Content Background for Module 3

The information given below is designed to address issues and topics related to regression that are not contained in typical high school textbooks or curriculum guides. Though some of the information on regression can be easily understood by reading about it below, other important information regarding regression is better presented in video format. With that said, we invite you to learn more about regression by viewing the video segments that accompany this module. The video available for viewing covers the following concept:

- Quadratic Regression – This video shows how to find the best fitting quadratic model for a set of data, using the TI-83 calculator.

What is regression analysis?
Regression analysis can be thought of as "finding the best fitting mathematical model". The primary objective of regression analysis is to determine the values of parameters for various functions based on a given data set. There are three common types of regression; linear regression, non-linear regression, and multivariable regression. Linear regression refers specifically to finding a best fitting line for a given data set, non-linear regression is finding a best fitting model that is not a line, and multivariable regression refers to regression that contains more than one independent variable. For pre-calculus, we stick to linear and non-linear regression where data sets contain one independent variable and one dependent variable. Multivariable regression is beyond the scope of a high school pre-calculus course. The most common types of functions that are used for regression in pre-calculus, along with their parameters, are listed below. (It is important to note that $a \neq 0$.)

Linear	$y = mx + b$
Quadratic	$y = ax^2 + bx + c$
Cubic	$y = ax^3 + bx^2 + cx + d$
Quartic	$y = ax^4 + bx^3 + cx^2 + dx + e$
Exponential	$y = a \cdot b^x$
Logarithmic	$y = b + a \cdot \ln x$
Power	$y = a \cdot x^b$
Sin	$y = a \cdot \sin(bx + c) + d$

Who uses regression?
Anyone who wants to make predictions or inferences based on a particular data set. Companies, for example, use regression to determine their profit, revenue, and cost functions and scientists use regression to analyze and predict all sorts of time series phenomenon including the population of certain bacteria, crop data, planetary orbits, and projectile tracking. These days, regression is the most used method for data analysis.

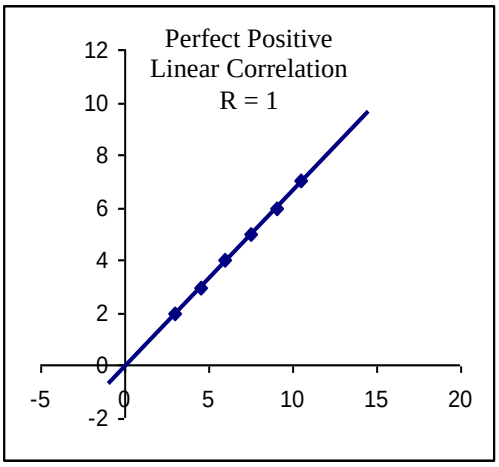
If you have a given data set, how do you find the parameters for the various mathematical models?
There are many types of calculators and computer programs that will find the different regression models. The videos that correspond to this module will give you the directions for finding regression models using the TI-83 calculator.

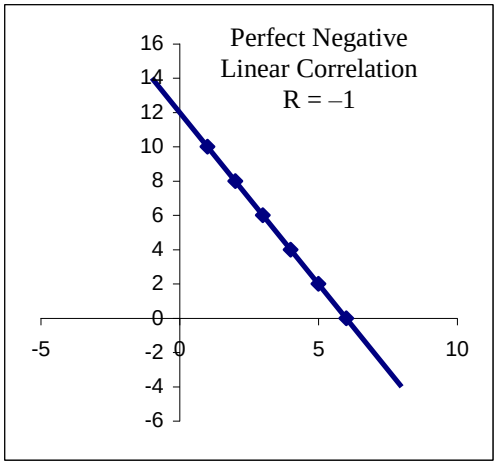
What type of data sets can be used for regression?
For a pre-calculus class you can use any type of data set that compares two quantities. When choosing a data set it is a good idea to keep the following points in mind.

