Chapter 6 – Curved Quantitative Relationships

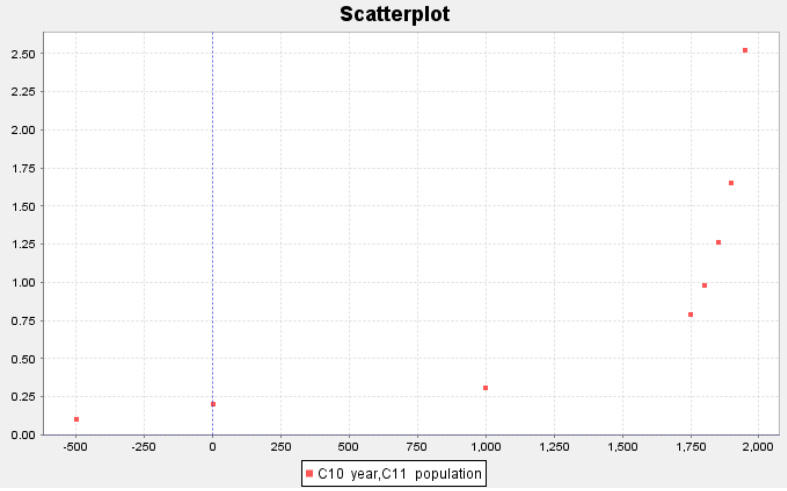
Introduction: In the last unit, we saw that when looking for relationships between quantitative data sets it is often useful to create a scatter plot of the data. Remember that the data should be quantitative and in paired form. The paired data should be quantitative, which means that it should be a measurable quantity with defined units, not categories.

We do this be designating one data set to be the explanatory variable (x) and one data sets 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. The variable that you want to make a prediction of, should be the response variable.

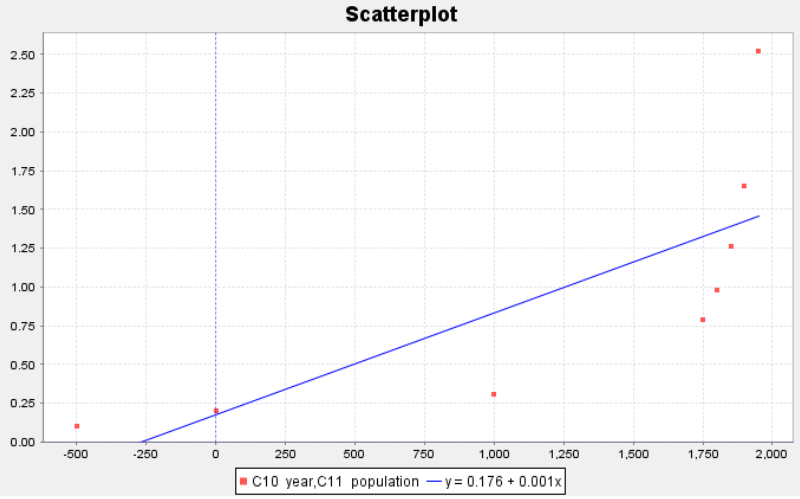
For example, let us look at the following paired data set giving the year and the world population in that year. Note that "-500" means 500 B.C.

Year	World Population in Billions
-500	0.1
1	0.2
1000	0.31
1750	0.791
1800	0.978
1850	1.262
1900	1.65
1950	2.519

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 to 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 a statistics software program like Statcato, we can generate the following scatterplot.



We notice right away that this graph does not have a linear shape. Using our statistics software, we can find our least squares regression line for the data and an *R-squared (linear) = 0.5882 (58.8%).* This tells us that approximately 58.8% of the variability in population can be explained by the linear relationship with time. The graph and the R-squared value confirm that this scatterplot does have some linear relationship (correlation), but it is not as strong as we might like. The statistics software 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 us look at this scatterplot again. Just because there does not 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 were to make predictions of what the world population will be, we would need to find a function that matches that curve.

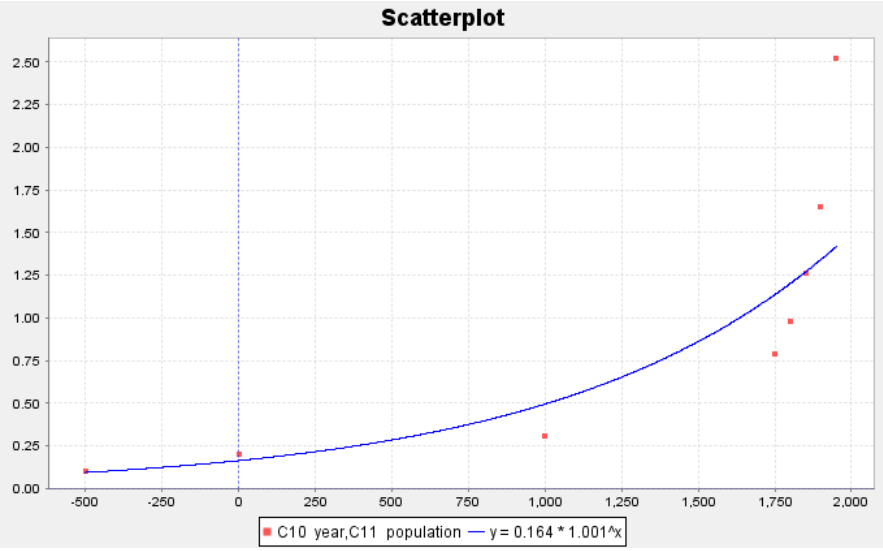
Hence, it is useful for anyone studying data to have some knowledge of curved functions.

Section 6A – Exponential Relationships with Technology

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

Let us look again at our world population data. Using a statistics software program, 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 in the introduction section? Not only have we found a better fit for the paired data, but the program also gives the equation for that exponential curve

$$\hat{y} = 0.16365(1.00111)^x .$$



We can see a few things from this graph. The first is the shape of an exponential curve. Do you see how the exponential curve has a backwards "L" shape to it? The curve increases as we go from left to right. This is very common with exponential curves, especially with population growth. That is why this pattern is called "exponential growth". Exponential decay functions tend to have a regular "L" shape and decrease from left to right. 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. This tells us that the y-values of exponential curves can only positive numbers.

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 someone say the following? "That is growing so fast that it is growing exponentially." That statement comes from the shape of the exponential growth function.

Now let us look at the equation of the exponential function $\hat{y} = 0.16365(1.00111)^x$ that the software found for us. Different software's can write this formula differently. 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. However, 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. However, 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%. Therefore, we would most definitely prefer the exponential model to 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 us 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.

$$\hat{y} = 1.00111\text{^}1000 \times 0.16365$$
$$\hat{y} = 0.496$$

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 residuals. 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 point. 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 us 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) as we did for lines. 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 give us the average prediction error should we use the curve to make a prediction. Let us look at the calculation of the Standard Deviation (S_e) below.

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

We first find the Sum of the Squares of the Residual Errors (SSE). Do not 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.

$$\text{SSE} = 0.0000362904 + 0.001308149 + \ldots + 1.199634697$$
$$SSE \approx 1.501530049$$

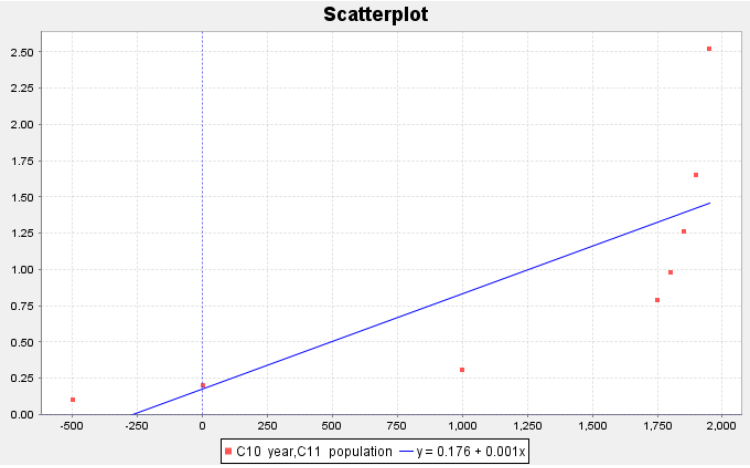
Now we can use the standard deviation formula $S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.501530049}{8-2}} \approx 0.5002549$ to calculate the standard deviation of the residual errors.

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. Therefore, for exponential curves $\hat{y} = a(b)^x$, 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 $\hat{y} = a(b)^x$ 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 does. 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 us 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. The software found that the regression line that best fits the population data was $\hat{y} = 0.1763 + 0.0007x$ and 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.

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. In addition, the residual plot verses the x-value should be evenly spread out. Another is that there should not be any significant outliers in the scatterplot. These are but a few.*

Statcato does not have a residual plot function or a histogram of the residuals for curves, so we will be focusing on analyzing and comparing the curves with R-squared and the Standard Deviation of the Residual Errors and using the models to make predictions. To make a residual plot with Statcato, you will need to calculate the residuals first and then make a scatterplot of the x values and the residuals. To

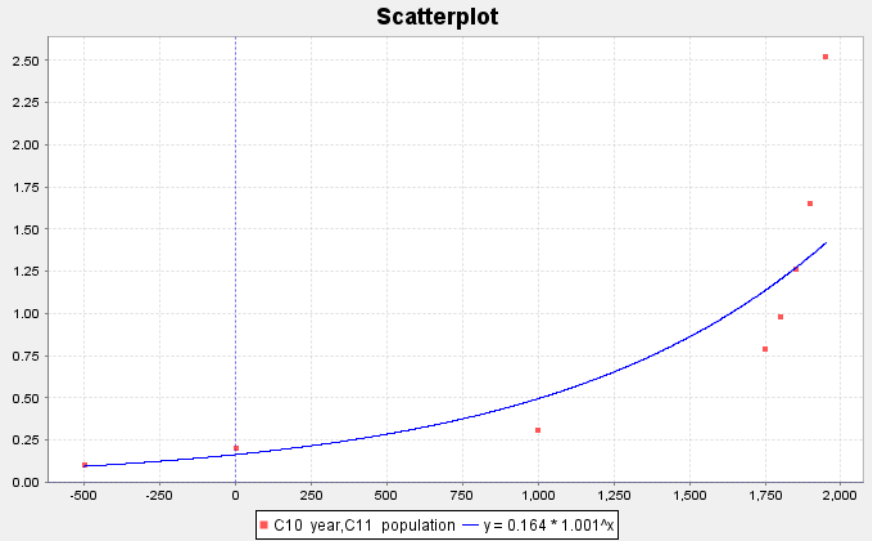
make a histogram of the residual with Statcato, you will need to calculate the residuals first and then make a histogram of just the residuals column.

