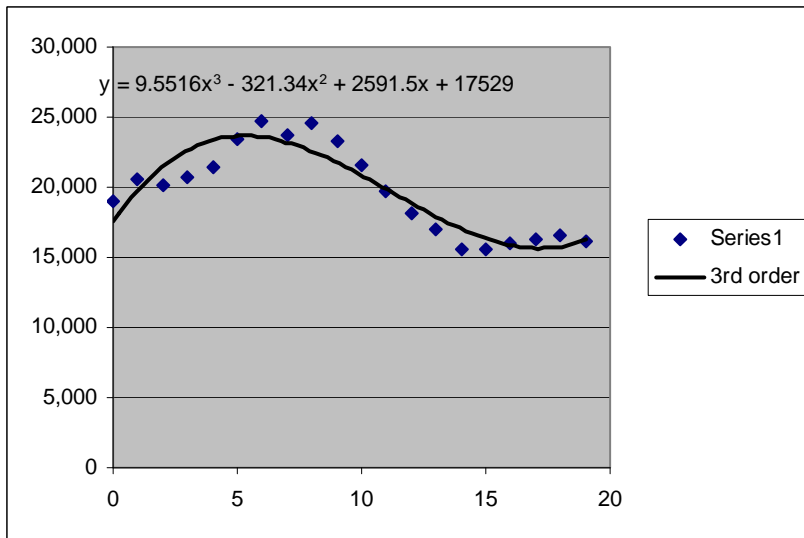


Name: _____

Polynomial Models Studio

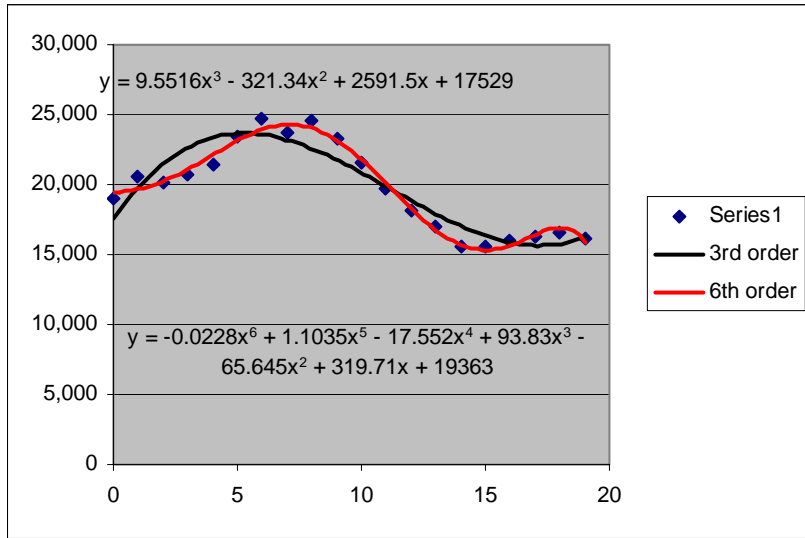
- A. Download the data spreadsheet, open it, and select the tab labeled Murder. This has the FBI Uniform Crime Statistics reports of “Murder and non-negligent manslaughter” in the United States for the years 1985-2004. Column A (Year) is the number of years since 1985 and column B is the number of murders reported that year. Select the data (block A2..B21) and graph it with an XY (Scatter) plot. Looking at the data, we can see several turning points, which is the usual sign that a polynomial fit might be in order. Lines, power curves, and exponentials have no turning points. Quadratics have one turning point and must be symmetric about that turning point.
- B. Click on one of the data points marked in the plot, then right-click to get a pop-up menu of options. Select “Add Trendline ...” This will give you a dialog box with a variety of different types of models to choose from. Select the Polynomial model and set Order to 3. Then click on the Options tab and check the “Display equation on chart” box. Also check the Custom option for Trendline name and give the name “3rd order.” Click OK and a polynomial curve will be drawn that is the “best fit” 3rd degree polynomial curve for the data provided. The equation will also appear on the chart. Note that you can click on the equation and move it around on the chart so that it isn’t obscured by the data points and the trendline. Your graph should now look something like this:



