www.simiode.org
SIMIODE
A SYSTEMIC INITIATIVE FOR MODELING INVESTIGATIONS
& OPPORTUNITIES WITH DIFFERENTIAL EQUATIONS

STUDENT VERSION
Population Modeling with Census Data

Jean Marie Linhart
Department of Mathematics
Central Washington University
Ellensburg, WA USA

STATEMENT

The United States conducts a census every 10 years, which gives us data on the population from 1790 to the present [4], [7], [6]. Table 1 and Figure 1, below, give the census data for the United States population. Population data for individual states and cities in the United States are also available through the census, and data on the population of foreign countries and cities are available on the world wide web from a variety of sources. We are going to explore mathematical models for populations, starting with the United States census data and following up by exploring one other data set: a US state or city, a foreign country or a foreign city, or possibly the population of the entire world. Choose your second population wisely; country borders are not always static with time and so finding appropriate data can be difficult. If your interest is in Germany, whose borders were rewritten several times before and at the conclusion World War II, you will have an easier time with data for a city like Berlin, Stuttgart, or Munich rather than trying to model the entire country. Each student must choose a different population to model, in addition to the United States population.

Year	Population
1790	3929214
1800	5308483
1810	7239861
1820	9638453
1830	12866020
1840	17069453
1850	23191876
1860	31443321
1870	39818449
1880	50155783
1890	62947714
1900	75994575
1910	91972266
1920	105710620
1930	122775046
1940	131669275
1950	151325798
1960	179323175
1970	203302031
1980	226545805
1990	248718302
2000	281424603
2010	308745538

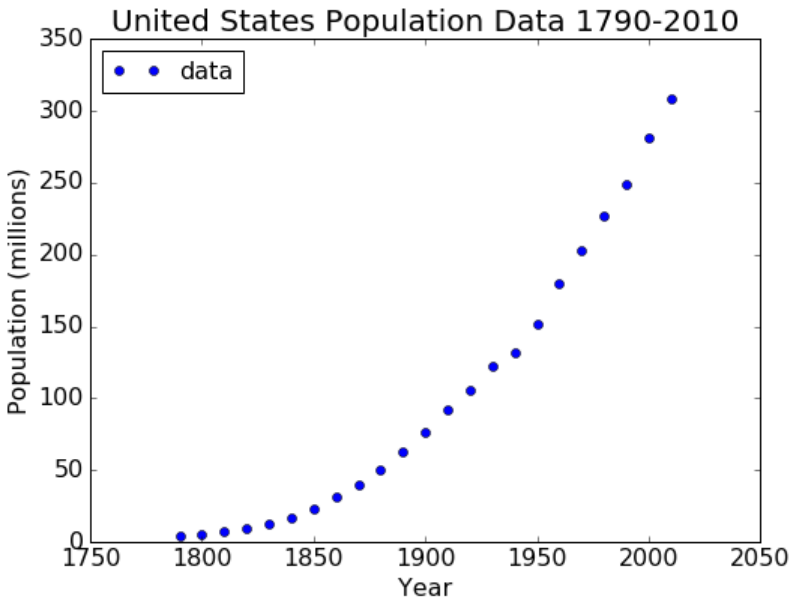

Table 1 (L) and Figure 1 (R): United States Population Data

You will investigate applying several population growth models to your data sets. The models are

- Exponential model

$$\frac{dP}{dt} = rP \qquad \text{with solution} \qquad P(t) = P_0 e^{rt}$$

- Logistic model (for an accessible history see [5])

$$\frac{dP}{dt} = r\left(1 - \frac{P}{L}\right)P \qquad \text{with solution} \qquad P(t) = \frac{L}{1 + \dfrac{(L - P_0)}{P_0} e^{-rt}}$$

- Gompertz model

$$\frac{dP}{dt} = r \log\left(\frac{L}{P}\right) P \qquad \text{with solution} \qquad P(t) = Le^{-\log\left(\frac{L}{P_0}\right)e^{-rt}}$$

You will determine the best parameter values in each model to fit your data. The error used to evaluate the model fit to the data is the R^2 value given by the **coefficient of determination**, or the **adjusted** R^2 value.

Your modeling goals are:

I. Determine which model is theoretically the best fit for your population data based on the assumptions of the models.

II. Determine which model actually has the best fit overall for the data sets. You cannot do this "by eye". You must have a numerical measure for best fit.

