

Biological Chemistry Laboratory
Biology 3515/Chemistry 3515
Spring 2017

Lecture 6

Curve Fitting, Part I

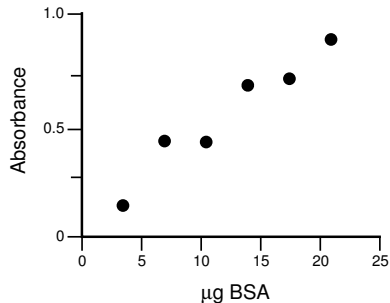
26 January 2017

©David P. Goldenberg

University of Utah

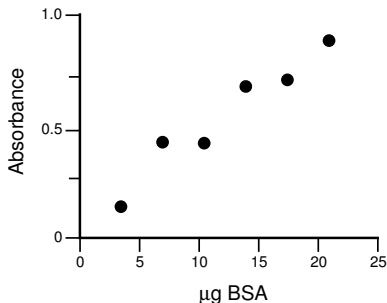
goldenberg@biology.utah.edu

The Curve-Fitting Problem



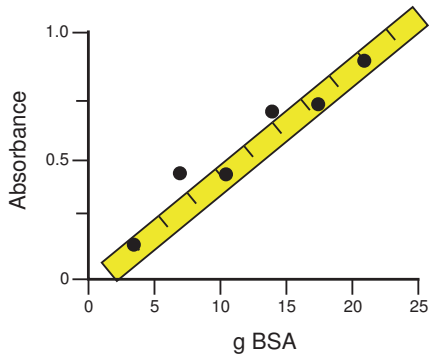
- How do we find the equation of the line (or other function) that best “fits” the experimental data?
- What assumptions do we make when fitting data to a function?
- Reminder: Graph axes should be clearly labeled and specify the correct units!

The Curve-Fitting Problem



- Two general kinds of situations:
 - We have a theoretical model that predicts a particular function, and we want to both test the model and estimate parameters that define the model.
 - We don't have a particular model in mind, but there is an empirical relationship that fits the data, and we want to estimate the parameters.
- Which category does the Bradford calibration fall into?

The Ruler Method



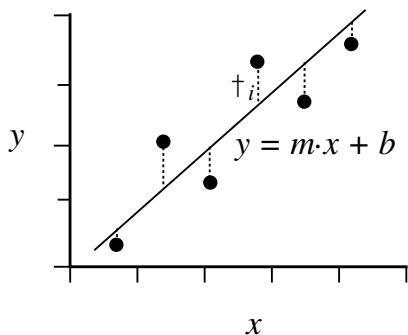
■ Advantages?

- It's easy and inexpensive!

■ Disadvantages?

- It's subjective. Different individuals will get different results.
The more scatter, the bigger the problem!
- Pretty much limited to the straight-line function.
- Doesn't provide a measure of how well the data fit the function.

The Method of Least Squares



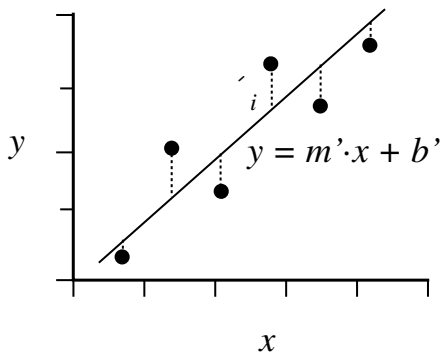
- Key Assumption: The experimental values of y are determined by a linear (or other) function of x and random error in the measurements.

$$y_i = m \cdot x_i + b + \epsilon_i$$

m and b are the “true” values of the parameters, and ϵ_i is the error in each measurement.

