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.

	Year	Population
1	2010	37319550
2	2011	37636311
3	2012	37944551
4	2013	38253768
5	2014	38586706
6	2015	38904296
7	2016	39149186
8	2017	39337785
9	2018	39437463
10	2019	39437610
11	2020	39368078

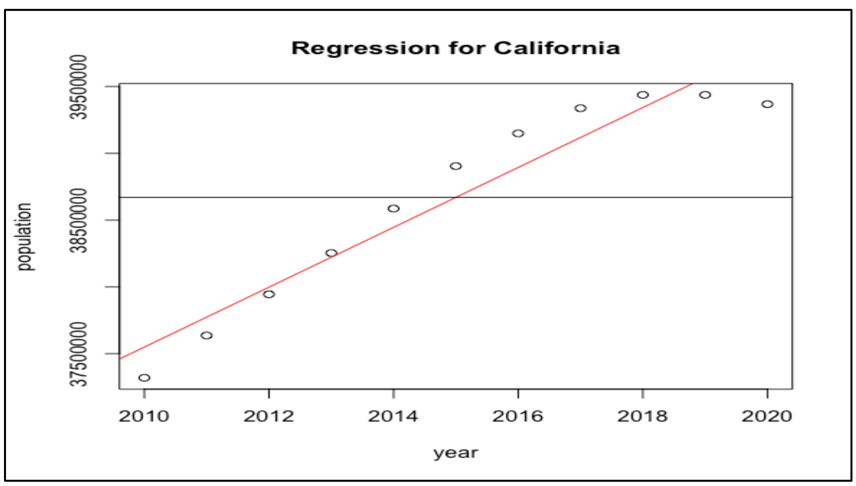

Figure 1: Linear Regression for California

From the linear model '***model***' (**Figure 2**), we see that $R^2 = 0.922$, 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 R^2 value.

```
> model<-lm(df3$Population ~ df3$Year, data = df3)
> summary(model)

Call:
lm(formula = df3$Population ~ df3$Year, data = df3)

Residuals:
    Min      1Q  Median      3Q     Max 
-423181 -133522   31596  179686  254548 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -413002502   43794844   -9.43 5.82e-06 ***
df3$Year        224155      21734   10.31 2.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 228000 on 9 degrees of freedom
Multiple R-squared:  0.922,	Adjusted R-squared:  0.9133 
F-statistic: 106.4 on 1 and 9 DF,  p-value: 2.766e-06
```

Figure 2: Population Summary (Model)

```
> cor(df3$Year, df3$Population)
[1] 0.9602018
```

Figure 3: R’s Correlation Script

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

```
> plot(df3$Year, df3$Population, main = "Regression for California", xlab = 'year', ylab = 'population')
> abline(lm(df3$Population ~ df3$Year, data = df3), col = 'red')
> 
> ## mean of population
> 
> abline(h=mean(df3$Population))
```

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**).

```
> plot(density(df3$Population), main="Density Plot: Population", ylab="Frequency", sub=paste("Skewness:", round
(e1071::skewness(df3$Population), 2)))
> polygon(density(df3$Population), col="red")
> boxplot(df3$Population, main="Population")
```

Figure 5: R’s Density Plot Script

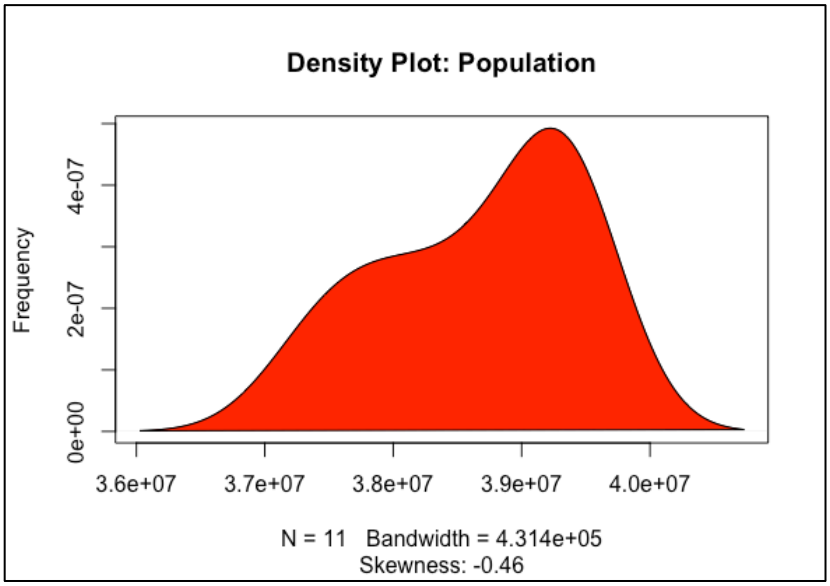

Figure 6: R's Density Graph

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* 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

```
library(readxl)
dataset <- read_excel(NULL)
View(dataset)
```