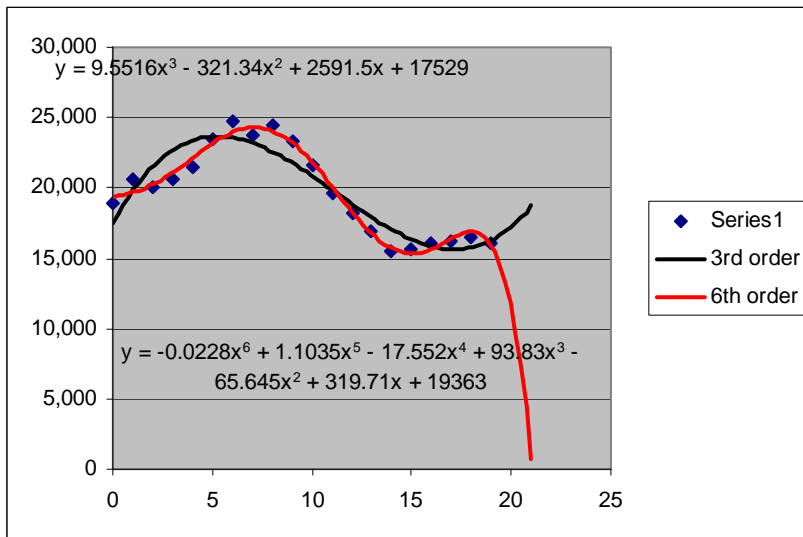
- C. This trendline doesn’t seem to fit the data very well. We might do better if we used a higher order polynomial, which could have more turning points. Click one of the data points and go through the same procedures as in step B again, except that this time set the Order to 6 (and set the name to 6th order of course). Once you have the trendline on the graph, you can right-click on it to get a pop-up menu of options. Select Format Trendline... and you will get the dialog box back. But this time you will have a Patterns tab. If you select the Patterns tab (Colors and Lines tab on a Mac), you will get options to change the color and weight (width) of the

Name: _____

trendline, so it is easier to distinguish the two trendlines. Your graph should now look something like this:



D. This trendline seems to do a much better job. But looks can sometimes be deceiving. A trendline is supposed to show the overall trend, and not necessarily follow all the little bumps of the data. Indeed, if your trendline follows the bumps too much, you end up “modeling the noise” and actually getting a less appropriate model. To get a sense of some of the difficulty, suppose we use both these models to predict how the murder rate will look in 2005 and 2006. Right-click on a trendline to bring up the pop-up menu and select Format Trendline. Click on the Options tab and then set the Forecast to Forward 2 units. Repeat this for the other trendline. Your graph should now look like



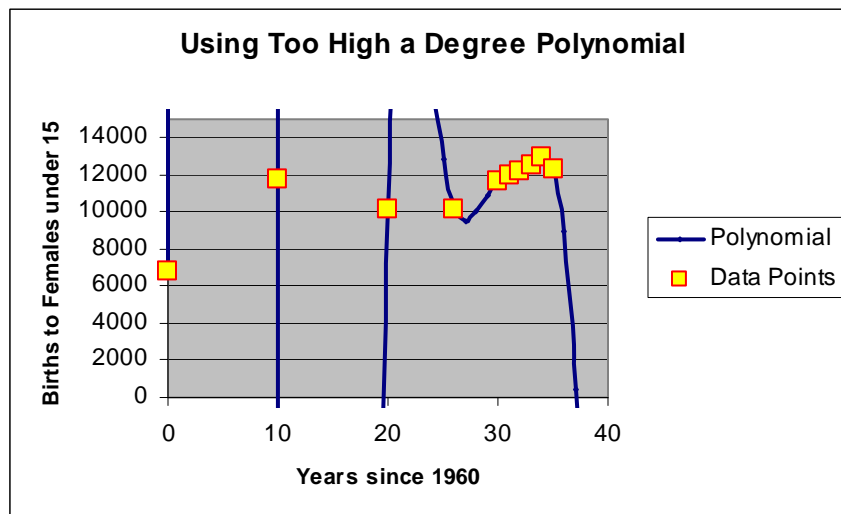
So if the 6th degree polynomial fit is accurate, in 2006 we should have solved the problem of murder in the United States. Unfortunately, we haven't. While final numbers aren't yet available, the FBI's preliminary report says murders went up from 16,137 in 2004 to

