Intermediate Lab

PHYS 3870

Lecture 4

Comparing Data and Models—Quantitatively

Linear Regression

References: Taylor Ch. 8 and 9
Also refer to "Glossary of Important Terms in Error Analysis"

Intermediate Lab

PHYS 3870

Errors in Measurements and Models

A Review of What We Know

Quantifying Precision and Random (Statistical) Errors

The "best" value for a group of measurements of the same quantity is the

Average

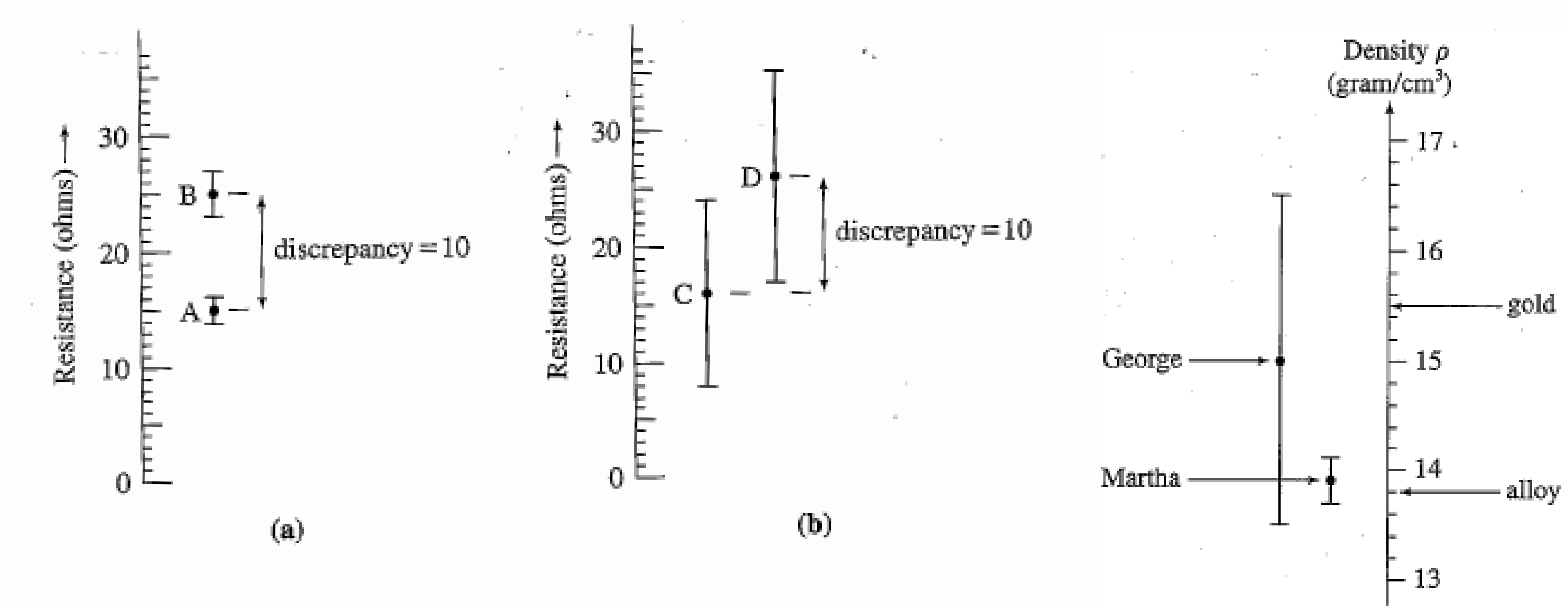
What is an estimate of the random error?

Deviations

A. If the average is the the best guess, then DEVIATIONS (or discrepancies) from best guess are an estimate of error

B. One estimate of error is the range of deviations.

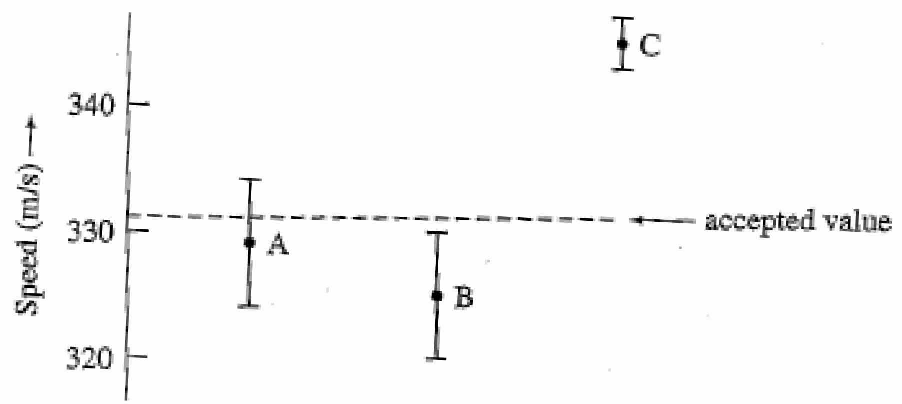
Single Measurement: Comparison with Other Data

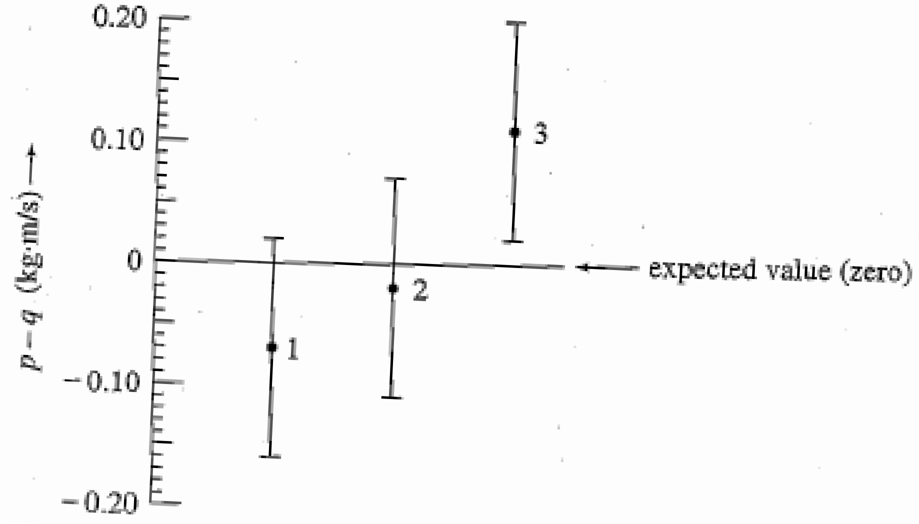


Comparison of precision or accuracy?

$$Precent\ Difference = \frac{(x_1 - x_2)}{\frac{1}{2}(x_1 + x_2)}$$

Single Measurement: Direct Comparison with Standard





Comparison of precision or accuracy?

$$Precent\ Error = \frac{x_{meansured} \quad - x_{Known}}{x_{Known}}$$

Multiple Measurements of the Same Quantity

Our statement of the best value and uncertainty is:

(<t> $\pm \sigma_t$) sec at the 68% confidence level for N measurements

1. **Note the precision of our measurement is reflected in the estimated error which states what values we would expect to get if we repeated the measurement**
2. **Precision is defined as a measure of the reproducibility of a measurement**
3. **Such errors are called random (statistical) errors**
4. **Accuracy is defined as a measure of who closely a measurement matches the true value**
5. **Such errors are called systematic errors**

Multiple Measurements of the Same Quantity

Standard Deviation

The best guess for the error in a group of N identical randomly distributed measurements is given by the standard deviation

$$\ldots\sigma = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N}(t_i - \bar{t})^2}$$

this is the rms (root mean squared deviation or (sample) standard deviation

It can be shown that (see Taylor Sec. 5.4) σ_t is a reasonable estimate of the uncertainty. In fact, for normal (Gaussian or purely random) data, it can be shown that

(1) 68% of measurements of t will fall within $\langle t \rangle \pm \sigma_t$
(2) 95% of measurements of t will fall within $\langle t \rangle \pm 2\sigma_t$
(3) 98% of measurements of t will fall within $\langle t \rangle \pm 3\sigma_t$
(4) this is referred to as the confidence limit

Summary: the standard format to report the best guess and the limits within which you expect 68% of subsequent (single) measurements of t to fall within is $\langle t \rangle \pm \sigma_t$

Multiple Sets of Measurements of the Same Quantity

Standard Deviation of the Mean

If we were to measure t again N times (not just once), we would be even more likely to find that the second average of N points would be close to <t>.