Chose an Arbitrary Line Given by the Equation:

$$y = m' \cdot x + b'$$

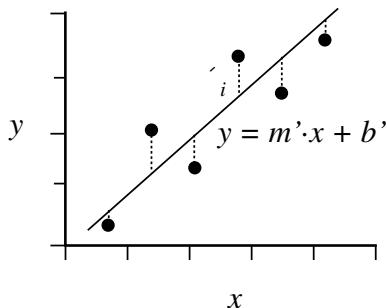


- The observed values of y are then described by the equation:

$$y_i = m' \cdot x + b' + \chi_i$$

- χ_i is referred to as the “residual” for each point, and is distinct from the error, ϵ_i . (We never really know the values of ϵ_i .)
- We want to choose m' and b' so that they represent the best estimates

χ^2 : The Sum of the Squares of the Residuals

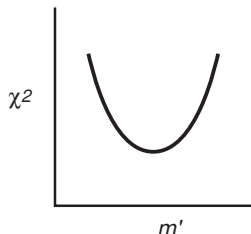
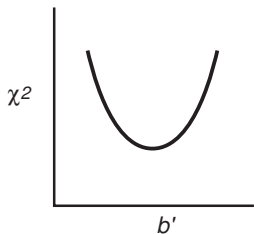


$$\begin{aligned}\chi^2 &= \sum \chi_i^2 \\ &= \sum (y_i - (m' \cdot x_i + b'))^2\end{aligned}$$

- Adjust m' and b' to minimize the value of χ^2 for the particular values of x_i and y_i in the experimental data set.
- Why are the residuals squared?

Minimization of χ^2

- χ^2 is a function of both b' and m'



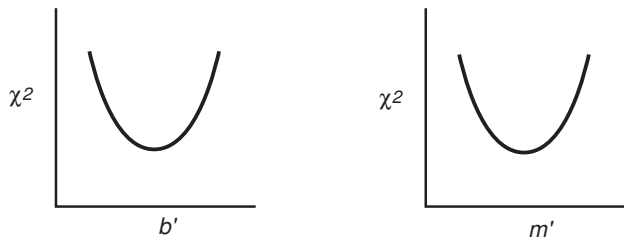
- How do we find values of b' and m' that minimize χ^2 ?
- Take derivatives of χ^2 with respect to b' and m' and set them to zero.

$$\frac{d}{db'} \sum (y_i - (m' \cdot x_i + b'))^2 = 0$$

$$\frac{d}{dm'} \sum (y_i - (m' \cdot x_i + b'))^2 = 0$$

Two equations, with unknowns b' and m' .

Minimization of χ^2



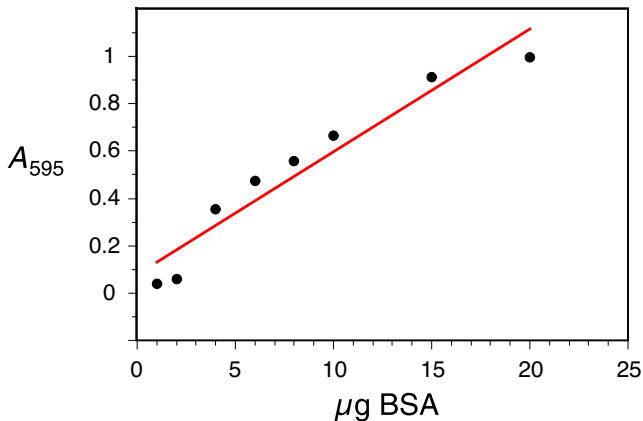
- For a linear function, there is a (relatively) simple solution to the two equations:

$$b' = \frac{\sum x_i^2 \sum y_i^2 - \sum x_i \sum x_i y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

$$m' = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - (\sum x_i)^2}$$

N is the number of experimental x, y pairs.

A Linear Least-squares Fit to Bradford Calibration Data



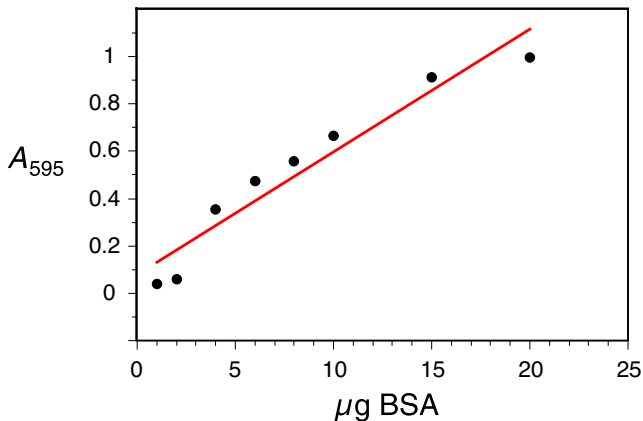
- The estimated parameters for $y = mx + b$:

$$m = 0.052 \pm 0.006$$

$$b = 0.08 \pm 0.06$$

- The uncertainties are analogous to the standard error of the mean.
- What are the units for the parameters?
- How “good” is the fit?

