BOM1 TASK 1: Estimating Population Size

Josh King

Introduction

This project results in the creation of a linear model to estimate future population in the **state of Georgia**. The data was collected from the United Stats Census Bureau website. Further information regarding the Census Bureau website can be found in the *References* section below, whereas the download link to the csv used in this project is: https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/national/totals/nst-est2018-alldata.csv.

This project submission contains the following files:

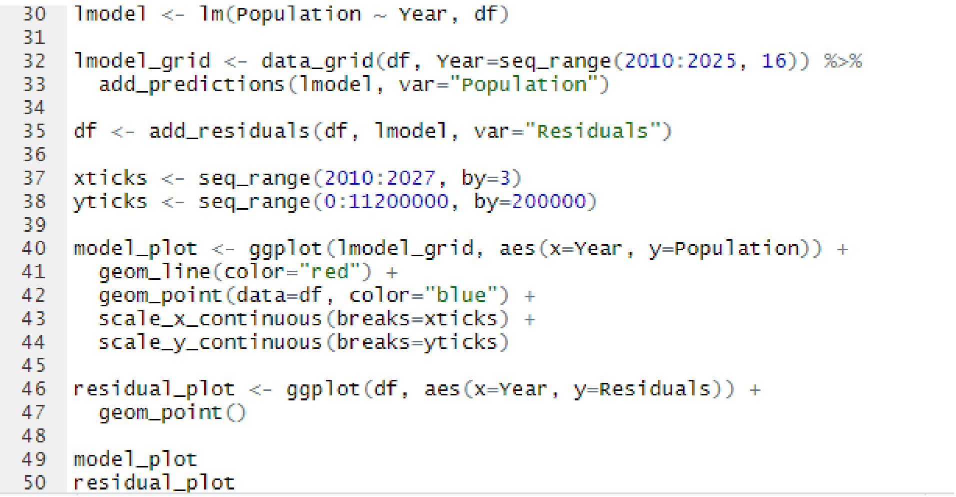
- "Josh King R Model Documentation.pdf" – This file, explaining the process.
- "LinearModel.R" – The R script containing the code used for this project.
- "nst-est2018-alldata.csv" – The downloaded data set from the United States Census Bureau website.

The R script created in this project utilizes the "tidyverse" package. Documentation for this package can be found in the *References* section.

Program Tasks

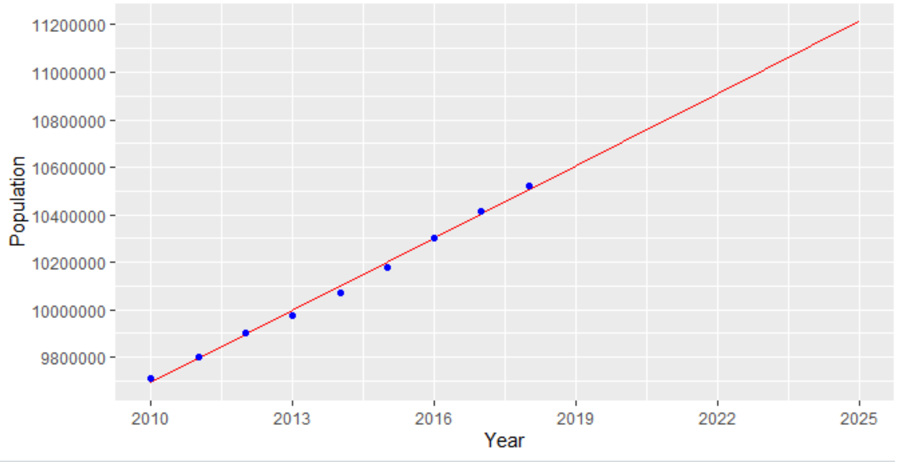
A) Linear Regression Model:

The following code was used to produce a linear model with year as the independent variable and population as the dependent:



