Lesson plan for *Curve Fit*: How well does the curve describe the data?
Time for activity: 30 minutes (not yet tested)

Learning Goals: Students will be able to:

- Explain how the range, uncertainty and number of data points affect correlation coefficient and Chi squared.
- Describe how correlation coefficient and chi squared can be used to indicate how well a curve describes the data relationship.
- Apply understanding of *Curve Fitting* to designing experiments

Advanced optional goals:

- Differentiate between correlation coefficient and chi squared
- Use the equation for Chi squared to explain why a curve fit with a value near one describes the data well.
- Given a set of data without the χ^2 and r^2 displayed, (**No curve fit** selected to get a screen capture), predict which data will have a better curve fit.

Background:

I will use this in my College Physics course at Evergreen high school. The class covers first semester algebra based physics, but there is an emphasis on inquiry and lab design based on data evaluation. My students use Excel and TI graphing calculators to curve fit their data. We do not determine uncertainty in their labs. In the text, there is a short introduction and a few problems determining uncertainty. I plan to use this activity during their semester projects where they design an experiment and then try to improve the design. The improvement is generally measured by an improved r^2. I want to introduce them to Chi squared so that they will see that there are more complex tools to better describe science than r^2, but I do not expect them to digest the calculation.

I do not intend to address the advanced learning goals in the student directions or in my course, but I included them for other teachers to use.

***Curve Fitting* Introduction:**

The bars on the data points are labeled "error bars", but I decided to call them "uncertainty bars." Technically, the half-length of the error bar is equal to one standard deviation.

You can zoom in on any Flash sim to show something well by right clicking. Clicking on the Help shows many of the features. The bucket of points can be dragged to any location; I used this feature as I made screen captures. In my lesson, I thought it would be better for the students to start their exploration using the default uncertainty bars to help constrain the variables that they are investigating.

Lesson:

I will go over the learning goals and remind the students what "range" means and "known shape of curve". *Range is the spread of the x values. Range as it applies to experimental design, is the spread of the independent variable. By" shape", I mean Linear- a line, quadratic-parabolic, etc*

Lesson plan for *Curve Fit*: How well does the curve describe the data?
Time for activity: 30 minutes (not yet tested)

Also, I will point out that the directions encourage them to constrain the variables that they are using to investigate curve fitting by not varying the uncertainty bars until question 5. *Just like any good experiment.*

My students work in pairs with a printed copy of the student instructions for guidance. The activity should take my college physics students about 30 minutes. I have not used this lesson in class yet.

My notes about the questions:

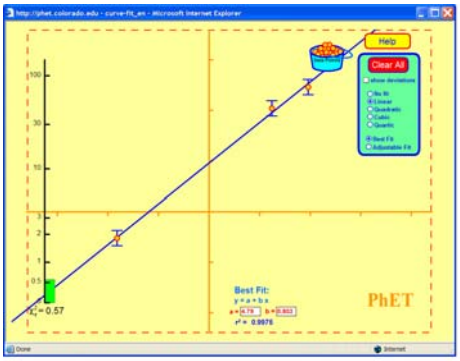
1: For each of the four types of curve fit, what is the minimum number of points *(using the default uncertainty bars)* to get a good correlation and a curve that demonstrates the known shape for the curve?

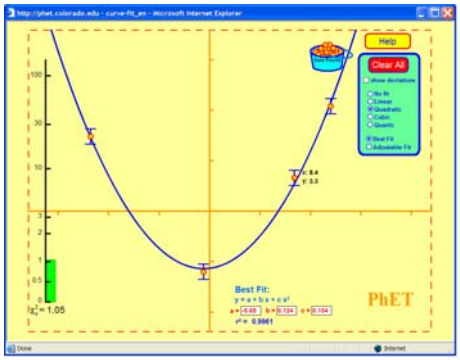
2 points give a perfect correlation for any curve, but the appropriate shape of the curve will not be see until you have one above the order. The correlation will still be one.

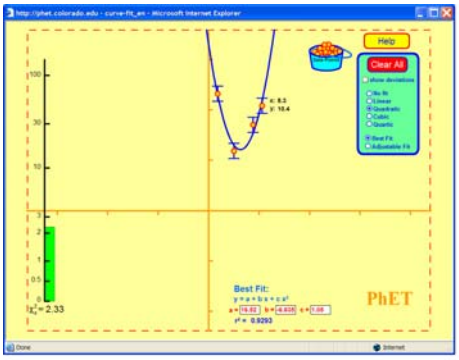
2: For each of the four types of curve fit, what is the minimum number of points*(using the default uncertainty bars)* to get in the green zone with a good correlation?

I can get in the green zone for linear with only three points if I have a wide range for the data.

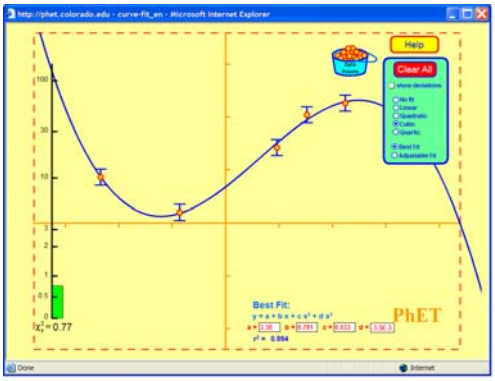
Quad 4 points. I started with wide range because it seemed reasonable, but I made another trial that worked with small range.

