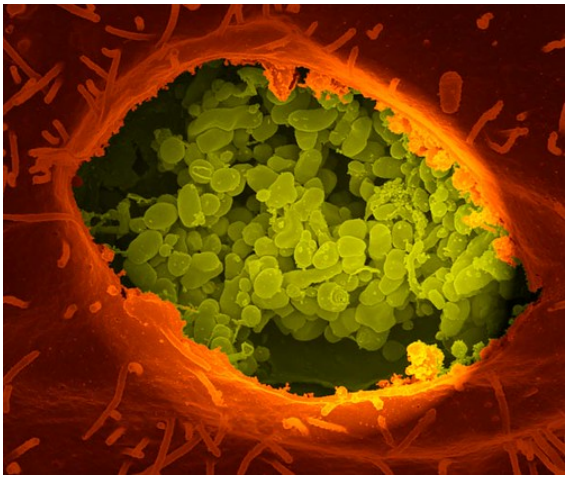
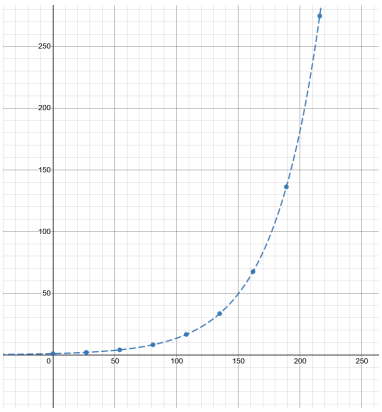
Fitting Exponential and Logistic Growth Models to Bacterial Cell Count Data



Introduction

Bacterial colonies are often used as a standard example for population growth models since they are easy to study under laboratory conditions. Colonies can be grown on an agar nutrient plate relatively free from external pressures, and population counts can be obtained by several laboratory methods (such as measuring the amount of light that can pass through the plate, which decreases in a predictable way as more bacteria accumulate).

Under ideal conditions bacteria reproduce at a fairly consistent rate. Assuming that there are no limitations due to nutrients or space, the population of bacteria is generally thought to grow *exponentially*, in which case a graph of the bacterial population versus time would follow an exponential curve.



Of course, in the real world, most things don't follow mathematical models exactly. Random noise due to fluctuations in individual cell behavior and measurement errors mean that our real-world measurements are unlikely to fit perfectly onto an exponential curve. However, it is still useful to be able to fit a model as closely as we can to the real-world data, since then the model can be used to make predictions or to better understand the mechanisms of the underlying system.

The goal of this project is to explore how to fit a simple mathematical model to describe a set of noisy real-world data. Along the way we will also explore the exponential growth model in more depth, and see how the method used to display data can help to better understand and analyze it.

Learning Objectives

After completing this project, you should be able to:

- Generate a semi-log plot of data, and understand when doing so might be useful.
- Apply your understanding of minimizing a function to find a mathematical model that minimizes an error.
- Use mathematical analysis in conjunction with computational technology, by developing theoretical results by hand and then implementing those results in a program to quickly process data.

Part 1: Exponential Growth and Computational Technology

The exponential growth model comes from the **Law of Natural Growth**, which is a differential equation of the form

$$\frac{dP}{dt} = rP \tag{1}$$

The Law of Natural Growth states that the rate of change in the population P with respect to time t is proportional to the current population, with a constant of proportionality $r > 0$ that describes how quickly the given population grows. The solution to this differential equation (i.e. the function $P(t)$ whose derivative satisfies the above property) is the **exponential growth model**

$$P = P_0 e^{rt} \tag{2}$$

where P_0 is the initial population (i.e. the population when $t = 0$). This equation explicitly gives the population as a function of time (in minutes).