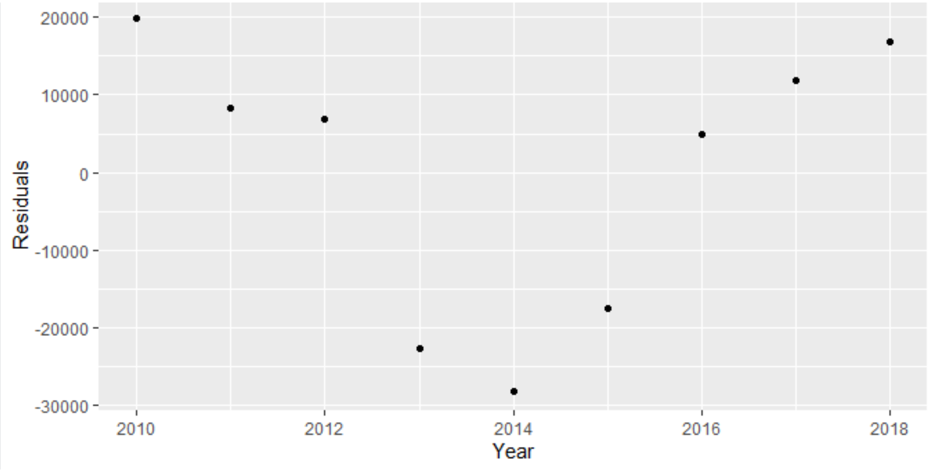
```r
lmodel <- lm(Population ~ Year, df)

lmodel_grid <- data_grid(df, Year=seq_range(2010:2025, 16)) %>%
  add_predictions(lmodel, var="Population")

df <- add_residuals(df, lmodel, var="Residuals")

xticks <- seq_range(2010:2027, by=3)
yticks <- seq_range(0:11200000, by=200000)

model_plot <- ggplot(lmodel_grid, aes(x=Year, y=Population)) +
  geom_line(color="red") +
  geom_point(data=df, color="blue") +
  scale_x_continuous(breaks=xticks) +
  scale_y_continuous(breaks=yticks)

residual_plot <- ggplot(df, aes(x=Year, y=Residuals)) +
  geom_point()

model_plot
residual_plot
```

By calling "model_plot" we retrieve a plot of the linear model (in red) overlaid against the original data points (in blue):



The linear model produces a fitted line closely approximating the original data without any obvious outliers. In the chart above, the model was plotted out to the year 2025 (original data was only provided from 2010 – 2018).

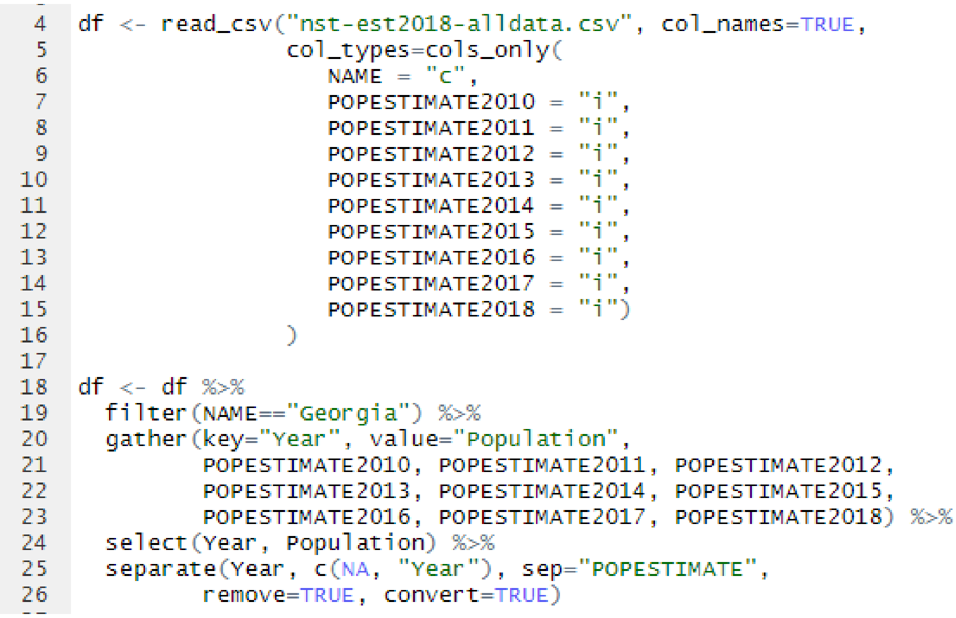
By calling "residual_plot", a plot of the residuals from the model was also produced:



This residual plot shows the errors from the predicted model for the provided data. The accuracy of the model however will be discussed in the *Model Summary Statistics* portion of this documentation.

B) Preparation of Data:

The data for this project was manually downloaded from the United States Census Bureau website in the form of the file “nst-est2018-alldata.csv” as noted above. Once downloaded, the following code was used to prepare the data:

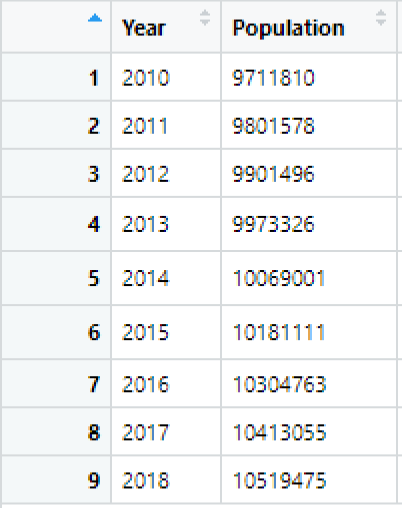


```
 4  df <- read_csv("nst-est2018-alldata.csv", col_names=TRUE,
 5                 col_types=cols_only(
 6                    NAME = "c",
 7                    POPESTIMATE2010 = "i",
 8                    POPESTIMATE2011 = "i",
 9                    POPESTIMATE2012 = "i",
10                    POPESTIMATE2013 = "i",
11                    POPESTIMATE2014 = "i",
12                    POPESTIMATE2015 = "i",
13                    POPESTIMATE2016 = "i",
14                    POPESTIMATE2017 = "i",
15                    POPESTIMATE2018 = "i")
16                 )
17
18  df <- df %>%
19    filter(NAME=="Georgia") %>%
20    gather(key="Year", value="Population",
21           POPESTIMATE2010, POPESTIMATE2011, POPESTIMATE2012,
22           POPESTIMATE2013, POPESTIMATE2014, POPESTIMATE2015,
23           POPESTIMATE2016, POPESTIMATE2017, POPESTIMATE2018) %>%
24    select(Year, Population) %>%
25    separate(Year, c(NA, "Year"), sep="POPESTIMATE",
26           remove=TRUE, convert=TRUE)
```

Lines 4 – 16 above represent the initial reading of the csv file into a tidyverse tibble. As the initial csv file had a number of columns not relevant to the intended analysis, these were down-selected to the noted “POPESTIMATE….” columns above. While there were multiple columns for the 2010 data, by reading through the Census Bureau website, I surmised that the “POPESTIMATE….” columns reflected the best estimates available for this project’s intent.

Lines 18 – 26 tackle the challenge of programmatically formatting the data pull in a tidy way and limiting the data to that of my state, Georgia. The “filter” function down-selects the tibble to just the Georgia data. The “gather” function then tidies up the data by making the “POPESTIMATE…” column names be values in a new column called “Year” with a corresponding “Population” value. By using “select”, only the “Year” and “Population” columns are kept. Finally, the “separate” function is used to parse the “Year” column, splitting out the actual year from the “POPESTIMATE” string.