How Do We Judge the Goodness of Fit?

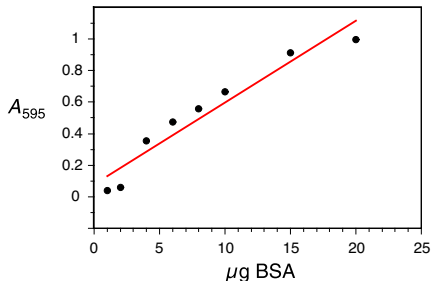


- The final, minimized, value of χ^2 .
For this fit:

$$\chi^2 = 0.062$$

- Can use this value to compare this fit to fits of the same data to other functions.
- But, actual value does not have a clear meaning.
- Adding more measurements will almost always increase χ^2 .

The Coefficient of Determination, R^2



- What fraction of the total variation of y -values is accounted for by the fit function?
- For this fit:

$$R^2 = 0.93$$

Compare χ^2 to the total variation in y -values.

- Define the total sum of squares:

$$SS_{\text{tot}} = \sum (y_i - \bar{y})^2$$

\bar{y} = mean of y -values.

- The ratio:

$$\frac{\chi^2}{SS_{\text{tot}}}$$

represents the fraction of the variation that is *not* accounted for by the fit function.

- The fraction of variation that *is* accounted for the function:

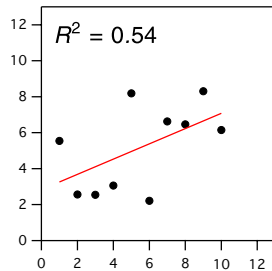
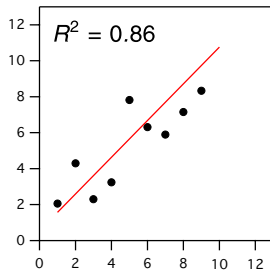
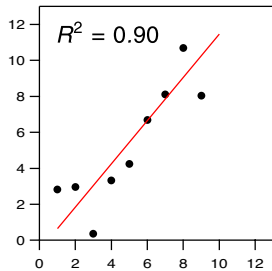
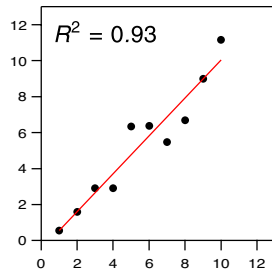
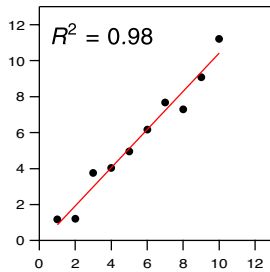
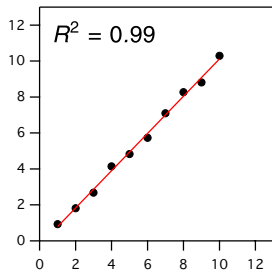
$$R^2 = 1 - \frac{\chi^2}{SS_{\text{tot}}}$$

R^2 should lie between 0 and 1.

Interpreting R^2

- $0 \leq R^2 \leq 1$
- $R^2 = 0$: No correlation between x and y in the experimental data.
- $R^2 = 1$: A “perfect” fit of the experimental data to the function.
- R^2 is the fraction of the total variation of the experimental y values that is accounted for by the linear function.

Some Examples



A Related Term: The Correlation Coefficient, r

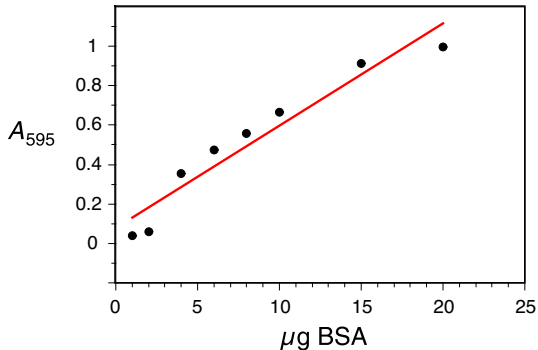
- For a fit to the linear function, $y = mx + b$:

$$R^2 = r^2$$

where r is called the correlation coefficient, or Pearson correlation coefficient, or linear correlation coefficient.