The standard error or standard deviation of the mean is given by…

$$\sigma_{SDOM} = \frac{\sigma_{SD}}{\sqrt{N}} = \sqrt{\frac{1}{N\,(N-1)} \sum_{i=1}^{N} (t_i - \bar{t})^2}$$

This is the limits within which you expect the average of N addition measurements to fall within at the 68% confidence limit

Errors Propagation—Error in Models or Derived Quantities

Define error propagation [Taylor, p. 45]

A method for determining the error inherent in a derived quantity from the errors in the measured quantities used to determine the derived quantity

Recall previous discussions [Taylor, p. 28-29]

I. Absolute error: $(\langle t \rangle \pm \sigma_t)$ sec
II. Relative (fractional) Error: $\langle t \rangle$ sec $\pm$ $(\sigma_t / \langle t \rangle)\%$
III. Percentage uncertainty: fractional error in % units

Specific Rules for Error Propagation (Worst Case)

Refer to **[Taylor, sec. 3.2]** for specific rules of error propagation:

1. Addition and Subtraction **[Taylor, p. 49]**
   For $q_{best}=x_{best} \pm y_{best}$ the error is $\delta q \approx \delta x + \delta y$
   Follows from $q_{best} \pm \delta q = (x_{best} \pm \delta x) \pm (y_{best} \pm \delta y) = (x_{best} \pm y_{best}) \pm (\delta x \pm \delta y)$

2. Multiplication and Division **[Taylor, p. 51]**
   For $q_{best}=x_{best} * y_{best}$ the error is $(\delta q / q_{best}) \approx (\delta x / x_{best}) + (\delta y / y_{best})$

3. Multiplication by a constant (exact number) **[Taylor, p. 54]**
   For $q_{best}= B(x_{best})$ the error is $(\delta q / q_{best}) \approx |B| (\delta x / x_{best})$
   Follows from 2 by setting $\delta B/B=0$

4. Exponentiation (powers) **[Taylor, p. 56]**
   For $q_{best}= (x_{best})^n$ the error is $(\delta q / q_{best}) \approx n (\delta x / x_{best})$
   Follows from 2 by setting $(\delta x / x_{best}) = (\delta y / y_{best})$

General Formula for Error Propagation

General formula for uncertainty of a function of one variable

$$\delta q = \left|\frac{\partial q}{\partial x}\right| \cdot \delta x$$ [Taylor, Eq. 3.23]

Can you now derive for specific rules of error propagation:

1. Addition and Subtraction **[Taylor, p. 49]**
2. Multiplication and Division **[Taylor, p. 51]**
3. Multiplication by a constant (exact number) **[Taylor, p. 54]**
4. Exponentiation (powers) **[Taylor, p. 56]**

General Formula for Multiple Variables

Uncertainty of a function of multiple variables [Taylor, Sec. 3.11]

1. It can easily (no, really) be shown that (see Taylor Sec. 3.11) for a function of several variables

Worst case

$$\delta q(x, y, z, \ldots) = \left|\frac{\partial q}{\partial x}\right| \cdot \delta x + \left|\frac{\partial q}{\partial y}\right| \cdot \delta y + \left|\frac{\partial q}{\partial z}\right| \cdot \delta z + \ldots$$

[Taylor, Eq. 3.47]

2. More correctly, it can be shown that (see Taylor Sec. 3.11) for a function of several variables

$$\delta q(x, y, z, \ldots) \leq \left|\frac{\partial q}{\partial x}\right| \cdot \delta x + \left|\frac{\partial q}{\partial y}\right| \cdot \delta y + \left|\frac{\partial q}{\partial z}\right| \cdot \delta z + \ldots$$

[Taylor, Eq. 3.47]

where the equals sign represents an upper bound, as discussed above.

3. For a function of several *independent and random* variables

Best case

$$\delta q(x, y, z, \ldots) = \sqrt{\left(\frac{\partial q}{\partial x} \cdot \delta x\right)^2 + \left(\frac{\partial q}{\partial y} \cdot \delta y\right)^2 + \left(\frac{\partial q}{\partial z} \cdot \delta z\right)^2 + \ldots}$$

[Taylor, Eq. 3.48]

Again, the proof is left for Ch. 5.

Error Propagation: General Case

Thus, if x and y are:

a) Independent (determining x does not affect measured y)

b) Random (equally likely for $+\delta x$ as $-\delta x$)

Then method the methods above overestimate the error

Consider the arbitrary derived quantity q(x,y) of two independent random variables x and y.

Expand q(x,y) in a Taylor series about the expected values of x and y (i.e., at points near X and Y).

Fixed, shifts peak of distribution

$$q(x,y) = q(X,Y) + \left(\frac{\partial q}{\partial x}\right)\bigg|_X (x - X) + \left(\frac{\partial q}{\partial y}\right)\bigg|_Y (y - Y)$$

Fixed **Distribution centered at X with width σ_X**

Error for a function of Two Variables: Addition in Quadrature

$$\delta q(x,y) = \sigma_q = \sqrt{\left[\left(\frac{\partial q}{\partial x}\right)\bigg|_X \sigma_x\right]^2 + \left[\left(\frac{\partial q}{\partial y}\right)\bigg|_Y \sigma_y\right]^2}$$

Best case

Independent (Random) Uncertaities and Gaussian Distributions

For Gaussian distribution of measured values which describe quantities with random uncertainties, it can be shown that (the dreaded ICBST), errors add in quadrature [see Taylor, Ch. 5]

$\delta q \neq \delta x + \delta y$

But, $\delta q = \sqrt{[(\delta x)2 + (\delta y)^2]}$

1. This is proved in [Taylor, Ch. 5]
2. ICBST [Taylor, Ch. 9] Method A provides an upper bound on the possible errors

Gaussian Distribution Function

Distribution Function

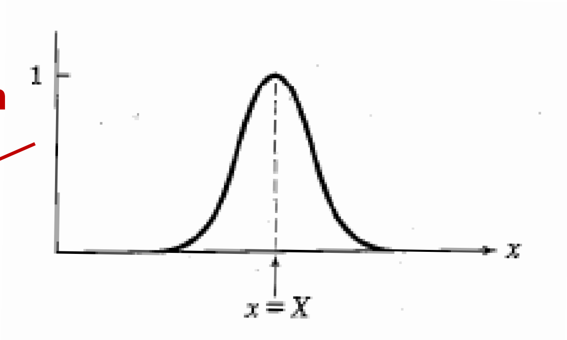
Independent Variable

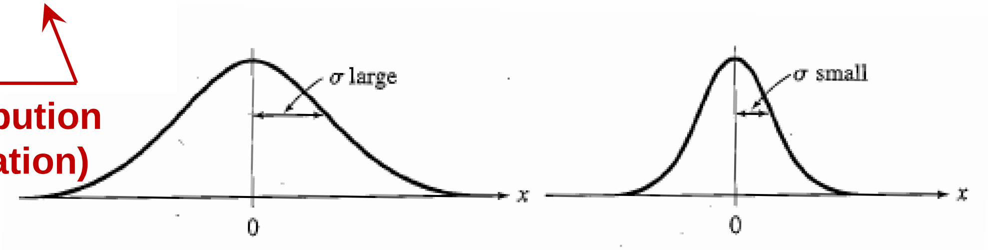
Center of Distribution (mean)

$$G_{X,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2}$$

Normalization Constant

Width of Distribution (standard deviation)





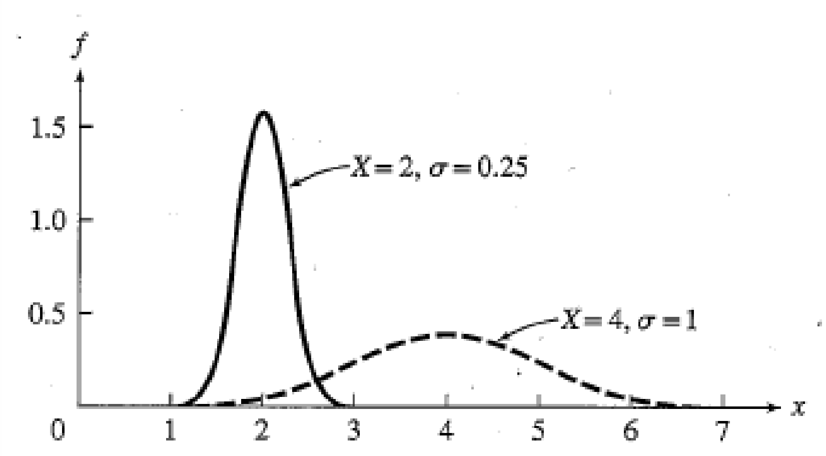


Figure 5.10. Two normal, or Gauss, distributions.

