BOM1 Task 1: Estimating Population Size

Sina Bozorgmehr

Western Governors University

BOM1 TASK 1: ESTIMATING POPULATION SIZE

Introduction

The United States Census Bureau's mission is "to serve as the nation's leading provider of quality data about its people and economy" (United States Census Bureau, 2020b, par. 1). The Census Bureau has three distinctive statistical programs: Decennial census, economic census, and census of governments. Every decade, the decennial census counts the population and housing in all 50 states, District of Columbia, Puerto Rico, and the Island Areas as required by the U.S. Constitution. This data extracted from this census will be used to revise the number of seats that each state gets in the U.S. House of Representatives and to allocate the federal funds each year (United States Census Bureau, 2020a). In this project, annual estimates of resident population for the United States, regions, states, and Puerto Rico: April 1st, 2020 to July 1st, 2019 was used to predict the population of California in the next five years: 2020 to 2024.

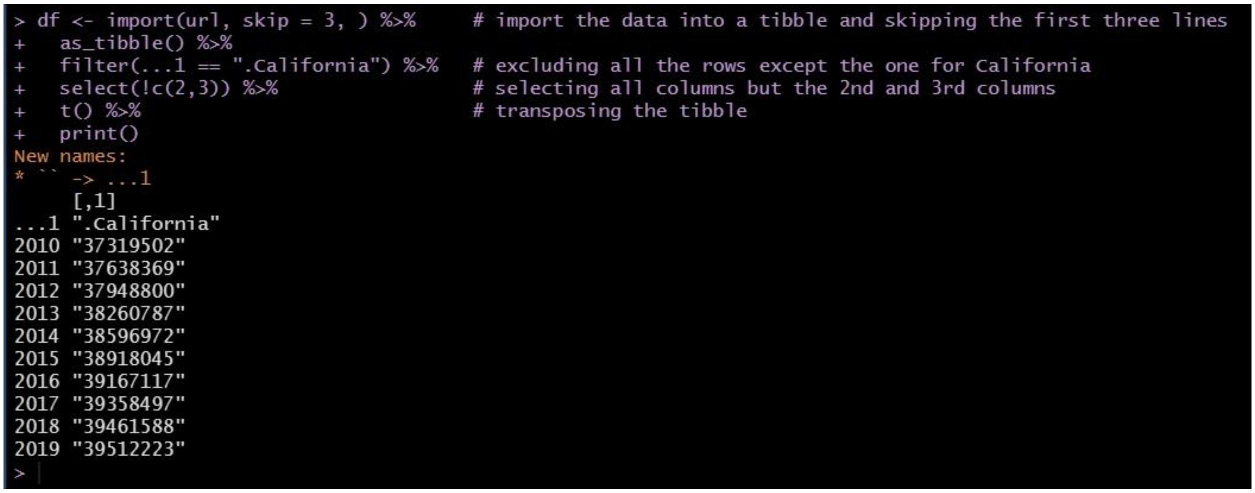
Data Collection and Cleaning

The Annual Estimates, 2019 was imported into the R using *rio::import()*. First three lines were skipped to exclude data title and subtitles. The row that contains the data points for the state of California were kept and all the remaining rows were excluded. Columns number 2 and 3 were removed to clean the data for linear regression modeling. Then, the data frame was transposed to have both Year and Population as variables in column. Figure 1 shows the script and the resulting tibble.

Next, the tibble was converted to a data frame and the row names were added as a new variable, column. The first row of this data frame was removed to get rid of the original column names and then the name of 2nd variable was changed to Population. The last step of data cleaning was converting the values types from character to integers to prepare the data for regression analysis (Figure 2).