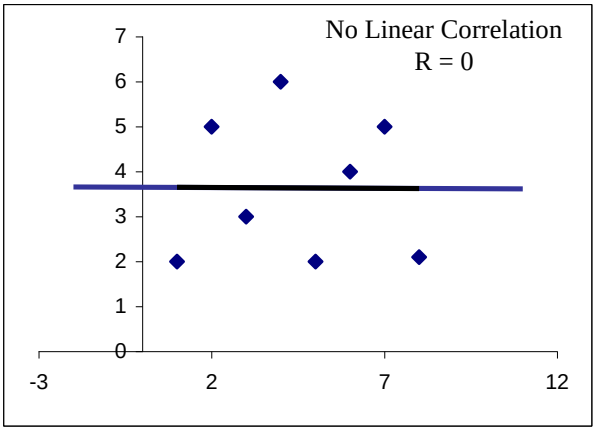
- The data set must contain a dependent and an independent variable.
- You must be able to standardize the data set.
- Two points determine a unique line, so you need more than 2 points to effectively use linear regression.
- Three points determine a unique quadratic function, so you need more than 3 points to effectively use quadratic regression.
- Four points determine a unique cubic function, so you need more than 4 points to effectively use cubic regression.

What is meant by "standardizing the data"?
Standardizing the data is a process that allows you to obtain the simplest symbolic expression for each type of function. To standardize the data, you need to choose reasonable x values to assign the independent variables and reasonable y values to assign the dependent variables. If, for example, your independent variables are the years from 1975 – 2000 and your dependent variables are the average salaries of the American presidents, you might want to let $x = 0$ represent the year 1970 and let your y values be in units of hundreds of thousands of dollars. Standardizing your data in this manner will yield smaller numbers for the parameters in each model. One factor you need to keep in mind when standardizing the data is the domain of the logarithmic and power functions. The logarithmic and (negative) power functions do not have zero in their domain, which means that you do not want to set your first data point to $x = 0$. This is why we would let $x = 0$ represent 1970 and not 1975, in the previous example. If you include zero as an x value in your standardized data and you try to perform logarithmic or power regression, the program you are using for regression will send a "domain error" and, consequently, will not be able to give you a logarithmic or power model.

What is a correlation coefficient?
The correlation coefficient, commonly known as the R value, is a number between –1 and 1 which measures the degree to which two variables are related. If there is a perfect, positive linear relationship between the two variables, we have a correlation coefficient of 1, R=1. If there is a perfect, negative linear relationship between the two variables, we have a correlation coefficient of –1, R = –1. If there is no correlation between the two variables then there is no correlation and consequently R = 0. These different correlations are shown in the figures below.







A quick way to interpret the R value is as follows:

R Value	Interpretation
$0 \le \lvert R \rvert < 0.2$	Very weak to negligible correlation
$0.2 \le \lvert R \rvert < 0.4$	Weak, low correlation (not very significant)
$0.4 \le \lvert R \rvert < 0.7$	Moderate correlation
$0.7 \le \lvert R \rvert < 0.9$	Strong, high correlation
$0.9 \le \lvert R \rvert \le 1$	Very strong correlation

Source: (http://www.bized.ac.uk/timeweb/crunching/crunch_relate_expl.htm)

How do we measure the correlation between two variables where the correlation is not always increasing or decreasing?

To determine the correlation between variables where the correlation is not always increasing or decreasing we need to look at the coefficient of determination, which is simply the square of the correlation coefficient, R^2. This squared result will give us a rough percentage for how much the

dependent variable is attributed to the independent variable. Since the coefficient of determination is available for all of the regression models used in pre-calculus, we use the R^2 when comparing and choosing regression models. You can use the same interpretation scale used above for the correlation coefficient to interpret the coefficient of determination (R^2). It is important to note that Sine regression will not generate a R^2 value, and we typically only use Sine regression if we see a periodic trend.

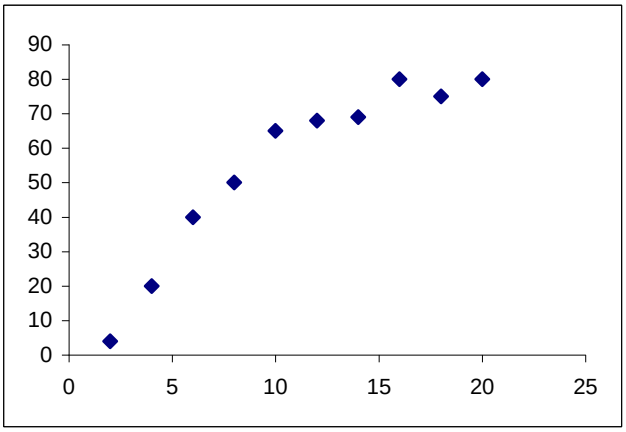
After you have the parameters for each mathematical model, how do you choose which model best fits a given data set?

