Matt Bartos & Jarod Cox

Dr. Fontenot CS 5/7394

02/07/2022

Project Report

Data Wrangling

The group gathered all of our data files from "World Bank" and stored the information inside different data frames. The first data frame contains the population of the Earth of each year from 1960 to 2020 which is used for our linear regression model. Each year, the population of every country was summed and paired with its correct year inside of a data frame. The next two data frames are used to store the death and birth rates of each year. Those data frames are then used to determine the average death rate while the birth rate is expected to decline so we use 2020's birthrate to initialize the model

Method #1 Linear Regression

The first model we used to predict the population was linear regression. This model utilized the given population data frame. The linear regression model used only the population data and in turn saw an increasing population that shows no sign of slowing down. The prediction of our linear regression model is shown below.

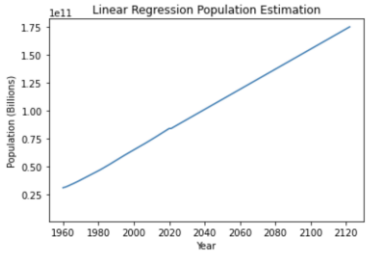


Figure 1: Linear Regression Model

The population estimated in 2122 for this model is around 17.5 billion people. This would mean that the population of the Earth is going to easily more than double within the next 100 years. The model works by simply predicting the trend line of the population linearly. Basic linear regression is not the best model in this case because it does not have enough information to accurately understand how the population is growing. The population data does not reveal any signs of slowing down. This model led to us wanting to attempt if there were better regression models that could account for the growth percentages changing per year.

To see if there were other types of regression that could work on this data, two other regression types were tested: Ridge Regression and Multi-task ElasticNet regression. Ridge regression is given a "bias term" hoping to lower variability and give a more accurate prediction while Multi-task ElasticNet combines ridge with lasso, which reduces the values of the weights [1]. We believed that trying these types of regression would allow us to gain different results than before. In order to capitalize on these techniques, multiple features are required in the model. The ridge regression has an alpha value that we tested changing to see if it gave a population prediction we thought was more realistic. We tested with very large alpha values which lowered the population prediction; however, ridge regression lowers the number of features that matter in a regression. We thought a high alpha value would provide better results, but because our regression model only has one input, we determined this model is inappropriate for our data. We determined Multi-task ElasticNet has a similar use case of models with multiple features.

Overall, these models produced similar results to basic linear regression and cemented the fact that more prediction variables are required to have a more accurate prediction. This

confirmation led to us wanting to try a model with birth and death rates added as those allow a more accurate representation of what the population is going through.

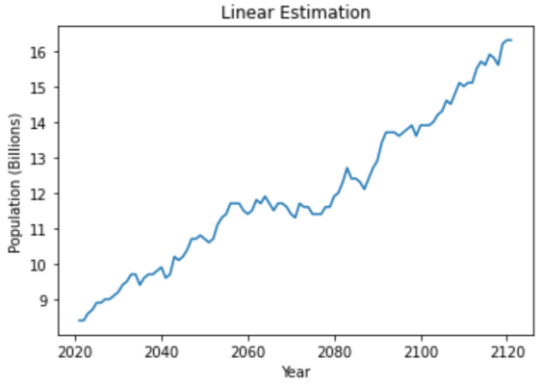
Method #2 Birth & Death Rate

When using linear regression, only the year number and population total was used to predict the earth's population in 2122. A model would produce more accurate results if it instead predicted the population with both the global birth and death rates. The Python package *BirDePy* can be utilized to predict discretely observed population sizes. The *BirDePy* package allows for easy simulation of non-trivial mathematical equations to predict the population [2]. Under the hood, *BirDePy* uses "Markov Chains" to discreetly simulate how a population changes over time in each state. Each new step or state is generated from the previous state information [3]. Meaning, each year's prediction is based on the previous year calculation. Each current state, *z*, is calculated by using the population birth rate function $\lambda(\theta)$ and population death rate $\mu(\theta)$ where θ is the population's statistical parameters. The parameters θ are essentially the input to the model, where *y*=birthrate and *v*= death rate. The population from 2020, the 2020 death rate, and the average birth rate from 1960-2019 (due to the inflated birth rate that year) was used to initialize the model. Then the bd.simulate.discrete function was utilized to predict the population each year. The *BirDePy* package allows many mathematical and statistical methods to be easily utilized to predict the population.