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 us 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 curve, we can get our prediction. Remember to follow the order of operations and perform the exponent first, then multiply.

$\hat{y} = 0.16365(1.00111)^{600}$

$\hat{y} \approx 0.3184$

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

Problem Set Section 6A

1. Open the Nonlinear Data Sets in Excel. Copy and paste the years since 1995 and wind power into Statcato. This data gives the number of years since 1995 and the worldwide wind power capacity in MW (megawatts). Let the number of years be the explanatory variable and the wind power be the response variable.

    a) Use Statcato to make a scatter plot of the ordered pairs. Does the scatterplot look like an exponential model might fit? If so, would it be exponential growth or exponential decay. Save or draw a rough sketch of the Statcato scatterplot with the exponential curve.

    b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to find the exponential model that best fits the curve. Do you think that the exponential curve fits the data well? Are the points close to the curve?

    c) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

    d) Give equation of the exponential curve that Statcato found.

    e) Most computer programs have a problem calculating the standard deviation of the residuals errors for exponential curves. Statcato says that the standard deviation for the exponential curve is 0.0788. This is wrong. The graph shows that the points are much farther from the curve than this. The real standard deviation is actually 4039.2 MW. Write two sentences explaining the meaning of the standard deviation of the residual errors in this context.

    f) What is the scope of the data (x values)? What years does the scope represent? Is this exponential growth or exponential decay?

g) Do you think Wind Power will continue to follow this pattern into the future? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

h) Predict the Wind Power in 2002 (year 7)? How far off could this prediction be on average?

i) Predict the Wind Power in 2008 (year 13)? How far off could this prediction be on average?

j) Do you think it would be all right to use this model to predict the worldwide wind power in 2065 (year 70)? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (c) and (d)?

2. Open the Nonlinear Data Sets in Excel. Copy and paste the month and retirement account balance into Statcato. The data gives the balance in a retirement account. The account started with $78,000 in their account in 2010, and have been slowly making withdrawals for their living expenses. The data gives the months since January 2010 and the retirement account balance in thousands of dollars. Let the months be the explanatory variable and the retirement account balance be the response variable.

   a) Use Statcato to make a scatter plot of the ordered pairs. Does the scatterplot look like an exponential model might fit? If so, would it be exponential growth or exponential decay. Save or draw a rough sketch of the Statcato scatterplot with the exponential curve.

   b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to find the exponential model that best fits the curve. Do you think that the exponential curve fits the data well? Are the points close to the curve?

   c) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

   d) Give equation of the exponential curve that Statcato found.

   e) Most computer programs have a problem calculating the standard deviation of the residuals errors for exponential curves. Statcato says that the standard deviation for the exponential curve is 0.0684. This is wrong. The graph shows that the points are much

farther from the curve than this. The real standard deviation is 4.182 thousand dollars. Write two sentences explaining the meaning of the standard deviation of the residual errors in this context.

f) What is the scope of the data (x values)? What months does the scope represent? Is this exponential growth or exponential decay?

g) Do you think the retirement account balance will continue to follow this pattern into the future? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

h) Predict the retirement account balance December 15th 2010 (month 11.5). How far off could this prediction be on average?

i) Predict the retirement account balance January 15th 2012 (month 24.5). How far off could this prediction be on average?

j) Do you think it would be all right to extrapolate a lot and use this model to predict the retirement account at the start of 2050 (month 480)? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (c) and (d)?

3. Open the Nonlinear Data Sets in Excel. Copy and paste the years since 1990 and the savings account balance into Statcato. The savings account was opened in 1990. The data gives the number of years since 1990 and the amount of money in a savings account. Let the number of years be the explanatory variable and the savings account balance be the response variable.

   a) Use Statcato to make a scatter plot of the ordered pairs. Does the scatterplot look like an exponential model might fit? If so, would it be exponential growth or exponential decay. Save or draw a rough sketch of the Statcato scatterplot with the exponential curve.

   b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to find the exponential model that best fits the curve. Do you think that the exponential curve fits the data well? Are the points close to the curve?

   c) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

d) Give equation of the exponential curve that Statcato found.

e) Most computer programs have a problem calculating the standard deviation of the residuals errors for exponential curves. Statcato says that the standard deviation for the exponential curve is 0.0801. This is wrong. The graph shows that the points are much farther from the curve than this. The real standard deviation is approximately $110.94 dollars. Write two sentences explaining the meaning of the standard deviation of the residual errors in this context.

f) What is the scope of the data (x values)? What years does the scope represent? Is this exponential growth or exponential decay?

g) Do you think the savings account balance will continue to follow this pattern into the future? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

h) Predict the savings account balance in 2006 (year 16). How far off could this prediction be on average?

i) Predict the savings account balance in 2011 (year 21). How far off could this prediction be on average?

j) Do you think it would be all right to extrapolate a lot and use this model to predict the savings account in 2040 (year 50)? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (c) and (d)?

4. Open the Nonlinear Data Sets in Excel. Copy and paste the metal distance and ultrasound response into Statcato. Ultrasound is used in a variety of applications. Let the explanatory variable be the metal distance in millimeters and the response variable be the ultrasound response.

   a) Use Statcato to make a scatter plot of the ordered pairs. Does the scatterplot look like an exponential model might fit? If so, would it be exponential growth or exponential decay. Save or draw a rough sketch of the Statcato scatterplot with the exponential curve.

   b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to find the exponential model that best fits the curve. Do you think that the

exponential curve fits the data well? Are the points close to the curve?

c) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

d) Give equation of the exponential curve that Statcato found.

e) Most computer programs have a problem calculating the standard deviation of the residuals errors for exponential curves. Statcato says that the standard deviation for the exponential curve is 0.2466. This is wrong. The graph shows that the points are much farther from the curve than this. The real standard deviation is 8.239 ultrasonic response units. Write two sentences explaining the meaning of the standard deviation of the residual errors in this context.

f) What is the scope of the data (x values)? Is this exponential growth or exponential decay?

g) Do you think the ultrasonic response will continue to follow this pattern outside the scope of the data? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

h) Predict the ultrasonic response if the metal is 2.83 mm away. How far off could this prediction be on average?

i) Predict the ultrasonic response if the metal is 4.51 mm away. How far off could this prediction be on average?

j) Do you think it would be all right to extrapolate some and use this model to predict ultrasonic response if the metal is 12.75 mm away? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (c) and (d)?

5. A student wanted to map an exponential function model to the atomic defect and energy data (See nonlinear data sets). Statcato said "Error" and "Not Available". The programs were unable to find an exponential function to fit the data. Explain why this happened. What does this tell us about data sets that cannot be modeled by exponential curves?

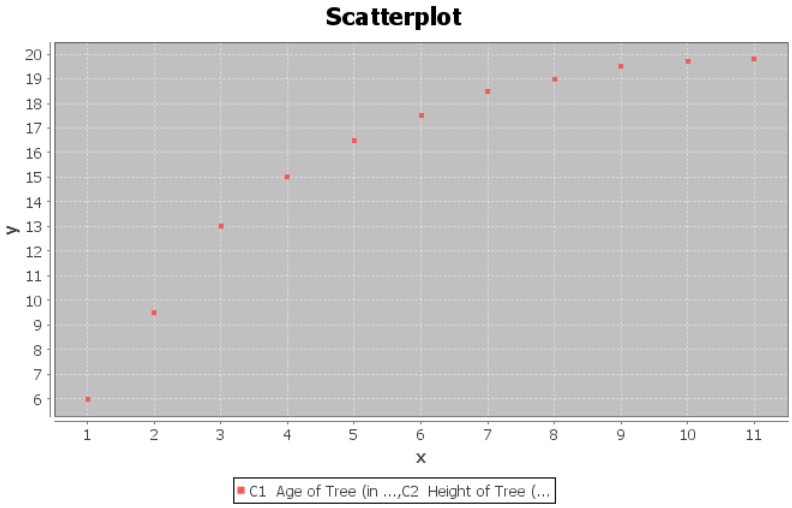
Section 6B – Logarithmic Relationships with Technology

Let us examine another data set. The following data set gives the height of a tree in feet and the age of the tree in years.