III. Using as much data as you can before 2010, find the best fit using only this data, and predict the population in 2010 (or another similar year). Determine which model(s) has (have) the best prediction(s). The error associated with the prediction is the (signed) relative error.

IV. Determine if one model is clearly better than the others.

V. Determine what relevant factors are omitted from the models.

VI. Determine if there are issues with the data that affects the suitability of these models.

Your goal is to write a project report which contains:

(a) A 150-250 word abstract summarizing your report. The abstract should include the main conclusions of your report.

(b) An introduction. In the introduction you should:

   i. Explain the theoretical differences between the models. To do this you will probably want to discuss growth rates, in particular the per-unit population growth rate. Make sure that you define your terms carefully and use them consistently for all models.

   ii. Determine if any model or models make more theoretical sense than the others for the populations you are looking at. Predict which you think will give the best fit and the reasons why. If your reasoning turns out to be incorrect, there is no penalty. In fact, this is good, as it reflects the scientific method at work. We make hypotheses, and sometimes they are not validated by the data.

(c) A methods section explaining how you will get parameters for the models, and how you will evaluate the models and their predictions.

   i. If you use nonlinear least squares curve fitting, explain how you linearize the models and get initial values for the parameters. Explain how you came up with your initial guess for L in the logistic and Gompertz models.

ii. Explain how you evaluate the fits.

iii. Describe how you make and evaluate the predictions.

(d) A results and conclusions section for each data set in which you include:

i. One graph of all the data and the model fits. A second graph of the data and the model fits made by omitting the most recent data point, along with the model prediction and the most recent data point. You should be able to clearly read the title, axis labels, and axis titles on your graphs in your report.

ii. Tables with the parameter values you find associated with your model fits. Example:

Fit	US population (millions)			Reduced Model Predictions		
Model	P_0	L	r	2010 Data	Prediction	Error %
Exponential	$P_0 =$		$r =$	308.7	341.4	10.6
Logistic	$P_0 =$	$L =$	$r =$	308.7		
Gompertz	$P_0 =$	$L =$	$r =$	308.7		
Fit	Bryan TX population (thousands)			Reduced Model Predictions		
Exponential	0.503		0.0264	65,660	76,394	16.4
Logistic	0.523	10.17	0.0421	65,660	65,480	-0.3
Gompertz	0.512	29.21	0.0122	65,660	68,036	3.6

iii. Evaluate which models performed best for the data set. Discuss any criticisms or comments about the population models based on your data, knowledge of history, and/or "uncommon sense". In other words, you should write conclusions about your data.

(e) A general conclusions section for the entire project which incorporates:

i. Table(s) with

- The error associated with each model for each data set (R^2 values).
- The signed relative error or percent signed relative error in predictions for each model and each data set.

Example:

	R^2			Prediction Error %		
Data	Exponential	Logistic	Gompertz	Exponential	Logistic	Gompertz
USA				10.6%		
Bryan, TX	0.950	0.998	0.998	16.35%	-0.27%	3.62%

ii. An evaluation of which model(s) performed best. The table should allow for an easy comparison.

iii. Analyze the meaning of the statement "these models have too much memory to produce accurate predictions."

iv. Criticisms and potential improvements for these population models. What do these models fail to take into account?

v. Criticisms and potential improvements for the population data. Are there aspects of the data or data collection that impact the appropriateness of the data for use with these models?

(f) A bibliography where you cite your sources of information. Where did you get your data? Where did you learn more about these models?

A STEP-BY-STEP GUIDE TO THE PROJECT

What follows is information, exercises, and questions to help you think about these models and prepare your report.

Picturing a good model

A graph of the data is provided in Figure 2. We will explore some mathematical models for this data. Questions and activities for you to complete are numbered below:

1) Draw a curve on the data that you think would be a reasonable model for this data.

2) Draw a curve that you think would be a poor model for this data.

3) Using complete sentences, explain why you picked your reasonable model and why you picked your poor model.

Possible Models

Figure 3 contains graphs of several functions that might be used to model populations. You are likely familiar with a linear model and an exponential model. The logistic and Gompertz models both have two horizontal asymptotes, one at zero, and one at $L > 0$.

4) Judging by shape characteristics alone, rank these models from best to worst for the United States data and explain your reasoning. Since the logistic and the Gompertz models have very similar shapes, you may put both in the same ranking.