Model label	$\lambda_z(\boldsymbol{\theta})$	$\mu_z(\boldsymbol{\theta})$	$\boldsymbol{\theta}$
"linear"	γz	νz	γ, ν
"linear-migration"	$\gamma z + \alpha$	νz	γ, ν, α
"pure-birth"	γz	0	γ
"pure-death"	0	νz	ν
"Poisson"	γ	0	γ
"Verhulst"	$\gamma\left(1-\alpha z\right)z$	$\nu\left(1+\beta z\right)z$	γ, ν, α, β
"Ricker"	$\gamma z \exp\left(-(\alpha z)^c\right)$	νz	γ, ν, α, c
"Hassell"	$\frac{\gamma z}{(1+\alpha z)^c}$	νz	γ, ν, α, c
"MS-S"	$\frac{\gamma z}{1+(\alpha z)^c}$	νz	γ, ν, α, c
"Moran"	$\frac{N-z}{N}\left(\frac{\alpha z(1-u)+\beta(N-z)v}{N}\right)$	$\frac{z}{N}\left(\frac{\beta(N-z)(1-v)+\alpha z u}{N}\right)$	α, β, u, v, N
"M/M/1"	γ	$\nu 1_{\{z>0\}}$	γ, ν
"M/M/inf"	γ	νz	γ, ν
"loss-system"	$\gamma 1_{\{z<c\}}$	νz	γ, ν, c

Figure 2: *BirDePy* supported models

First the "linear" model utilizing linear birth and linear death rate was used to estimate the earth's population in 2122. The bd.simulate.discrete function was called. The birth and death rates described above were utilized as parameters to initialize the model as well as the initial population of the earth in 2020.. The figure belows shows a linear population growth rate for our simulation reaching a total of 16.3 billion.

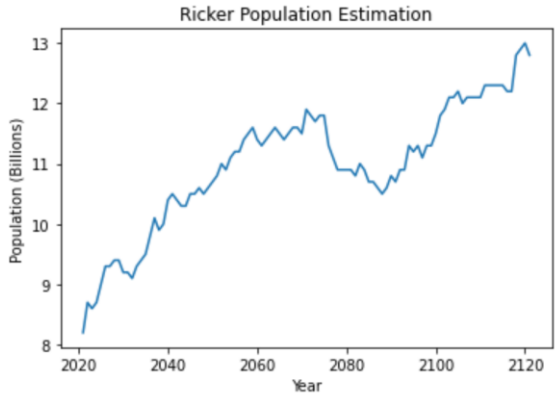



```
Final population estimation: 16.3 billion
```

Figure 3: Linear Population Estimation

The linear birth and death rate estimation was satisfactory. However, linearly increasing birth and death rates seemed unrealistic for the next hundred years even when observing data from the past 60 years. Instead the classic logistic growth "*Ricker model*" may produce a more accurate estimation. In this model, the birth rate decreases exponentially while the death rate

increases linearly still. The Ricker model seemed like the best to estimate the population due to its exponentially decreasing birth rate growth. Looking at the figure below, the population quickly increases due to the birth rate being higher than the death rate. Around 2060, the birth and death rate start to even out, due to the Ricker growth rates, and the population starts to decrease. In response to the rising death rate, the birth rate starts to increase around 2100. The population reaches a total of 12.8 billion with the Ricker model. The Ricker model final estimation is 12.8 billion. Also please note, the Ricker population estimation varies when run multiple times due to the simulation state changing randomly and unpredictably.




```
Final population estimation: 12.8 billion
```

Figure 4: Ricker Population Estimation

Conclusion

As obtained using the Ricker model, the population of the earth in 2122 will be 12.8 billion people. The Ricker model produced our most reasonable number estimation while utilizing the most logical math. The Ricker model predicts based on birth and death rates and utilizes an exponentially decreasing birth rate. The exponentially decreasing birth rate seems logical after examining the decreasing human birth rate over the last 60 years. Overall, the Ricker model produces a logical final population estimation while utilizing a logical mathematical method to arrive at the conclusion.

Works Cited

[1] Xu, W. (2021, June 17). *What's the difference between linear regression, Lasso, Ridge, and ElasticNet?* Medium. Retrieved February 4, 2022, from https://towardsdatascience.com/whats-the-difference-between-linear-regression-lasso-ridge-and-elasticnet-8f997c60cf29

[2] BirDePy documentation https://birdepy.github.io/

[3] Norris, James R., and James Robert Norris. Markov chains. No. 2. Cambridge university press, 1998.