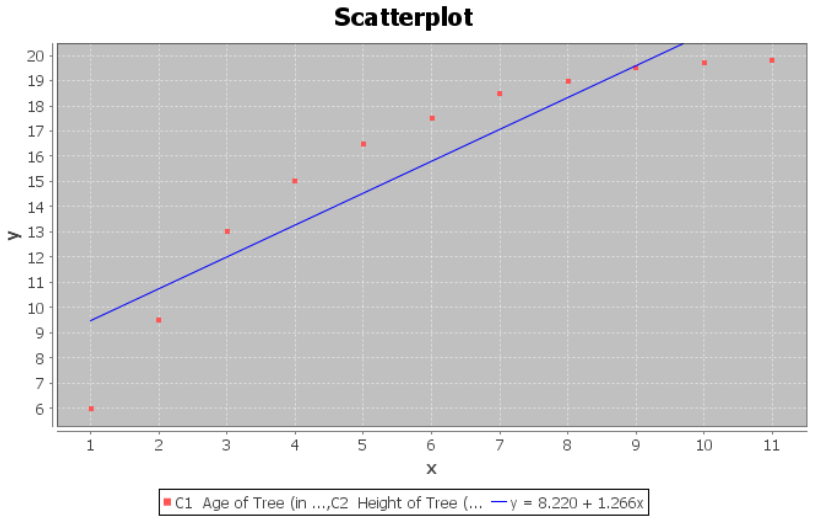
Age of Tree (in years)	Height of Tree (in feet)
1	6.0
2	9.5

3	13.0
4	15.0
5	16.5
6	17.5
7	18.5
8	19.0
9	19.5
10	19.7
11	19.8

As with the population data, we wonder if there is a relationship between the age of the tree and the height of the tree, but which variable should be the explanatory and which should be the response? It seems logical that the height of the tree responds to its age, so we will make the year the explanatory variable (x) and the height the response variable (y). It also makes sense to make the height the response variable since we may want to predict the height of the tree from knowing the age of the tree. Plugging the data into a statistics software, we get the following scatterplot.



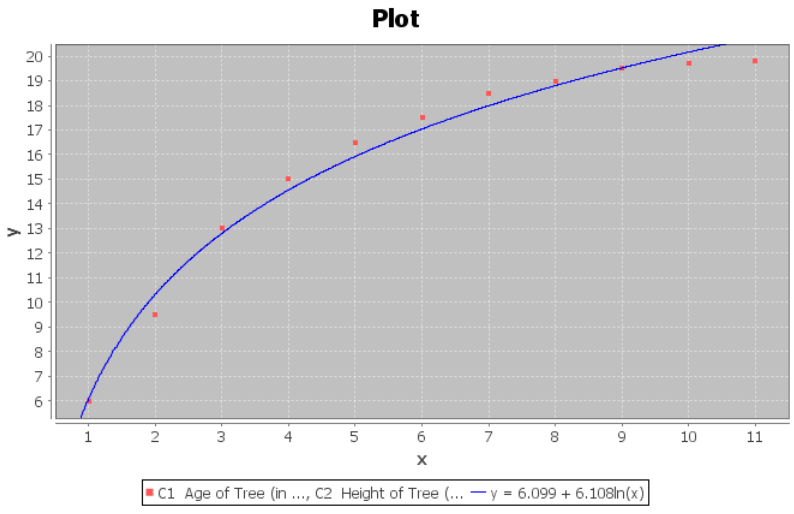
Let us start by seeing how well a line will fit the data. Creating a scatterplot with the regression line $\hat{y} = 8.2200 + 1.2664x$ drawn and we see that the data fits the line reasonably well.



Let us see how well the line really fits. The software calculated the R-squared to be 0.8400 and the Standard Deviation of the Residual Errors to be 1.9325 feet. Therefore, we see that the regression line fits the data reasonably well. Points in the scatterplot are an average of 1.9325 feet from the regression line and predictions with the regression line formula will have an average error of 1.9325 feet.

The line does seem to fit reasonably, but if we were able to draw a curve, do you think we could fit the data even better? Do you notice how after 8 years, the trees start to approach a maximum height of about 20 feet. This causes the scatterplot to take on an upside down L shape. The curve seems to be increasing from left to right, but it is increasing very slowly. This is the shape of another type of curve, the Logarithmic curve. Logarithmic curves (or Log curves for short) have a shape that frequently occurs when we analyze data sets.

For example, we can find out how many years it will take money to grow in your bank account with a Log function. So let us try to graph a Log curve with statistics software that approximates this data set and see what happens.



We can see right away that the Log curve appears to fit the data better than the line. This software uses the Natural Log or (LN) for short. The function came out to be $\hat{y} = 6.09934 + 6.10818LN(x)$. The found the function in terms of the Natural Log because it is one of the few types of logarithms you can find on your calculator. Some programs use Log base 10 (LOG). After all, isn't the purpose again of finding this function to use it to predict the height of a tree? So how does logarithms work and in particular the Natural Logarithm function?

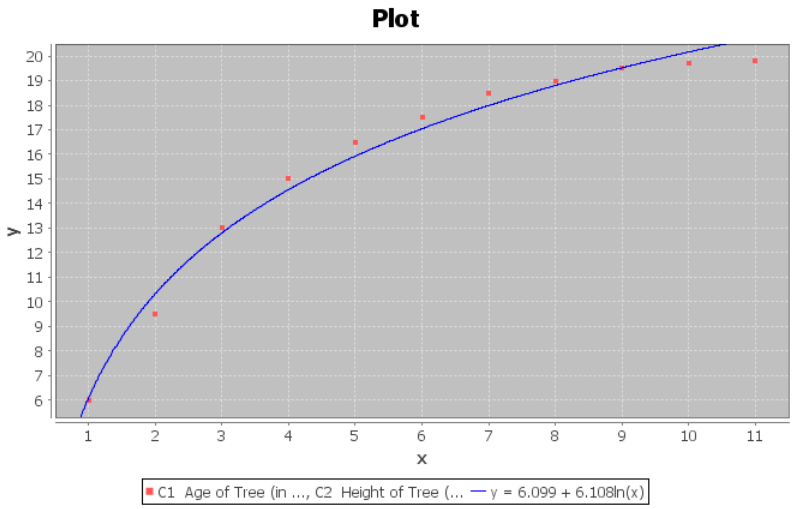
About Logarithms

Logarithms are really the inverse of exponentials. Logs in fact are exponents. When you find the LN(8) for example, on your calculator you are finding an exponent on a particular base that when evaluated gives you an answer of 8. However, what is the base for the Natural Log function? The answer to that question is the number "e". "e" is an irrational number (infinite non-repeating decimal) that is approximately 2.718. Again, "e" is not exactly 2.718 but that is pretty close. So let us see if we can understand this. When we find the LN(8) on our calculator we are really finding the following exponent ($e^{??} = 8$). If we replace e with 2.718, we get $2.718^{??} = 8$. See if you can find the LN(8) on your calculator? Every calculator is different. For most calculators, you will push the "LN" key then the number 8 and then enter or =. For a few calculators, you may have to push the 8 first, then the LN key. You should have gotten an answer of 2.079. Therefore, this implies that $e^{2.079} \approx 8$.

Let us plug in some other numbers into the LN function. Find the LN(0) or LN(-5). What does your calculator tell you? It probably said "ERROR" or "UNDEFINED". There is a reason for this. The number you plug into the LN function is equal to 2.718 to some power. 2.718 to some power will always be a positive number. Hence, we can only plug in positive numbers into the LN function. Do you remember the name for the values of x we are allowed to plug into a formula? You are right. It is called the Domain. So what is the Domain of the natural Log function? Since we can only plug in positive numbers for x, our Domain is all positive numbers $(x > 0)$. This Domain is very common in most basic logarithms. This also implies that if we have negative numbers in our explanatory data set (x values) we should not use a Log function as our model.

Assessing the fit of a Log function

Let us go back now to our tree data. Statcato found that the natural Log function that best fit the data was $\hat{y} = 6.099 + 6.108LN(x)$. Notice the distinctive upside down "L" shape with a slow growth.



But how well does this log curve fit the tree data? One way to measure this is with the standard deviation of the residual errors (S_e). As we did with the standard deviation calculation in the line above, we will plug in all the ages (x values) into $\hat{y} = 6.09934 + 6.10818LN(x)$ and get our predicted $\hat{y}$ values. Let us try to calculate the predicted height $\hat{y}$ for a tree that is 2 years old. Plugging in 2 for x in the natural Log equation gives the following.

$$\hat{y} = 6.09934 + 6.10818LN(2)$$
$$\hat{y} = 6.09934 + 6.10818 \times 0.69314718$$
$$\hat{y} = 6.09934 + 4.233867745$$
$$\hat{y} \approx 10.3 \text{ feet}$$

When doing the calculation on a calculator, be sure to follow the order of operations. So be sure to do the LN(2) first, then the multiplication, and lastly the addition. How far off was this predicted value? Recall that a two-year-old tree in the data set had an actual height of 9.5 feet. By subtracting the actual y value minus the predicted $\hat{y}$, we get that $y - \hat{y} = 9.5 - 10.33 \approx -0.83$. If you remember, this number is often called a "residual" and tells us that the ordered pair in the data set $(2, 9.5)$ was 0.83 feet below the natural Log curve. If we calculate all the predicted values $\hat{y}$ and the residuals $y - \hat{y}$, we will get the following table.

Age of tree (years)	Height of tree (feet)	pred y	Residual
1	6	6.099	-0.099
2	9.5	10.33274	-0.83274

3	13	12.80932	0.190676
4	15	14.56649	0.433514
5	16.5	15.92945	0.570553
6	17.5	17.04307	0.456933
7	18.5	17.98462	0.515381
8	19	18.80023	0.199771
9	19.5	19.51965	-0.01965
10	19.7	20.16319	-0.46319
11	19.8	20.74534	-0.94534

Notice again that a positive residual means that the ordered pair was above the natural Log curve and a negative residuals means that the ordered pair was below the natural Log curve. To calculate the Standard Deviation of the Residual Errors square all the residuals and add them. If you recall this is called the sum of squares. Now divide by n-2 and take the square root. The statistics software calculated the Standard deviation for us and found it to be 0.5653 feet. Therefore, if we predict the height with the Natural Log curve, we will have an average error of 0.5653 feet.

This is much better than the standard deviation for the regression line of 1.9325 feet we calculated earlier. So not only is the natural log curve a much better fit, but if we use it to make predictions, we will have a much smaller average error. Recall also that the R-squared for the regression line was 0.84. The R-squared for the natural log curve is 0.9863 and is much better than the regression line. Only 84% of the variability in height can be explained by the regression line, while 98.6% of the variability can be explained by the natural log curve. This also re-enforces that the natural log is a much better overall fit.