Name: _____

16,910 in 2005, which is not all that far off from the 3rd degree model's prediction (which was 17,236). It is typical of polynomial models that in making extrapolations you can have trouble when the model starts rapidly heading off to either positive or negative infinity. Summarizing what we've seen so far, we have the following two points.

- Polynomial models can be useful when the data you are modeling has turning points.
- Polynomials will only forecast accurately for a limited range at the end of your data before they take off to plus or minus infinity.

You can also have trouble, particularly if the data is not evenly spaced, with the model making big curves between data points. All these problems are magnified as you use higher degree approximations (which is one reason why Excel limits you to at most an order 6 approximation). For example, if you click on the "Birth" tab, which contains data from problem 24 in section 4.2 for births to females under 15 in the United States, we could use a 9th degree polynomial which allows for enough flexibility to pass exactly through every data point. But this wouldn't produce a useful model, as the graph below illustrates. The model predicts wild gyrations in the birth rate peaking at over 500,000 in 1968 and falling to -315,000 in 1999, which are obviously absurd values. By the way, the teenage birth rate did fall sharply in the late 90's with a strong turning point just at the end of this data set.



So we want to avoid using a polynomial of too high a degree to avoid problems with wild gyrations like this. In fact, using too high a degree polynomial leads to other problems as well. If your data has a lot of random variation (noise) masking the underlying trend (think of the data we had for tornadoes in the second studio), using a high degree polynomial will tend to build a model which reflects the random noise rather than the underlying trend. We will look at this more below.

Name: _____

When using polynomial models, you have to decide what degree polynomial to use, and as noted above, this process can be tricky. To get a better understanding of how to make this decision, we will start with an artificial situation where we know exactly what the true situation is. The usual assumption in building such a model is that there is some underlying trend to the data. However, the year to year values deviate from the underlying trend due to a variety of random factors, which are called noise. The observed data is the sum of the trend and the noise. Look at the Trend tab on the spreadsheet. The first data column (A) has the x values from 0 to 12. The second column (B) has the underlying trend values, the third column (C) has noise values, and the fourth column (D) has their sum, which will be the observed data (we will worry about column E later). The top graph shows the actual trend line in purple, the best 3rd degree polynomial approximation to the data in black, and the best 6th degree polynomial approximation to the data in red (we will worry about the bottom graph later). The equation for the 3rd degree approximation and its R^2 value (which measures how well it fits the data) is printed at the bottom of the graph while the equation for the 6th degree approximation and its R^2 value is printed at the top of the graph. You will all have slightly different graphs at this point, since the noise values are random.

1. If you press the F9 key on a PC or hold down the Apple key and press = on a Mac, you will cause the spreadsheet to compute a new set of random values for the noise and update all the data and graphs. Fill in the chart below for 6 different sets of random noise values.

Attempt	Which model (3 rd or 6 th) had the better R^2 ?	Which model (3 rd or 6 th) was closer to the actual trendline?
1		
2		
3		
4		
5		
6		

In general, which model fits the data better (has the higher R^2)? Which model fits the trend line better?

Name: _____

This is an example of “fitting the noise.” The data has both trend and noise components and if what you want is to know the underlying trend, you want to be careful that you don’t use too high a degree polynomial that will catch the random ups and downs of the noise at the cost of being less true to the underlying trend. This is the idea of the *principle of parsimony*.

- It is important to choose as low a degree polynomial as reasonable to capture just the trends and avoid “modeling the noise.”