Introducing PPGR

One way to evaluate potential competing models is to examine the assumptions they give us about our data. With a population model, we often look at the Per (unit) Population Growth Rate (PPGR). If the population as a function of time is $P(t)$, then the PPGR is

$$\text{PPGR} = \frac{1}{P}\frac{dP}{dt} \tag{1}$$

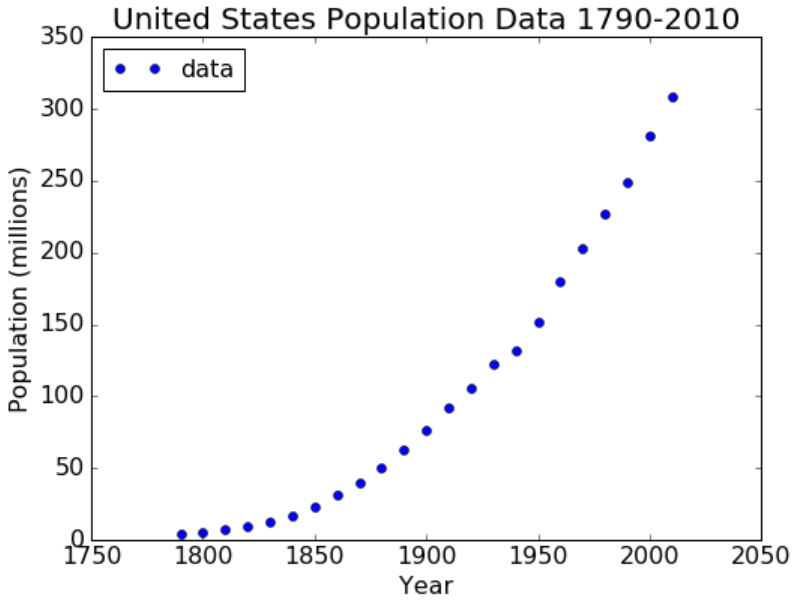

Figure 2: Draw a curve on the data that you think would be a reasonable model for the data. Draw a curve on the data that you think would be a poor model for the data.

The PPGR measures how much one unit of population, in our case, one human, contributes to the growth (or reduction) in the population per unit time.

5) Analyze (1) and find the units on PPGR.

6) Say you are given an estimate that human PPGR is equal to $\frac{1}{2}$ per year with reasoning that women are half the population, and each woman can give birth once a year, contributing a new member to the population. Do you think this PPGR is too high or too low? Why?

7) Think about the factors this estimate does not take into account, and generate a better estimate (including units) for how big you think the PPGR should be. What did you take into account? Explain how you came up with your estimate.

8) Should the PPGR be constant and unchanging, or are there situations or scenarios in which you would expect it to increase or decrease?

PPGR and modeling assumptions

Each of the possible models given above can be represented as both a function and the differential equation whose solution is that function. Thus far we have considered the following models

- Exponential model

$$\frac{dP}{dt} = rP \qquad \text{with solution} \qquad P(t) = P_0 e^{rt}$$

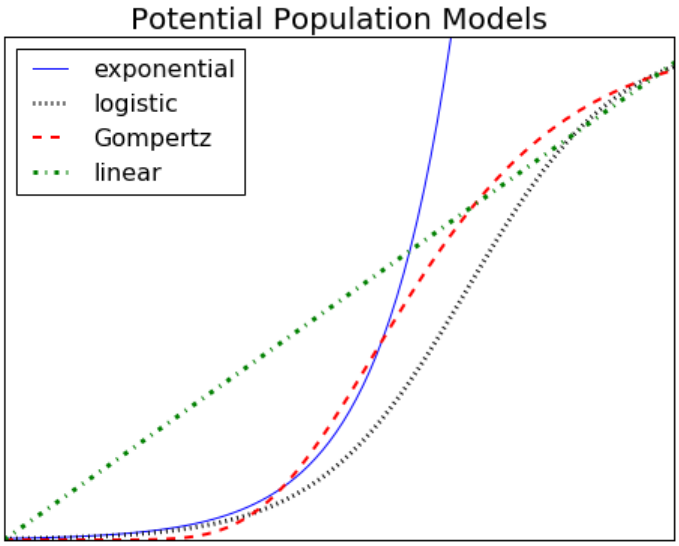

Figure 3: Possible models for the United States population data.