Typically students think that they can find the best fitting model by finding the model with the highest coefficient of determination. This is not always true. There are other factors involved in finding the best model. Some points to consider are:

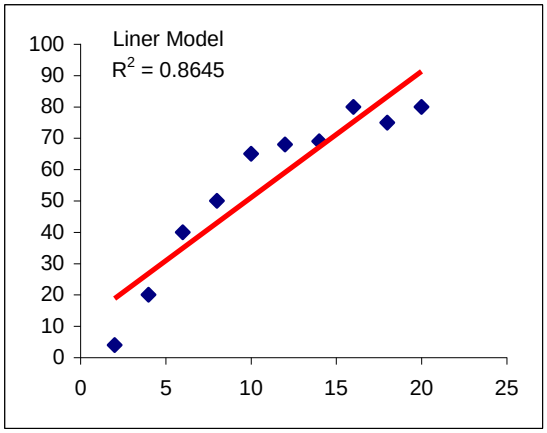
- What are you using the model for? Will you make predictions regarding values within the given data points (interpolate) or will you make predictions regarding values outside the given data points (extrapolate)?
- What is the value of the coefficient of determination, the R^2 value?
- Are there constraints? Can the independent variable be negative?
- Which of the mathematical models is the simplest?

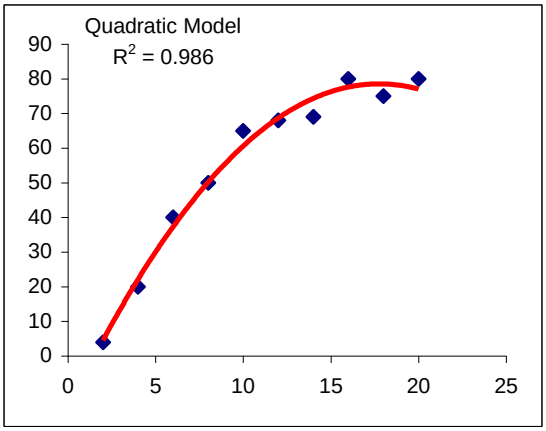
The most important point to consider is interpolation versus extrapolation. If you are using the mathematical model for interpolation, that is you want to use the model to answer questions concerning points within the given data set, you can generally choose from the models that have high R^2 values. If, however, you are using the mathematical model for extrapolation, that is you want to use the model to answer questions concerning points outside the given data set, you need to look at the graph of the model versus the overall trend of the data. To better demonstrate how to find the best model let's use the following set of random data and its corresponding scatter plot.

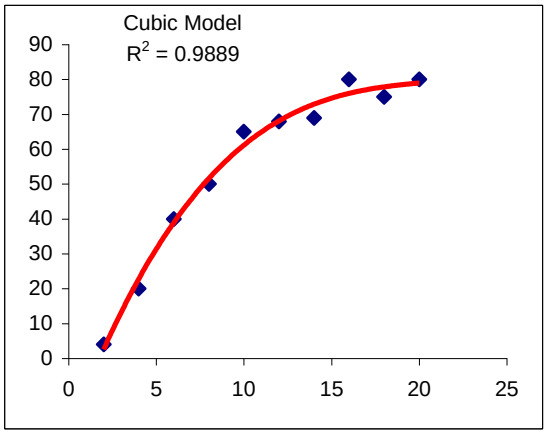
x	2	4	6	8	10	12	14	16	18	20
y	4	20	40	50	65	68	69	80	75	80