Figure 7: Code to Load Dataset

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

	Annual Estimates of the Resident Population for the United States, Regions, States, the District of Columbia, and Puerto Rico: April 1, 2010 to July 1, 2020										
3	Geographic Area	April 1, 2010		Population Estimate (as of July 1)							
4		Census	Estimates Base	2010	2011	2012	2013	2014	2015	2016	
5	**United States**	308,745,538	308,758,105	309,327,143	311,583,481	313,877,662	316,059,947	318,386,329	320,738,994	323,071,755	
6	**Northeast**	55,317,240	55,318,414	55,380,764	55,608,318	55,782,661	55,912,775	56,021,339	56,052,790	56,063,777	
7	**Midwest**	66,927,001	66,929,737	66,975,328	67,164,092	67,348,275	67,576,524	67,765,576	67,885,682	68,018,175	
8	**South**	114,555,744	114,563,042	114,869,421	116,019,483	117,264,196	118,397,213	119,666,248	121,049,223	122,419,547	
9	**West**	71,945,553	71,946,912	72,101,630	72,791,588	73,482,530	74,173,435	74,933,166	75,751,299	76,570,256	
10	Alabama	4,779,736	4,780,118	4,785,514	4,799,642	4,816,632	4,831,586	4,843,737	4,854,803	4,866,824	
11	Alaska	710,231	710,246	713,982	722,349	730,810	737,626	737,075	738,430	742,575	
12	Arizona	6,392,017	6,392,292	6,407,342	6,473,416	6,556,344	6,634,690	6,732,873	6,832,810	6,944,767	
13	Arkansas	2,915,918	2,916,029	2,921,998	2,941,038	2,952,876	2,960,459	2,968,759	2,979,732	2,991,815	
14	California	37,253,956	37,254,522	37,319,550	37,636,311	37,944,551	38,253,768	38,586,706	38,904,296	39,149,186	
15	Colorado	5,029,196	5,029,319	5,047,539	5,121,900	5,193,660	5,270,774	5,352,637	5,454,328	5,543,844	
16	Connecticut	3,574,097	3,574,151	3,579,173	3,588,632	3,595,211	3,595,792	3,595,697	3,588,561	3,579,830	
17	Delaware	897,934	897,947	899,647	907,590	915,518	924,062	933,131	942,065	949,989	
18	District of Columbia	601,723	601,767	605,282	620,290	635,737	651,559	663,603	677,014	687,576	
19	Florida	18,801,310	18,804,589	18,846,143	19,055,607	19,302,016	19,551,678	19,853,880	20,219,111	20,627,237	
20	Georgia	9,687,653	9,688,737	9,712,209	9,803,630	9,903,580	9,975,592	10,071,204	10,183,353	10,308,442	
21	Hawaii	1,360,301	1,360,304	1,364,004	1,379,562	1,395,199	1,408,822	1,415,335	1,422,999	1,428,885	
22	Idaho	1,567,582	1,567,658	1,570,819	1,584,272	1,595,910	1,612,053	1,632,248	1,652,495	1,684,036	
23	Illinois	12,830,632	12,831,572	12,840,545	12,867,783	12,883,029	12,895,778	12,885,092	12,859,585	12,821,709	
24	Indiana	6,483,802	6,484,050	6,490,555	6,517,250	6,538,989	6,570,575	6,596,019	6,611,442	6,637,898	

Figure 8: Raw Population Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	District of Co	Florida	Georgia	Hawaii	Idaho
2	**2010**	4,785,514	713,982	6,407,342	2,921,998	37,319,550	5,047,539	3,579,173	899,647	605,282	18,846,143	9,712,209	1,364,004	1,570,819
3	**2011**	4,799,642	722,349	6,473,416	2,941,038	37,636,311	5,121,900	3,588,632	907,590	620,290	19,055,607	9,803,630	1,379,562	1,584,272
4	**2012**	4,816,632	730,810	6,556,344	2,952,876	37,944,551	5,193,660	3,595,211	915,518	635,737	19,302,016	9,903,580	1,395,199	1,595,910
5	**2013**	4,831,586	737,626	6,634,690	2,960,459	38,253,768	5,270,774	3,595,792	924,062	651,559	19,551,678	9,975,592	1,408,822	1,612,053
6	**2014**	4,843,737	737,075	6,732,873	2,968,759	38,586,706	5,352,637	3,595,697	933,131	663,603	19,853,880	10,071,204	1,415,335	1,632,248
7	**2015**	4,854,803	738,430	6,832,810	2,979,732	38,904,296	5,454,328	3,588,561	942,065	677,014	20,219,111	10,183,353	1,422,999	1,652,495
8	**2016**	4,866,824	742,575	6,944,767	2,991,815	39,149,186	5,543,844	3,579,830	949,989	687,576	20,627,237	10,308,442	1,428,885	1,684,036
9	**2017**	4,877,989	740,983	7,048,088	3,003,855	39,337,785	5,617,421	3,575,324	957,942	697,079	20,977,089	10,417,031	1,425,763	1,719,745
10	**2018**	4,891,628	736,624	7,164,228	3,012,161	39,437,463	5,697,155	3,574,561	966,985	704,147	21,254,926	10,519,389	1,423,102	1,752,074
11	**2019**	4,907,965	733,603	7,291,843	3,020,985	39,437,610	5,758,486	3,566,022	976,668	708,253	21,492,056	10,628,020	1,415,615	1,789,060
12	**2020**	4,921,532	731,158	7,421,401	3,030,522	39,368,078	5,807,719	3,557,006	986,809	712,816	21,733,312	10,710,017	1,407,006	1,826,913
13														

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.

```
> model<-lm(df3$Population ~ df3$Year, data = df3)
> summary(model)

Call:
lm(formula = df3$Population ~ df3$Year, data = df3)

Residuals:
    Min      1Q  Median      3Q     Max 
-423181 -133522   31596  179686  254548 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -413002502   43794844   -9.43 5.82e-06 ***
df3$Year        224155      21734   10.31 2.77e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 228000 on 9 degrees of freedom
Multiple R-squared:  0.922,	Adjusted R-squared:  0.9133 
F-statistic: 106.4 on 1 and 9 DF,  p-value: 2.766e-06
```

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.

```
> Year<-2022:2026
```

```
> Predictions <- (Year * model$coef[2] + model$coef[1])
> population_predictions<-data.frame(Year, Predictions)
> View(population_predictions)
> |
```

Figure 10: R Script for Predictions

	Year	Predictions
1	2022	40239569
2	2023	40463725
3	2024	40687880
4	2025	40912035
5	2026	41136191

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.