- Logistic model (for an accessible history see [5])

$$\frac{dP}{dt} = r\left(1 - \frac{P}{L}\right)P \qquad \text{with solution} \qquad P(t) = \frac{L}{1 + \dfrac{(L - P_0)}{P_0}e^{-rt}}$$

- Gompertz model

$$\frac{dP}{dt} = r\log\left(\frac{L}{P}\right)P \qquad \text{with solution} \qquad P(t) = Le^{-\log\left(\frac{L}{P_0}\right)e^{-rt}}$$

9) Calculate the PPGR for each of these models.

10) What happens to the PPGR in each of these models when there are very few people? If there aren't very many people, there should be plenty of resources or raw materials in the environment to support human life, so you can also think of this as a situation in which resources are abundant or unlimited. Explain what your result means for humans in each model. Do some of these limits make more sense for human populations than others?

In both the Gompertz and in the logistic model, L represents the *limiting population*, meaning largest population the environment can support, which also goes by the name *carrying capacity*. In the graph of the logistic and Gompertz equation, you can see that $P(t)$ grows to a horizontal asymptote at L.

11) Analyze the equations and determine the units on L.

12) What happens to the PPGR in the logistic and Gompertz models as the population gets close to having L people?

13) How does this relate to the behavior you see in the graphs in Figure 3?

14) Does this make sense for human populations?

15) In the exponential and linear model, the population can keep getting larger and larger. In both of these models what happens to the PPGR as the population gets bigger and bigger? Does this make sense for human populations?

16) Based on your analysis of PPGR, rank the models from those that make the most sense for human populations to those that make the least, and explain the reasoning behind your rankings.

Both the shape and the modeling assumptions for the linear model make it a bad match for our population data. Thus, we drop it from our discussion for the remainder of the project.

Solving ordinary differential equations

17) Show how to use separation of variables to solve the differential equation models to get the exponential, logistic and Gompertz modeling functions. Assume an initial condition $P(0) = P_0$. You may also need integration by substitution and/or partial fractions to integrate the logistic and Gompertz differential equations.

Parameter Estimation Using PPGR

The United States population data [4], [7], [9] is given in Table 1. This same data is also in the file `USPopulation.csv`, which is a text file tabular data in the form of"comma separated values" that is read by most spreadsheets and some computer algebra systems. The first parameter we need to do population modeling is $P_0 = P(0)$ which is the population at time zero.

18) What year should we pick to call 0? We will refer to this as our zero year.

19) Why did you make this choice?

Figure 4 shows a plot of a calculation for the PPGR vs. the population data. Most of this discussion mimics the discussion in [6]. The furthest (blue) data point to the right represents a PPGR of approximately 0.010 at a population of 2.8×10^8.

20) Describe a way to calculate PPGR from the data, and calculate a few values by hand. Can you think of different ways to calculate PPGR from the data?

21) Use a computer algebra system to read in the United States population data from `USPopulation.csv` and use it to generate a plot of PPGR vs. P, like that in Figure 4. Check your mathematics by checking against the values you calculated by hand. Your graph should look similar to that in Figure 4, but may not be exactly the same.

22) How can you use the average PPGR and the best line approximation to the PPGR data to find parameters for the exponential and logistic models?

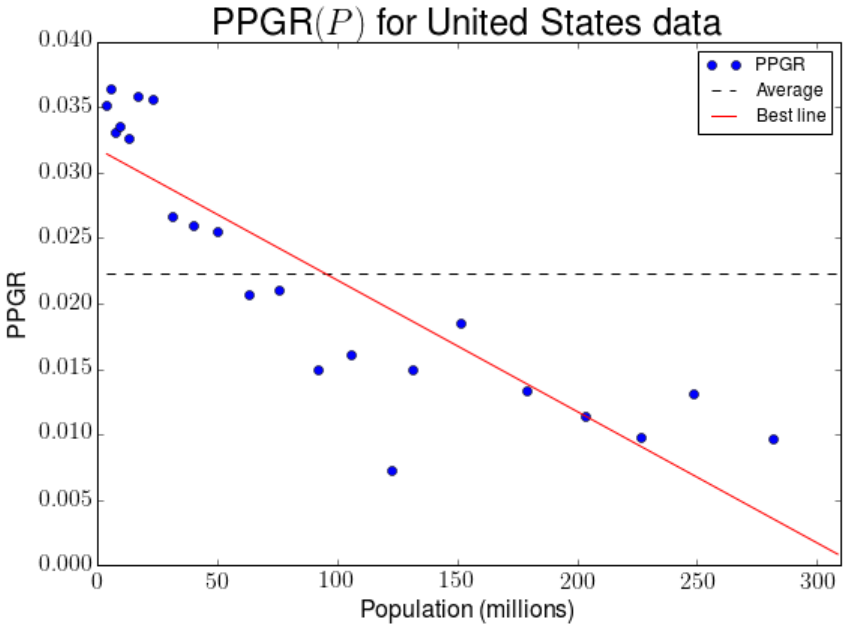

Figure 4: A graph of PPGR vs P for the United States population data.