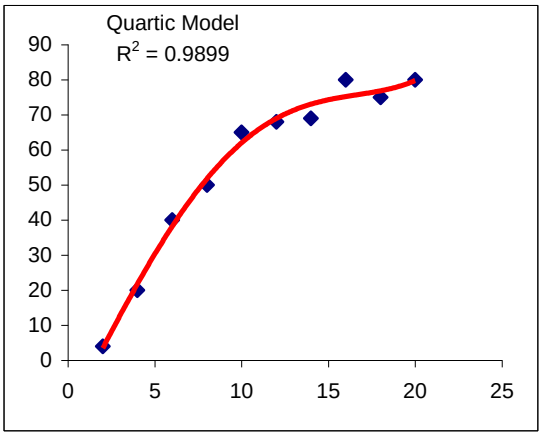


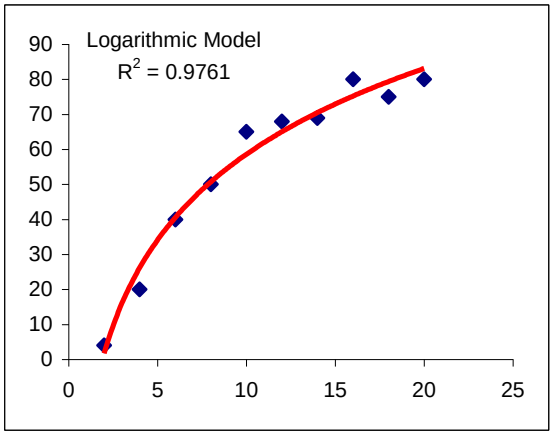
(Using Excel to generate the best models, we get the following regression curves along with their coefficients of determination.)