Making Predictions with the Log Equations

Since the natural Log equation was a good fit for the tree data. Let us see if we can use it to make predictions. Remember, we should only make predictions in the scope of the data. Since our x values were between 1 year and 11 years, we should only make predictions for $1 \le x \le 11$. If we make a prediction for an x value out of the scope of the data, we should expect more error in the prediction.

Use the natural Log equation $\hat{y} = 6.099 + 6.108LN(x)$ to predict the height of a tree that is 10.5 years old. Plugging in 10.5 for x in the equation and using the order of operations to simplify we get the following:

$$\hat{y} = 6.099 + 6.108LN(10.5)$$
$$= 6.099 + 6.108 \times 2.351375$$
$$= 6.099 + 14.3622$$
$$\approx 20.5\,ft$$

Therefore, we expect a tree that is 10.5 years old to be about 20.5 ft. Since we found earlier that the standard deviation was 0.5653 feet, we know that our prediction of 20.5 ft. could have an approximate error of 0.5653 feet.

Problem Set Section 6B

1. Open the Nonlinear Data Sets in Excel. Copy and paste the number of years since 1980 and the number of drunk driving fatal accidents. Let the number of years be the explanatory variable and the number of drunk driving fatal accidents be the response variable.

    a) Make a scatter plot of the ordered pairs. Does the scatterplot look like a logarithmic model might fit? If so, would it be logarithmic growth or logarithmic decay.

    b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to find the Logarithmic curve that that best fits data. Do you think that the logarithmic curve fits the data well? Are the points close to the curve? Save or draw a rough sketch of the scatterplot with the log curve.

    c) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

    d) The standard deviation of the residual errors in Statcato is accurate for log curves. What was the standard deviation of the residual errors? Write two sentences explaining the standard deviation.

    e) Give the equation of the log curve.

    f) What is the scope of the data (x values)?

    g) Do you think the number of drunk driving fatal car accidents will continue to follow this pattern into the future? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

    h) Use the logarithmic equation to predict the number of fatal drunk driving accidents that may occur in year 12.5 (half way through the year 1992). How far off could this prediction be on average?

    i) Use the logarithmic equation to predict the number of fatal drunk driving accidents that may occur in year 23.75 (three quarters of the way through the year 2003). How far off

could this prediction be on average?

j) Do you think it would be all right to extrapolate a lot and use this model to predict the number of fatal car accidents in the year 2050 (year 70)? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (c) and (d)?

2. Open the Nonlinear Data Sets in Excel. Copy and paste the bear age and bear length columns into Statcato. Let the age be the explanatory variable and the length be the response variable.

a) Use Statcato to make a scatter plot of the ordered pairs. Does the scatterplot look like a logarithmic model might fit? If so, would it be logarithmic growth or logarithmic decay.

b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to find the equation of the Logarithmic curve that that best fits data. Do you think that the logarithmic curve fits the data well? Are the points close to the curve? Save or draw a rough sketch of the scatterplot with the log curve.

c) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

d) The standard deviation of the residual errors in Statcato is accurate for log curves. What was the standard deviation of the residual errors? Write two sentences explaining the standard deviation.

e) Give the equation of the log curve.

f) What is the scope of the data (x values)?

g) Do you think the length of the bear will continue to follow this pattern into the future? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

h) Use the logarithmic function to predict the length of a black bear that is four years (48 months) old. How far off could this prediction be on average?

i) Use the logarithmic function to predict the length of a black bear that is 10 years (120 months) old. How far off could this prediction be on average?

j) Do you think it would be all right to extrapolate a lot and use this model to predict the length of a bear that is 50 years (600 months) old? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (c) and (d)?

3. Open the Nonlinear Data Sets in Excel. Copy and paste the temperature and copper expansion data into Statcato. This data set shows the relationship between temperature of copper in degrees Kelvin and how much the volume of the copper expands in cubic centimeters. Let temperature be the explanatory variable (x) and the copper expansion be the response variable (y).

   a) Use Statcato to make a scatter plot of the ordered pairs. Does the scatterplot look like a logarithmic model might fit? If so, would it be logarithmic growth or logarithmic decay.

   b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to find the equation of the Logarithmic curve that that best fits data. Do you think that the logarithmic curve fits the data well? Are the points close to the curve? Save or draw a rough sketch of the scatterplot with the log curve.

   c) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

   d) The standard deviation of the residual errors in Statcato is accurate for log curves. What was the standard deviation of the residual errors? Write two sentences explaining the standard deviation.

   e) Give the equation of the log curve.

   f) What is the scope of the data (x values)?

   g) Do you think copper expansion will continue to expand outside the scope of the data? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

   h) Use the logarithmic function to predict the amount of expansion when the temperature is 400 degrees Kelvin. How far off could this prediction be on average?

i) Use the logarithmic function to predict the amount of expansion when the temperature is 600 degrees Kelvin. How far off could this prediction be on average?

j) Do you think it would be all right to extrapolate some and use this model to predict how much copper would expand when the temperature is 1000 degrees Kelvin? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (c) and (d)?

4. Open the Nonlinear Data Sets in Excel. Copy and paste the atom defects and energy data into Statcato. This data set shows the relationship between atoms and the energy they can release. The energy numbers are all negative, denoting the loss of electrons.

a) Use Statcato to make a scatter plot of the ordered pairs. Does the scatterplot look like a logarithmic model might fit? If so, would it be logarithmic growth or logarithmic decay.

b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to find the equation of the Logarithmic curve that that best fits data. Do you think that the logarithmic curve fits the data well? Are the points close to the curve? Save or draw a rough sketch of the scatterplot with the log curve.

c) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

d) The standard deviation of the residual errors in Statcato is accurate for log curves. What was the standard deviation of the residual errors? Write two sentences explaining the standard deviation.

e) Give the equation of the log curve.

f) What is the scope of the data (x values)?

g) Do you think the atomic energy released will continue to follow this pattern outside the scope of the data? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

h) Use the logarithmic function to predict the amount of energy released if the atom has a defect of 0.37. How far off could this prediction be on average?

i) Use the logarithmic function to predict the amount of energy released if the atom has a defect of 0.56. How far off could this prediction be on average?

j) Do you think it would be all right to extrapolate some and use this model to predict the energy released if the atom defect is 0.9? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (c) and (d)?

5. A student wanted to map a logarithmic function model to the years since 1995 and worldwide wind power data (see nonlinear data sets). Statcato said "Error" and was unable to find the log function. Explain why this happened. What does this tell us about data sets that cannot be modeled with logarithmic models? Can you think of a way to change the data set so that we would be able to find a logarithmic model?

6. Draw an exponential growth curve, exponential decay curve, logarithmic growth curve, and logarithmic decay curve and discuss the key features of each curve. What is the relationship between logarithmic functions and exponential functions?

Section 6C – Quadratic Relationships with Technology

Another type of curve seen in scatterplots is the quadratic curve. Quadratic curves have a distinctive "U" shape. This U shape is often called a "parabola". Because of their parabolic shape, quadratic curves are commonly used to map airplane flights or missile launches. A scatterplot does not have to have a U shape to use a quadratic curve. Quadratic curves can be used to model many different patterns in the scope of the x-values. Think of it as using a piece of the parabola instead of the whole thing. It only has to match in the scope of the x-values.

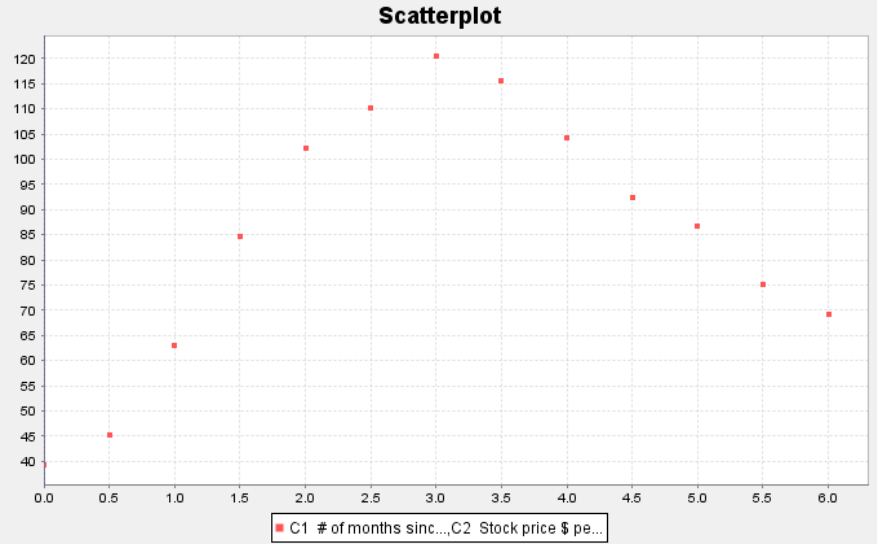
However, what makes a curve Quadratic? What does the equation look like? Whereas lines have the form $\hat{y} = mx + b$, exponential curves are known for the variable exponents, and log curves have "LN(x)" in the formula, Quadratic curves are known for their squared variables. The standard form for a quadratic function is $\hat{y} = c + bx + ax^2$ where *a, b* and *c* are real numbers.