We can use least squares curve fitting to find the best fit line to the PPGR vs. P data. Let's start with the situation where you have data of the form (x_i, y_i), where x is the independent variable, and you want to model with a linear function $y = f(x, a, b)$ that depends on two parameters, a (the slope of the line), and b (the y-intercept) as well as x. Our problem is to minimize the sum of the squared errors, also called the *sum of squared residuals* (SS_{res})

$$\text{SS}_{\text{res}} = \sum_i (y_i - f(x_i, a, b))^2$$

with respect to a and b. The method of ordinary least squares uses linear algebra to minimize the sum of the squared residuals. Linear algebra methods for least-squares minimization work not only for lines, but also for functions f that are polynomials in x; $f(x, a_0, \ldots, a_n) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$. Ordinary least squares returns unique values for the parameters $(a_0, \ldots, a_n)$ that minimize the sum of the squared residuals.

23) In Figure 4, what is x, the independent variable? What are its units? What is y? What are its units?

24) In Figure 4, draw and label the distance $y_i - f(x_i, a, b)$ for a single data point, where $f(x_i, a, b)$ is the function in represented by the solid line (best line).

25) What is the formula for the function $f(x, a, b)$ in terms of x, a and b?

26) Use your computer algebra system to find the best line to the PPGR(P) vs. P data, as well as

the average of PPGR(P), and plot both, as in Figure 4. There will be a built-in routine to do this.

27) Find the values of the parameters for the exponential model, and those for the logistic model using your PPGR(P) vs. P graph, average, and best line. For exponential and logistic, you need P_0 and r. For logistic, you will also need L, the limiting population.

28) Use your parameter values to make exponential and logistic models for the data, and plot these with the USA population data. Are the models a good match for the data?

29) What would you expect a graph of PPGR(P) vs P to look like if the Gompertz model was best suited to the data?

Nonlinear least squares parameter estimation

There are many nonlinear functions, such as the exponential, logistic, and Gompertz functions for which ordinary least squares, if you can find a way to use it at all, does not give the parameter values that minimize the sum of the squared residuals. For these functions, we turn to nonlinear least squares estimation. Nonlinear least squares estimation generally uses a stepping method to approximate the minimum of the sum of the squared residuals and the associated parameter values. It may not succeed because there may not be a unique minimum of the sum of the squared residuals, or the region around the minimum might be very flat, making it difficult to find the location of the minimum.

Similar to ordinary least squares, nonlinear least squares methods also require you to provide your data (x_i's and y_i's), and the function $f(x, a, b, c)$, in which we assume there are three unknown parameters to be estimated, a, b, and c. Often, in order to get a satisfactory answer, you must also provide a good set of starting values for the parameters, so that the routine starts somewhere near the minimum of the sum of squared residuals.

Fortunately, for our models, we can get a good set of starting values for the parameters using ordinary least squares, but to do this we have to rewrite the equations in the form of a line, $y = ax + b$, where a and b are are the unknown parameters for slope and y-intercept. This is easy for the exponential function, where taking a logarithm of both sides transforms it to a linear equation

$$\begin{aligned} P(t) &= P_0 e^{rt}, \\ \log(P(t)) &= \log\left(P_0 e^{rt}\right), \\ \log(P(t)) &= \log(P_0) + rt. \end{aligned}$$

The last equation is the linearization of the exponential equation.

23) What are the parameters to be estimated in the exponential equation?

24) What do you use for your data if you want to use ordinary least squares estimation to find the parameters in the exponential model? What are your y_i's and x_i's?

25) Using ordinary least squares estimation, you will get the slope, a, and y-intercept, b. What are P_0 and r for the exponential equation in terms of a and b?

26) Use your computer algebra system to get values for P_0 and r using ordinary least squares estimation.

