Population Models I

Agenda and Checklist

The main goals of this worksheet are:

- To explore a couple of growth models in the context of modeling the world's population.
- To practice reading and using data to set model parameters.
- To use data and appropriate metrics to validate a model.

- ☑ Write your name here: Lilo Heinrich
- ☑ Write the name(s) of your studio partner(s) here: Marion Madanguit
- ☐ By midnight Thursday, September 12th: Scan this worksheet and submit it on Canvas.
- ☐ Read Chapters 7-8 for next week.

Reading Reflection Questions (15 minutes)

1. What are the two growth models seen in Chapters 5 and 6? What are the underlying assumptions of each model?

constant growth model - assumes population will continue growing at a constant rate

proportional growth model - assumes the birth and death rates are constant

2. How many model parameters are there for each of these models? What are they?

constant - initial & final time, initial population, annual growth

proportional - initial population, initial & final time, birth rate & death rate (alpha)

3. How did the book use data to set these parameters? Also, how are data used to validate the models?

they used data to input initial population and to calculate the annual growth rate and birth & death rate using best fit.

comparing the census population data collected to the predicted population in the model is how data is used to validate a model.

Chapter Notebooks

Reading Data

Work through the first section of the Chapter 5 notebook (*Reading data*) and look at the data that was loaded.

Q1: What is a `DataFrame`?

a table is contained in a DataFrame

Q2: We chose to index the data using the year (`index_col = 0`). What does it mean to have year as the index, and why did we choose to use year?

If we're looking at population over time, we don't know the population, and the population can be the same for multiple years. Years are sequential and unique and make a better index value.

Series

Work through the second section of the Chapter 5 notebook (*Series*).

Q3: Why is relative error a good metric to use to validate the model?

It shows the % difference from the actual result

Constant Growth

Work through the the next sections of the Chapter 5 notebook (*Constant growth, TimeSeries*). Then, complete the exercise at the end of the notebook.

Q4: Comment on the differences between this model and the one fit on 1950-2016. For each model, give one argument in support of using that model to estimate future population growth.

1950 predicted population = census [1970] - annual · 20y

the difference between the 1950 & the 1970 graph is the growth rate. The 1950 has a smaller growth rate than the 1970, to account for slower growth period from 1950-1970, while the 1970 is steeper and more linear

1950 - this model could be more useful because it uses a more complete set of data to determine growth

1970 - this model could be more useful because the growth rate significantly changes around 1970, so only using data from after should create a better prediction

System Objects

Work through the *System objects* section of the Chapter 6 notebook.

Q5: What's the difference between a `System` object and a `State` object? Why are these two separate types?

system variables don't change during simulation while state variables do

Proportional Growth

Work through the next few sections of the Chapter 6 notebook (*Proportional growth, Factoring out the update function, Combining birth and death*).

Q6: In comparing the proportional growth model to the data, what do you notice? What do you expect about the model's predictions for future years?

Since constant growth is always linear, it matched well from 1970 - 2016 and not on 1950-1970, while the proportional model fit the data for 1950-1970 well since the population growth rate was increasing, but didn't fit well from 1970 - 2016. Also since the proportional grows in a sort of exponential way, I predict it will have a large error on future data.

Q7: Why is it useful to factor out the update function? How might this be useful for the process of refining models in general?

This makes the update function easily interchangeable, reducing the amount of code you have to rewrite to make changes to this part of the code by taking it and making it a variable instead of hard-coded.

Proportional Growth with Two Rates

Work through the exercise at the end of the Chapter 6 notebook.

Q8: We could set three values for `alpha`, or five, or twenty .... What are the advantages of adding extra parameters to the model? The disadvantages?

Advantages - allows to account for changing birth rate over time, when things happen in society that affects the data

flexibility

Disadvantages - code becomes more complicated, but allows room for user interference and interpretation error

overfitting

Queueing Proto-Project (20 minutes)

Read and run the sample project about queueing theory that draws on the models you've seen in the book. The project can be found in CoCalc under **/soln/queue_soln.ipynb**. Answer the following questions about it.

$$\text{customers} = \frac{\text{customer}}{\text{min}} \cdot \text{min}$$

1. What is the motivating question? Is it an ~~explanatory~~, ~~predictive~~, or design question?

$$\text{min} = \frac{1}{\left(\frac{\text{cust}}{\text{min}}\text{ leaving} - \frac{\text{cust}}{\text{min}}\text{ arr}\right)}$$

What checkout configuration has the lowest customer wait time?

2. How could the answer to this question be used?

The answer can be used to determine whether the difference between different checkout configurations is significant enough to choose one over another for the store design.

3. What conclusions did the project draw? Based on what evidence?

single queue single checkout – 6.85 min

single queue mult checkout – 6.35 min

mult queue mult checkout – 6.75 min

By doing a sweep through different lambdas, or rates of customer arrival, they determined what the average customer wait time is in each scenario, factoring in different amounts of business. This showed that the single line double checkout was best.

However, from a practical point of view, the reason why it's better in the sim is that customers in multiple lines can't switch, but in the real world, they can, making single & mult line equivalent

Worksheet Reflection Questions (15 minutes)

1. In these chapters the same data was used to fit the model parameters and to validate the model. Do you think this is a better or worse approach than using a separate data set for model validation? Why?

This internal validation is worse than external validation because if you are making a model based on a set of data, of course you can make it fit that set of data, but it is a better test of validity to check it against new, separate data.

2. So far, what strategies have been most helpful in debugging your code? What strategies have been most helpful in avoiding errors in the first place?

- reading the error messages
- remembering python syntax
- using previous code blocks as a starting point to modify

3. Now that you have seen several types of models, are there any questions or applications that have popped into your mind or sparked an interest? Project 1 is coming up so you'll have a chance to explore some of these!

(data cross validation)

- money: interest gained, income, avg amount spent/year (similar to proportional population growth model)
- spread of diseases? # people infected etc
- most efficient way to board an airplane
- how many days is it likely to rain this month based on prev years' data for this region?
-

already covered:

- grocery checkout -queues
- population growth
- bikeshare/transportation