Figure 1

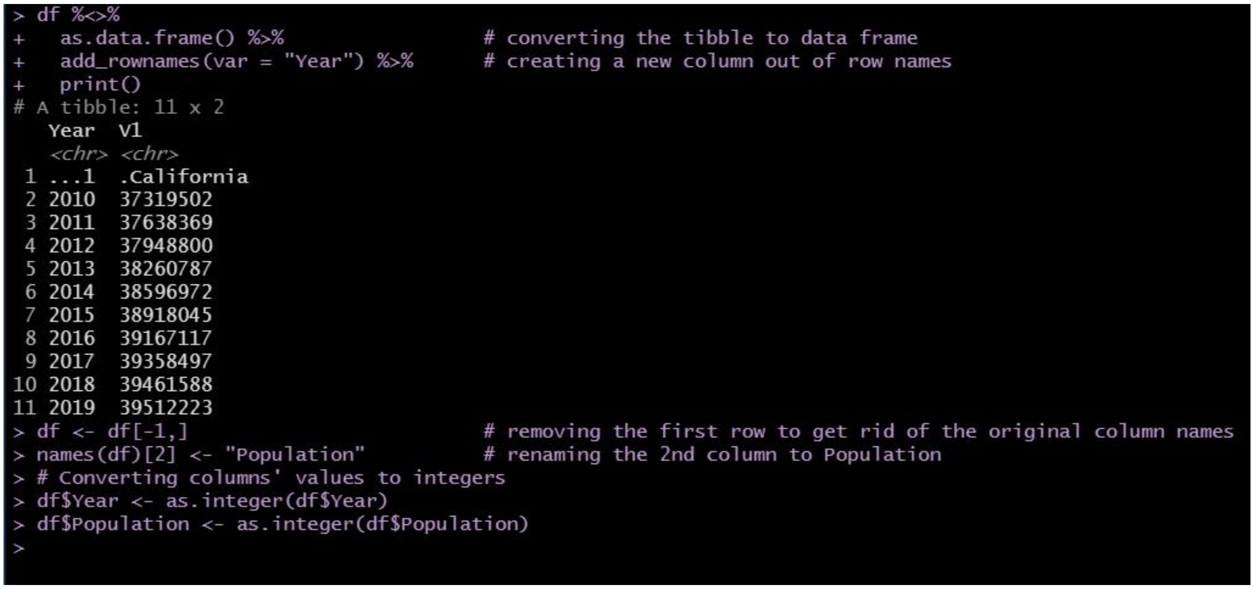
Data importing and cleaning; first part.



```
> df <- import(url, skip = 3, ) %>%     # import the data into a tibble and skipping the first three lines
+   as_tibble() %>%
+   filter(...1 == ".California") %>%   # excluding all the rows except the one for California
+   select(!c(2,3)) %>%                 # selecting all columns but the 2nd and 3rd columns
+   t() %>%                             # transposing the tibble
+   print()
New names:
* `` -> ...1
     [,1]
...1 ".California"
2010 "37319502"
2011 "37638369"
2012 "37948800"
2013 "38260787"
2014 "38596972"
2015 "38918045"
2016 "39167117"
2017 "39358497"
2018 "39461588"
2019 "39512223"
> 
```

Figure 2

Data importing and cleaning; second part.



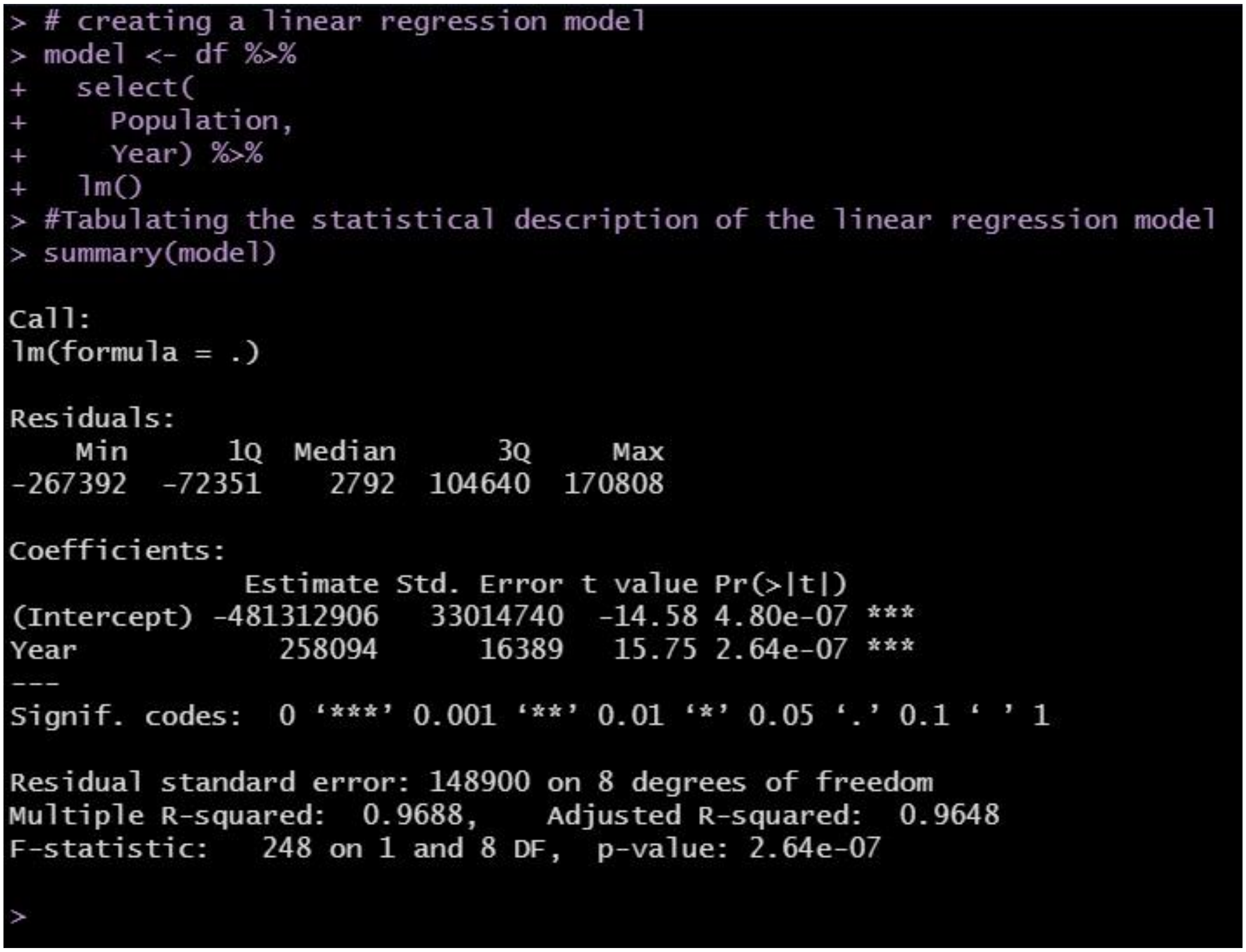
```
> df %<>%
+   as.data.frame() %>%                  # converting the tibble to data frame
+   add_rownames(var = "Year") %>%       # creating a new column out of row names
+   print()
# A tibble: 11 x 2
   Year  V1
   <chr> <chr>
 1 ...1  .California
 2 2010  37319502
 3 2011  37638369
 4 2012  37948800
 5 2013  38260787
 6 2014  38596972
 7 2015  38918045
 8 2016  39167117
 9 2017  39358497
10 2018  39461588
11 2019  39512223
> df <- df[-1,]                          # removing the first row to get rid of the original column names
> names(df)[2] <- "Population"           # renaming the 2nd column to Population
> # Converting columns' values to integers
> df$Year <- as.integer(df$Year)
> df$Population <- as.integer(df$Population)
> 
```