It is this need to be able to look at data and recognize what type of polynomial it looks like that leads us to study graphs of polynomials (and to problems like the online homework this week). If you pick the polynomial of the minimum degree necessary to catch the turning points in the data, you are more likely to find a good approximation to the underlying trend line. Typically, you look at the data and try to decide on the number of turning points, then pick the minimum degree that will allow for that number of turning points. Of course, since you only know the data and not the underlying trend when you work with real data, it can sometimes be difficult to decide exactly how many real turning points there are. Usually you ignore turning points involving just a couple of values and look for broader turns. On the other hand, since a quadratic must be symmetric around its turning point, if you have asymmetric data with a single turning point, you may want to use a cubic rather than a quadratic. Also, if your data seems to have relatively little noise, then it may be safe to use a slightly higher degree polynomial to try to capture smaller bumps in the trend line. If you want to learn more about how to pick the right model, you should take a course or two in statistics.

Another problem with using high degree polynomials is that they are less “robust” than lower degree polynomials. If making a change in a small number of values will cause a large change in the model, then the model is not robust. In practice, it is unfortunately not uncommon for a small number of values to be incorrectly recorded, and hence non-robust models should be avoided. To compare the robustness of different types of models, look at the column E on the spreadsheet. This has exactly the same values as column D, except that there is a (deliberate) error in cell D3, where the value 0 has been put in place of the correct value. The bottom graph shows the best-fit 3rd and 6th degree models for this data with one bad value.

2. As in problem 1, you can use F9 (or Apple =) to recompute the random noise. Look at a minimum of 6 more
Which model changes more when there is a bad value?

Name: _____

Comparing the models in the top graph (where there is no bad value) to the models in the bottom graph which reflect the bad value, *can a bad value at $x=1$ add extra turning points to the model after $x=8$ for the 6th degree model and/or the 3rd degree model (specify which, both, or neither)?*

Which model is more robust?

3. It's time to apply what you've learned to picking a model for real data, where you don't know the actual trend line. Click on the GTA tab. This sheet has the motor vehicle theft data for the last 20 years (where x represents years since 1985). Graph the data.
How many turning points do you see in the data (look for broad turns, not for quick up and downs that result from random variations)?

Based on that number of turning points, what degree polynomial do you want to use to model the data?

Use the spreadsheet to find the best-fit model of the degree you've chosen. *Use the model to predict motor vehicle thefts for 2005.*

Name: _____

4. Repeat problem 3 for the data on Aggravated Assault found under the Assault tab on the spreadsheet.

How many turning points do you see in the data (look for broad turns, not for quick up and downs that result from random variations)?

Based on that number of turning points, what degree polynomial do you want to use to model the data?

Use the spreadsheet to find the best-fit model of the degree you've chosen. *Use the model to predict the number of aggravated assaults for 2005.*

Since there is one turning point in this data, you should have used a quadratic model above. However, that doesn't seem to fit the data particularly well. The data doesn't seem too noisy (though it also doesn't seem too asymmetric). Try a cubic model instead of a quadratic. You should see it appears to fit the data much better. *Use the cubic model to predict the number of aggravated assaults for 2005.*

Note that this leads to a prediction that assaults will rise, while the quadratic model predicts they will fall. In many situations deciding which model is appropriate is not a clear-cut decision, but different models may lead to quite different predictions. That's part of the fun of social science (and business too).

Name: _____

5. In problem 31, part d, of the written homework, you are to compare the **model** values to get the ratio of the juvenile crime rate in 1999 to the juvenile crime rate in 1990. This might seem odd. Since we know the exact values for those years, why would we want to compare model values rather than observed values? To get a sense of why we might want to do this, go back to the Trend tab. Compare which is closer to the trend value (in purple) at $x = 8$, the observed value (shown in blue) or the model value (the black line). Press F9 (or Apple =) to see how the trend, observed and model data compare for 10 different sets of random noise values. *Out of the 10 attempts, how many times is the model value closer to the trend value and how many times is observed value closer at $x=8$?* Note that it is possible for them to be essentially the same distance apart.

Write a short paragraph about the advantages and disadvantages of comparing model values vs. observed values. Note that there are both advantages and disadvantages.