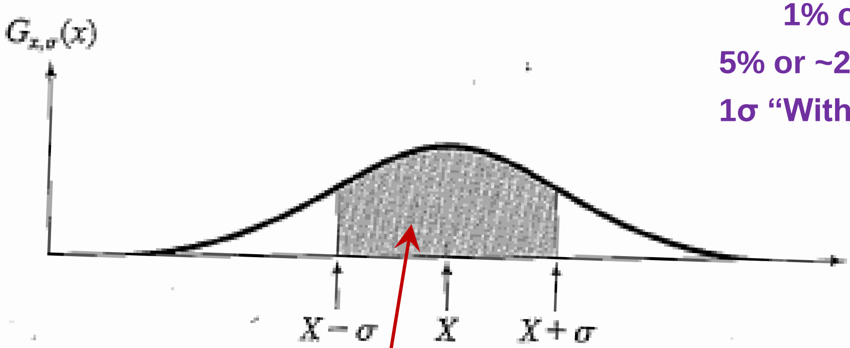
Gaussian Distribution Function

Standard Deviation of Gaussian Distribution

$$Prob(\text{within } \sigma) = \int_{X-\sigma}^{X+\sigma} G_{X,\sigma}(x)\,dx \qquad (5.32)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{X-\sigma}^{X+\sigma} e^{-(x-X)^2/2\sigma^2}\,dx. \qquad (5.33)$$

See Sec. 10.6: Testing of Hypotheses



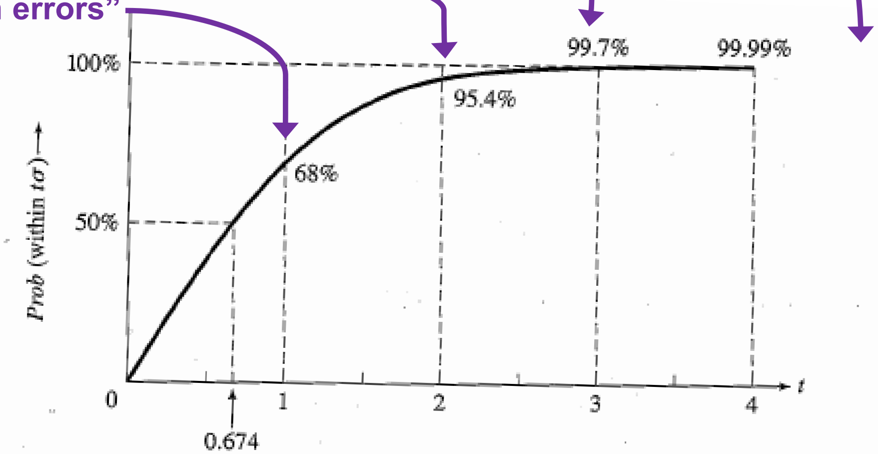
Area under curve (probability that –σ<x<+σ) is 68%

5 ppm or ~5σ “Valid for HEP”

1% or ~3σ “Highly Significant”

5% or ~2σ “Significant”

1σ “Within errors”



t	0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.5	3.0	3.5	4.0
Prob (%)	0	20	38	55	68	79	87	92	95.4	98.8	99.7	99.95	99.99

Figure 5.13. The probability *Prob*(within $t\sigma$) that a measurement of x will fall within t standard deviations of the true value $x = X$. Two common names for this function are the *normal error integral* and the *error function*, erf(t).

More complete Table in App. A and B

Mean of Gaussian Distribution as "Best Estimate"

Principle of Maximum Likelihood

To find the most likely value of the mean (the best estimate of ẋ),
find X that yields the highest probability for the data set.

Consider a data set $\{x_1, x_2, x_3 \ldots x_N\}$

Each randomly distributed with

$$Prob_{X,\sigma}(x_i) = G_{X,\sigma}(x_i) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-X)^2/2\sigma} \propto \frac{1}{\sigma} e^{-(x_i-X)^2/2\sigma}$$

The combined probability for the full data set is the product

$$Prob_{X,\sigma}(x_1, x_2 \ldots x_N) = Prob_{X,\sigma}(x_1) \times Prob_{X,\sigma}(x_2) \times \ldots \times Prob_{X,\sigma}(x_N)$$

$$\propto \frac{1}{\sigma} e^{-(x_1-X)^2/2\sigma} \times \frac{1}{\sigma} e^{-(x_2-X)^2/2\sigma} \times \ldots \times \frac{1}{\sigma} e^{-(x_N-X)^2/2\sigma} = \frac{1}{\sigma^N} e^{\sum -(x_i-X)^2/2\sigma}$$

Best Estimate of X is from maximum probability or minimum summation

Minimize Sum $\sum_{i=1}^{N} (x_i - X)^2/\sigma$ **Solve for derivative set to 0** $\sum_{i=1}^{N} (x_i - X) = 0$ **Best estimate of X** $X_{best} = \sum x_i / N$

Uncertaity of "Best Estimates" of Gaussian Distribution

Principle of Maximum Likelihood

To find the most likely value of the mean (the best estimate of ẋ), find X that yields the highest probability for the data set.

Consider a data set $\{x_1, x_2, x_3 \ldots x_N\}$

The combined probability for the full data set is the product

$$Prob_{X,\sigma}(x_1, x_2 \ldots x_N) = Prob_{X,\sigma}(x_1) \times Prob_{X,\sigma}(x_2) \times \ldots \times Prob_{X,\sigma}(x_N)$$

$$\propto \frac{1}{\sigma} e^{-(x_1 - X)^2/2\sigma} \times \frac{1}{\sigma} e^{-(x_2 - X)^2/2\sigma} \times \ldots \times \frac{1}{\sigma} e^{-(x_N - X)^2/2\sigma} = \frac{1}{\sigma^N} e^{\Sigma -(x_i - X)^2/2\sigma}$$

Best Estimate of X is from maximum probability or minimum summation

Minimize Sum $\sum_{i=1}^{N} (x_i - X)^2/\sigma$ **Solve for derivative wrst X set to 0** $\sum_{i=1}^{N} (x_i - X) = 0$ **Best estimate of X** $X_{best} = \sum x_i / N$

Best Estimate of σ is from maximum probability or minimum summation

Minimize Sum $\sum_{i=1}^{N} (x_i - X)^2/\sigma$ **Solve for derivative wrst σ set to 0** See Prob. 5.26 **Best estimate of σ** $\sigma_{best} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - X)^2/\sigma}$

Weighted Averages

Question: How can we properly combine two or more separate independent measurements of the same randomly distributed quantity to determine a best combined value with uncertainty?

Weighted Averages

The probability of measuring two such measurements is

$$Prob_x(x_1, x_2) = Prob_x(x_1)\, Prob_x(x_2)$$

$$= \frac{1}{\sigma_1 \sigma_2} e^{-\chi^2/2} \quad where \;\; \chi^2 \equiv \left[\frac{(x_1 - X)}{\sigma_1}\right]^2 + \left[\frac{(x_2 - X)}{\sigma_2}\right]^2$$

To find the best value for X, find the maximum Prob or minimum X^2

Best Estimate of χ is from maximum probability or minimum summation

Minimize Sum

$$\chi^2 \equiv \left[\frac{(x_1 - X)}{\sigma_1}\right]^2 + \left[\frac{(x_2 - X)}{\sigma_2}\right]^2$$

Solve for derivative wrst χ set to 0

$$2\left[\frac{(x_1 - X)}{\sigma_1}\right] + 2\left[\frac{(x_2 - X)}{\sigma_2}\right] = 0$$

Solve for best estimate of χ

$$X_{best} = \left(\frac{x_1}{\sigma_1{}^2} + \frac{x_2}{\sigma_2{}^2}\right) \Big/ \left(\frac{1}{\sigma_1{}^2} + \frac{1}{\sigma_2{}^2}\right)$$

This leads to

$$\boxed{x_{W\_avg} = \frac{w_1 x_1 + w_2 x_2}{w_1 + w_2} = \frac{\sum w_i\, x_i}{\sum w_i} \quad where\; w_i = {}^{1}\!/_{(\sigma_i)^2}}$$

Note: If w_1=w_2, we recover the standard result X_{wavg}= (1/2) (x_1+x_2)

Finally, the width of a weighted average distribution is $\sigma_{weighted\ avg} = \frac{1}{\sum_i w_i}$

