Fitting to a Model

How can we quantitatively determine if a model describes our data?

Is this cubic fit America’s next top model?

No. No it’s not.
Learning objectives

After this lab, you should be able to...

- Understand the meaning of reduced $\chi^2$
- Calculate the reduced $\chi^2$ for a given model and data
- Evaluate if a fit is good or bad - this is, determine if it is a good description of the data
- Use Python to calculate best-fit values for a model given a set of data
- Fit real data to draw conclusions!
What do scientists do??
(I ask myself every day…)

Does “doing science” look like this?

Or does it look like this?
Good science depends on both!!
In the scientific community, progress is made by two separate, but equally important groups:

The theorists, who devise novel theories...

But how do these two branches of science come together? *How do we quantitatively test predictions?* This is the story:

...and the experimentalists, who test their predictions*

* Of course, there are some scientists who do both theory and experiment.
Intro to:

Curve Fitting

- How can we quantitatively say if a fit is "good" or "bad"?

CURVE FITTING is a quantitative technique to determine if measurements are consistent with a chosen model.

Below, we see a set of measured data points (in blue) and a mathematical model that describes those points (in red). By eye, it is pretty obvious that the model on the left plot does the best job of describing the data, and the model on the right does the worst job. But what if we had two models that seemed to describe the data equally well? How can we be quantitative about this?
CURVE FITTING is a quantitative technique to determine if measurements are consistent with a chosen model.

Below, we see a set of measured data points (in blue) and a mathematical model that describes those points (in red). By eye, it is pretty obvious that the model on the left plot does the best job of describing the data, and the model on the right does the worst job. But what if we had two models that seemed to describe the data equally well? **How can we be quantitative about this?**

Picking the right fit isn’t always obvious…..
In your group, discuss ways you could calculate the “goodness” of a fit. Which of the following pieces of information do you think you would need?

(a) Your measured data (black markers on the plot to the right)

(b) The uncertainties on your measured data (black error bars on the data shown at right)

(c) Your model’s prediction of what you “should” measure (orange points shown right)

(d) The number of data points

(e) The number of parameters in your model

(f) The derivative of your model

(g) The integral of your model
Introducing... Reduced $\chi^2$

- Motivating the reduced $\chi^2$ statistic.

$\chi^2 = \frac{1}{N-d} \sum_{i=1}^{N} \frac{(Y_{\text{measured},i} - Y_{\text{model},i})^2}{\sigma_i^2}$

- Data points further from the model contribute more!
- Positive and negative variations have the same contribution! (And don’t cancel out!)
- Penalty for having fewer data points!
- Points with smaller uncertainties have a larger contribution!

Chi-Square Checklist
- $X$
- $Y$
- $\sigma$
- Model Predictions
Subskill 2

- Using reduced $\chi^2$

The two plots below show the same data (in black) with two different models - one in green (left), and one in purple (right).

(a) Which of the models (left or right) would give a larger value of chi squared for the data shown?
   (i) Left  
   (ii) Right

(b) By eye, which would you say is a better fit, the model on the left or the model on the right?
   (i) Left  
   (ii) Right

(c) What signifies a better fit - a higher value of chi squared, or a lower value of chi squared?
   (i) Higher  
   (ii) Lower

$$\chi^2 = \frac{1}{N - d} \sum_{i=1}^{N} \left( \frac{y_{\text{measured},i} - y_{\text{model},i}}{\sigma_i^2} \right)^2$$
Subskill 3: What’s a “good” $\chi^2$?

- Develop an intuition for $\chi^2$
- Investigate the limits of $\chi^2$ fitting
- Why does our choice of model matter?

Open the following PHET in your browser (https://phet.colorado.edu/sims/html/curve-fitting/latest/curve-fitting_en.html), and play around a bit. Drag four data points from the bin onto the plot, like so:

Plot the best-fit line for the data. What's the $\chi^2$? Try plotting the best-fit quadratic, and then the best-fit cubic function. What happens to $\chi^2$ for each fit function? Why? Do you think this is physical?

Try adjusting the error bars on each plot. What happens to your fit when you do so?

Try fitting to a quadratic and then a cubic - what happens to your fit and the value of $\chi^2$?

How does reducing the errors on the data points change the $\chi^2$?

Find the best model and the corresponding chi2 value - be prepared to discuss this in the main room.
Picking a physical model

- Why even bother?

As we saw on the PHET, just because a fit passes through your data doesn’t mean it’s the right model! Physics is about deriving quantitative predictions from known principles - the model you use in your fit should also be derived from known physics laws.

For instance, if you measured the force exerted by a spring as function of its displacement from equilibrium, you would expect a linear relationship. Fitting to a higher-order polynomial might give you a lower chi2 value, but doing so isn’t physically meaningful.

Below is a sillier example: what physically motivated model would you expect in this example? Discuss amongst your group:

a) A linear function
b) An exponential function
c) A polynomial function
d) A step function
Beware of Overfitting!

- Just because your curve passes through all your points doesn’t mean it’s right!
χ² fitting in Python is pretty straightforward, but there are some little things that can be tricky. Let’s do our first curve fit as a group - load the notebook below, make a copy, and follow along!

Make a copy of this notebook and follow along!
Using the `mycurvefit()` function

Data:
- **XX**: Independent Variable
- **YY**: Dependent Variable
- **sigma**: Measurement / Modelled Error

**Func (eg: Modelling function)**

\[ y = mx + c \]

**mycurvefit**

- **fitparams**: Best Fit Parameters:
  - \( P1 = 4.74 \pm 0.24 \)
  - \( P2 = 1.82 \pm 0.07 \)

**fiterrs**: Fit Metrics:
- Degrees of freedom: 3
- Chi Squared = 2.0
- Reduced Chi Squared = 0.7
Let’s get some more practice! Open the below notebook, and use mycurvefit() to find the best-fit curves for three other data sets:

Make a copy of this notebook and try it!
Activity 3: Let’s model a pandemic

A certain data set has been in the news quite a lot lately… successfully modelling the number of new COVID-19 cases as a function of time is crucial for enacting effective policy!

In this activity, we will see if we can fit how the number of COVID-19 cases changes with time in a few different countries!

However, we first need to pick a physically-motivated model…
A Mathematical Model of a Pandemic

All models are wrong, but some are useful.
~ George Box (attrib.)

The Exponential Growth Curve

\[ n(t) = n(0)R_0^{\frac{t}{\tau}} \]

Assumptions:
- If \( n_t \) people contracted the disease on day \( t \), they infect exactly \( R_0 n_t \) people after \( \tau \) days.
- No limitations in the spread of the disease are considered i.e. an infected person will always infect \( R_0 \) more people, irrespective of when they contracted the disease.
- We assume that only the new cases detected on a day \( t \), will contribute to the \( R_0 n_t \) infections \( \tau \) days hence.

What are we modelling?
New cases daily (Y) over time (X)
\( \sigma \) : Check out Subskill 2!

Chi-Square Checklist
☑ \( \chi \)
☑ \( Y \)
☑ \( \sigma \)
☑ Model Predictions
Now that we know what model to expect, let’s try fitting some real COVID-19 data! Try it below:

Make a copy of this notebook and try it!
As always, we care about your feedback! Please take the Mid-Semester Lab Survey:

https://canvas.harvard.edu/courses/76554/quizzes/187631