Example 1

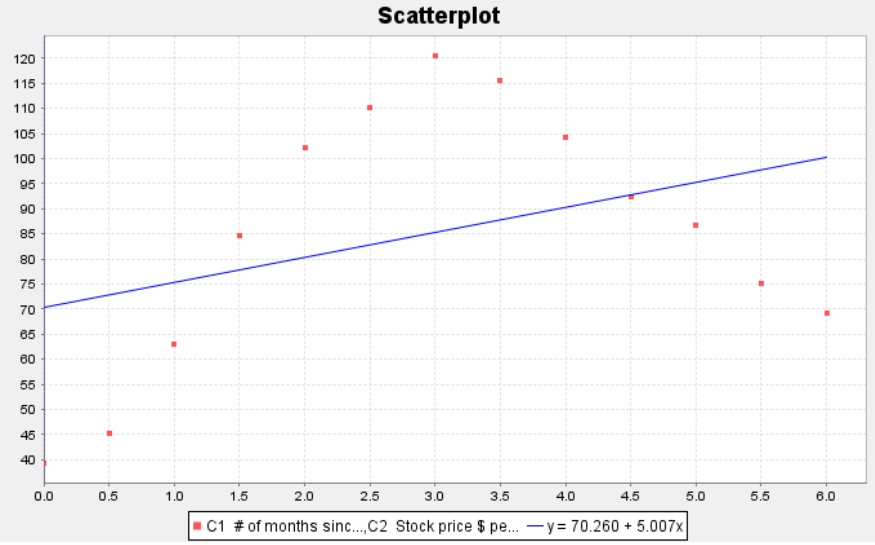
Let us look at the following data set. This gives the value of a stock over a 6-month period. Stocks are notorious for going up and down in value depending on the state of the economy at the time. We let the explanatory variable be the number of months since January 1st, and the response variable is the stock price per share in dollars.

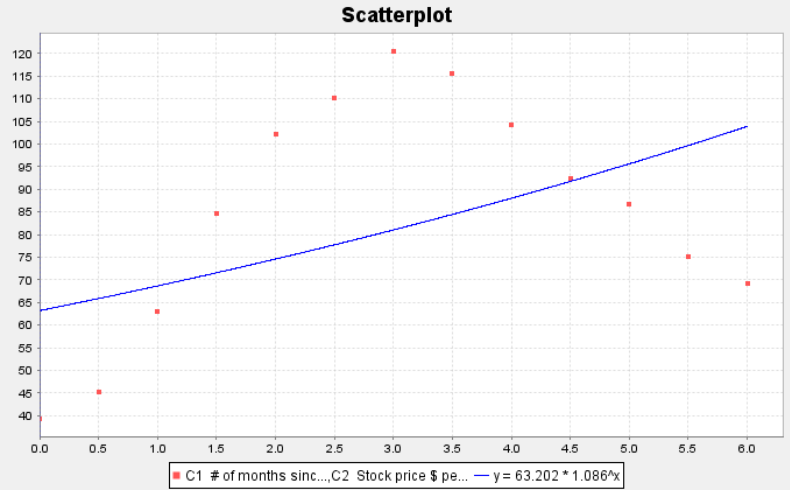
# of months since January 1	**Stock price $ per share**
0	39.4
0.5	45.35
1	62.91
1.5	84.71
2	102.31
2.5	110.19
3	120.4
3.5	115.63
4	104.23
4.5	92.5
5	86.61
5.5	75.12
6	69.29

If we make a scatterplot of this data on Statcato, we get the following graph. Notice the distinctive upside down U shape.

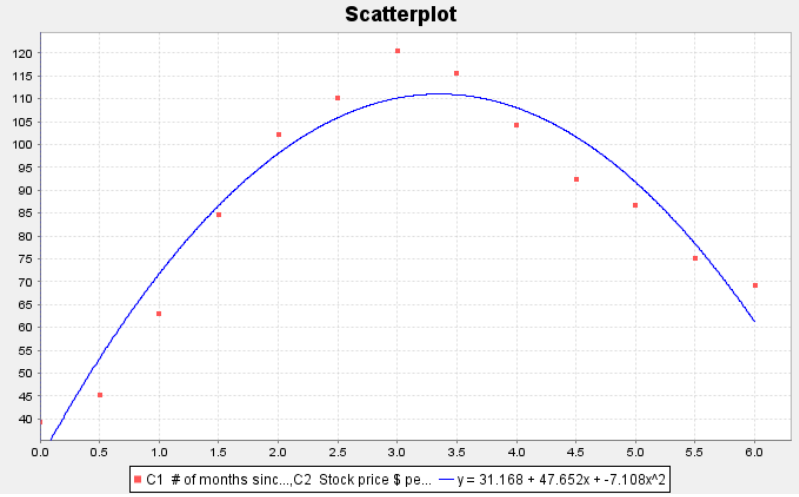


So far, we have discussed the linear, exponential, and log curves. Having Statcato find the least squares regression line and the best-fit exponential function, we obtain the following graphs.





Notice that neither the regression line nor the exponential curve seem to fit the data very well. In fact, the exponential looks almost the same as the regression line. We will now find the quadratic curve that best fits the data.



The quadratic curve seems to fit the data very well. Let us look at the quadratic equation that the software found.

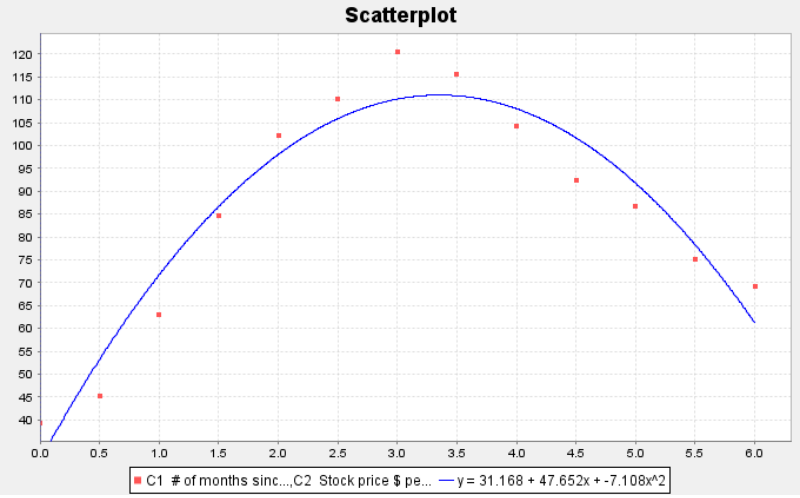
$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

Notice that the number in front of the x^2 term is negative (-7.108). This is called the leading coefficient. When the leading coefficient is positive, the quadratic will have an opening up "U" shape, but as in this function, when the leading coefficient is negative, the quadratic will have an upside down "U" shape.

Finding the Vertex

Parabolas that open up have a minimum Y value and parabolas that open down have a maximum Y value. This can be important information to businesses trying to find an approximate minimum cost or maximum profits.

In the last example, we looked at some stock data that seemed to take on a parabolic shape. We found the equation of the quadratic curve that fit the data and confirmed that the curve does fit the data by looking at R-squared and the standard deviation of the residual errors.



$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

One might ask during what month the stock reached a maximum price and what was that maximum price. First notice that since the parabola opens down, the maximum occurred at the top of the parabola. We call this point the vertex. It is important to keep in mind that any point has an x coordinate and a y coordinate. In this problem x represented the number of months since January 1st and the y represented the stock price in dollars per share.

So to find the point in time (months) when the stock reached a maximum, we will need to find the x coordinate of the vertex. To find the maximum predicted price (dollars per share), we will need to find the y coordinate of the vertex.

Fortunately, algebra can help us. There is formula for finding the x coordinate of the vertex.

X coordinate of the vertex = -1b / 2a

The "b" is the number in front of x and the "a" is the number in front of x-squared. Let us use the formula to calculate the x coordinate of the vertex for the stock price data.

X coordinate of the vertex = -1b / 2a = -1(47.652) / 2(-7.108) = -47.652 / -14.216 ≈ 3.352

What does this tell us? Remember the units. The explanatory (X) variable in this problem was the number of months since January. Therefore, the model predicts that the maximum stock price occurred about 3.352 months after January 1st.

What is the predicted maximum stock price? For this, we will need the Y coordinate of the vertex.

Y coordinate of the vertex: Plug in the x coordinate of the vertex into the equation of the quadratic curve and compute the Y value. Make sure to follow the order of operations.

$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

Y = 31.168 + 47.652(3.352) – 7.108 (3.352)^2

= 31.168 + 47.652(3.352) – 7.108 (11.235904)

≈ 31.168 + 159.7295 – 79.8648

≈ 111.03

Therefore, the predicted maximum stock price is about $111.03 per share.

Making Predictions with the Quadratic Curve

Since the quadratic curve fits the stock price data pretty well, let us use the curve to make a prediction.

Let us predict the stock price in mid-February (month 2.5). We would need to plug in 2.5 for x in the equation of the quadratic curve and compute. Remember to follow the order of operations.

$$\hat{y} = 31.168 + 47.652x - 7.108x^2$$

Y = 31.168 + 47.652(2.5) – 7.108 (2.5)^2

= 31.168 + 47.652(2.5) – 7.108 (6.25)

≈ 31.168 + 119.13 – 44.425

≈ $105.87 (Negative!)

Therefore, the predicted stock price in mid-February is about $105.87 per share.

Remember to be careful with extrapolation.

The scope of the x values for this data are between 0 and 6. So making predictions out of the scope is called extrapolation and may lead to large prediction errors.

Extrapolation: Let us predict the stock price in mid-September (month 9.5). Notice this is not in the scope of the x values. Let us plug in 9.5 for x in the equation of the quadratic curve and see what happens. Remember to follow the order of operations.

$\hat{y} = 31.168 + 47.652x - 7.108x^2$

Y = 31.168 + 47.652(9.5) – 7.108 (9.5)^2

= 31.168 + 47.652(9.5) – 7.108 (90.25)

≈ 31.168 + 452.694 – 641.497

≈ -$157.64 (Negative!)