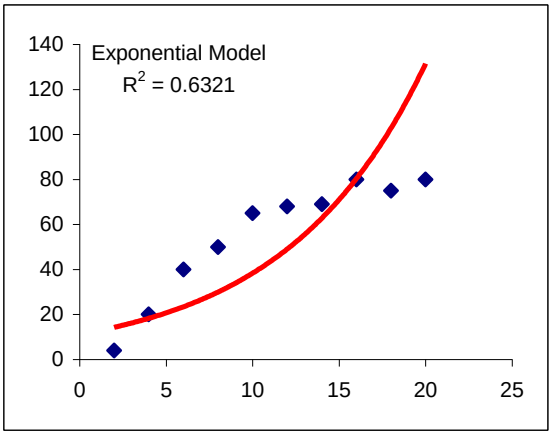

Liner Model
R2 = 0.8645

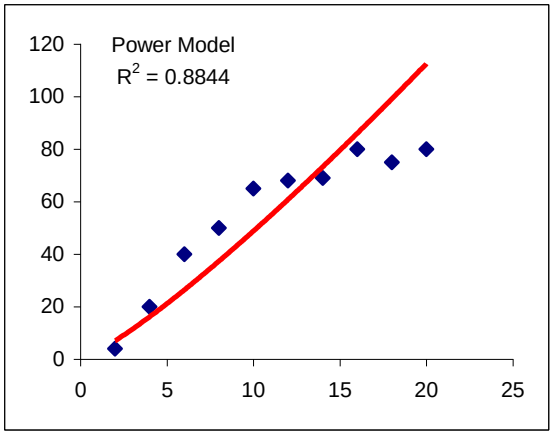

Quadratic Model
R2 = 0.986


Cubic Model
R2 = 0.9889


Quartic Model
R2 = 0.9899

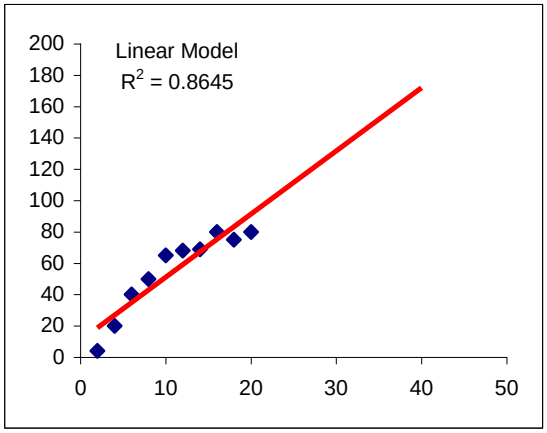

Logarithmic Model
R2 = 0.9761

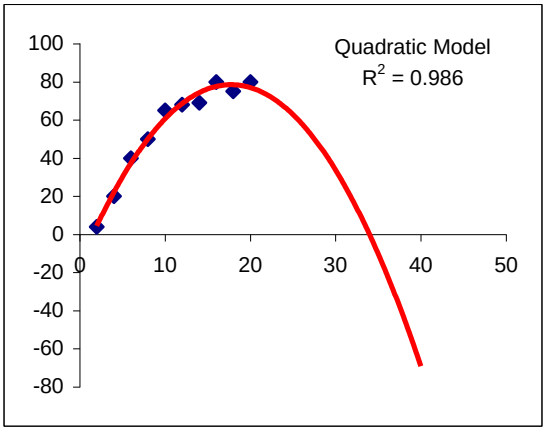

Exponential Model
R2 = 0.6321

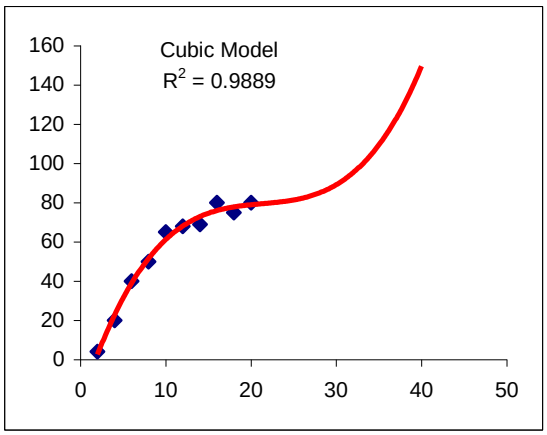


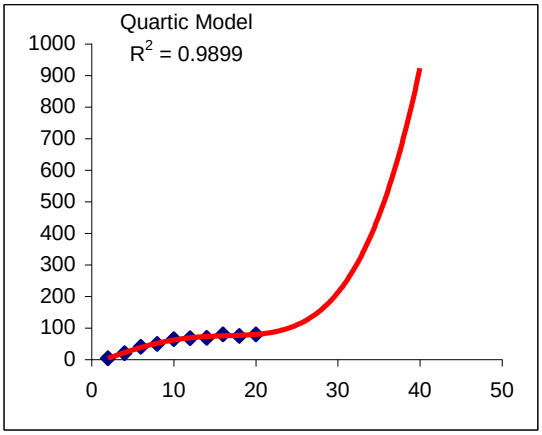
If we wanted to interpolate this data then we may choose from the quadratic (R^2 = 0.986), cubic (R^2 = 0.98896), and quartic (R^2 = 0.9899) model because they have the higher coefficients of determination. In general, for polynomial regression, the higher the degree, the greater the number of parameters, and the higher the R^2 value will be. However, in comparing the different R^2 values of these polynomial models we see that there is little difference between the values, 0.0039 at most, so in this case we would choose the simplest model, which would be the quadratic because it is the polynomial of lesser degree.

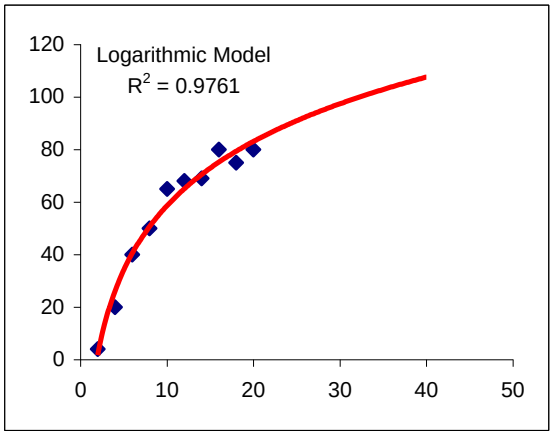
If, however, we wanted to extrapolate this data and look at the model when x = 40, then we need to extend the graph of our models to cover this particular x value.