Linear Regression Analysis

After the data was prepared for analysis, the linear regression model was created using *lm()*. Population was given to the function as the response and Year as predictor. Using the *summary()* the statistical description of the model was tabulated as you can see in the Figure 3. As the result shows, the Y-intercept is -481312906 and the slope is 258094. Also, the R squared of this model is 0.9688 which proves a strong correlation between variables.

Figure 3

Linear regression analysis and model summaries



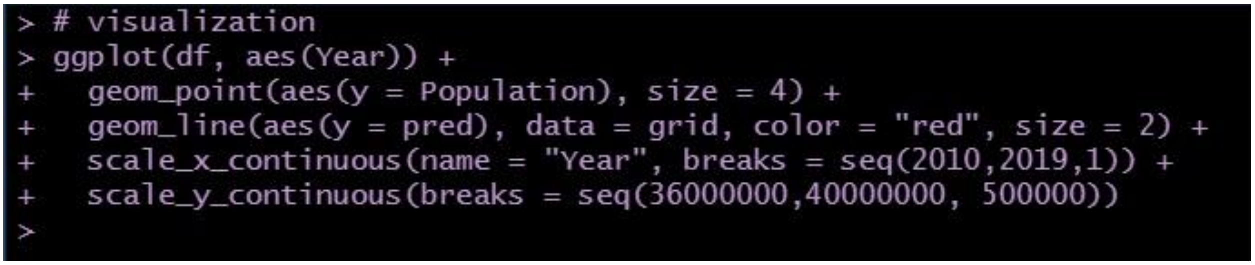
```
> # creating a linear regression model
> model <- df %>%
+   select(
+     Population,
+     Year) %>%
+   lm()
> #Tabulating the statistical description of the linear regression model
> summary(model)

Call:
lm(formula = .)

Residuals:
    Min      1Q  Median      3Q     Max 
-267392  -72351    2792  104640  170808 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -481312906   33014740  -14.58 4.80e-07 ***
Year            258094      16389   15.75 2.64e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 148900 on 8 degrees of freedom
Multiple R-squared:  0.9688,	Adjusted R-squared:  0.9648 
F-statistic:   248 on 1 and 8 DF,  p-value: 2.64e-07

>
```

Visualization

To demonstrate the data points and the regression line, a data grid was made of the predictor, Year, and the fitted values. Then, using ggplot, the observed data points were visualized on a coordinate systema and the regression line that was modeled earlier, was seated on the top layer (Figures 4 and 5).

