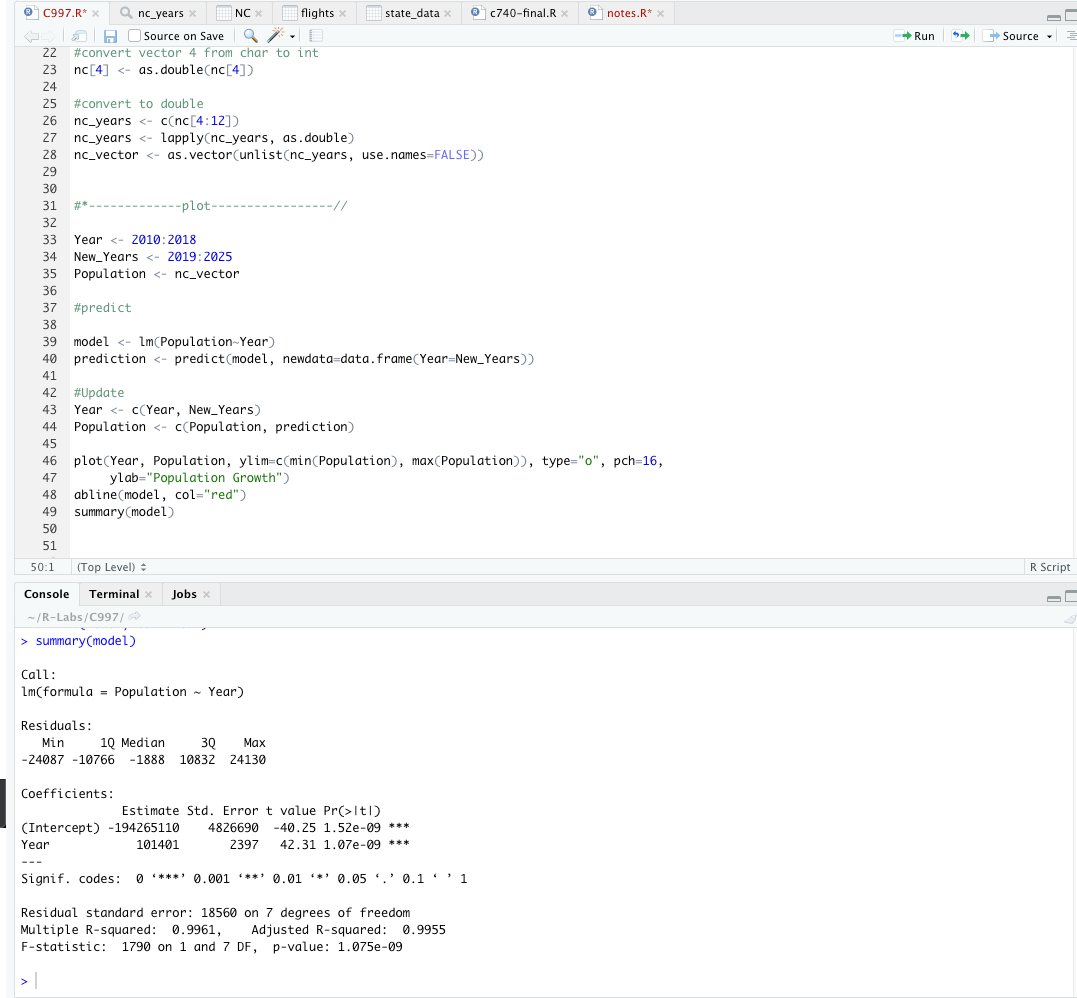
**B. Explain how you prepared the data from part A and how the dataset was imported into R, including screenshots of your results.**

The data for the linear regression was imported into r via the readxl package. The file was downloaded to working directory which was set with the setwd() method, and imported as a tibble using the deployr. Deployr is included in the tidyverse package (Wickham & Grolemund, 2014, para 11.1). The structure of the tibble was not ideal for modeling since it consisted of multiple features of yearly data that would have been difficult to work with. By converting the data to a vector it was much easier to model, predict, and plot. One of the elements of the vector had to be converted to a numeric double due to being of string type.



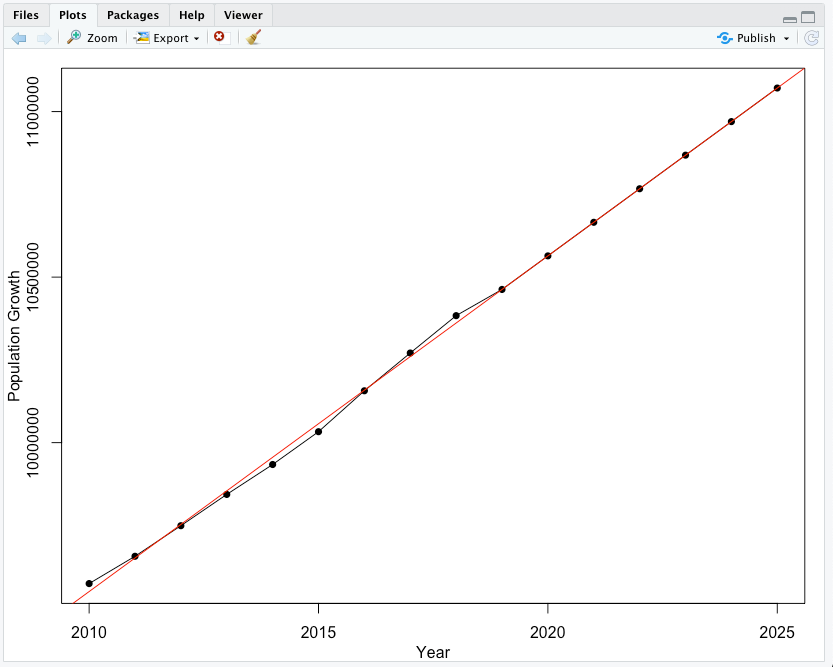
**C. Create an R script that will tabulate a statistical description of the model using R’s summary() function and provide a screenshot of your results.**

The following screenshot of R studio demonstrates both the code to produce a summary and the result. To get the summary the data points for year was entered as the Y value and the data points for population was entered for the X value. The lm() function will create a model based on that data and can be passed into the summary() function to give a summary of the model.



**D. Predict the population size of your state in five years using a linear regression from part A and provide a screenshot of your results.**

From the provided screenshot the predicted population will be 11,071,173 in the year 2025 for the state of North Carolina, this prediction was made using the built-in predict() function, that takes a model and vector as arguments and predicts a value for each element in the vector. The results of the lm() function returns a model, and was passed into the predict() function with a vector containing the years 2019 - 2025. This returned a vector of populations that correspond to each year, and was then plotted to give a visual representation of the prediction.



**E. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.**

The citations for this paper were generated using the APA citation wizard from the writing center.

References

Wickham, H., & Grolemund, G. (2017). R for Data Science. Sebastopl, CA: O'Reilly.