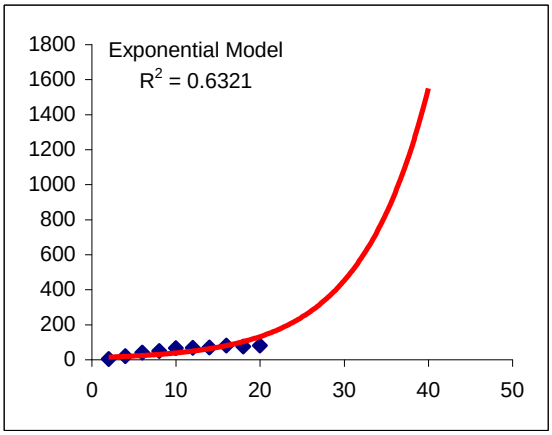


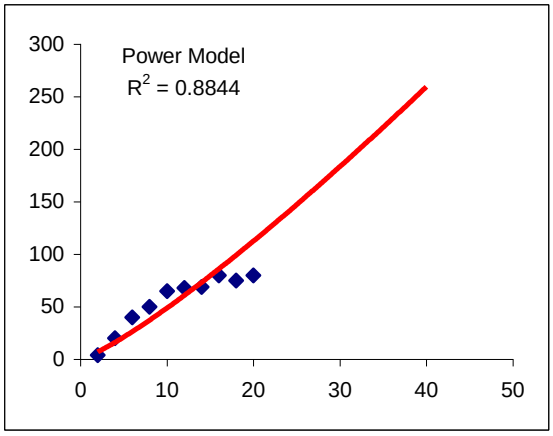












When we look at the general trend of the data, and if we assume this trend will continue, then it seems that the logarithmic model would best describe the data at $x = 40$. Notice that the linear and power models show a steady increase from $x = 20$ to $x = 40$, which does not follow the general trend of the data and the cubic, quartic, and exponential models show a dramatic increase after $x = 30$. The quadratic model begins to decrease around $x = 20$, and at about $x = 34$ the quadratic model gives negative values.

Though this data is randomly generated and does not really have any real world meaning, it still clearly shows the importance of knowing what you are using the model for. In this example, we chose one model for interpolation and a different model for extrapolation, but the data set remained the same. When

teaching how to choose the best model, it is extremely important to stress to the students that the coefficient of determination is just one factor when choosing a best fitting model.

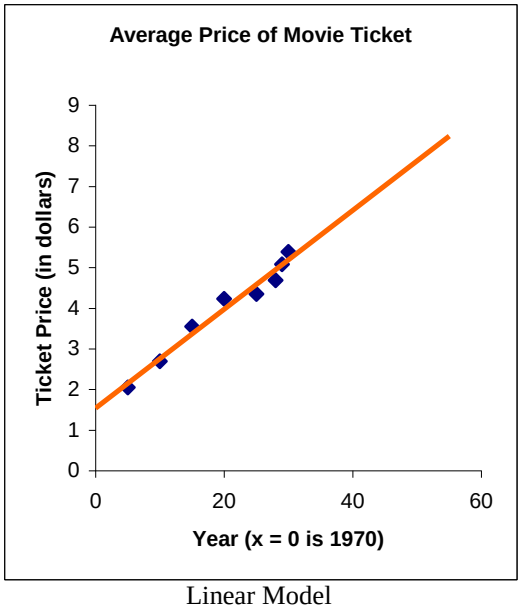
Can I see a sample problem along with the discussion the teacher would have with the students? The data below gives the average motion picture ticket price from 1975 – 1980. Use regression to find the best fitting mathematical model if you want to predict the average motion picture price in 2020.

Ticket price (in dollars)	2.05	2.69	3.55	4.23	4.35	4.69	5.08	5.39
Year	1975	1980	1985	1990	1995	1998	1999	2000

Since the ticket price depends on the year, we make the year the independent variable (x) and the ticket price the dependent variable (y). It is important to note here that using the actual year for the x values can create a symbolic expression that contains large numbers and consequently a more difficult model to use. With all of this said, let's standardize our data such that $x = 0$ represents the year 1970. Once this decision has been made, it is a good idea to make your students make an additional row in the table for the x values that correspond to each year, as shown below.

Ticket price (in dollars)	2.05	2.69	3.55	4.23	4.35	4.69	5.08	5.39
Year	1975	1980	1985	1990	1995	1998	1999	2000
x	5	10	15	20	25	28	29	30

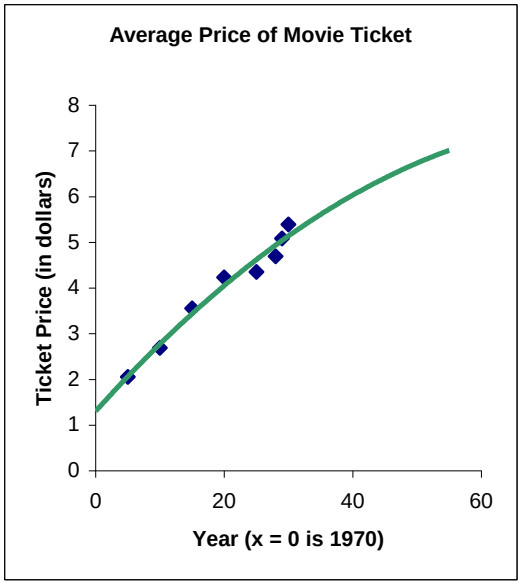
At this point our data is ready for regression. Using a TI-83 calculator or Excel we can generate the following scatter plots and corresponding regression models. Since our question asks about the year 2020, we need to be sure the scatter plots extend beyond $x = 50$. Note that we will not find a model for the Sine function. This is because the data is not showing a periodic trend.



