C997 – Performance Assessment of R for Data Analyst

Morrell J. Parrish

Western Governors University

Abstract

# “Use R to create a linear regression model of the population dynamics of California and predict the size of its population” (BOM1 TASK 1: Estimating Population Size).

*keywords: R Supply, US Census Bureau*

C997 – Performance Assessment of R for Data Analyst

**A. Create a linear regression analysis with R to predict the size of the population for the state of California; provide a screenshot of the results.**

“Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion” (Brownlee, 2020); and in this case the dependent variable would be the ***population*** and the independent variable would be the ***year.***

From the linear regression analysis performed in R (***Figure 5***), we can see that the that there is as strong positive growth trend up to 2019; however, the trend seems to drop slightly in 2020. **Figure 1** shows that linear regression model and the dataset it was derived from.

Table

Description automatically generated Chart, scatter chart

Description automatically generated

F**igure 1:** Linear Regression for California

From the linear model ‘***model***’ (**Figure 2**), we see that R2 = 0.922, which shows a strong correlation and best fit of the linear regression data; however, I wanted to verify the correlation between the year and population so I used the ***cor()*** function (**Figure 3**) within R; which actually has a higher value than R2 value.

Text

Description automatically generated

**Figure 2:** Population Summary (Model)



**Figure 3:** R’s Correlation Script

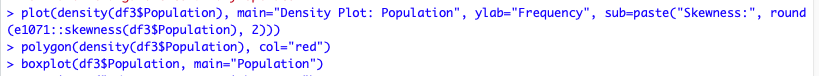
In order to create the scatter plot and regression line the below code was used.

A picture containing text

Description automatically generated

**Figure 4:** R’s Regression Plot Script

In order to check for normalcy, the below script was used to generate a density plot; upon looking at the graph you can see that there is a slight skewness (**Figure 6**).



**Figure 5:** R’s Density Plot Script

Chart, histogram

Description automatically generated

**Figure 6:** R’s Density Graph

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

First the data was downloaded from the [*US Census Bureau*](https://www.census.gov/programs-surveys/popest.html)website. The data needed to be cleaned prior to import; below are the steps taken to prepare the data for analysis.

* opened/imported the file into excel
* removed columns B and C, and rows 2 and 3, as well as the rows for the United States, Northeast, Midwest, South and West
* removed all special characters - the dot or period (.) from within the states fields
* copy and pasted the data into excel using the special paste (transpose) function, this allowed the years to be on the vertical axis (in columns) and the states to be on the horizontal axis (in rows).
* imported the dataset into R



**Figure 7:** Code to Load Dataset

Below is a screenshot of the raw dataset (***Figure 3***) and the clean data set **(*Figure 4*)**.­­

Graphical user interface, application, table, Excel

Description automatically generated

**Figure 8**: Raw Population Dataset

Graphical user interface, table

Description automatically generated

**Figure 9:** Clean Dataset

**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.**

I created a variable called ***model*** and then used R’s ***lm()***function to create the linear model; once the model was created I was able to use the ***summary()*** function to call for the calculations stored within the ***model***variable.

Text

Description automatically generated

**Figure 2:** Population Summary (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.**

Within 5 years the population for California is predicted to be about 41,136,191.



**Text

Description automatically generated with medium confidence**

**Figure 10:** R Script for Predictions

Table

Description automatically generated

**Figure 11:** Prediction Results

References

Brownlee, Jason. “Linear Regression for Machine Learning.” *Machine Learning Mastery*, 14 Aug. 2020, machinelearningmastery.com/linear-regression-for-machine-learning/.

*WGU Performance Assessment*, BOM1 TASK 1: Estimating Population Size tasks.wgu.edu/student/000194226/course/15940008/task/1148/overview.