Nonlinear Regression

Exponential, Quadratic & Logarithmic Modeling

Introduction

In Chapter 3, we saw that when looking for relationships between data sets it is often useful to create a scatter plot of the data. Remember that the data should be quantitative and in paired form. We do this by designating one data set to be the explanatory variable (x) and one data set to be the response variable (y). The choosing of the explanatory and response variables is very important. Remember to choose the response variable to the one that might naturally responds to changes in the explanatory variable. Every case is different and in some paired data either data set might be the response. Remember that if you plan to find a function which can describe the relationship between the variables, then we will probably use that function to make a prediction. The variable that you want to make a prediction of, should be the response variable. The paired data should also be quantitative, which means that it should be a measurable quantity with defined units, not categories.

For example, let’s take a look at the following paired data set giving the year and the world population in that year. Note that ‘-500’ means 500 B.C.

<table>
<thead>
<tr>
<th>Year</th>
<th>World Population in Billions</th>
</tr>
</thead>
<tbody>
<tr>
<td>-500</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>1000</td>
<td>0.31</td>
</tr>
<tr>
<td>1750</td>
<td>0.791</td>
</tr>
<tr>
<td>1800</td>
<td>0.978</td>
</tr>
<tr>
<td>1850</td>
<td>1.262</td>
</tr>
<tr>
<td>1900</td>
<td>1.65</td>
</tr>
<tr>
<td>1950</td>
<td>2.519</td>
</tr>
</tbody>
</table>

In thinking about this paired data we wonder if there is a relationship, but which variable should be the explanatory and which should be the response? It seems logical that the population might change or respond depending on the year, so we will make the year be the explanatory variable (x) and the world population be the response variable (y). Plugging the data into the statistics software program “Statcato”, we can generate the following scatterplot.
From our training in chapter 3, we notice right away that this graph does not have a linear shape. Using our statistics software, we find that the correlation coefficient for the graph is $r = 0.767$. Graphing the least squares regression line, the graph confirms that this scatterplot does not match the regression line very well. Statcato also gives us the equation of the regression line $\hat{y} = 0.1763 + 0.0007x$, but it is clear that this line is not a good model for the data.

Let’s look at this scatterplot again. Just because there doesn’t seem to be a linear relationship, does not mean there is no relationship at all. In fact the scatterplot shows a very strong curved relationship in the data. If our goal is to make predictions of what the world population will be, we would need to find a function that matches that curve.

**Hence, it is useful for anyone studying data to have some knowledge of curved functions.**
Nonlinear Regression Section 1
Exponential Modeling

Probably the most common type of curved function is the exponential function. This function is seen in a variety of data analysis settings including growth and decay, compound interest, and most especially in population data.

Let’s look again at our world population data. Using a statistics software program called Statcato; we have the program plot an exponential curve over the scatter plot. Does the curve seem to fit the data better than the regression line we found above? Not only have we found a better fit for the paired data, but Statcato also gives the equation for that exponential function \( y = 0.16365(1.00111)^x \).

We can see a few things from this graph. The first is the shape of an exponential function. Do you see how the exponential function has a backwards “L” shape to it. This is very common with exponential functions especially with population and growth functions. Decay functions tend to have a regular “L” shape. Did you also notice how the exponential function does not cross the x-axis, but simply approaches it? This is also very common in exponential functions. If you think about it, the response variable (y) is describing the world population. Of course the y values must be positive. If the y-value was zero or negative, I would not be here talking to you. The world population cannot be zero or a negative number. Remember that the y-values make up the range of the function, so we would say that the range of this exponential function is only positive numbers. This is also very common in exponential functions. What else can we learn from this graph? Did you notice that the curve tends to get close to the x-axis for the first years in the data set (500 BC, 1 AD and 1000 AD)? When a curve gets close to a line for certain x-values, we call this line an asymptote. In this graph, the x-axis is an asymptote. Did you also notice that the graph starts to get very big, very quickly? From 1000 AD to 1950 AD, the population has risen from
approximately 310 million to over 2.5 billion people!! That is an incredible increase if you think about it. Have you ever heard the term “that is growing so fast that it is growing exponentially”? That statement comes from the shape of the exponential function.

Now let’s look at the equation of the exponential function \( \hat{y} = 0.16365 \times (1.00111)^x \) that Statcato found for us. Do you notice how the x variable is actually the exponent in the equation? That is how “exponential” functions get their name. Recall that the number an exponent is attached to is called the “base”. In this equation the number (1.00111) is the base. Notice also that 0.16365 is the number that the exponential expression (exponent and base) is multiplied by. This number is usually called the “initial value” or in this case the “initial population”. In this data set our “initial” ordered pair was 500 BC. Unfortunately this is not the initial value described in the equation. When we say our initial value, we mean the y-value when the x = 0. If you have studied any algebra, you may remember that this would be the y-intercept. Look at the graph of the exponential curve. Approximately where does the curve cross the y-axis? Did you notice that the curve seems to cross the y-axis at 0.16365? This is the same initial value as given in the equation?

