ERROR ANALYSIS 2: A USER'S GUIDE TO LEAST-SQUARES FITTING

INTRODUCTION

This activity is a "user's guide" to least-squares fitting and to determining the goodness of your fits. It doesn't derive many results. There are good textbooks listed in the references.

LEARNING GOALS

At the end of the activity you will be able to:

1. Explain why we minimize the sum of squares to get the best fit
2. Carry out a least-squares minimization graphically
3. Plot residuals to visually inspect the goodness of a fit
4. Be able to interpret the uncertainty in fit parameters that Mathematica's fit routines output
5. Be able to compute χ^2 for a fit and use it to determine if a fit is "good"

WHY DO WE MINIMIZE THE SUM OF SQUARES?

Question: Why do we call it "least-squares" fitting?

Answer: Because the best fit is determined by minimizing the weighted sum of squares of the deviation between the data and the fit. Properly speaking this "sum of squares" is called "chi-squared" and is given by

$$\chi^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(y_i - y(x_i, a, b, c, \dots)\right)^2 \tag{1}$$

where there are N data points (x_i, y_i), and the fit function is given by $y(x_i, a, b, c, \dots)$ where *a, b, etc.* are the fit parameters.

Question: What assumptions are made for the method to be valid?

Answer: The two assumptions are

(1) **Gaussian distributed.** The random fluctuations in each data point y_i are Gaussian distributed with standard deviation σ_i.

(2) **Uncorrelated.** The random fluctuations in any one data point are uncorrelated with those in another data point.

Question: Why does minimizing the sum of squares give us the best fit?

Answer: Given those two assumptions, the fit that minimizes the sum of squares is the ***most likely*** function to produce the observed data. This can be proven using a little calculus and probability. A more detailed explanation is found in Taylor's *Introduction to Error Analysis* Sec. 5.5 "Justification of the Mean as Best Estimate" or Bevington and Robinson's *Data Reduction* Sec. 4.1 "Method of Least-Squares."

MINIMIZING χ^2 GRAPHICALLY:

Question 1

Graphically minimizing χ^2

You will rarely minimize χ^2 graphically in a lab. However, this exercise will help you better understand what the fitting routines in Mathematica actually do to find the best fit.

a. **Import and plot** the data set from the website. It was generated by inserting a razor blade into the x columns is microcrometer (razor) position in meters. Y column is photodetector voltage in Volts.
   - `Import["http://www.colorado.edu/physics/phys3340/phys3340_sp12/sample_data/profile_data_without_errors_activity_2.csv"]`

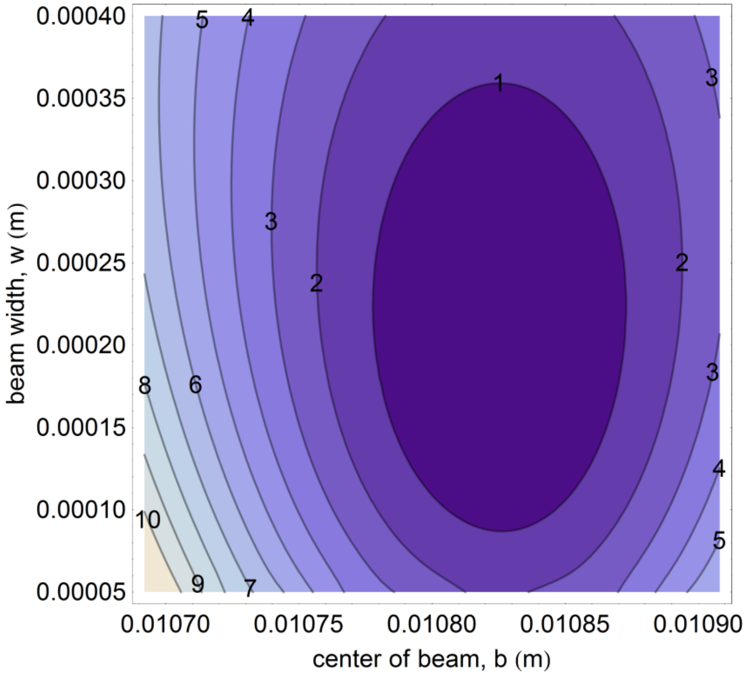
b. **Define the same fit function** many of you used in the Gaussian laser beams lab:

