Analyzing Two-Dimensional Data

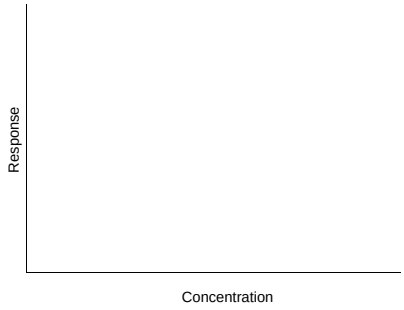
The most common analytical measurements involve the determination of an unknown concentration based on the response of an analytical procedure (usually instrumental).

Such a measurement requires ***calibration***, or the preparation of a ***calibration curve***.

- Determination of the response of the method to solutions of known concentration (**standards**).
- Once the response for the standards is known, the concentration of an unknown can be determined IF the concentration/response relationship is well defined.
- Ideally prefer a linear relationship
  - doesn't have to be linear as long as you know what it is, can often "force" nonlinear relationships to appear linear by appropriate experiment design

1

Analyzing Two-Dimensional Data



Important questions to ask:

1. How do we define the "best" line?
2. How do errors in our data affect this line?
3. How confident can we be of the unknown concentration that we calculate from our calibration curve?

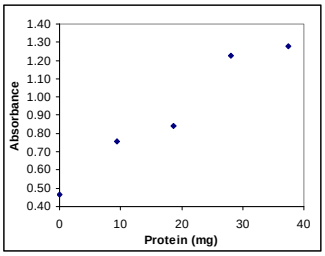
2

Analyzing Two-Dimensional Data

Example: Protein determination using spectrophotometry.

IMPORTANT: Absorbance ∝ Protein mass

Protein (μg)	0.00	9.36	18.72	28.08	37.44
Absorbance	0.466	0.756	0.843	1.226	1.280



Our objective is to draw the "best fit" line through the data, but how?

- Minimize deviation (spread) of the data around the line

Mathematically, this is a "*least squares*" analysis

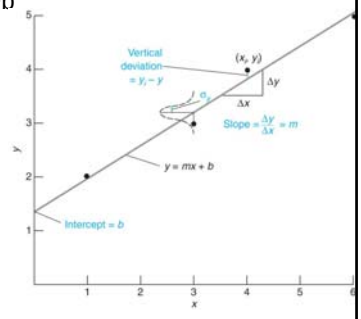
- Work to minimize the square of the deviation (to remove effects of sign) from our calculated line.
- Qualitatively this is easy, quantitatively; things are a little more challenging.

3

Method of Least Squares

Typically working to define a straight line, $y = mx + b$

- Assume that values for x have little error, but more error is associated with values for y.
- Since our data have some scatter, each datum may deviate from the line in the *y-direction.*
  - *This is also called a residual (d_i)*



$$d_i = y_i - y_{line} = y_i - (mx_i + b)$$

We really want to *minimize* the *square* of the deviations (actually the SUM of the squares):

$$(d_i)^2 = (y_i - mx_i - b)^2$$

$$(d_i)^2 = y_i^2 - 2mx_iy_i - 2by_i + 2mx_ib + m^2x_i^2 + b^2$$

$$\Sigma(d_i)^2 = \Sigma y_i^2 - \Sigma 2mx_iy_i - \Sigma 2by_i + \Sigma 2mx_ib + \Sigma m^2x_i^2 + \Sigma b^2$$

How do we do this?

4

Method of Least Squares

$$\Sigma(d_i)^2 = \Sigma y_i^2 - \Sigma 2mx_iy_i - \Sigma 2by_i + \Sigma 2mx_ib + \Sigma m^2x_i^2 + \Sigma b^2$$

Two parameters, so two partial derivatives to set equal to zero:

$$\frac{\partial(\sum d^2)}{\partial m} = -\Sigma 2x_iy_i + \Sigma 2x_ib + 2\Sigma mx_i^2 = -\Sigma x_iy_i + b\Sigma x_i + m\Sigma x_i^2 = 0$$

$$\frac{\partial(\sum d^2)}{\partial b} = -\Sigma 2y_i + \Sigma 2mx_i + 2\Sigma b = -\Sigma y_i + m\Sigma x_i + bn = 0$$

This produces two equations and two unknowns (m, and b)…
we should be able to solve this system!

5

Method of Least Squares

$$(d_i)^2 = y_i^2 - 2mx_iy_i - 2by_i + 2mx_ib + m^2x_i^2 + b^2$$

With a little hand-waving (and the magic of calculus and linear algebra), we are able to minimize the equation above and solve for m and b, when we do, we get:

$$m = \begin{vmatrix} \sum(x_iy_i) & \sum x_i \\ \sum y_i & n \end{vmatrix} \div D \qquad b = \begin{vmatrix} \sum x_i^2 & \sum(x_iy_i) \\ \sum x_i & \sum y_i \end{vmatrix} \div D \qquad D = \begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}$$

Each operation involves taking the determinant of a matrix

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = (a \cdot d) - (b \cdot c)$$

There is only **one solution** to the system of equations
So only **one** least squares line!