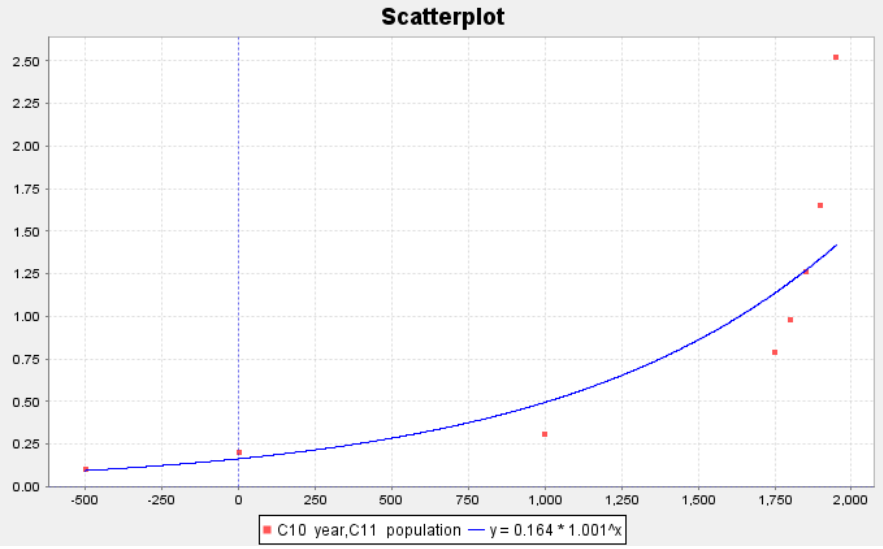
So the predicted stock price is about -$157.64 per share. Notice this does not make much sense and seems to have a huge error in the prediction. That is what can happen when you extrapolate.

Example 2

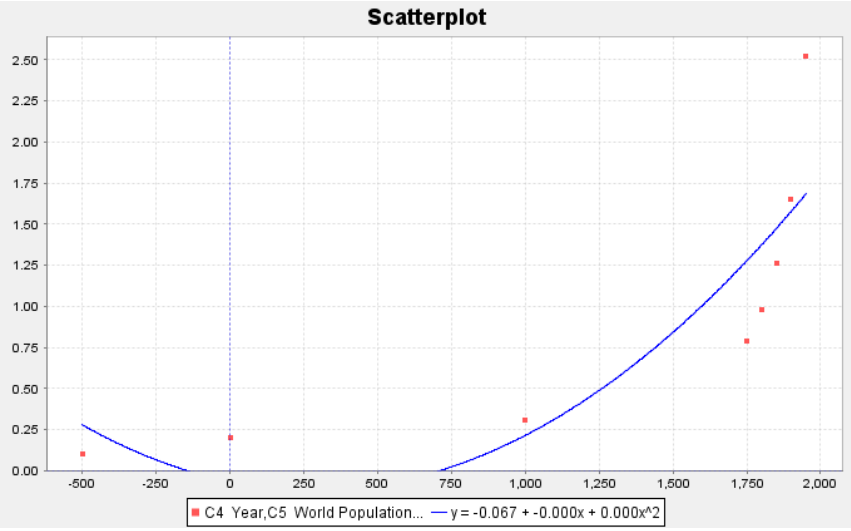
As I said earlier, do not make the mistake of thinking that quadratic functions are only useful when your scatterplot has a U shape. On the contrary, quadratic functions can be a good model for many curves.

For an example of this, let us look again at our world population data. Recall that this data gives the year and the world population from the year 500 BC to the year 1950 AD. The statistics software found that the equation for that exponential curve that fits the data was $\hat{y} = 0.164(1.001)^x$. Here is the data and a scatterplot of the data with the exponential function.

Year	World Population in Billions
-500	0.1
1	0.2
1000	0.31
1750	0.791
1800	0.978
1850	1.262
1900	1.65
1950	2.519



We said that the exponential function fits the data set better than the regression line, but still not perfectly. We may think about whether another function might fit the data better than the exponential. Plugging in this data into the statistics software, we have the program find the quadratic curve that best fits the data. Here is the scatterplot.

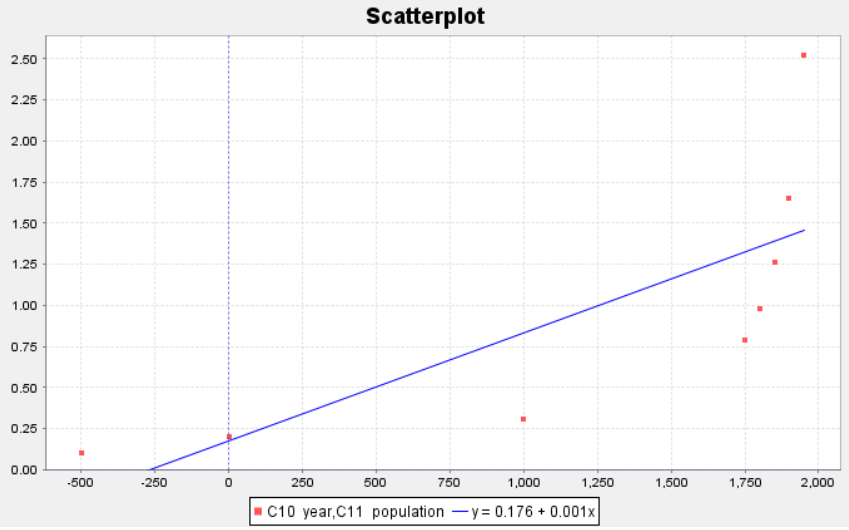


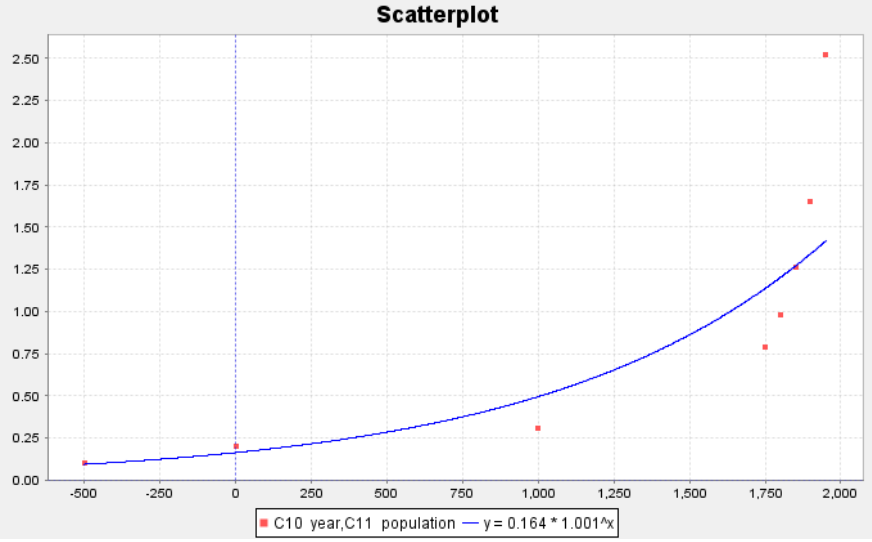
Notice that the quadratic function fits the data reasonably well. Looking at the graph, you may see that the leading coefficient says "0.000". This does not mean that the leading coefficient is zero. (If that were the case, this function would not be quadratic.) It just means that the number when rounded to three decimal places is zero. We can get these numbers with better accuracy. We found the curve to be $\hat{y} = -0.6674 - 0.00037x + 0.0000006479x^2$. Notice the leading coefficient is positive, which corresponds to the graphs U shape.

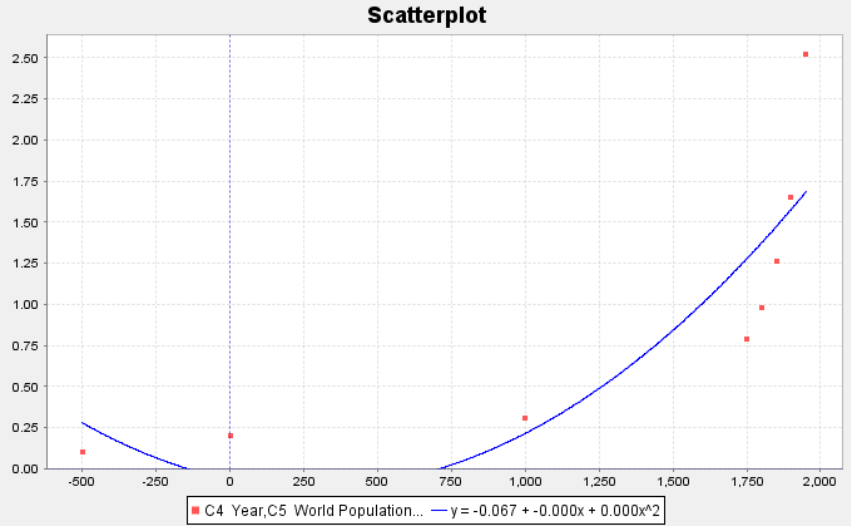
Now we have a quandary. We found the linear and exponential functions that fit this data. Now we also have the quadratic function to think about. So which is the best fit for the data?

Assessing the fit of a quadratic function

Let us start be looking again at the scatterplots. Which curve or line looks like it fits the data the best, the line, the exponential curve, or the quadratic curve?







We definitely can see that the curves fit the data better than the line, but it is hard to tell which of the curves fit the data better. Recall that one way to answer the question of best fit is to look at the R-squared values for the line and each of the two curves.

R-Squared (line) = 0.5878
R-Squared (Exponential curve) = 0.9078
R-Squared (Quadratic curve) = 0.7369

Remember that we do not like to use a more complicated model unless there is a significant improvement. We see from the R-squared values that the quadratic (73.7%) is significantly better than the line (58.8%) but not nearly as good as the exponential (90.8%). For this data set, it seems the exponential is the best model.

We also like to look at the standard deviation of the residual errors (S_e). Recall that the curve with the smallest standard deviation is also an indication of best fit. We can use the statistics software to calculate the standard deviations for the line and the two curves.

S_e (line) = 0.5721 Billion people

S_e (exponential) = 0.3673 Billion people

S_e (quadratic) = 0.5007 Billion people.