$$y(x,a,b,c,w) = a\,\mathrm{Erf}\left(\frac{\sqrt{2}}{w}(x-b)\right) + c$$

c. **Reduce the fit to two free parameters. This step is only necessary because you it is hard to visualize more than 3 dimensions.** Assume $a_{fit} = (V_{max} - V_{min})/2 = 1.4375$ and $c_{fit} = (V_{max} + V_{min})/2 = 1.45195$. These were determined by averaging the first 6 data points to get V_{min} and the last 5 to get V_{max}.

d. **Use Equation 1 to write an expression for χ^2** in terms of your w and c parameters, and the x (position) data, y (voltage) data. Since you don't have any estimate for the uncertainties σ_i, do what Mathematica does, and assume they are all unity so $\sigma_i = 1$.

e. **Make a contour plot** of $\chi^2(w,c)$ and tweak the plot range until you see the minimum. Just like with `NonlinearModelFit`, it will help to have a good initial guess for your fit parameters. You can iteratively improve the plot range your plot to zoom in on the parameter values that minimize χ^2, you should get a plot kind of like:



-

f. **Graphically determine the best fit parameters** to 3 significant digits.

g. **Compare with the best fit result from `NonlinearModelFit`** (allow all 4 parameters to vary). Do the fits agree for those three digits of precision?

UNCERTAINTY IN THE FIT PARAMETERS

Question: Where does the uncertainty in the fit parameters come from?.

Answer: The optimal fit parameters depend on the data points (x_i, y_i). The uncertainty σ_i in the y_i means there is a propagated uncertainty in the calculation of the fit parameters. The error propagation calculation is explained in detail in the references, especially Bevington and Robinson.

Question: How does Mathematica calculate the uncertainty in the fit parameters when no error estimate for the σ_i is provided?

Answer: Mathematica (and other programs) estimate the uncertainty in the data σ_y^2 using the "residuals" of the best fit:

$$\sigma_y^2 = \frac{1}{N-n}\sum_{i=1}^{N}\left(y_i - y(x_i, a_0, b_0, c_0, \dots)\right)^2 \tag{2}$$

where there are N data points and n best fit parameters $a_0, b_0, c_0, \dots$. It is very similar to how you would estimate the standard deviation of a repeated measurement, which for comparison's sake is given by

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{3}$$

Question 2	Estimating the uncertainty in the data
	a. Use Equation 2 and your best fit parameters to estimate σ_y^2, the random error of each data point given by your data. b. Compare your result with Mathematica's `NonlinearModelFit`, which can also output this estimate of the random error. If `nlm` is the `NonlinearModelFit` output, the estimate of σ_y^2 is given by `nlm[“EstimatedVariance”]` c. Do the estimates agree? Why or why not?

GOODNESS OF FIT

This section covers two ways to analyze if a fit is good.

1. Plotting the residuals
2. Doing a χ^2 test

PLOTTING THE FIT RESIDUALS

The first step is to look at the residuals. The residuals r_i are defined to be the difference between the data and the fit.

$$r_i = y_i - y(x_i, a, b, c, \dots)$$

Question 3	**Making a plot of the residuals and interpreting it** a. Make a `ListPlot` of the residuals. If `nlm` is the `NonlinearModelFit` output, the list of residuals is given by `nlm[“FitResiduals”]` b. Since we didn’t provide any estimates of the uncertainties, Mathematica assumed the uncertainty of every point is the same. Based on the plot of residuals, was this a good assumption? c. Do the residuals look randomly scattered about zero or do you notice any systematic error sources? d. Is the distribution of residuals Gaussian? e. What is the most likely source of the large uncertainty as the beam is cut near the center of the beam?