Linear Model

$y = 0.1218x + 1.5364$

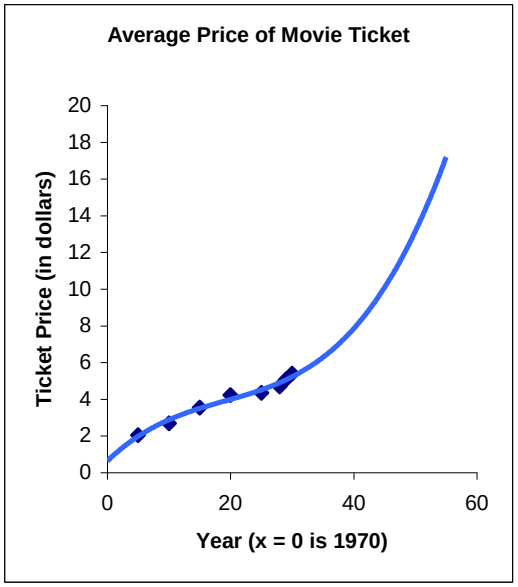
$R^2 = 0.9711$



Quadratic Model

$y = -0.001x^2 + 0.1568x + 1.2982$

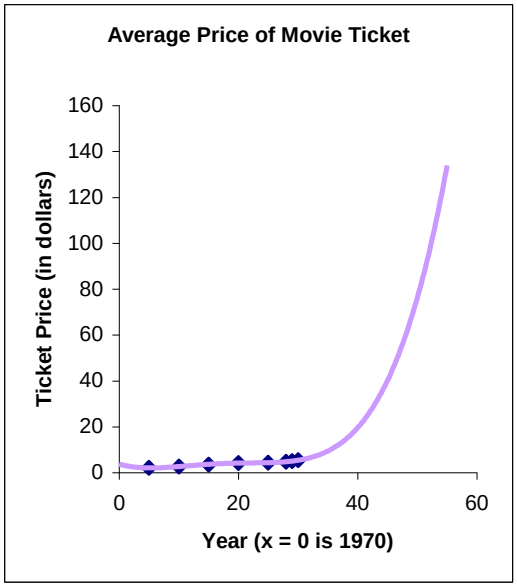
$R^2 = 0.9736$



Cubic Model

$y = 0.0002x^3 - 0.0121x^2 + 0.3248x + 0.629$

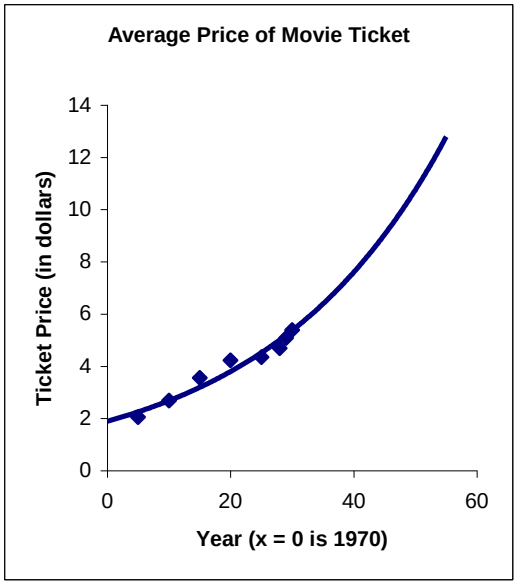
$R^2 = 0.9786$



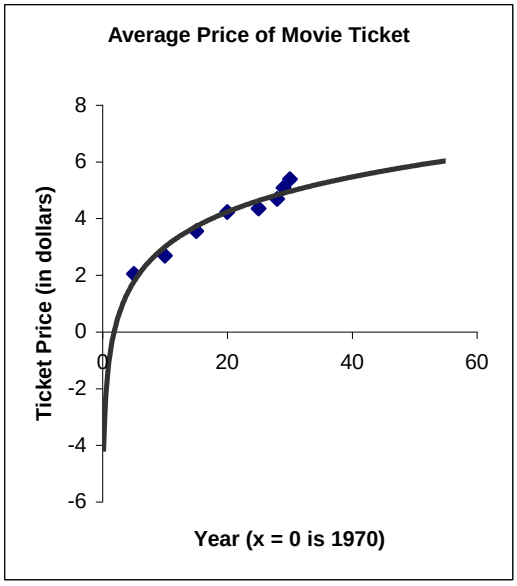
Quartic Model

$y = 0.00007x^4 - 0.0045x^3 + 0.1014x^2 - 0.7403x + 3.7466$

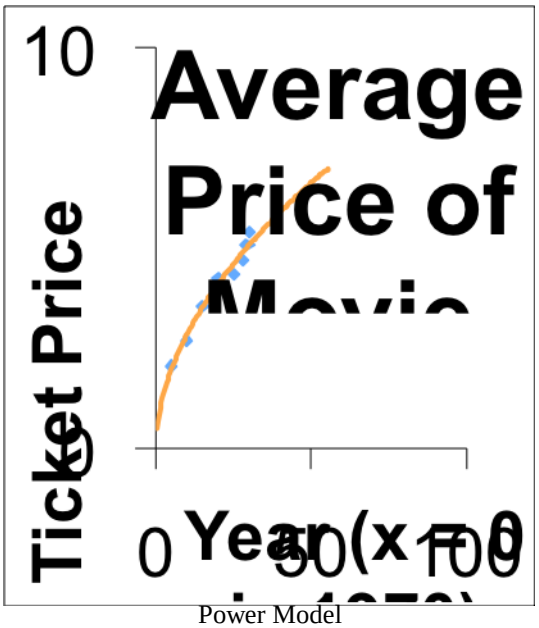
$R2 = 0.9976$



Exponential Model
$y = 1.8961e^{0.0347x}$
$R^2 = 0.9511$



Logarithmic Model
$y = 1.7815\ln(x) - 1.1036$
$R^2 = 0.9442$



Power Model
$y = 0.8555x^{0.5228}$
$R^2 = 0.9813$

In examining the graphs from above, we see that the quartic model has the highest R^2 value, but the graph does not appear to follow the same trend as the data around $x = 50$, so we would eliminate this as a possible model. For similar reasons, we would eliminate the exponential, logarithmic, and cubic models. Thus, we must decide whether to choose the linear, quadratic, or power model. It appears that all models follow the trend of the data even when it is extrapolated to $x = 50$ and their R^2 values show a very strong correlation. Since the power model has a noticeably higher R^2 value, we would choose the power model. We can now predict the average price of a movie ticket by substituting $x = 50$ into $y = 0.8555x^{0.5228}$ and obtain $6.61365 \approx \$6.61$.

You could change this problem by asking your students to choose which polynomial model best predicts the average motion picture price in 2010. For the same reasons listed above, the students should eliminate the cubic and quartic models and therefore choose between the linear or quadratic models. It appears that both models follow the trend of the data even when it is extrapolated to $x = 50$ and their R^2 values differ by only 0.0025. Since there is no strong argument to keep the quadratic model, we would choose the linear model because it is a simpler model than the quadratic. Thus, the students should find that using the linear model, the average price of a movie ticket in 2010 would be approximately $7.63.

What are some important things to consider when constructing questions that involve regression?
There are a number of things to keep in mind when constructing questions that involve regression, especially when it comes to making an exam.

1) Keep your data sets manageable by giving the students the smallest number of data points without distorting the trend of the data.
2) Keep your data sets "real". Search the internet, go to the library, or look in newspapers and magazines and find data sets that are found in the real world. Try to avoid making up data sets.
3) Be sure that your students have some knowledge about or can relate to the data set. If, for example, you found data on the amount of money spent by consumers on the internet from the years 1990 – 2004, some students may not know that the internet was not in mainstream America in the early 1990's and consequently they may have difficulty finding a model that would predict the amount of money spent by consumers in 1989.
4) Often times it is helpful to give your students a few models to select from instead of having them try every type of model on every problem. In the movie ticket example above we could have asked the students to determine which model best fits the data set; the quadratic, linear, logarithmic, or cubic model. This is especially helpful on exams when time is a factor.
5) Implement numerous concepts regarding functions previously learned by your students, like finding the roots, maximum and minimum values, intercepts, and asymptotes. In the movie ticket example from above we could have asked the student to use the linear model to find the year when the average movie ticket price would be $8.