Activity

During this project you will be making use of computational technology for quickly processing large amounts of data. For simplicity this activity uses Microsoft Excel (these same commands should also work in free alternatives like Google Sheets and OpenOffice Spreadsheets, but the instructions below will refer to "Excel"). As an initial exercise, we will use Excel to create a plot for the exponential growth function

$$P = 40e^{0.015t} \tag{3}$$

1a. Open the file `first_exercise.xlsx`. You should see a spreadsheet with the first two columns labeled "t" and "P", with the first column containing a list of times between 0 and 600 minutes in 30-minute increments.

1b. You will be using the second column to generate the population values for each of these times. In Excel, calculations are performed within cells by beginning the cell entry with an equals sign (`=`) followed by whatever mathematical operations or functions are needed. The contents of other cells can be included in these formulas by referring to their column/row coordinate (for example, the cell in the fourth column and third row would be D3).

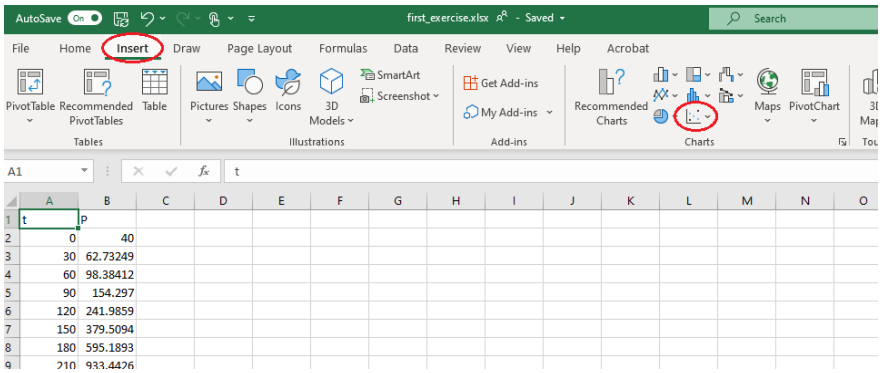
According to the model in equation (3), the first population value (in cell B2) should be defined as the value of the function $40e^{0.015t}$ evaluated at the first time value (in cell A2). To do this, enter the following formula into cell B2:

```
=40*EXP(0.015*A2)
```

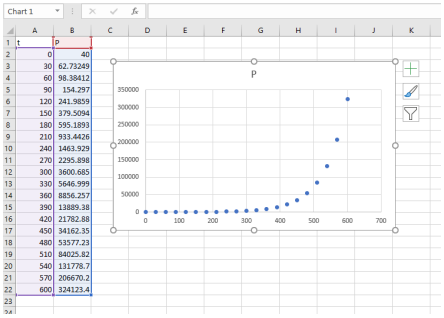
(Note that you can also click on the cell A2 to insert it into the formula rather than typing "A2" manually.) Multiplication in Excel is performed with the star `*` operator, and `EXP( )` is Excel's version of the exponential function e^x. After you enter the formula and press [Enter] you should see the number 40 appear in the cell, which is indeed the correct population at time $t = 0$.

1c. In order to quickly copy this same formula into all of the remaining rows of column 2, click on cell B2 to highlight it, then drag the bottom right corner down to copy it into the cells below. You should see numbers appear in cells B2 through B22. If you click on any of them and look at the "fx" row above the table you'll see that their formulas have automatically incremented the time cell that they refer to, so for example the formula in cell B16 is `=40*EXP(0.015*A16)`.

1d. Finally, use these data to generate a plot by highlighting all of the filled cells (cells A1 through B22), clicking Insert on the ribbon, clicking the scatter plot button, and choosing an appropriate type of scatter plot (in Google Sheets, plots can be inserted by selecting the Insert drop-down menu and clicking Chart to access the chart editor).



You should see a scatter plot of the exponential growth curve appear. If you wish, you can double-click on various elements of the plot to select different style options for them from the format menu on the righthand side.



1e. You can save this image for your submission either by taking a screenshot or by right clicking it and selecting "Save as Picture...".