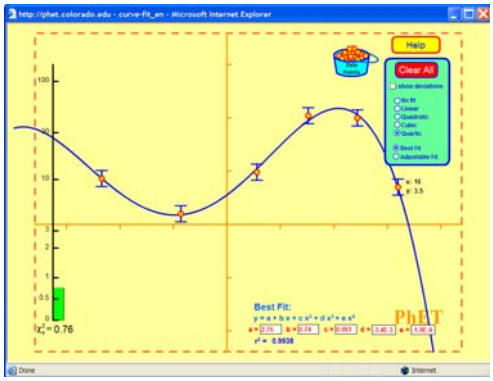






I made a good cubic easily with wide range with 5 points. And quartic with 6.





Lesson plan for *Curve Fit*: How well does the curve describe the data?
Time for activity: 30 minutes (not yet tested)

3: How can you desensitize the coefficient to change? In other words, if you collect data that would not fit on the line, under what conditions would r^2 stay nearly the same? Is χ^2 desensitized the same way?
Use more points with a wide range but χ^2 is still very sensitive with the default uncertainty bars; you would need to use larger uncertainty to desensitize χ^2 .

4: What is the relationship between the χ^2 red zone and the correlation coefficient?
There is not a mathematical relationship, but in general, if you are looking at the χ^2 and make changes to the data that increases the red zone, r^2 decreases. ie χ^2 getting larger when it is greater than one decreases r^2

5: Given two situations of data points, identify which is better and explain how the data variations changed the quality of the fit. (May need to add text boxes with r^2 if the values on question 5 are not readable)
If all the points lie on the curve there will be a poor χ^2 no matter what. But if the points do not lie on the curve, then increasing the uncertainty bars can improve χ^2 . I am hoping that the students will be able to discover this relationship without using the equation. [Using the equation: $y(x_i)- y_i$ and σ (or Δy), the uncertainty, are more similar so more terms in the sum are closer to 1, if there are no deviations (points all lie on line), then the sum will be zero.]

a. Better fit on left. Same data points, so same r^2 but different uncertainty bars. Since the bars touch the curve the χ^2 will improve.
b. Better fit on right. The r^2 is improved because more data points lie near the line and χ^2 improved because more uncertainty bars touch the curve.
c. Better fit on right. The r^2 is improved because more data points lie near the line and χ^2 improved because more uncertainty bars touch the curve.

6: How does your of understanding χ^2 and r^2 help you with your experimental design?
We don't calculate data uncertainty in our labs, but the students should recognize that range should be maximized as well as the number of data points. Also that they can get a great r^2, but not really have determined an appropriate curve; it is important for them to consider how known physics can be used to decide on the order of the equation. Sophisticated experimental design would require determining uncertainty.

Advanced Learning goals:
1. Differentiate between correlation coefficient and chi squared
On a very basic level, both include deviation from the predicted y value, but Chi squared includes uncertainty of the data. The correlation coefficient depends heavily on the deviation. When the uncertainty and the deviations are similar, then Chi squared is near one and indicates a high quality of fit.

Lesson plan for *Curve Fit*: How well does the curve describe the data?
Time for activity: 30 minutes (not yet tested)

2. Use the equation for Chi squared to explain why a curve fit with a value near one describes the data well.

The reduced chi squared statistic is

$$\chi_r^2 = \frac{1}{N-f}\sum_i \frac{[y(x_i)-y_i]^2}{\sigma_i^2}$$

In an experiment, one measures a series of x-values: x_1 , x_2, x_3..x_i .. x_N. and corresponding y-values: y_1 , y_2, y_3... The i^{th} measurement is the pair (x_i , y_i). We assume that the uncertainties in the x-values are negligibly small; the uncertainty in y_i , due to uncertainties in the measurement, is given by σ_i (or Δy_i). The best-fit curve to the data is given by a function y = y(x) , so the quantity $y(x_i)$ is the predicted value of y (predicted by the best fit curve) at the x-value x_i. The difference $[y(x_i) - y_i]$ is the discrepancy between the measured y_i and the predicted value $y(x_i)$. This difference is called the *deviation*. For one point, we expect that the deviation is roughly equal to the uncertainty σ_i, so the ratio $[y(x_i) - y_i]/\sigma_i$ is about 1 and so also is the ratio-squared: $[y(x_i) - y_i]^2/\sigma_i^2$. If all N points were like this (ratio-squared $\approx$ 1) and you added up the N ratios for the N data points, and then divided by N : you would get $\frac{1}{N}\sum_i\left(\frac{y(x_i)-y_i}{\sigma_i}\right)^2 \approx 1$.

This situation is complicated by the fact that the function y = y(x) is not the *true* curve, but instead, it is the *best-fit* curve to the data. If there is very little data, the best fit curve will always fit the data well (it's the *best fit*, after all), regardless of the errors in the data. Because of this, the ratios $[y(x_i) - y_i]/\sigma_i$ are generally smaller than they would be if y(x) were the *true* curve. For instance, if there are only 3 data points (N = 3), and we perform a 3-parameter (quadratic) polynomial fit (f = 3), then the fit will always pass through the 3 points exactly, all the deviations will be zero, and the quantity $\sum_i\left(\frac{y(x_i)-y_i}{\sigma_i}\right)^2$ will be exactly zero. To compensate for the fact that y = y(x) is not the true curve, we must perform a fix: instead of dividing by N we divide by (N – f) where f is the number of degrees of freedom in the fit. The resulting quantity $\chi_r^2 = \frac{1}{(N-f)}\sum_i\left(\frac{y(x_i)-y_i}{\sigma_i}\right)^2$ is called the *reduce chi-squared*. Because we divide by (N-f) instead of N, the reduced chi-squared is undefined unless N > f. For the case N>f, it turns out that the reduced chi-square is about one, as we expect.

Lesson plan for *Curve Fit*: How well does the curve describe the data?
Time for activity: 30 minutes (not yet tested)