Figure 4

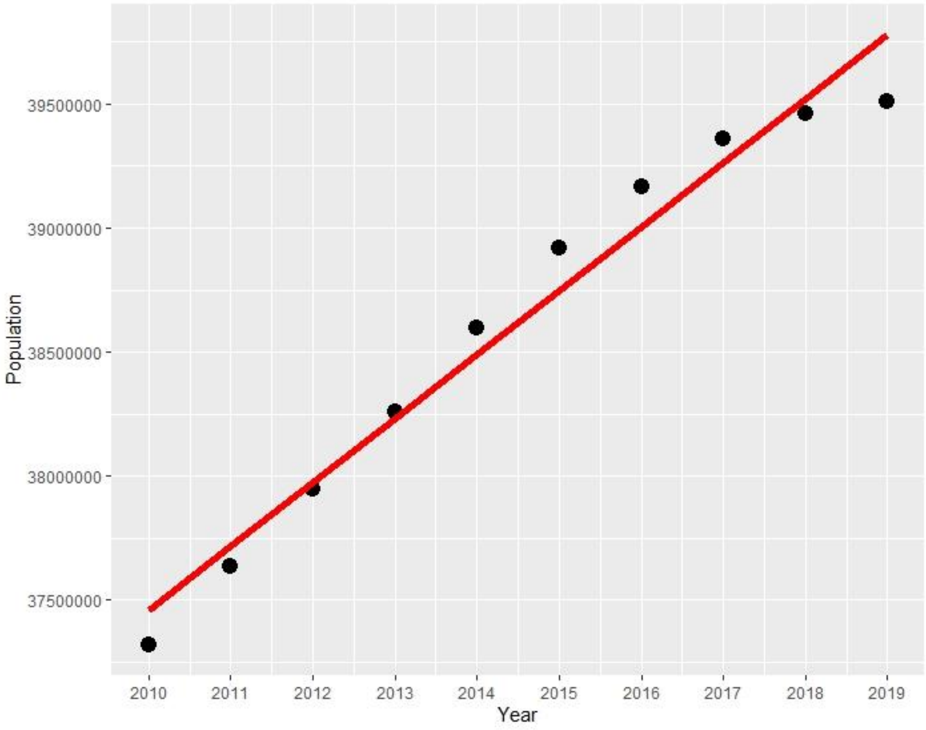
Code for regression line visualization.



```
> # visualization
> ggplot(df, aes(Year)) +
+   geom_point(aes(y = Population), size = 4) +
+   geom_line(aes(y = pred), data = grid, color = "red", size = 2) +
+   scale_x_continuous(name = "Year", breaks = seq(2010,2019,1)) +
+   scale_y_continuous(breaks = seq(36000000,40000000, 500000))
>
```

Figure 5

Linear regression line plot

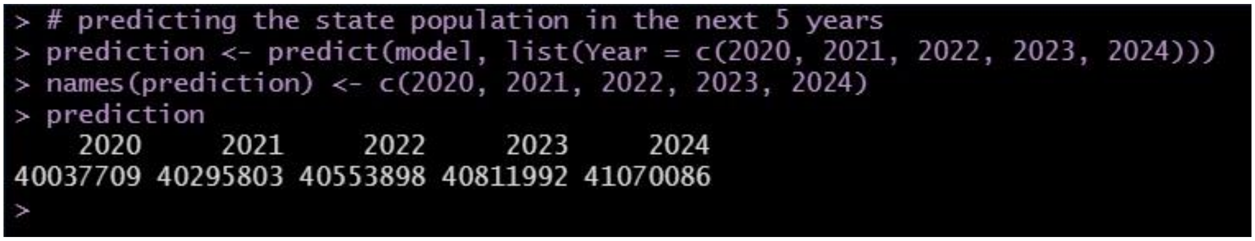


Prediction

Using the linear regression model made in previous section, the estimated population of California in years 2020 to 2024 were predicted. First, a vector of years given to the *predict()* and then the results were printed. Based on the prediction, the estimated population of California in year 2024 will surpass 41 million people (Figure 6).

Figure 6

Prediction of California population estimate in the next 5 years.



```
> # predicting the state population in the next 5 years
> prediction <- predict(model, list(Year = c(2020, 2021, 2022, 2023, 2024)))
> names(prediction) <- c(2020, 2021, 2022, 2023, 2024)
> prediction
    2020     2021     2022     2023     2024
40037709 40295803 40553898 40811992 41070086
>
```

Conclusion

Based on the Census Bureau population estimates, the population of California on July 1st, 2019 was 39,512,223 persons (United States Census Bureau, 2019). The linear regression analysis showed that this number will increase annually by a 258,094 and in 2024 will surpass 41 million people.

References

United States Census Bureau. (2019). *Table 1. Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019.* U.S. Department of Commerce. https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/totals/nst-est2019-01.xlsx

United States Census Bureau. (2020a, October). *Our Censuses*. U.S. Department of Commerce. https://www.census.gov/programs-surveys/censuses.html

United States Census Bureau. (2020b, October). *What We Do*. U.S. Department of Commerce. https://www.census.gov/about/what.html