As with the R-squared, the exponential model seems to be the best fit for this data. It has not only the highest R-squared value, but also the lowest standard deviation of the residual errors.

Problem Set Section 6C

1. Open the Nonlinear Data Sets in Excel. Copy and paste the number of seconds and height of a rock data into Statcato. A rock was thrown off a 273-foot cliff and this data set gives the number of seconds and the corresponding height in feet of the rock. Let the number of seconds be the explanatory variable and the height be the response variable.

   a) Make a scatter plot of the ordered pairs. Does the scatterplot have a parabolic shape? If so, would the parabola open up or down?

   b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to create the scatterplot with the quadratic curve. Save or draw a rough sketch of the scatterplot and curve. Do you think that the quadratic function fits the data well? Are the points close to the curve?

   c) What is the equation for the quadratic curve?

   d) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

   e) What was the standard deviation of the residual errors s_e? Write two sentences explaining the two meanings of the standard deviation in this context.

   f) Use the formula $\frac{-b}{2a}$ to predict the number of seconds will elapse before the rock reaches a maximum height. What is the maximum height?

   g) What is the scope of the data (x values)?

   h) Do you think the height of the rock will continue to follow this pattern for a long time? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

   i) Use the quadratic function to predict the height of the rock after 1.8 seconds. How far off could this prediction be on average?

   j) Use the quadratic function to predict the height of the rock after 3.2 seconds. How far off could this prediction be on average?

k) Do you think it would be all right to extrapolate a lot and use this model to predict the height of the rock after 20 seconds? Why or why not? If a person did make this prediction, would the prediction even make sense?

2. Open the month and solar energy data in Statcato. The college kept track of how much solar energy was made by their solar panels in kilowatt-hours (kWh) for every month in 2009. Let the explanatory variable be the month and the solar energy be the response variable.

   a) Make a scatter plot of the ordered pairs. Does the scatterplot have a parabolic shape? If so, would the parabola open up or down?

   b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to create the scatterplot with the quadratic curve. Save or draw a rough sketch of the scatterplot and curve. Do you think that the quadratic function fits the data well? Are the points close to the curve?

   c) What is the equation for the quadratic curve?

   d) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

   e) What was the standard deviation of the residual errors s_e? Write two sentences explaining the two meanings of the standard deviation in this context.

   f) Use the formula $\frac{-b}{2a}$ to predict what month will have the maximum solar energy. What is the maximum energy?

   g) What is the scope of the data (x values)?

   h) Do you think the solar energy will follow this pattern into the future? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

   i) Use the quadratic function to predict the amount of solar energy in mid-March (month 3.5). How far off could this prediction be on average?

j) Use the quadratic function to predict the amount of solar energy in mid-October (month 10.5). How far off could this prediction be on average?

k) Do you think it would be all right to extrapolate a lot and use this model to predict the solar energy in January of 2029 (month 240)? Why or why not? If a person did make this prediction, would the prediction even make sense?

3. A company that manufactures transmissions wants to minimize their costs. They think there may be a relationship between their monthly costs and the average number of hours their employees work per week. Nonlinear Data Sets in Excel. Copy and paste the average hours employees work and the transmission company costs per month into Statcato. Let the hours worked be the explanatory variable and the costs be the response variable.

a) Make a scatter plot of the ordered pairs. Does the scatterplot have a parabolic shape? If so, would the parabola open up or down?

b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to create the scatterplot with the quadratic curve. Save or draw a rough sketch of the scatterplot and curve. Do you think that the quadratic function fits the data well? Are the points close to the curve?

c) What is the equation for the quadratic curve?

d) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

e) What was the standard deviation of the residual errors s_e? Write two sentences explaining the two meanings of the standard deviation in this context.

f) Use the formula $\frac{-b}{2a}$ to determine how much their employees should work in order to minimize costs. What is the minimum cost?

g) What is the scope of the data (x values)?

h) Do you think monthly costs will continue to follow this pattern outside the scope of the data? Why or why not? Discuss the implications on extrapolation. How far do you think you

can extrapolate before the prediction becomes really bad? Why?

i) Use the quadratic function to predict the monthly costs when the employees work an average of 44 hours per week. How far off could this prediction be on average?

j) Use the quadratic function to predict the monthly costs when the employees work an average of 35 hours per week. How far off could this prediction be on average?

k) Do you think it would be all right to extrapolate a lot and use this model to predict monthly costs when employees work an average of 120 hours per week? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (i) and (j)?

4. Open the Nonlinear Data Sets in Excel. Copy and paste the black bear age in months and length in inches into Statcato. Let the bear age be the explanatory variable and the length be the response variable.

a) Make a scatter plot of the ordered pairs. Does the scatterplot have a parabolic shape? If so, would the parabola open up or down?

b) Now use the nonlinear model function in Statcato (under the correlation and regression menu) to create the scatterplot with the quadratic curve. Save or draw a rough sketch of the scatterplot and curve. Do you think that the quadratic function fits the data well? Are the points close to the curve?

c) What is the equation for the quadratic curve?

d) What is r^2? Write a sentence explaining the meaning of r^2 in this context.

e) What was the standard deviation of the residual errors s_e? Write two sentences explaining the two meanings of the standard deviation in this context.

f) Use the formula $\frac{-b}{2a}$ to determine the age of a bear when it reaches its maximum length. What is the maximum length?

g) What is the scope of the data (x values)?

h) Do you think the bear lengths will continue to follow this pattern outside the scope of the data? Why or why not? Discuss the implications on extrapolation. How far do you think you can extrapolate before the prediction becomes really bad? Why?

i) Use the quadratic function to predict the length of a bear that is four years (48 months) old. How far off could this prediction be on average?

j) Use the quadratic function to predict the length of a bear that is ten years (150 months) old. How far off could this prediction be on average?

k) Do you think it would be all right to extrapolate a lot and use this model to predict the length of a bear that is 30 years (360 months) old? Why or why not? If a person did make this prediction, would it have the same prediction error as parts (i) and (j)?

5. How can we identify a quadratic function if we only see the equation? How can we know from just the equation of the quadratic function whether it opens up or down?

6. How do we know if the quadratic function has a maximum or minimum point? Where does the maximum or minimum value occur? What are some applications where knowing the maximum or minimum will be important to know?

Chapter 6 Review

Here are the important topics to remember from this chapter.

- An exponential growth pattern looks like a backward L shape and increases very quickly from left to right.
- A logarithmic growth pattern looks like an upside down L shape and increases very slowly as the graph goes from left to right.
- Exponential and logarithmic decay patterns both look L shaped and decrease from left to right. The main difference is that a logarithmic decay curve can cross the x-axis but not the y-axis, while the exponential curve can cross the y-axis but not the x-axis.

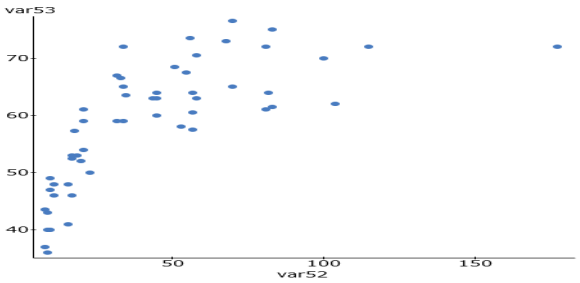
- Exponential Curves have equation where the x is an exponent. The equation looks like $y = a \bullet b^x$ where "a" is the y-intercept and "b" is the base. If the base is greater than 1, you will have an exponential growth curve. If the base is less than 1 you will have an exponential decay curve.
- You cannot use the exponential curve if the response variable (Y) has zero or negative numbers in the data set. (There may be ways to adjust the data though.)
- Logarithmic Curves have an "LN (X) in the equation. The equation looks like $y = a + b \bullet LN(x)$. If the number in front of the LN(x) is positive, you will have a logarithmic growth curve. If the number in front of the LN(x) is negative, you will have a logarithmic decay curve.
- You cannot use the logarithmic curve if the explanatory variable (X) has zero or negative numbers in the data set. (There may be ways to adjust the data though.)
- A traditional quadratic pattern has a parabolic "U" shape. The "U" may be facing up or down. A quadratic curve may still be a good model even if the shape is not "U" shaped since we can use a piece of the curve.
- The quadratic curve y = c + bx + ax^2. The quadratic curve opens up and has a minimum Y value if the leading coefficient "a" is positive. The quadratic curve opens down and has a maximum Y value if the leading coefficient "a" is negative. The quadratic curve works well with positive numbers, negative numbers and zero.
- The quadratic curve y = c + bx + ax^2 has a maximum or minimum point at the vertex. The x coordinate of the vertex can be calculated with -1b/2a. The y coordinate of the vertex can be calculated by plugging in -1b/2a in for x in the formula. The vertex may not always make sense, especially if the vertex is out of the scope of the x values.
- R-squared is the percent of variability in the response variable (Y) that can be explained by the (exponential, logarithmic, or quadratic) relationship with the explanatory variable (X). R-squared is a very useful number to judge how well the curve is fitting the data. The higher the r-squared percentage the better the fit.
- The standard deviation of the residual errors measures how far the points in the scatterplot are from the (exponential, logarithmic, or quadratic) curve on average. It also tells us the average prediction error if we use the equation to make a prediction in the scope of the x values.
- The curve with the highest r-squared and lowest standard deviation is generally the best-fit curve. However, statisticians also look at things like outliers, residual plots, and histograms of the residuals when judging the fit of a curve.