“CHI BY EYE” – EYEBALLING THE GOODNESS OF FIT

Question: If I have a good fit, should every data point lie within an error bar?

Answer: No. Most should, but we wouldn’t expect every data point to lie within an error bar. If the uncertainty is Gaussian distributed with a standard deviation σ_i for each data point y_i, then we expect roughly 68% of the data points to lie within their error bar. This is because 68% of the probability in a Gaussian distribution lies within one standard deviation of the mean.

χ^2 AND χ^2_{red} FOR TESTING THE “GOODNESS OF A FIT”

This section answers the question “What should χ^2 be for a good fit?”

Suppose the only uncertainty in the data is statistical (i.e., random) error, with a known standard deviation σ_i, then the on average each term in the sum is

$$\frac{1}{\sigma_i^2}\left(y_i - y(x_i, a, b, c, \dots)\right)^2 \approx 1 \tag{4}$$

and the full χ^2 sum of squares is approximately

$$\chi^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left(y_i - y(x_i, a, b, c, \ldots)\right)^2 \approx N - n \tag{5}$$

So a good fit has

$$\chi^2_{red} \equiv \frac{\chi^2}{N - n} \approx 1 \tag{6}$$

Question 4	**Fact: In order to goodness of fit test, you must first estimate the uncertainties.** How would you briefly explain the reason for this in your own words?
Question 5	**Choosing a strategy to estimate the uncertainty** Considering your answers to question 4, especially 4.d, which method would give you the best estimate of the uncertainty in each data point? • Eyeballing the fluctuations in each data point. • Taking N measurements at each razor position and then going onto the next position. • Taking the entire data set N times

WEIGHTED BEST FITS IN MATHEMATICA

When you have estimated the uncertainty σ_i of each data point y_i you would like to use this information when fitting to correctly evaluate the χ^2 expression in Equation 1. The points with high uncertainty contribute less information when choosing the best fit parameters. If you have a list of uncertainties

```
σlist = { σ1, σ2, σ3,…}
```

then the weights for the fit are

weightslist = 1/ σlist^{2} = {1/ σ1^{2}, 1/ σ2^{2}, …}

Add the `Weights->weightslist` option to the `LinearModelFit` or `NonlinearModelFit`. For example,

```
NonlinearModelFit[data,fit[x,a,b,c],{a,b,c},x, Weights->weightslist]
```

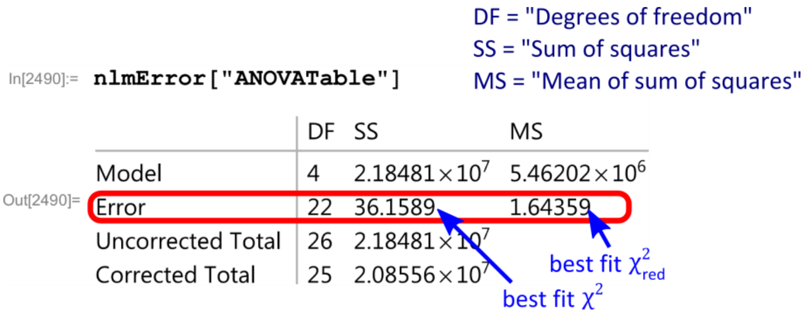
Question 6

a. Import the data set for a beam width measurement with uncertainties from the website. The first column is razor position in meters, the second column is photodetector output voltage, and the third column is the uncertainty in each measurement.

-
  ```
  Import["http://www.colorado.edu/physics/phys3340/
  phys3340_sp12/sample_data/
  profile_data_without_errors_activity_2.csv"]
  ```

b. Do a weighted fit using the same fit function as in question 1. Use the uncertainty estimates in the third column to calculate the weights.

c. Calculate χ^2. You can obtain χ^2 from the fit returned by Mathematica. Supposing the fit was called `nlmError`, use `nlmError["ANOVATable"]`. For the curious, ANOVA stands for ANalysis Of VAriance.