6

Method of Least Squares

Lets apply this to our example data:

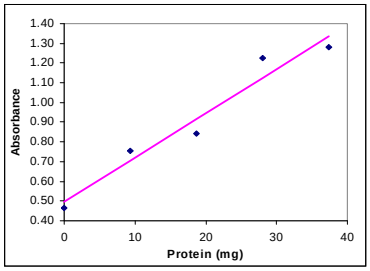
$$D = \begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix} = 4380.48$$

	Protein (µg) x	Absorbance y	xy	x^2
	0.00	0.466	0.000	0.000
	9.36	0.756	7.076	87.609
	18.72	0.843	15.780	350.438
	28.08	1.226	34.426	788.486
	37.44	1.280	47.920	1401.75
Σ	93.60	4.571	105.206	2628.288

$$m = \begin{vmatrix} \sum (x_i y_i) & \sum x_i \\ \sum y_i & n \end{vmatrix} \div D = 0.022415 \qquad b = \begin{vmatrix} \sum x_i^2 & \sum (x_i y_i) \\ \sum x_i & \sum y_i \end{vmatrix} \div D = 0.4946$$

Now lets calculate some points based on our line:

	Protein (µg) x	Absorbance y	y_{calc}	d	d^2
	0.00	0.466	0.495	-0.029	0.000818
	9.36	0.756	0.704	0.052	0.002663
	18.72	0.843	0.914	-0.071	0.005069
	28.08	1.226	1.124	0.102	0.010404
	37.44	1.280	1.334	-0.054	0.002894
Σ	93.60	4.571	4.571	0.000	0.022



7

How reliable are m, b, and values we determine based on our calibration curve?

The majority of our confidence depends on the scatter of y values about the line, or the standard deviation in y, s_y (also called s_r, st. dev. about regression).

$$s_y = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-2}} = \sqrt{\frac{\sum d_i^2}{n-2}}$$

- Like usual, the number of degrees of freedom is in the denominator.
- Why *n-2* degrees of freedom?

Other std. devs. depend on s_y

$$s_m^2 = \frac{s_y^2 \times n}{D} \qquad s_b^2 = \frac{s_y^2 \sum x_i^2}{D}$$

$$s_x = \frac{s_y}{|m|}\sqrt{\frac{1}{k} + \frac{x^2 n}{D} + \frac{\sum x_i^2}{D} - \frac{2x\sum x_i}{D}} = \frac{s_y}{|m|}\sqrt{\frac{1}{k} + \frac{1}{n} + \frac{(y-\bar{y})^2}{m^2 \sum (x_i - \bar{x})^2}}$$

where k is the number of replicate measurements of the unknown and n is the number of calibration points.

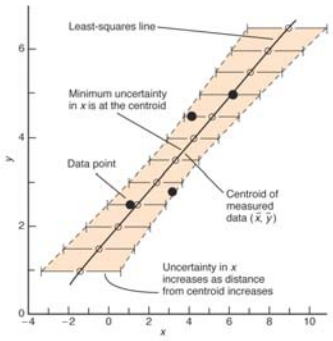
8

Confidence Limits for m, b, x_{calc}

- **Confidence Limits for m, b, x_{calc}**

$$m \pm ts_m$$
$$b \pm ts_b$$
$$x_{calc} \pm ts_x$$

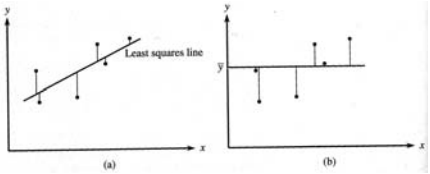
t is for n-2 degrees of freedom



9

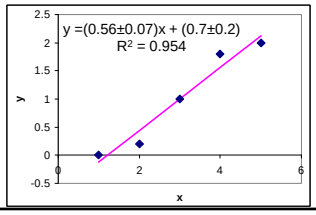
R^2 and Such

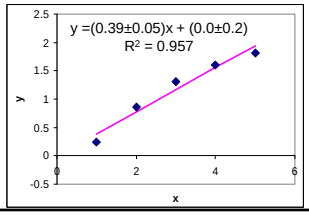
- Plotting in Excel (or on my calculator) gives me R^2 (or R) values. What the #$%@ do these mean?
  - R^2 (or r^2): **Coefficient of Determination** is the fraction of the scatter in the data that can be described by the linear relationship.
  - R^2 compares the variation of the data from the least-squares line to that due to random scatter:



$$R^2 = 1 - \frac{\sum (y_i - y_{line})^2}{\sum (y_i - \bar{y})^2}$$

  - An R^2 close to 1 doesn't guarantee good precision in m and b





10