**Assessing the fit of an exponential function**

This is fine for Statcato to give us an exponential function, but how well does this curve really fit the data? We again see that the exponential curve does not fit the data perfectly, but it does seem to fit better than the line. If we are going to use this exponential function to maybe make predictions, then we need to have some way of assessing how well the curve fits the data.

**R-Squared**

One of the first things to look at when assessing the fit of a curve to a scatterplot is the “R-squared” value. Remember that “R” is the correlation coefficient. But when we square R it gives the percent of variability in y that can be explained by the relationship with x. For our exponential function and the population growth data Statcato determined that \( R^2 \approx 0.9078 \) This tells us that approximately 90.8% of the variability in population can be explained by the exponential relationship with time. This is a very high percentage and indicates the exponential function fits pretty well.

Another use of R-squared is to determine which model is a better fit. For example, suppose I want to know if the exponential model is a better fit than the linear model. We can determine this by comparing the R-squared values.

**Regression Line:** \( R^2 \approx 0.5878 \)

**Exponential Regression Curve:** \( R^2 \approx 0.9078 \)

While only 58.8% of the variability in population can be explained by the linear relationship, almost 91% of the variability can be explained by the exponential relationship. The model with the higher R-squared is the better fit.
**Note:** We prefer to use the simpler formula (linear) when possible. If there is a significant increase in the \( r \)-squared value for the curved function, we will use the curve. But, if the curve has only a slightly better \( r \)-squared, we prefer to use the simpler model. In the last example the \( R \)-squared value for the exponential was 90.8\%. This is significantly higher than the regression line’s \( R \)-squared value of 58.8\%. So we would most definitely prefer the exponential model over the linear model. Let us suppose in a different problem, the \( R \)-squared value for a curve is 84\% and the \( R \)-squared for the linear model is 83\%, then we would stay with simpler model (linear) because there is not a significant increase.

**Standard Deviation of the Residual Errors** \( (S_e) \)

Let’s look at one of the ordered pairs in our data set, say (1000 year, 0.31 billion people). Can you find the ordered pair on the curve with that same x-value (1000 AD)? If we plug in 1000 into the regression equation we can get the y value on the curve. This is often called our predicted value \( \hat{y} \). Let us calculate the predicted value for the year 1000 AD.

\[
\hat{y} = 1.00111^1000\times0.16365 \\
\hat{y} = 0.496
\]

So the regression line predicted that the population in the year 1000 AD would be approximately 0.496 billion people. The actual observed population in the year 1000 AD was 0.31 billion. So how much error was in our prediction? One way to measure error is through the Residual. Recall that a residual is the difference between the observed ordered pair (y) and the predicted value \( \hat{y} \) if the original x value is plugged into the function. Another way to explain the residual is that it is the vertical distance from the curve to the line. For example, for the year 1000 AD, the residual would be calculated as follows:

\[
y - \hat{y} = 0.31 - 0.496 \\
y - \hat{y} = -0.186
\]

Notice this gives us a residual (error) of -0.186. This means that the ordered pair is 0.186 below the curve when \( x = 1000 \) AD. Let’s now make a table of the residuals. For each x value in the data set we will plug the x value into the regression curve from Statcato \( \hat{y} = 0.16365(1.00111)^x \). This will give us our predicted \( \hat{y} \) values. Subtracting the actual y value minus the predicted \( \hat{y} \) value gives us the residual.

Notice that when the curve is too high the residuals are negative and when the curve is too low, the residuals are positive. We still have the problem of assessing how well the curve fits the data set. One possibility would be to find the **Standard Deviation of the Residual Errors** \( (S_e) \) like we did in Module 3. By squaring the residuals, we are able to eliminate the negative residuals. Now we add up the squares, divide by \( n-2 \) and take the square root of the answer. Recall that the **Standard Deviation of the**
Residual Errors \( (S_e) \) will give us how far on average the points are from the curve and will also give us the average prediction error should we use the curve to make a prediction. Let’s look at the calculation of the Standard Deviation \( (S_e) \) below.

<table>
<thead>
<tr>
<th>Year (x)</th>
<th>World Pop (y) in Billions</th>
<th>pred y from Exp curve</th>
<th>Residual Error</th>
<th>Residuals Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>-500</td>
<td>0.1</td>
<td>0.093975847</td>
<td>0.006024153</td>
<td>0.0000362904</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.163831652</td>
<td>0.036168349</td>
<td>0.001308149</td>
</tr>
<tr>
<td>1000</td>
<td>0.31</td>
<td>0.496267158</td>
<td>-0.186267158</td>
<td>0.034695454</td>
</tr>
<tr>
<td>1750</td>
<td>0.791</td>
<td>1.140420931</td>
<td>-0.349420931</td>
<td>0.122094987</td>
</tr>
<tr>
<td>1800</td>
<td>0.978</td>
<td>1.205466528</td>
<td>-0.227466528</td>
<td>0.051741021</td>
</tr>
<tr>
<td>1850</td>
<td>1.262</td>
<td>1.274222097</td>
<td>-0.012222097</td>
<td>0.00014938</td>
</tr>
<tr>
<td>1900</td>
<td>1.65</td>
<td>1.346899241</td>
<td>0.303100759</td>
<td>0.09187007</td>
</tr>
<tr>
<td>1950</td>
<td>2.519</td>
<td>1.423721635</td>
<td>1.095278365</td>
<td>1.199634697</td>
</tr>
</tbody>
</table>