Part 2: Preliminary Analysis and Semi-Log Plots

Suppose you are studying a colony of *Staphylococcus equorum*, a species of bacteria found in the skin of horses (and some smoked sausages[1]). Suppose that a culture of *S. equorum* has been isolated and grown in a petri dish, and that population measurements have been collected at 30-minute intervals (see the Excel spreadsheet `data_table.xlsx`). From a preliminary observation of the data the population of *S. equorum* seems to grow exponentially, so it would be reasonable to begin by attempting to model the population using the exponential growth model.

Activity

2a. Open the file `data_table.xlsx` (which includes population measurements at times ranging from 0 to 300 minutes) and create a scatter plot of the data.

2b. The population appears to roughly follow the exponential growth model given in equation (2). The data table gives the initial population as $P_0 = 16$, so in particular the population might follow the model

$$P = 16e^{rt} \tag{4}$$

for some growth rate r.

Go to the Desmos link labeled *Fitting an Exponential Model*[2], which displays a scatter plot of the data alongside the exponential growth curve $P = 16e^{rt}$. Play with the slider to see how r affects the shape of the graph. What value of r seems to cause the model to most closely match the data?

2c. Your eventual goal will be to find a mathematically precise way to choose the "best" value for r, in the sense of producing the model that most closely represents the real-world data. That turns out to be hard to do for the exponential model, itself, in part because the rapid increase in the function values causes minor differences in the first few population values to appear insignificant next to the larger values. This issue can be addressed by transforming the data in some way that will display all of the values on roughly the same scale.

In STEM applications, data that encompass a wide range of scales are often displayed on a **semi-log plot**, where the vertical axis shows the *logarithm* of the dependent measurement while the horizontal axis remains unmodified. You can generate a semi-log plot of the data by evaluating the natural logarithm of each population value.

[1] S. Leroy, I. Lebert, J.-P. Chacornac, P. Chavant, T. Bernardi, and R. Talon. Genetic diversity and biofilm formation of *Staphylococcus equorum* isolated from naturally fermented sausages and their manufacturing environment. *International Journal of Food Microbiology*, 134(1–2):46–51, 2009. `https://doi.org/10.1016/j.ijfoodmicro.2008.12.012`.

[2] Fitting an Exponential Model: `https://www.desmos.com/calculator/fm6rsv7mmp`

The Excel function for natural logarithm is `LN( )`. In column C of `data_table.xlsx`, labeled "$\ln(P)$", calculate the natural logarithms of the population measurements in each row, and then display $\ln(P)$ versus t on a scatter plot. This can be accomplished by holding [Ctrl] while selecting two separate data ranges (the times in cells A3 through A13 and the log-populations in cells C3 through C13), and then inserting a scatter plot as in Part 1.

2d. The plot of $\ln(P)$ versus t seems to roughly follow a straight line, and this can be explained using the properties of logarithms. Taking logarithms of both sides of equation (4) produces

$$\ln(P) = \ln(16e^{rt})$$

Use the properties of logarithms to show that the righthand side of this equation is equivalent to a linear function of t. What is the slope of this linear function? What is the intercept?

2e. The previous problem shows that, if P grows exponentially with respect to t, then $\ln(P)$ grows linearly with respect to t. It's far easier to fit a linear function to data than an exponential function, so the next goal will be to try to find the linear model that most closely matches the data. From the previous problem you should have found that the slope is r while the intercept is $b \approx 2.772589$, so the linear model has the form

$$\ln(P) = rt + 2.772589 \tag{5}$$

Go to the Desmos link labeled *Fitting a Linear Model*[3], which displays a scatter plot of the log-transformed data alongside the linear curve $y = mx + 2.772589$. Play with the slider to see how the slope affects the shape of the graph. What slope seems to cause the model to most closely match the data?