- r can lie between -1 and 1 .
 - If $r > 0$, y increases when x increases.
 - If $r < 0$, y decreases when x increases.
 - If $r = 0$, x and y are uncorrelated.
- r really only makes sense for the linear fit, but some programs will report it for other types of fits.
- R^2 can be calculated, and interpreted for a fit to any function:

What If the Fit isn't as Good as We'd Like?



$$R^2 = 0.93$$

- We could delete some points.
Only if there is a very good reason!
- Is there a function that better represents the data?
- Is there a systematic deviation of the data from the linear function?
- Maybe a curvy function?

Polynomials as Fitting Functions

- General form of a polynomial function:

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n$$

- A polynomial in which the largest power of x is x^n is called an n^{th} -order polynomial.

- A first-order polynomial is a straight line: $y = a_0 + a_1x$
- A second-order polynomial is also called a quadratic function:

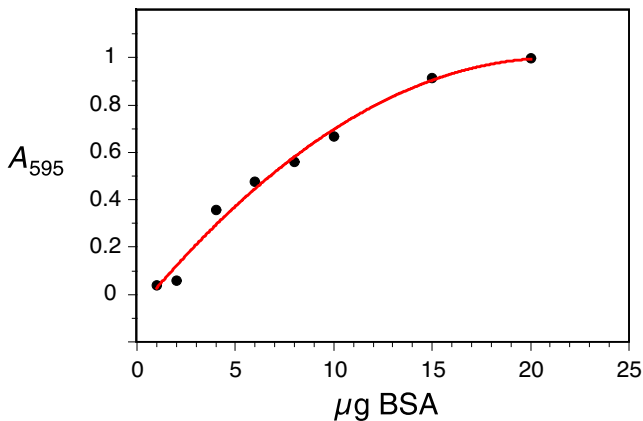
$$y = a_0 + a_1x + a_2x^2$$

- A third-order polynomial is also called a cubic function:

$$y = a_0 + a_1x + a_2x^2 + a_3x^3$$

- An n^{th} -order polynomial contains $n + 1$ coefficients ($a_0, a_1, a_2, \dots, a_n$).
- A **minimum** of $n + 1$ data points are required to fit an n^{th} -order polynomial.

A 2nd-order Polynomial Least-squares Fit to Bradford Calibration Data



- For 2nd-order polynomial fit:

$$\chi^2 = 0.012$$

$$R^2 = 0.988$$

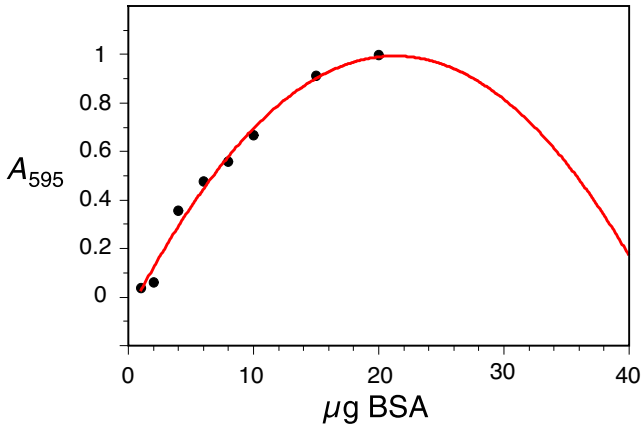
- For linear fit:

$$\chi^2 = 0.062$$

$$R^2 = 0.93$$

- Increasing the number of parameters almost always improves the fit!
- Is it justified here?

Does the Fit Function Make Sense Physically?



- Should the absorbance decrease as the amount of BSA increases beyond $20 \mu\text{g}$?
Probably not!
- The function serves as a calibration curve over the range used to fit it, but not beyond.