DF = "Degrees of freedom"
SS = "Sum of squares"
MS = "Mean of sum of squares"

```
In[2490]:= nlmError["ANOVATable"]
```

Out[2490]=

	DF	SS	MS
Model	4	2.18481×10^7	5.46202×10^6
Error	22	36.1589	1.64359
Uncorrected Total	26	2.18481×10^7	
Corrected Total	25	2.08556×10^7	

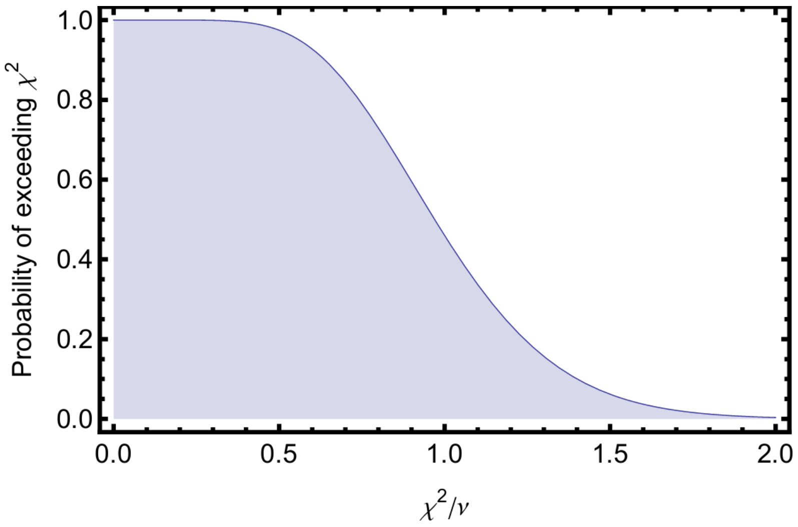
best fit χ^2_{red}

best fit χ^2

d. How close is the reduced chi-squared to 1?

e. **The "chi-squared test"**. This part helps us understand if the value of χ^2 is statistically likely or not. The following graph gives the probability of exceeding a particular value of χ^2 for $\nu = N - n = 22$ degrees of freedom. It can be calculated using the Cumulative Density Function (CDF) for the chi-squared distribution. Use the graph to estimate the likelihood this value of χ^2 occurred by chance.

```
ν = 22; (*Degrees of freedom*)
Plot[1 - CDF[ChiSquareDistribution[ν], x* ν, {x, 0, 2}]
```



WHY IS IT OFTEN BAD TO OVERESTIMATE UNCERTAINTIES?

Question 7	Why can overestimating the uncertainty make your fit appear good (i.e., $\frac{\chi^2}{N-n} \approx 1$)?

Overestimating the uncertainties makes the fit seem good (according to a χ^2 test), even when it might be obviously a bad fit. It is best to do the χ^2 test using an honest estimate of your uncertainties. If the χ^2 is larger than expected $(\chi^2 > N - n)$, then you should consider both the possibility of systematic error sources and the quality of your estimates of the uncertainties. On the other hand, if the χ^2 test is good $(\chi^2 \approx N - n)$, then it shows you have a good handle on the model of your system, and your sources of uncertainty.

REFERENCES

1. Taylor, J. R. (1997). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* (p. 327). University Science Books. This is the standard undergraduate text for measurement and uncertainty.
2. Bevington, P. R., & Robinson, K. D. (2003). *Data Reduction and Error Analysis for the Physical Sciences* Third Edition (3rd ed.). New York: McGraw-Hill. Great for advanced undergrad error analysis. Professional physicists use it too.