Part 3: Fitting the Linear Model to the Data

From Part 2 you might suspect that the population of *S. equorum* follows the exponential growth model, in which case its log-transformed population should follow a linear growth model. Moreover, the slope of the linear model corresponds to the growth rate of the population model. While you could try to fit a linear model to the data by eye, that's not something that a computer is equipped to do, and it wouldn't generalize well to data sets with more than one independent measurement (which would need to be displayed in more than two dimensions). As a result, the next goal will be to develop a mathematically precise, programmable way of calculating the "best" slope for the line.

There are many possible definitions for what "best" means in this case, but in general the goal is to choose our model's parameters in a way that minimizes some kind of *error*, or the difference between what the model would predict for a certain time and the actual value at that time. Since this error depends on the model's parameters, finding the model that minimizes the error is an *optimization* problem. In the next activity you will use your knowledge of optimization techniques to find the slope that minimizes the model's error.

[3]Fitting a Linear Model: `https://www.desmos.com/calculator/hwcsxo7oyc`

Activity

3a. For any given time t, the linear model in equation (5) provides a prediction for the log-population. Like most real-world data, the semi-log plot from Part 2 was not a perfectly straight line, and so some of these predictions will be slightly wrong. For example, if you attempted to use the model with $r = 0.015$,

$$\ln(P) = 0.015t + 2.772589$$

for the log-transformed data, then for time $t = 180$ minutes the model would predict a log-population of

$$0.015 \cdot 180 + 2.772589 \approx 5.472589$$

but the actual log-population at $t = 180$ should be $\ln(675) \approx 6.514713$, for an *absolute error* of

$$|5.472589 - 6.514713| \approx 1.042124$$

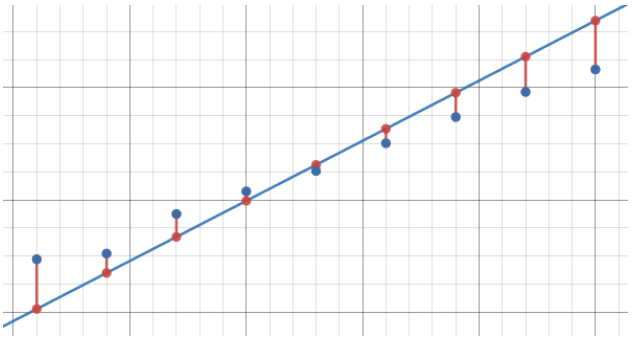
For a variety of computational and analytical reasons, statisticians usually prefer to study **squared error**, which is the *squared* difference between a predicted value and the measured value

$$(5.472589 - 6.514713)^2 \approx 1.086022$$

The squared error at a particular point clearly depends on the model being used. For example, choosing $r = 0.02$ instead of $r = 0.015$ would produce a squared error of

$$[(0.02 \cdot 180 + 2.772589) - 6.514713]^2 \approx (6.372589 - 6.514713)^2 \approx 0.020199$$

which is much smaller. A reasonable criterion to use in choosing the parameters for the model would be to try to minimize the sum of squared errors over all 11 data points.



Go to the Desmos link labeled *Linear Model Error*[4], which displays the linear model from before, now with the errors drawn as lines between the predicted and actual log-populations, and with a numerical readout of the sum of squared errors. Play with the slider to see how the slope affects the sum of squared errors. What value of r seems to cause error to be minimized? How does this relate to your estimate from Part 2?

[4]Linear Model Error: `https://www.desmos.com/calculator/jzecmsidnt`

3b. Setting aside the bacterial growth model for a moment, consider a more general problem where the goal is to fit a linear function of the form

$$f(x) = mx + b$$

to a set of three data points: (x_1, y_1), (x_2, y_2), and (x_3, y_3). Suppose that you've already decided on a value of b, so the only task left is to choose the slope m.