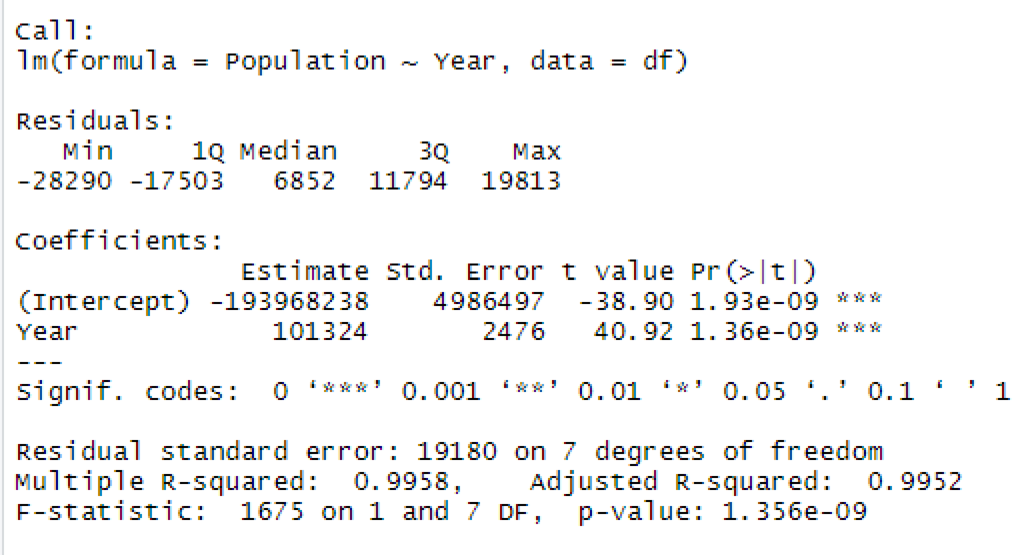
These steps ultimately result in the creation of the tibble “df”, seen below, which is used to build the linear model:



	Year	Population
1	2010	9711810
2	2011	9801578
3	2012	9901496
4	2013	9973326
5	2014	10069001
6	2015	10181111
7	2016	10304763
8	2017	10413055
9	2018	10519475

C) Model Summary Statistics:

By calling “summary()” on the created linear model (“lmodel”), the following summary statistics were created:



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

Residuals:
   Min     1Q Median     3Q    Max 
-28290 -17503   6852  11794  19813 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -193968238    4986497  -38.90 1.93e-09 ***
Year            101324       2476   40.92 1.36e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19180 on 7 degrees of freedom
Multiple R-squared:  0.9958,	Adjusted R-squared:  0.9952 
F-statistic:  1675 on 1 and 7 DF,  p-value: 1.356e-09
```

From the summary statistics above, we can glean considerable useful information about the model. Of particular note, we can find:

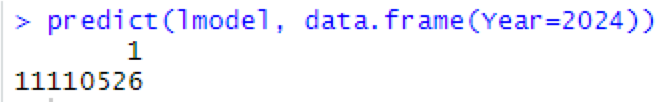
- The intercept and slope values utilized can in the “Estimate” column.
- The p value for the intercept and slope, found in the “Pr(>|t|)” column. Of note, both values are significantly below 0.05 suggesting strong significance in this model.

- The adjusted R^2 value of 0.9952, suggesting a reasonable fit.

Within the summary statistics we can also find information about the residuals. At an initial glance of the residuals plotted in section *A*, we might notice a seemingly non-random distribution (as seen by the "dip" towards the middle of the plot). However, given the small sample size, the low value of the residuals compared to the slope, and the generally positive summary statistics seen above, we can be fairly confident in assuming reasonable validity of the model for the scope of this project.

D) Five Year Prediction:

To predict the population of the state of Georgia in five years (2024), we can use the "predict" function. The output of this in R is seen below:



```
> predict(lmodel, data.frame(Year=2024))
       1
11110526
```

This code takes in the model ("lmodel") and a dataframe to produce its results. Since we are only looking for one year, a make-shift dataframe with just the year value of 2024 was passed in. The result is an estimated population in 2024 of 11,110,526.

References

State Population Totals: 2010-2018. (2019). Retrieved 24 August 2019, from https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html

tidyverse package | R Documentation. (2019). Retrieved 24 August 2019, from https://www.rdocumentation.org/packages/tidyverse/versions/1.2.1