Intermediate Lab

PHYS 3870

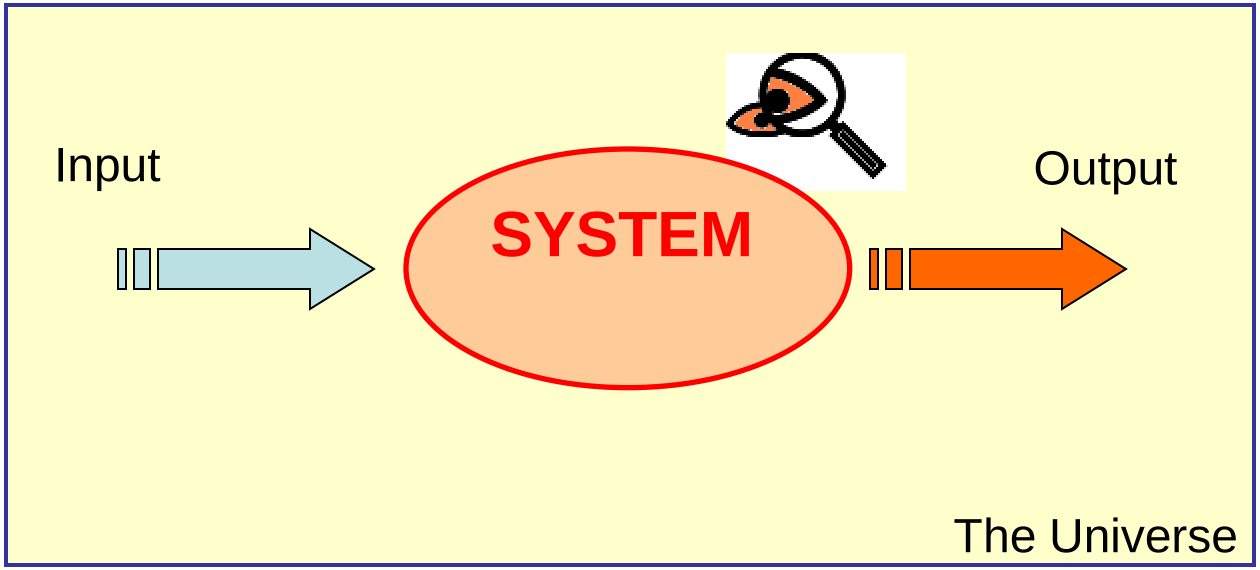
Comparing Measurements to Models

Linear Regression

Motivating Regression Analysis

Question: Consider now what happens to the output of a nearly ideal experiment, if we vary how hard we poke the system (vary the input).

Uncertainties in Observations



The simplest model of response (short of the trivial constant response) is a linear response model

$$y(x) = A + B\,x$$

Questions for Regression Analysis

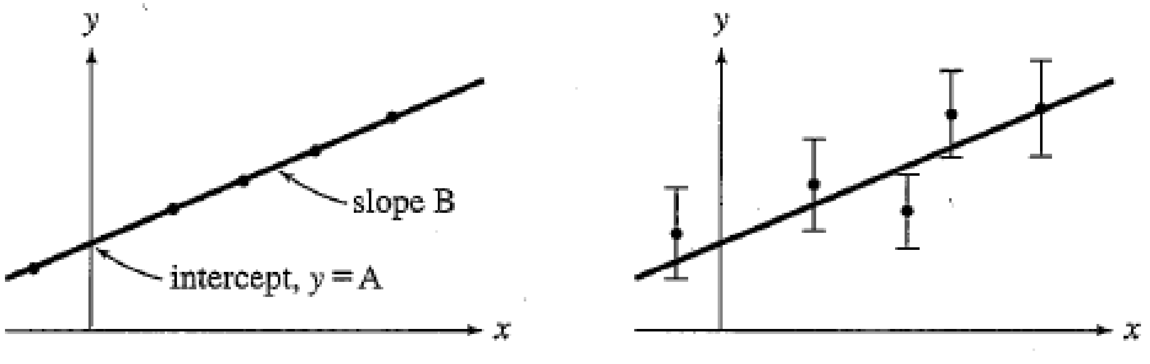
The simplest model of response (short of the trivial constant response) is a linear response model

$$y(x) = A + B\,x$$

Three principle questions:

What are the best values of A and B, for: **(see Taylor Ch. 8)**

- **A perfect data set, where A and B are exact?**
- **A data set with uncertainties?**



(a) (b)

What are the errors in the fitting parameters A and B?
What confidence can we place in how well a linear model fits the data?
(see Taylor Ch. 9)

Intermediate Lab

PHYS 3870

Review of

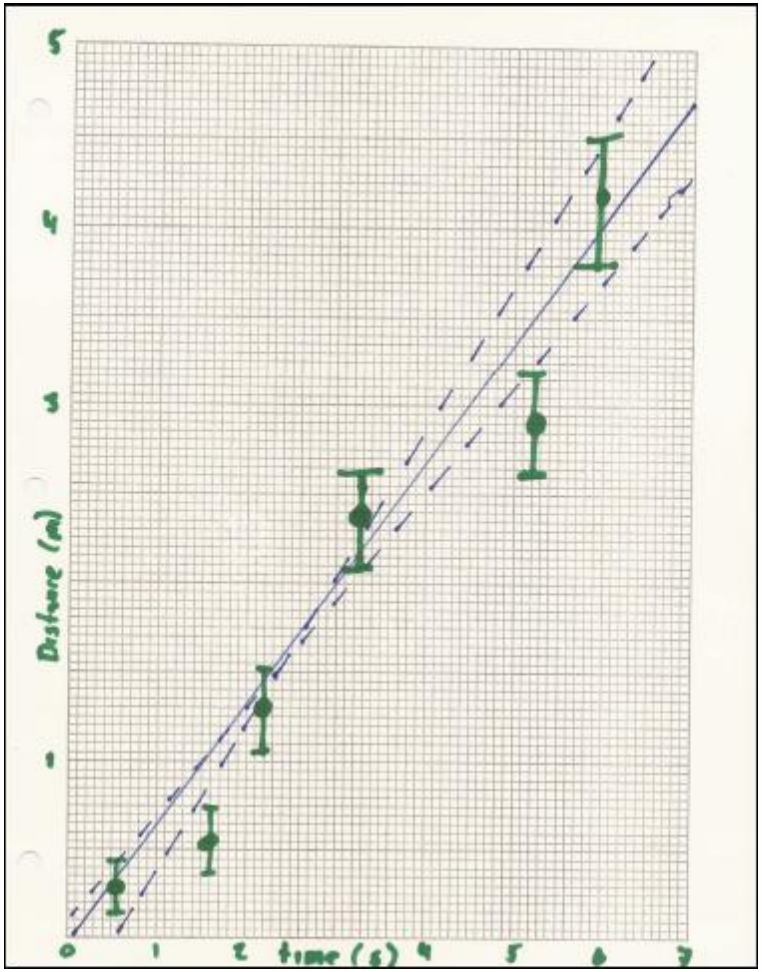
Graphical Analysis

Graphical Analysis

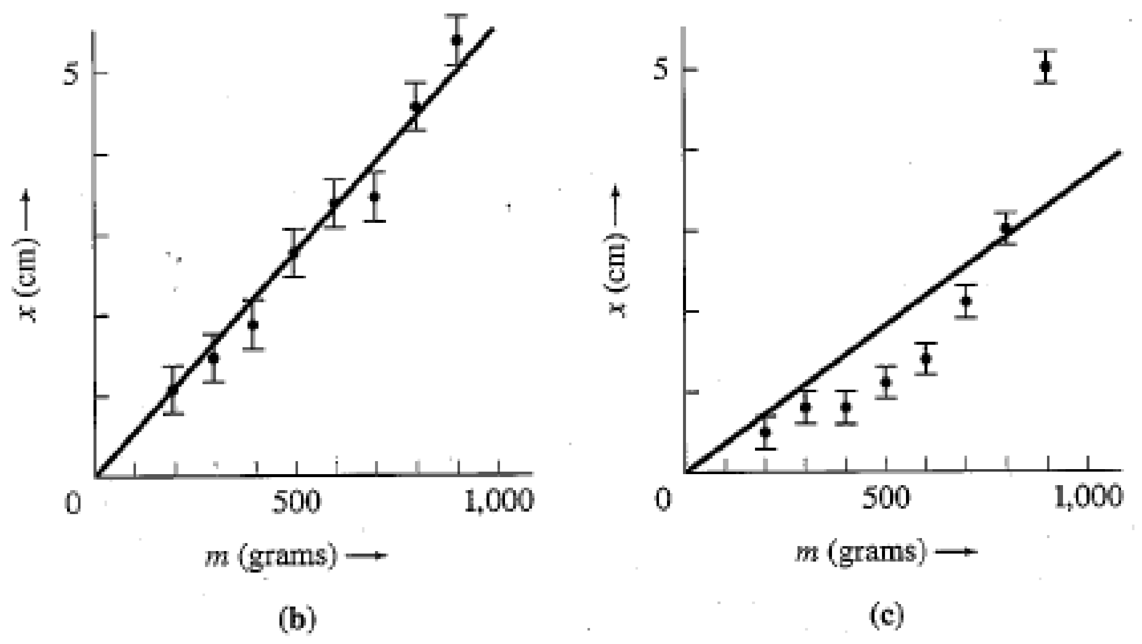
An “old School” approach to linear fits.