Problem Set Chapter 6 Review

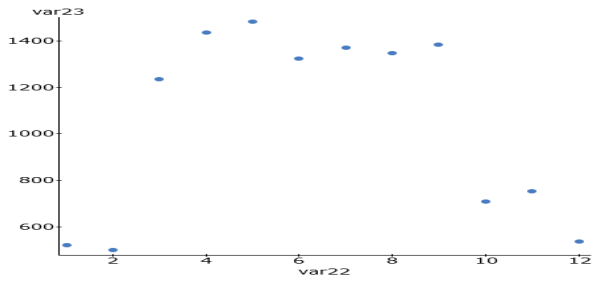
(For #1-4) Multiple Choice: Match each of the following scatterplots with one of the following patterns:

a) Exponential Growth
b) Logarithmic Growth
c) Exponential/Log Decay
d) Open Up Quadratic
e) Open Down Quadratic)

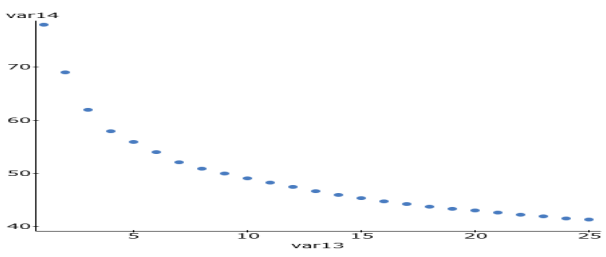
1.



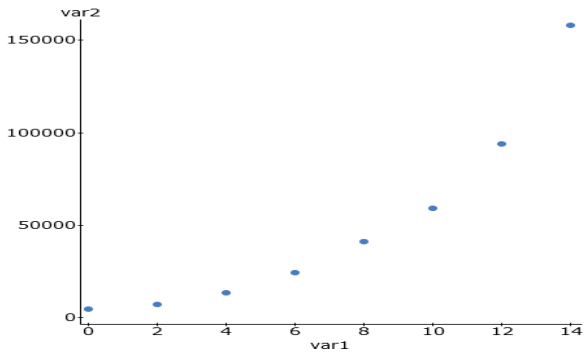
2.



3.



4.



5. A local business, decided to do an experiment. They wanted to see if there is a relationship between the number of lunch and snack breaks they gave their employees and how efficient their employees worked. Each week, a computer randomly selected how many breaks each employee would get, and then measured how efficient the employees were. The explanatory variable X was the number of breaks and the response variable Y was the efficiency rating percentage. After analyzing the data, we found that a quadratic curve fit the data pretty well and the following formula was found with statistics software.

$$Y = c + b\,x + a\,x^2$$

$$Y = 41.800 + 5.868\,x + {}^{-}0.163\,x^2$$

a) Use the formula $\frac{-1b}{2a}$ to find the number of breaks X that the company should give its employees per week in order to maximize their efficiency rating. *(Follow the order of operations and show your work.)*

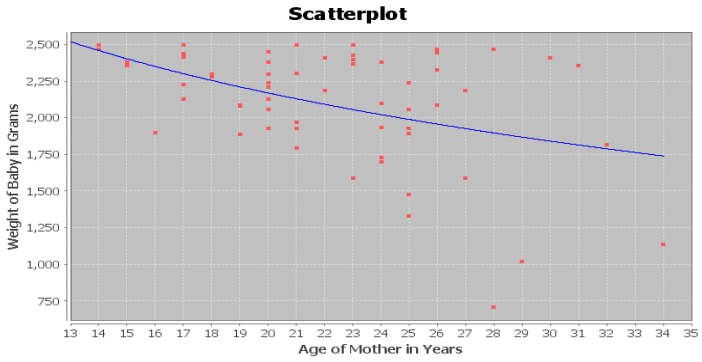
b) What is the predicted maximum efficiency rating Y if the company gives the employees the recommended number of breaks from part (a)? *(Hint: Plug in your answer in part (a) into the formula for x (and x-squared) and work it out with your calculator. Follow the order of operations and show your work. Round your rating to the tenths place.)*

(For #6-12) The following data describes the relationship between the age of a mother in years and the weight of underweight babies in grams. The age of the mother was the explanatory variable X and the weight of the underweight baby was the response variable Y. We used statistics software to find a Natural Logarithmic function that may fit the data.

Regression Equation: y = 4596.59332 + -810.36250 LN (x)

$r^2 = 0.1808$

Standard Deviation of the Residual Errors = 356.8787 grams



6. What percent of the variability in baby weight can be explained by the logarithmic relationship to the age of the mother?

7. How far are the points from the log curve on average?

8. If we use the log curve and the age of the mother to predict the weight of the baby, how far off might that prediction be?

9. Use the formula y = 4596.59332 + -810.36250 LN (x) to predict the babies weight if the mother was 33 years old. *(Hint: Plug in 33 for x and work it out with your calculator. Follow the order of operations and round your answer to the ones place. Don't round during the calculation)*

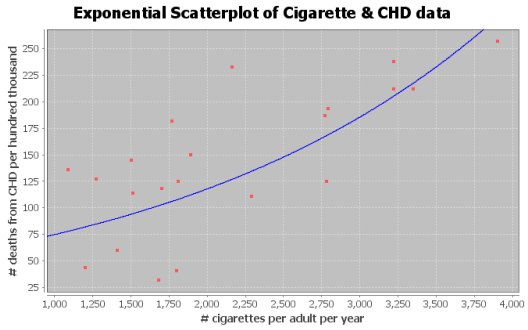
10. How well do you think the Log curve fits the data? Explain your answer with the scatterplot, r-squared, and standard deviation of the residual errors.

11. Does this study prove that a mother's age causes a baby to more underweight? Explain why or why not.

12. List some possible confounding variables that might influence a baby's weight other than just the age of the mother.

(For #13-22) The following data describes the relationship between smoking (# cigarettes per adult per year) and congestive heart disease (CHD) (# deaths per hundred thousand). The number of cigarettes was the explanatory variable X and the deaths by CHD was the response variable Y. Plugging the data into a statistics software, we tried both an exponential curve and a quadratic curve.

Exponential Scatterplot, Regression Equation, R-squared, Standard Deviation of Residuals

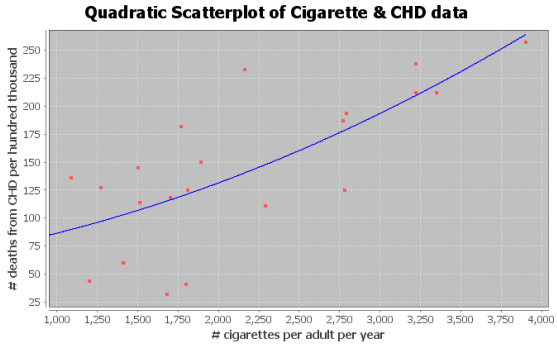


Exponential Equation: Y = 47.49274 (1.00045^X)

(Exponential) R-squared = 0.3746

(Exponential) Standard Deviation of the Residual Errors = 48.297 CHD deaths (*The computer said 0.4865 deaths but this is a mistake.)*

Quadratic Scatterplot, Regression Equation, R-squared, Standard Deviation of Residuals



Quadratic Equation: $Y = 58.59448 + 0.01939\ X + 0.00000852869\ X^2$

(Quadratic) R-squared = 0.5397

(Quadratic) Standard Deviation of the Residual Errors = 47.5847 CHD deaths

13. What percent of the variability in CHD deaths can be explained by the exponential relationship to the number of cigarettes?

14. What percent of the variability in CHD deaths can be explained by the quadratic relationship to the number of cigarettes?

15. Which curve (quadratic or exponential) had the strongest relationship? Explain your answer using R-squared.

16. How far are the points from the exponential curve on average?

17. How far are the points from the quadratic curve on average?

18. Which curve (quadratic or exponential) were the points in the scatterplot closer to?

19. If we use the exponential curve and the number of cigarettes per adult per year to predict the number of deaths by CHD, how far off might we be in that prediction.

20. If we use the quadratic curve and the number of cigarettes per adult per year to predict the number of deaths by CHD, how far off might we be in that prediction.

21. Which curve (quadratic or exponential) has less prediction error?

22. Which curve (quadratic or exponential) was the better fit for the cigarette and CHD data? Explain why using the R-squared and the standard deviation of the residuals.

Project Chapter 6 - Curved Quantitative Relationships

Directions: Create a Poster on the following topic. Make sure your poster has the following items. You will be presenting your poster to the other students in the class. Pick a team name for your group. Then chose one of the following paired data sets to analyze from the "nonlinear data sets" (updated summer 2017). Each group should have different data to analyze.

Data	Group #	Team Name	Exponential Se Fix
Age of Mother / Low Birth Weight			360.6 grams
Ave Cigarette / Deaths			48.1 deaths
Percent and cost of cleaning Lake			$17813.22
Work Hours / Transmission Company Cost			$2401.79
Month / Solar Energy			434.34 kWh
Year (Adjusted) / World Population			0.08052 Billion People
Year (Adjusted) / House Prices			$7145.94
Temperature / Copper Expansion			6.699 cubic cent.

- Pick two quantitative variables and pick which should be X and which should be Y. The poster should give the explanatory variable (x) and response variables (y), what the units are for x and y.

- Use Statcato to create a scatterplot, R-squared, standard deviation of the residuals, and the equation of the curve (formulas) for the exponential curve, the logarithmic curve, and the quadratic curve. There should be three scatterplots, three equations (formulas), three r-squared values, and three standard deviations on your poster. *(Note: The standard deviation for the exponential curve will be wrong in Statcato. The correct standard deviation is given above. The quadratic and the log curves have the correct standard deviation in Statcato.)*

- Write a sentence for each of the three R-squared values. (Three total sentences)

- Write two sentences for each of the three standard deviations. (Six total sentences)

- List the r-squared values and standard deviations on your poster and use them to decide which of the three curves the best fit for the data is? Explain your choice.

- What is the scope of the x-values? *(May differ if using "adjusted" data.)*

- Choose any x value in the scope and plug it into the equation of your best-fit curve to predict the y value. *(Use only your best-fit curve.)* How far off could your prediction be on average?

- Decorate your poster!!