27) Now use a nonlinear least-squares estimator to find improved values for P_0 and r. Graph your data with the ordinary least squares estimate for P_0 and r and with the new values from the nonlinear least-squares estimator.

28) With the United States data you should clearly see an improvement in model fit from the original parameters found by ordinary least squares and the parameters from nonlinear least-squares. Describe how the nonlinear least-squares model appears to be better.

We can also linearize the logistic and the Gompertz equations. Recall that they are

$$P(t) = \frac{L}{1 + \frac{(L-P_0)}{P_0}e^{-rt}} \quad \text{(logistic)} \qquad \text{and} \qquad P(t) = Le^{-\log\left(\frac{L}{P_0}\right)e^{-rt}} \quad \text{(Gompertz).} \qquad (2)$$

For both, we first must look at the data and from the data come up with a reasonable starting guess for the value of L. Then, in both equations, we solve for the part that is in the form Ae^{-rt}. We know from our work with the exponential equation that we can take logarithms to get a linearization. In the logistic equation, the term of the form Ae^{-rt} is $\dfrac{(L-P_0)}{P_0}e^{-rt}$, and so our logistic equation becomes in linearized form

$$\frac{L}{P(t)} - 1 = \frac{(L-P_0)}{P_0}e^{-rt},$$
$$\log\left(\frac{L}{P(t)} - 1\right) = \log\left(\frac{(L-P_0)}{P_0}\right) - rt.$$

29) Look at your data and choose a value for L that is too small, one that is too big, and one that is reasonable. How do you know the value that is too small is really too small? How do you know the value that is too big is really too big? Why is your reasonable choice for L reasonable?

30) Using your reasonable value for L, what are the y_i data values in ordinary least squares estimation with the linearized logistic equation? What are the x_i data values?

31) Once you have a and b from ordinary least squares with the linearization of the logistic equation, how do you get initial values for P_0 and r from these?

32) Use your computer algebra system to get values for P_0 and r using ordinary least squares and your reasonable value for L.

33) Now use nonlinear least squares estimation to find improved values for L, P_0, and r. Graph the data with the logistic model created from the reasonable value for L, and the ordinary least squares estimate for P_0 and r, and the logistic model created from the nonlinear least squares estimation for L, P_0, and r. Which looks like the better model? Why does it look like the better model?

34) Now, find the linearization of the Gompertz model using your reasonable value for L, and repeat the work in questions .30 to .33 for the Gompertz model.

35) Why do you think that nonlinear least squares gets us better parameter values than the values for using the linearization and ordinary least squares? Hint: what does the logarithm do to large data values and large errors in the linearized models?

Goodness of fit: coefficient of determination

Sometimes we can see from a graph that one set of model parameters gives us a better fit to data than another, but there are many cases in which it is not clear which model gives a better fit to the data. We prefer to use a quantitative measure for goodness of fit, such as the *coefficient of determination* known as R^2. The coefficient of determination tells us how well a model does compared to a linear model with an intercept only. Our data is a set of n values indexed (x_i, y_i). Let $\bar{y} = \frac{1}{n}\sum_i y_i$ be the mean of the dependent variable, and let $f(x_i)$ be the values obtained by the model, then ([8], [2])

$$R^2 = 1 - \frac{\sum_i (y_i - f(x_i))^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

An R^2 of 1 indicates the model perfectly fits the data, and generally, the closer R^2 is to 1, the better the model fit. But since R^2 is a comparison between a given model and a linear model with an intercept only, you cannot directly compare two models with their R^2 values [1]. Another problem is that increasing the number of parameters in the model implicitly increases the power of the model to explain the data, and so comparing the exponential model which has two unknown parameters to the logistic or Gompertz with three unknown parameters, is inherently unfair. It is known that R^2 will never decrease as parameters are included in a regression model.

A related measure is the adjusted R^2 value which takes the number of parameters, p, in the model compared to the number of data points, n, into account. Adjusted R^2 is denoted $\bar{R}^2$ and a formula for it is [8], [2]

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = R^2 - (1 - R^2)\frac{p}{n-p-1}$$

where n is the number of data points in the model, and p is the number of parameters in the model.

Interpretation of adjusted R^2 is similar to that for R^2, in that values closer to 1 are considered better. But adjusted R^2 also cannot be used compare two models directly, since adjusted R^2 is a measure of suitability comparing the given model with a linear model with only an intercept [1]. One model might do slightly better than a second in comparison to the linear model with only an intercept, but that does not mean that it does better when directly compared to the second model.