• Rough plot of data
• Estimate of uncertainties with error bars
• A “best” linear fit with a straight edge
• Estimates of uncertainties in slope and intercept from the error bars

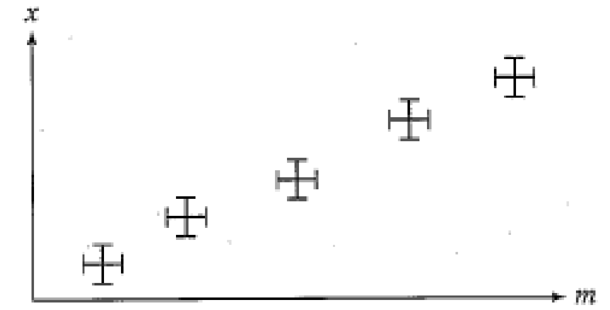
This is a great practice to get into as you are developing an experiment!



Is it Linear?

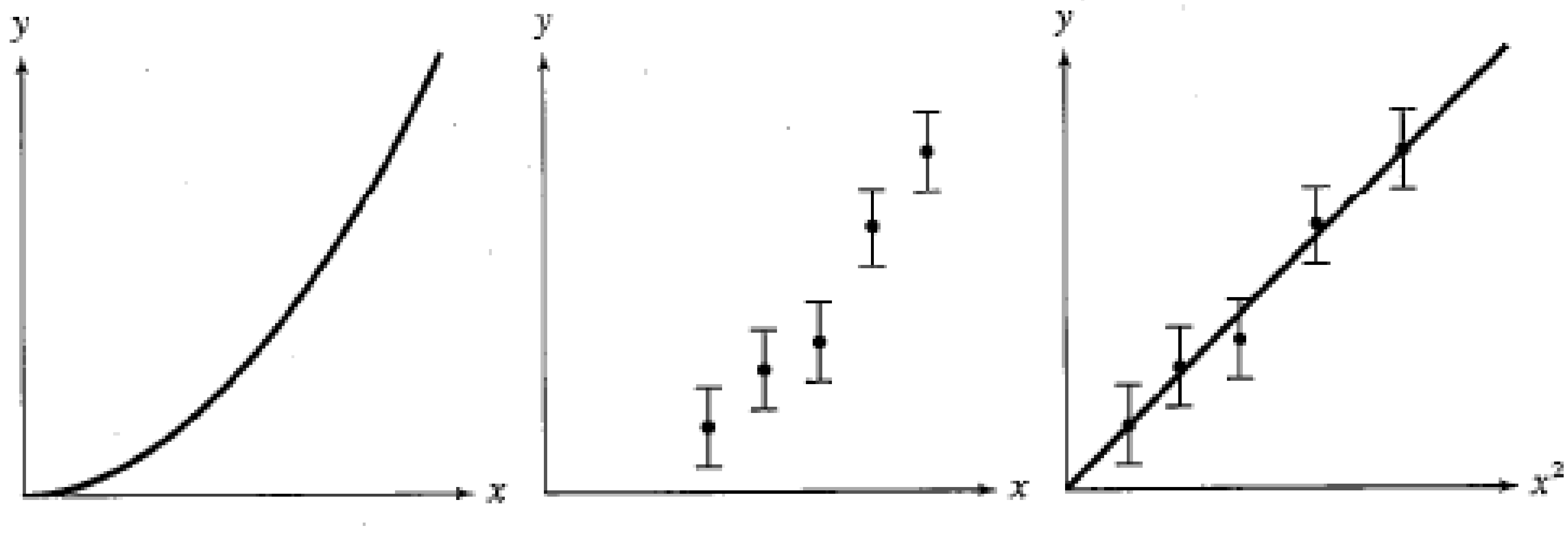


- A simple model is a linear model
- You know it when you see it (qualitatively)
- Tested with a straight edge
- Error bar are a first step in gauging the "goodness of fit"



Adding 2D error bars is sometimes helpful.

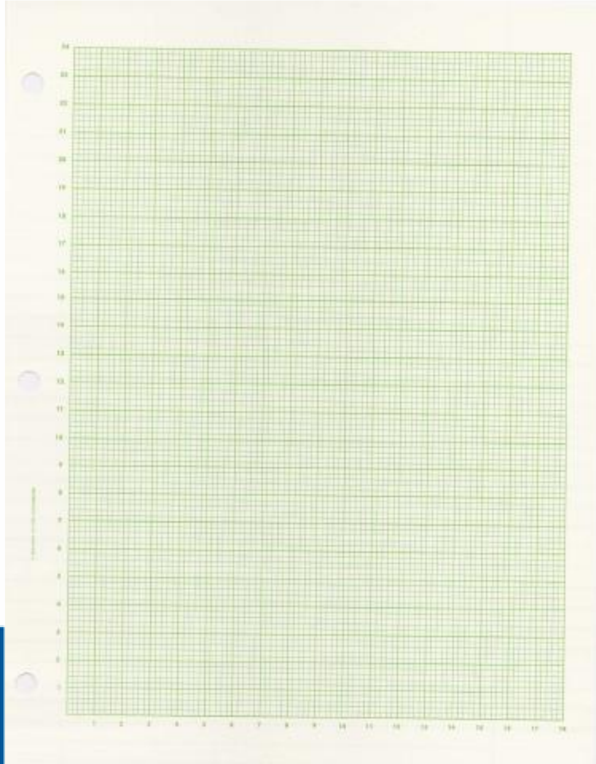
Making It Linear or Linearization



(a) (b) (c)

- A simple trick for many models is to linearize the model in the independent variable.
- Refer to Baird Ch.5 and the associated homework problems.

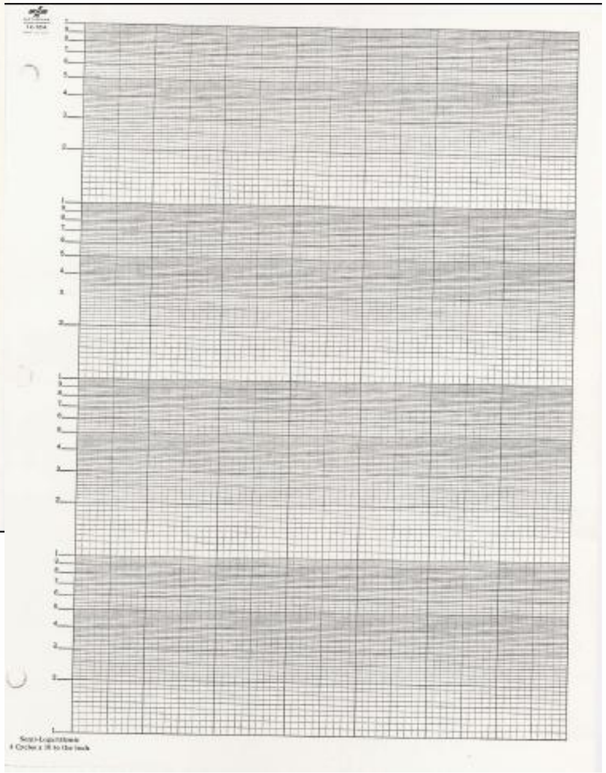
Special Graph Paper



Linear

"Old School" graph paper is still a useful tool, especially for reality checking during the experimental design process.