For any of the measured x-values, say x_2, the model would predict the corresponding y-value to be $f(x_2) = mx_2 + b$, while the true y-value should actually be y_2. The squared difference between the predicted value and the measured value is

$$(mx_2 + b - y_2)^2$$

and the sum of all squared errors over the entire data set would be

$$S(m) = (mx_1 + b - y_1)^2 + (mx_2 + b - y_2)^2 + (mx_3 + b - y_3)^2$$

This sum $S(m)$ is a quadratic function of m. Use the optimization techniques you've learned in class to find the value of m that minimizes S. Show that the optimal slope is

$$m = \frac{(x_1y_1 + x_2y_2 + x_3y_3) - (bx_1 + bx_2 + bx_3)}{x_1^2 + x_2^2 + x_3^2}$$

and use the second derivative test to verify that this corresponds to a local minimum of S.

(Note: Since S is quadratic, this also implies that it is the global minimum.)

3c. Returning now to the bacterial growth model, the formula for the optimal slope from the previous problem can be generalized to work for any number of data points. If you wanted to fit a linear model to n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_{n-1}, y_{n-1}), (x_n, y_n)$, then the method used in the previous problem could be used to obtain an optimal slope of

$$m = \frac{(x_1y_1 + x_2y_2 + \cdots + x_{n-1}y_{n-1} + x_ny_n) - (bx_1 + bx_2 + \cdots + bx_{n-1} + bx_n)}{x_1^2 + x_2^2 + \cdots + x_{n-1}^2 + x_n^2} \tag{6}$$

This formula can be used to fit the linear model in equation (5) to the log-transformed population data. The x-values are the time measurements t, the y-values are the log-population measurements $\ln(P)$, the intercept is $b \approx 2.772589$, and the slope you're trying to find is $m = r$. The three sums included in equation (6) can be evaluated quickly using Excel, but some setup will be required first.

Use columns E, F, and G of `data_table.xlsx`, labeled "$t\ln(P)$", "bt", and "t^2", to compute the values of $t\ln(P)$, bt, and t^2, respectively, for each row. Note that exponentiation in Excel is performed with the caret `^` operator.

3d. Next you will need to compute the sums of the values in these columns. In Excel, ranges of cells are specified by placing the colon `:` operator between two cell coordinates, so for example `A3:A13` specifies all of the cells from A3 through A13. Summation is performed with the `SUM( )` function on a range of cells, so for example `=SUM(A3:A13)` would produce the sum of all values in cells A3 through A13.

Use Excel to compute the sums of the $t\ln(P)$, bt, and t^2 columns. Leave the calculated sums underneath the corresponding columns, in the row labeled "Sums:".

3e. With all three sums calculated you can finally apply equation (6) to obtain the optimal slope. Perform the computation at the top of column H of `data_table.xlsx`, labeled "r". How does this estimate compare to your estimate from problem 3a?

3f. With the slope r and the intercept $b \approx 2.772589$ now set, you've finally defined a linear model for the log-transformed population. In column I of `data_table.xlsx`, labeled "Log Model", use the slope you've just computed in the linear model from equation (5) to compute an estimated log-population for each time. This can be accomplished quickly by using the formula

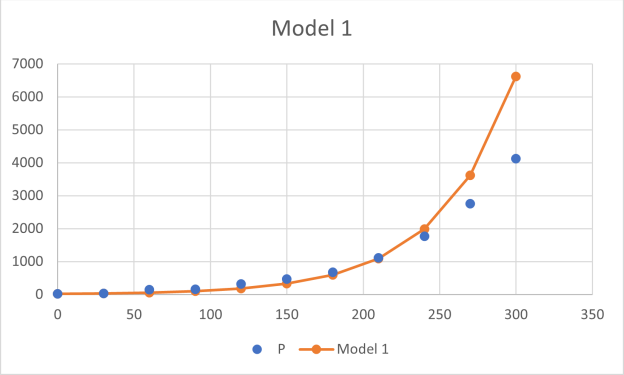
```
=H$3*A3+2.772589
```

in cell I3 and then dragging the corner of the cell down to copy the formula into the remaining rows. The dollar sign in `H$3` prevents the row coordinate from being automatically incremented when the formula is copied into the lower cells, which is needed here since every cell in every row of column I should use the value in H3 as the slope.