We first find the Sum of the Squares of the Residual Errors (SSE). Don’t be confused. The SSE is not the standard deviation. SSE and \( S_e \) are completely different. In a sense we need to use the sum of squares to get the standard deviation.

\[
SSE = 0.0000362904 + 0.001308149 + ... + 1.199634697
\]

\[
SSE \approx 1.501530049
\]

Now we can use the standard deviation formula \( S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.501530049}{8-2}} \approx 0.5002549 \) to calculate the standard deviation of the residual errors.

Remember were-as one data set has degrees of freedom n-1, ordered pair data has a degrees of freedom n-2. This is why we divide by n-2 instead of n-1.

So \( S_e \approx 0.5 \) billion. As with chapter 3, the standard deviation of the residuals tells us how well the data fits the regression curve. A regression curve tries to minimize this vertical distance. So for exponential curves of the form \( \hat{y} = ab^x \), the curve \( \hat{y} = 0.16365(1.00111)^x \) was the best fit. This again means that it minimized the vertical distance to the curve \( (S_e) \). So no other function of the form \( \hat{y} = ab^x \) will have a smaller \( S_e \) than the function \( \hat{y} = 0.16365(1.00111)^x \).

Sometimes we may want to know if one curve or line fits the data better than another. The Standard Deviation of the Residual Errors can be used for this purpose. The curve that has the smallest Standard Deviation of the Residual Errors will be the one that fits the data best.
Let’s explore this a little bit. We just found that the Standard Deviation for the Exponential Curve Residuals \( S_e \) was about 0.500 billion. Earlier we said that we thought the exponential curve fit the data better than the regression line. Can we confirm what our eyes are telling us? Look at the scatterplot below. Statcato found that the regression line that best fits the population data was \( \hat{y} = 0.1763 + 0.0007x \). In Module 3, we also calculated the Standard Deviation of the Residual Errors \( S_e \).

First we plug in each year \( (x) \) into the regression line \( \hat{y} = 0.1763 + 0.0007x \) and obtain our predicted \( \hat{y} \) values. Subtracting the observed population \( y \) values minus the predicted \( \hat{y} \) gives us the residuals. We had Statcato calculate the Standard Deviation this time.

For the regression line, \( S_e \approx 0.572 \). Notice this is larger than the standard deviation for the exponential curve (0.500). Since there is much less error when we use the exponential function verses the linear function, this implies that the exponential curve is a much better fit to this population data than the regression line from Module 3.

**Key Idea:** A linear or curved model with a larger R-squared and smaller Standard Deviation of the Residual Errors gives evidence of a better fit.

**Note:** The study of regression is broad and complicated branch of Statistics. It would be wrong to suggest that all of regression can be summarized into the highest R-squared and the lowest Standard Deviation. We often study many factors before deciding on a particular model. For example the histogram of the residuals should be bell shaped and centered at zero. Also the residual plot vs the variable should not be fan shaped. Another is that there should not be any significant outliers in the scatterplot. These are but a few. The study of these can be left for a later more in depth study.
Making Predictions from an Exponential Function

Remember the goal of finding the exponential curve for the Population data was to hopefully use it to make predictions. So now that we have assessed that the exponential curve does fit the population data reasonably well, let us use the function to make a prediction.

Caution!! Remember in chapter 3 that we should only make predictions within the scope of the data. The x-values for our population paired data were between -500 (500 BC) and the year 1950. We should not try to make predictions outside of this range. If you recall making predictions out of the scope of the data is called Extrapolation. People that extrapolate tend to have a lot of error in their predictions because there is no guarantee that the data will follow the curve outside the scope of the data. Remember the Standard Deviation of the Residual Errors only applies in the scope of the data. Once you go outside the scope of the data, there is no telling how much error there may be.

So let us predict the world population in Billions for a given year. Let’s look at the scatterplot below. Try to estimate the world population in the year 600 AD?

By plugging in 600 for x in our Exponential Regression Function we can get our prediction. Remember to follow the order of operations and perform the exponent first, then multiply.

\[ \hat{y} = 0.16365 \times (1.00111)^{600} \]

\[ \hat{y} \approx 0.3184 \]

Hence in 600 AD, the exponential function predicts that the world population was 0.318 Billion (318 Million) people.