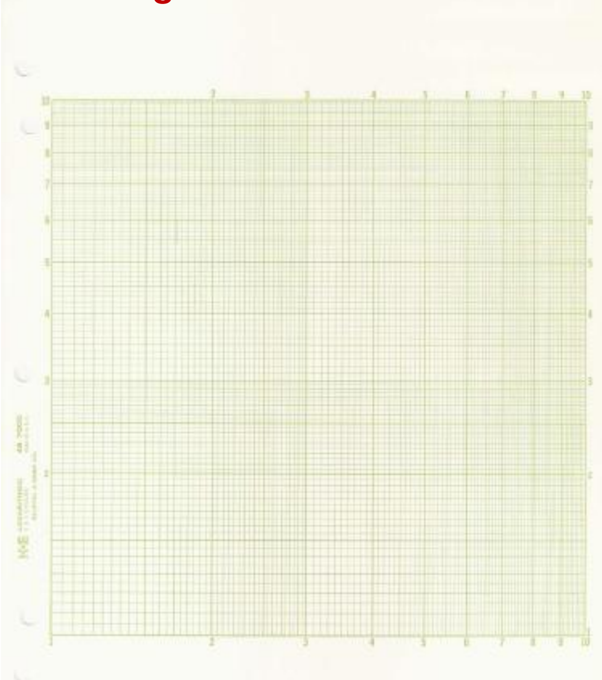
Semi-log paper tests for exponential models.



Semilog

Log-log paper tests for power law models.

Both Semi-log and log -log paper are handy for displaying details of data spread over many orders of magnitude.



Log-Log

Intermediate Lab

PHYS 3870

Linear Regression

References: Taylor Ch. 8

Basic Assumptions for Regression Analysis

We will (initially) assume:

- **The errors in how hard you poke something (in the input) are negligible compared with errors in the response (see discussion in Taylor Sec. 9.3)**

- **The errors in y are constant (see Problem 8.9 for weighted errors analysis)**

- **The measurements of y_i are governed by a Gaussian distribution with constant width σ_y**

Question 1: What is the Best Linear Fit (A and B)?

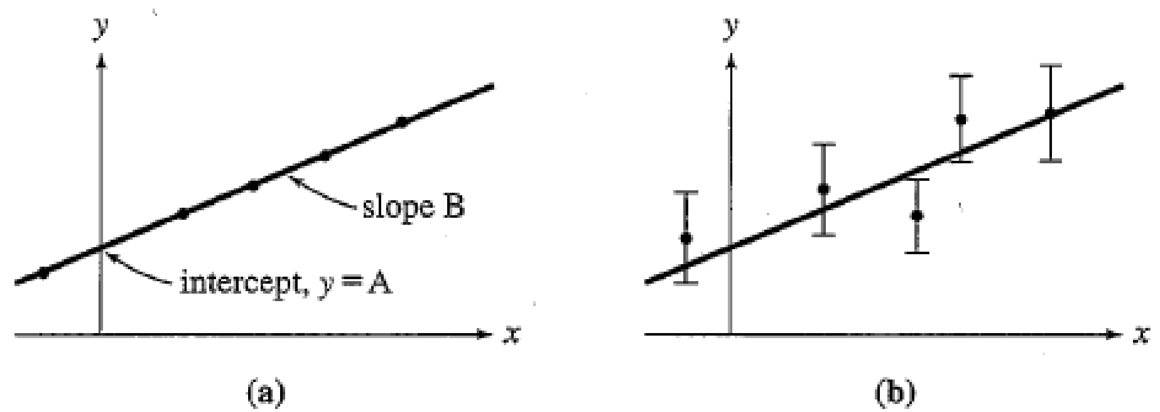


Figure 8.1. (a) If the two variables y and x are linearly related as in Equation (8.1), and if there were no experimental uncertainties, then the measured points (x_i, y_i) would all lie exactly on the line $y = A + Bx$. (b) In practice, there always are uncertainties, which can be shown by error bars, and the points (x_i, y_i) can be expected only to lie reasonably close to the line. Here, only y is shown as subject to appreciable uncertainties.

Best Estimate of intercept, A , and slope, B,
for
Linear Regression or Least Squares-Fit for Line

For the linear model y = A + B x

Intercept: $A = \frac{\sum x^2 \sum y - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}$ $\qquad \sigma_A = \sigma_y \frac{\sum x^2}{N \sum x^2 - (\sum x)^2}$

Slope $B = \frac{N \sum xy - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}$ $\qquad \sigma_B = \sigma_y \frac{N}{N \sum x^2 - (\sum x)^2}$

where $\sigma_y = \sqrt{\frac{1}{N-2} \sum [y_i - (A + Bx_i)]^2}$

"Best Estimates" of Linear Fit

Consider a linear model for y_i, $y_i=A+Bx_i$

The probability of obtaining an observed value of y_i is

$$Prob_{A,B}(y_1 \ldots y_N) = Prob_{A,B}(y_1) \times \ldots \times Prob_{A,B}(y_N)$$

$$= \frac{1}{{\sigma_y}^N} e^{-\chi^2/2} \quad where \;\; \chi^2 \equiv \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i)]^2}{{\sigma_y}^2}$$

To find the best simultaneous values for A and B, find the maximum Prob or minimum X^2

Best Estimates of A and B are from maximum probibility or minimum summation

Minimize Sum

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i)]^2}{{\sigma_y}^2}$$

Solve for derivative wrst A and B set to 0

$$\frac{\partial \chi^2}{\partial A} \equiv \frac{-2}{{\sigma_y}^2} \sum_{i=1}^{N} [y_i - (A + Bx_i)] = 0$$

$$\frac{\partial \chi^2}{\partial B} \equiv \frac{-2}{{\sigma_y}^2} \sum_{i=1}^{N} x_i [y_i - (A + Bx_i)] = 0$$

Best estimate of A and B

$$AN + B \sum x_i = \sum y_i$$

$$A \sum x_i + A \sum {x_i}^2 = \sum x_i y_i$$

"Best Estimates" of Linear Fit

Best Estimates of A and B are from maximum probibility or minimum summation

Minimize Sum

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i)]^2}{{\sigma_y}^2}$$

Solve for derivative wrst A and B set to 0

$$\frac{\partial \chi^2}{\partial A} \equiv \frac{-2}{{\sigma_y}^2} \sum_{i=1}^{N} [y_i - (A + Bx_i)] = 0$$

$$\frac{\partial \chi^2}{\partial B} \equiv \frac{-2}{{\sigma_y}^2} \sum_{i=1}^{N} x_i [y_i - (A + Bx_i)] = 0$$

Best estimate of A and B

$$AN + B\sum x_i = \sum y_i$$

$$A\sum x_i + B\sum {x_i}^2 = \sum x_i y_i$$

In a linear algebraic form

$$\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{bmatrix} \cdot \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

This is a standard eigenvalue problem

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \cdot \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}$$

With solutions

$$A = \frac{E_1 M_{22} - E_2 M_{12}}{M_{11} M_{22} - M_{12} M_{21}}$$

$$B = \frac{E_1 M_{21} - E_2 M_{11}}{M_{11} M_{22} - M_{12} M_{21}}$$

For the linear model y = A + B x

Intercept:

$$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{N\sum x^2 - (\sum x)^2} \qquad \sigma_A = \sigma_y \frac{\sum x^2}{N\sum x^2 - (\sum x)^2}$$

(Prob (8.16)

Slope

$$B = \frac{N\sum xy - \sum x \sum xy}{N\sum x^2 - (\sum x)^2} \qquad \sigma_B = \sigma_y \frac{N}{N\sum x^2 - (\sum x)^2}$$

where $\sigma_y = \sqrt{\frac{1}{N-2} \sum [y_i - (A + Bx_i)]^2}$ **is the uncertainty in the measurement of y, or the *rms* deviation of the measured to predicted value of y**

Least Squares Fits to Other Curves

a) Approaches

(1) Mathematical manipulation of equations to "linearize"

(2) Resort to probabilistic treatment on "least squares" approach used to find A and B

b) Straight line through origin, y=Bx

(1) Useful when you know definitely that y(x=0) = 0

(2) Probabilistic approach

(3) Taylor p. 198 and Problems 8.5 and 8.18

Slope $B = \frac{\sum xy}{\sum x^2}$ with $\sigma_B = \frac{\sigma_y}{\sum x^2}$

uncertainty of measurements in y

where $\sigma_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - Bx_i)^2}$

Least Squares Fits to Other Curves

1. Variations on linear regression
   a) Weighted fit for straight line
      (1) Useful when data point have different relative uncertainties
      (2) Probabilistic approach
      (3) Taylor pp. 196, 198 and Problems 8.9 and 8.19

Intercept: $$A = \frac{\sum wx^2 \sum wy - \sum wx \sum wxy}{N \sum wx^2 - (\sum wx)^2} \qquad \sigma_A = \sigma_y \frac{\sum wx^2}{N \sum wx^2 - (\sum wx)^2}$$