36) Is R^2 or adjusted R^2 a better measure for comparing the goodness of fit between the exponential, logistic and Gompertz models? Why?

37) Calculate R^2 or adjusted R^2 for the exponential, logistic and Gompertz models, and interpret your findings. What does this tell you about which model performs best?

Predictions

Another way to evaluate how well the model performs is to investigate how good it is at making predictions. Since we do not want to wait for the next census to be taken, this is evaluated by a "Leave One Out" (LOO) method of omitting the final data point, creating a model based on the remaining data (called the reduced data), then comparing the prediction of the reduced/LOO model and the most recent data point. Comparison is usually done by signed relative error:

$$\text{Error} = \frac{\text{predicted} - \text{actual}}{\text{actual}}$$

The relative error always compares to the actual data value, not to the model predicted value.

38) When is the signed relative error, as defined above, positive? When is it negative?

39) Why do we calculate the signed relative error as

$$\text{Error} = \frac{\text{predicted} - \text{actual}}{\text{actual}} \qquad \text{instead of} \qquad \text{Error} = \frac{\text{predicted} - \text{actual}}{\text{predicted}}$$

40) Why do we omit that final data point in finding the model parameters before evaluating the correctness of the prediction of the final data point? Why not just use the model based on all of the data?

41) Find model parameters using the reduced/LOO data set, leaving out the most recent data point. Do you need to linearize again, or do you already have good starting values for the parameters? What starting values for the parameters can you use?

42) Use the reduced/LOO model parameters to predict the most recent data point, and calculate the signed relative error.

A last way to evaluate a model's performance, at least for the logistic and Gompertz models, is to look at the estimate or prediction of L, the limiting population, and see if what you know about the population gives you any insight as to which is more reasonable.

43) Do the predicted L values from the logistic and Gompertz models give you any insight as to which is more reasonable?

Conclusions

44) With all the information and analysis available to you, which model do you think performed the best, and why? Is this the one you predicted would be the best from your analysis of the model hypotheses?

45) Someone said, "These models have too much memory to produce accurate predictions." Figure out what is meant by that and use it as part of your criticism of these models.

46) What factors relevant to population size are omitted from these population models?

47) Consider the population data sets. Are there any aspects of the data or the method of collecting the data that impact the appropriateness of the data for modeling?

REFERENCES

[1] Anderson-Sprecher, Richard. 1994. Model Comparison and R^2. *The American Statistician*. 48:2, May 1994. pp. 113-117.

[2] Cannon, Ann R., et. al. 2013. *STAT2: Building Models for a World of Data*. New York: W. H. Freeman and Company.

[3] Cohan, C. L. and Cole, S. W. 2002. Life course transitions and natural disaster: marriage, birth and divorce following Hurricane Hugo. *Journal of Family Psychology*. 16:1, Mar 2002. pp. 14-25.

[4] Gibson, Campbell and Jung, Kay. 2002. *Population Division: Historical Census Statistics on Population Totals by Race 1790-1990....* U.S. Census Bureau, Washington DC 20233. Available on the world wide web at `https://www.census.gov/content/dam/Census/library/working-papers/2002/demo/POP-twps0056.pdf` Accessed 24 August 2017.

[5] Kingsland, Sharon. 1982. The Refractory Model: The Logistic Curve and the History of Population Ecology. *The Quarterly Review of Biology*. 57:1, Mar 1982. pp. 29-52.

[6] Mesterton-Gibbons, Mike. 2007. *A Concrete Approach to Mathematical Modeling*. Hoboken, NJ: John Wiley & Sons, Inc.

[7] U. S. Census Bureau. 2010. *Resident Population Data (Text Version)*. U. S. Census Bureau, Washington DC 20233. Available on the world wide web at `https://www.census.gov/2010census/data/apportionment-pop-text.php`

[8] Wikipedia Contributors. 2017. *Coefficient of Determination*. Wikipedia, The Free Encyclopedia, 31 August 2017, `https://en.wikipedia.org/wiki/Coefficient_of_determination`. Accessed 31 August 2017.

[9] Wikipedia Contributors. 2017. *Demography of the United States*. Wikipedia, The Free Encyclopedia, 24 August 2017, `https://en.wikipedia.org/wiki/Demography_of_the_United_States`. Accessed 24 August 2017.