3g. Plot the log model estimates on the same set of axes as the actual log-population from the $\ln(P)$ column. This can be done by holding [Ctrl] while selecting cells A3 through A13 (for t), cells C3 through C13 (for $\ln(P)$), and cells I3 through I13 (for the Log Model) all at the same time and then inserting a scatter plot.

3h. In Part 2 the slope of the linear model r came from evaluating the natural logarithm of the exponential model in equation (4). In column J of `data_table.xlsx`, labeled "Model 1", use this same value of r to compute an estimated population $P = 16e^{rt}$ for each time.

3i. Create a scatter plot of the actual data P along with the Model 1 on the same scatter plot. How do the estimates compare to the real data?



Part 4: Extending to a Logistic Growth Model

Suppose that additional data has come in from the culture of *S. equorum*, with population measurements now extending to 600 minutes (see the Excel spreadsheet `data_table_revised.xlsx`). Now that more time has passed the population growth rate seems to be starting to trail off, possibly due to overcrowding or limited nutrients on the plate. In order to account for this added complexity, you've decided to try modeling the population using the **logistic growth model**.

The logistic growth model is based on the differential equation

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right) \tag{7}$$

which describes the rate of change in the population P with respect to the time t. Like the exponential growth model there is a growth rate constant $r > 0$, but now there is also a new parameter $K > 0$. The solution to this differential equation, which gives the population as a function of time (in minutes), is

$$P = \frac{KP_0}{P_0 + (K - P_0)e^{-rt}} \tag{8}$$

where P_0 is again the initial population. Your next goal will be to analyze this more complicated population model, and to utilize the techniques developed in Part 3 to find the logistic growth model that most closely matches this new set of population data.

Activity

4a. Open `data_table_revised.xlsx` (which includes population measurements at times from 0 to 600 minutes) and create a scatter plot of the data.

4b. The first goal is to investigate the significance of the parameter K from the logistic growth model, which you will do in two ways. First consider the function P of t in equation (8). Determine the long-term behavior of this function by evaluating the limit

$$\lim_{t\to\infty} \frac{KP_0}{P_0 + (K - P_0)e^{-rt}}$$

4c. Next consider the definition of the derivative $\frac{dP}{dt}$ from equation (7). For what populations P is the rate of population growth positive? Negative? Zero? What does this imply will happen to the population in the long term?

4d. The exponential and logistic growth models are closely related, which can be seen by looking at their derivatives. When the population is very small compared to K, the factor of $(1 - \frac{P}{K})$ in equation (7) is close to 1. In this case the logistic growth derivative (7) is essentially the same as the exponential growth derivative rP from equation (1), which indicates that the logistic growth model behaves almost like exponential growth when the population is small. As a result, for the purposes of this activity you will treat r as known, using $r = 0.02$ (which should be similar to the value you estimated at the end of Part 3). The initial population is still $P_0 = 16$, which leaves K as the only parameter left to estimate.

With the values of r and P_0 set, for the remainder of this activity your main goal will be to model the revised data set using the differential equation

$$\frac{dP}{dt} = 0.02P\left(1 - \frac{P}{K}\right) \tag{9}$$

and its resulting population function

$$P = \frac{16K}{16 + (K - 16)e^{-0.02t}} \tag{10}$$

Go to the Desmos link labeled *Fitting a Logistic Model*[5], which displays a scatter plot of the data alongside the logistic growth curve from equation (10). Play with the slider to see how K affects the shape of the graph. What value of K seems to cause the model to most closely match the data?

4e. The next goal is to find the value of K that minimizes the error between the model's predictions and the data. Unfortunately, unlike the exponential growth model, a semi-log plot is not much help here since evaluating the natural logarithm of both sides of the equation (10) does not lead to anything linear with respect to t. However the differential equation (9) *can* be manipulated to obtain a linear relationship.