Slope $$B = \frac{N \sum wxy - \sum wx \sum wxy}{N \sum wx^2 - (\sum wx)^2} \qquad \sigma_B = \frac{\sum wx}{N \sum wx^2 - (\sum wx)^2}$$

Least Squares Fits to Other Curves

a) Polynomial
   (1) Useful when
      (a) formula involves more than one power of independent variable
      (e.g., $x(t) = (1/2)a\, t^2 + v_o \cdot t + x_o$)
      (b) as a power law expansion to unkown models
   (2) References
      (a) Taylor pp. 193-194
      (b) [Baird 6-11]

Fitting a Polynomial

FITTING A POLYNOMIAL

Often, one variable, y, is expected to be expressible as a polynomial in a second variable, x,

$$y = A + Bx + Cx^2 + \cdots + Hx^n. \tag{8.23}$$

For example, the height y of a falling body is expected to be quadratic in the time t,

$$y = y_0 + v_0 t - \tfrac{1}{2}gt^2,$$

where y_0 and v_0 are the initial height and velocity, and g is the acceleration of gravity. Given a set of observations of the two variables, we can find best estimates for the constants $A, B, \ldots, H$ in (8.23) by an argument that exactly parallels that of Section 8.2, as I now outline.

To simplify matters, we suppose that the polynomial (8.23) is actually a quadratic,

$$y = A + Bx + Cx^2. \tag{8.24}$$

(You can easily extend the analysis to the general case if you wish.) We suppose, as before, that we have a series of measurements (x_i, y_i), $i = 1, \ldots, N$, with the y_i all equally uncertain and the x_i all exact. For each x_i, the corresponding true value of y_i is given by (8.24), with A, B, and C as yet unknown. We assume that the measurements of the y_i are governed by normal distributions, each centered on the appropriate true value and all with the same width σ_y. This assumption lets us compute the probability of obtaining our observed values $y_1, \ldots, y_N$ in the familiar form

$$Prob(y_1, \ldots, y_N) \propto e^{-\chi^2/2}, \tag{8.25}$$

where now

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i - Cx_i^2)^2}{\sigma_y^2}. \tag{8.26}$$

[This equation corresponds to Equation (8.5) for the linear case.] The best estimates for A, B, and C are those values for which $Prob(y_1, \ldots, y_N)$ is largest, or χ^2 is smallest. Differentiating χ^2 with respect to A, B, and C and setting these derivatives equal to zero, we obtain the three equations (as you should check; see Problem 8.21):

$$\begin{aligned} AN + B\sum x + C\sum x^2 &= \sum y, \\ A\sum x + B\sum x^2 + C\sum x^3 &= \sum xy, \\ A\sum x^2 + B\sum x^3 + C\sum x^4 &= \sum x^2 y. \end{aligned} \tag{8.27}$$

Extend the linear solution to include on more tem, for a second order polynomial

This looks just like our linear problem, with the deviation in the summation replace by

$$[y_i - (A + Bx_i)] \to [y_i - (A + Bx_i + Cx_i{}^2)] \equiv [y_i - f_{fit}(x_i)]$$

This leads to a standard 3x3 eigenvalue problem,

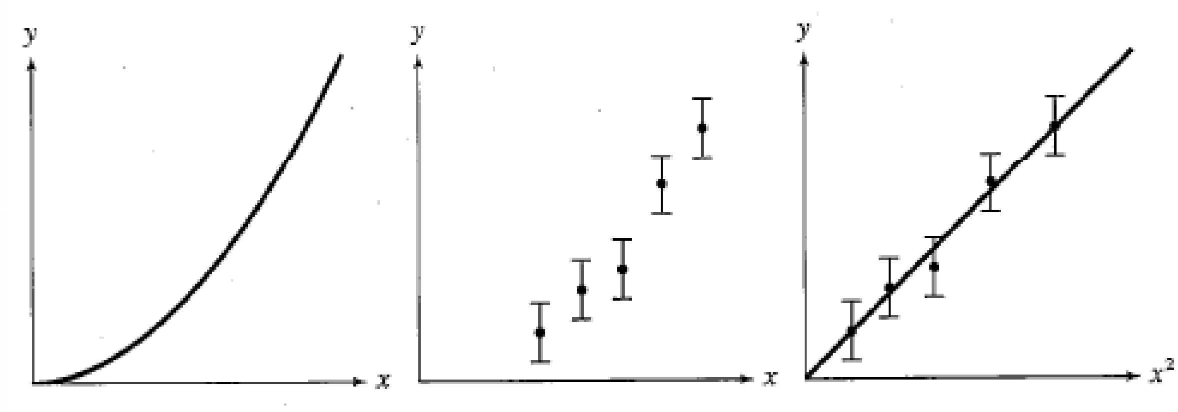
Which can easily be generalized to any order polynomial

Least Squares Fits to Other Curves

a) Exponential function
   (1) Useful for exponential models
   (2) "linearized" approach
      (a) recall semilog paper – a great way to quickly test model
      (b) recall linearizaion
         (i) $y = A\,e^{Bx}$
         (ii) $z = \ln(y) = \ln A + B \cdot x = A' + B \cdot x$
   (3) References
      (a) Taylor pp 194-196
      (b) Baird p. 137

Least Squares Fits to Other Curves

C. Power law
1. Useful for variable power law models
2. “linearized” approach
   a) recall log paper – a great way to quickly test model
   b) recall linearizaion
      (1) $y = A\,x^{B}$
      (2) $z = \ln A + B \cdot \ln(x) = A' + B \cdot w$
         (a) $z = \ln(y)$
         (b) $w = \ln(x)$
3. References
   a) Baird p. 136-137



(a) (b) (c)

Least Squares Fits to Other Curves

D. Sum of Trig functions
1. Useful when
    a) More than one trig function involved
    b) Used with trig identities to find other models
        (1) $A \sin(w \cdot t+b) = A \sin(wt)+B \cos(wt)$
2. References
    a) Taylor p.194 and Problems 8.23 and 8.24
    b) See detailed solution below

E. Multiple regression
1. Useful when there are two or more independent variables
2. References
    a) Brief introduction: Taylor pp. 196-197
3. More advanced texts: e.g., Bevington

Problem 8.24

8.24. ★★ A weight oscillating on a vertical spring should have height given by

$$y = A \cos \omega t + B \sin \omega t.$$

A student measures ω to be 10 rad/s with negligible uncertainty. Using a multiflash photograph, she then finds y for five equally spaced times, as shown in Table 8.10.

Table 8.10. Positions (in cm) and times (in tenths of a second) for an oscillating mass; for Problem 8.24.

"x": Time t	−4	−2	0	2	4
"y": Position y	3	−16	6	9	−8

Use Equations (8.41) to find best estimates for A and B. Plot the data and your best fit. (If you plot the data first, you will have the opportunity to consider how hard it would be to choose a best fit without the least-squares method.) If the student judges that her measured values of y were uncertain by "a couple of centimeters," would you say the data are an acceptable fit to the expected curve?

8.25 ★★ Th

Problem 8.24

Problem 8.24 Enter the data: Number of data points: $N := 5$ $n := 0..(N-1)$

$t_n :=$ $y_n :=$

t_n	y_n
$-4\cdot sec$	$3\cdot cm$
$-2\cdot sec$	$-16\cdot cm$
$0\cdot sec$	$6\cdot cm$
$2\cdot sec$	$9\cdot cm$
$4\cdot sec$	$-8\cdot cm$

$\omega := 1\cdot rad\cdot sec^{-1}$

To solve equations (8.41) for the coefficients A and B, where y = A·f(x) + B·g(x), we rewrite them in matrix format.

$$\begin{bmatrix} \sum_n \left(y_n \cdot f(t_n)\right) \\ \sum_n \left(y_n \cdot g(t_n)\right) \end{bmatrix} = \begin{bmatrix} \sum_n \left(f(t_n)\right)^2 & \sum_n \left(f(t_n)\cdot g(t_n)\right) \\ \sum_n \left(f(t_n)\cdot g(t_n)\right) & \sum_i \left(g(t_n)\right)^2 \end{bmatrix} \cdot \begin{pmatrix} A \\ B \end{pmatrix}$$

In shorthand notation, this is simply an eigen function equation of the form:

$$\begin{pmatrix} E1 \\ E2 \end{pmatrix} = \begin{pmatrix} M11\cdot A + M12\cdot B \\ M21\cdot A + M22\cdot B \end{pmatrix}$$