Dividing both sides of (9) by P gives $\frac{dP/dt}{P}$ on the lefthand side, and a linear function of P on the righthand side. What is the slope of this linear function? What is the intercept?

4f. The value $\frac{dP/dt}{P}$ actually has a biological interpretation on its own. Considering that $\frac{dP}{dt}$ is the rate of change in the total population size, what biological measurement does $\frac{dP/dt}{P}$ represent?

Part 5: Fitting the Logistic Growth Model to the Data

The observations from Part 4 show that, if P grows logistically with respect to t, then $\frac{dP/dt}{P}$ changes linearly with respect to P. Moreover, the slope and intercept of this linear relationship can be used to compute K, and from Part 3 you know how to fit linear functions to data. In this particular case the intercept is 0.02, so the linear function has the form

$$\frac{dP/dt}{P} = mP + 0.02 \tag{11}$$

for some slope m. In this final activity you will use the techniques built up over the previous parts to fit this linear model to the revised data set.

Activity

5a. Fitting the model from equation (11) requires knowing both P and $\frac{dP}{dt}$ at each time t, but the derivatives are not directly given by the data. You will have to instead estimate them using *average* rates of change (i.e. difference quotients) between pairs of nearby data points. Use column D of `data_table_revised.xlsx`, labeled "dP/dt", to compute an estimate for the derivative of P at each time value t. Then use column E, labeled, "$(dP/dt)/P$", to compute the approximate value of $\frac{dP/dt}{P}$ at teach time value t.

5b. Go to the Desmos link labeled *Fitting a Second Linear Model*[6], which displays a scatter plot of $\frac{dP/dt}{P}$ versus P, alongside the linear curve $\frac{dP/dt}{P} = mP + 0.02$. Play with the slider to see how m affects the shape of the graph. What value of m seems to cause the model to most closely match the data? How does this relate to the value of K that you found in Part 4?

[5]Fitting a Logistic Model: `https://www.desmos.com/calculator/iria3vrzp2`
[6]Fitting a Second Linear Model: `https://www.desmos.com/calculator/selvf1cyro`

5c. In Part 3 you derived a formula for finding the slope that minimizes the sum of squared errors between a linear model and a set of data. To apply that to this linear model $\frac{dP/dt}{P} = mP + 0.02$ you will need to compute the three sums in equation (6). Bearing in mind that the data points in this case have the form $(x, y) = (P, \frac{dP/dt}{P})$, one of these sums includes terms of the form $xy = P\frac{dP/dt}{P}$, one includes terms of the form $bx = bP$, and the third includes terms of the form $x^2 = P^2$. Since $P\frac{dP/dt}{P} = \frac{dP}{dt}$, the "$dP/dt$" column already contains one of the terms that you need.

For the other two, use columns F and G of `data_table_revised.xlsx` labeled "bP" and "$P\hat{}2$" to compute the values of $0.02P$ and P^2, respectively, for each row.

5d. Compute the sums of the $\frac{dP}{dt}$, bP, and P^2 columns. Leave the calculated sums underneath the corresponding columns.

5e. Apply these three sums in formula (6) to obtain the optimal slope. Perform the computation at the top of column H of `data_table_revised.xlsx`, labeled "m". How does this slope compare to your estimate from problem 5b?

5f. In the previous problem you found that the slope of the linear function for $\frac{dP/dt}{P}$ is related to the parameters r and K from the logistic growth model. Use the slope you just calculated along with $r = 0.02$ to compute the value of K. Perform the computation at the top of column I of `data_table_revised.xlsx`.

5g. In column J of `data_table_revised.xlsx`, labeled "Model", use this value of K to compute an estimated population

$$P = \frac{16K}{16 + (K - 16)e^{-0.02t}}$$

for each time.

5h. Create a scatter plot of the actual data P along with the Model 1 estimates on the same scatter plot. How do the estimates compare to the real data?