Problem 8.24

In shorthand notation, this is simply an eigen function equation of the form:

$$\begin{pmatrix} E1 \\ E2 \end{pmatrix} = \begin{pmatrix} M11 \cdot A + M12 \cdot B \\ M21 \cdot A + M22 \cdot B \end{pmatrix}$$

This can be solved symbollically to find A and B:

Given

$$M11 \cdot A + M12 \cdot B = E1$$

$$M21 \cdot A + M22 \cdot B = E2$$

$$\mathrm{Find}(A,B) \rightarrow \begin{pmatrix} \dfrac{E1 \cdot M22 - E2 \cdot M12}{M11 \cdot M22 - M12 \cdot M21} \\ -\dfrac{E1 \cdot M21 - E2 \cdot M11}{M11 \cdot M22 - M12 \cdot M21} \end{pmatrix}$$

That is, for A

$$A = \frac{(M22 \cdot E1 - E2 \cdot M12)}{(M11 \cdot M22 - M21 \cdot M12)}$$

Problem 8.24

$$A = \frac{(M22 \cdot E1 - E2 \cdot M12)}{(M11 \cdot M22 - M21 \cdot M12)}$$

Now, substituting the expressions for the shorthand notation, we arrive at the general form:

$$A = \frac{\left[\sum_n \left(g\left(t_n\right)\right)^2\right] \cdot \left[\sum_n \left(y_n \cdot f\left(t_n\right)\right)\right] - \left[\sum_n \left(y_n \cdot g\left(t_n\right)\right)\right] \cdot \left[\sum_n \left(f\left(t_n\right) \cdot g\left(t_n\right)\right)\right]}{\left[\sum_n \left(f\left(t_n\right)\right)^2\right] \cdot \left[\sum_n \left(g\left(t_n\right)\right)^2\right] - \left[\sum_n \left(f\left(t_n\right) \cdot g\left(t_n\right)\right)\right] \cdot \left[\sum_n \left(f\left(t_n\right) \cdot g\left(t_n\right)\right)\right]}$$

Finally, inserting the expressions f(x) = cos(ωt) and g(x) = sin(ωt) , we get:

$$A := \frac{\left[\sum_n \left[\sin\left[(\omega \cdot t)_n\right]\right]^2\right] \cdot \left[\sum_n \left[y_n \cdot \cos\left[(\omega \cdot t)_n\right]\right]\right] - \left[\sum_n \left[y_n \cdot \sin\left(\left(\omega \cdot t_n\right)\right)\right]\right] \cdot \left[\sum_n \left[\cos\left[(\omega \cdot t)_n\right] \cdot \sin\left[(\omega \cdot t)_n\right]\right]\right]}{\left[\sum_n \left[\cos\left[(\omega \cdot t)_n\right]\right]^2\right] \cdot \left[\sum_n \left(\sin\left(\omega \cdot t_n\right)\right)^2\right] - \left[\sum_n \left[\cos\left[(\omega \cdot t)_n\right] \cdot \sin\left[(\omega \cdot t)_n\right]\right]\right] \cdot \left[\sum_n \left[\cos\left[(\omega \cdot t)_n\right] \cdot \sin\left[(\omega \cdot t)_n\right]\right]\right]}$$

which yields a value for A of

$$A = 5.535\ \text{cm}$$

Problem 8.24

Likewise, for B we have

$$B = \frac{(M11 \cdot E2 - E1 \cdot M21)}{(M11 \cdot M22 - M21 \cdot M12)}$$

Now, substituting the expressions for the shorthand notation, we arrive at the general form:

$$B = \frac{\left[\sum_n \left(f(t_n)\right)^2\right] \cdot \left[\sum_n \left(y_n \cdot g(t_n)\right)\right] - \left[\sum_n \left(y_n \cdot f(t_n)\right)\right] \cdot \left[\sum_n \left(f(t_n) \cdot g(t_n)\right)\right]}{\left[\sum_n \left(f(t_n)\right)^2\right] \cdot \left[\sum_i \left(g(t_n)\right)^2\right] - \left[\sum_n \left(f(t_n) \cdot g(t_n)\right)\right] \cdot \left[\sum_n \left(f(t_n) \cdot g(t_n)\right)\right]}$$

Finally, inserting the expressions f(x) = cos(ωt) and g(x) = sin(ωt) , we get:

$$B := \frac{\left[\sum_n \left(\cos(\omega \cdot t_n)\right)^2\right] \cdot \left[\sum_n \left(y_n \cdot \sin(\omega \cdot t_n)\right)\right] - \left[\sum_n \left(y_n \cdot \cos(\omega \cdot t_n)\right)\right] \cdot \left[\sum_n \left(\cos(\omega \cdot t_n) \cdot \sin(\omega \cdot t_n)\right)\right]}{\left[\sum_n \left(\cos(\omega \cdot t_n)\right)^2\right] \cdot \left[\sum_n \left(\sin(\omega \cdot t_n)\right)^2\right] - \left[\sum_n \left(\cos(\omega \cdot t_n) \cdot \sin(\omega \cdot t_n)\right)\right] \cdot \left[\sum_n \left(\cos(\omega \cdot t_n) \cdot \sin(\omega \cdot t_n)\right)\right]}$$

which yields a value for B of

$$B = 11.095 \text{ cm}$$

Intermediate Lab

PHYS 3870

Correlations

Uncertaitiies in a Function of Variables

Consider an arbitrary function q with variables x,y,z and others.
Expanding the uncertainty in y in terms of partial derivatives, we have

$$\delta q(x,y,z,...) \approx \left|\frac{\partial q}{\partial x}\right| \cdot \delta x + \left|\frac{\partial q}{\partial y}\right| \cdot \delta y + \left|\frac{\partial q}{\partial z}\right| \cdot \delta z + ...$$

If x,y,z and others are independent and random variables, we have

$$\delta q(x,y,z,...) = \sqrt{\left(\frac{\partial q}{\partial x} \cdot \delta x\right)^2 + \left(\frac{\partial q}{\partial y} \cdot \delta x\right)^2 + \left(\frac{\partial q}{\partial z} \cdot \delta x\right)^2 + ...}$$

If x,y,z and others are independent and random variables governed by normal distributions, we have

$$\sigma_x = \sqrt{\left(\frac{\partial q}{\partial x} \cdot \sigma_x\right)^2 + \left(\frac{\partial q}{\partial y} \cdot \sigma_y\right)^2 + \left(\frac{\partial q}{\partial z} \cdot \sigma_z\right)^2 + ...}$$

We now consider the case when x,y,z and others are not independent and random variables governed by normal distributions.

Covariance of a Function of Variables

We now consider the case when x and y are not independent and random variables governed by normal distributions.

Assume we measure N pairs of data (x_i,y_i), with small uncertaities so that all x_i and y_i are close to their mean values X and Y.

Expanding in a Taylor series about the means, the value q_i for (x_i,y_i),

$$q_i = q(x_i, y_i)$$

$$q_i \approx q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y})$$

Note partial derivatives are all taken at X or Y and are hence the same for each i

We then find the simple result for the mean of q

$$\bar{q} = \frac{1}{N}\sum_{i=1}^{N} q_i = \frac{1}{N}\sum_{i=1}^{N}\left[q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}\cancelto{0}{(x_i - \bar{x})} + \frac{\partial q}{\partial y}\cancelto{0}{(y_i - \bar{y})}\right] \xrightarrow{yields} \bar{q} = q(\bar{x}, \bar{y})$$

The standard deviation of the N values of q_i is

$$\sigma_q{}^2 = \frac{1}{N}\sum_{i=1}^{N}[q_i - \bar{q}]^2$$

$$\sigma_q{}^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y})\right]^2$$

Covariance of a Function of Variables

The standard deviation of the N values of q_i is

$$\sigma_q{}^2 = \frac{1}{N}\sum_{i=1}^{N}[q_i - \bar{q}]^2$$

$$\sigma_q{}^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y})\right]^2$$

$$\sigma_q{}^2 = \left(\frac{\partial q}{\partial x}\right)^2 \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 + \left(\frac{\partial q}{\partial y}\right)^2 \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 + 2\left(\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right)\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_q{}^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x{}^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y{}^2 + 2\left(\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right)\sigma_{xy}$$

with $\sigma_{xy} \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$

If x and y are independent

$$\sigma_{xy} \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \rightarrow 0$$

Schwartz Inequality

Show that

$$|\sigma_{xy}| \leq \sigma_x \sigma_y$$

See problem 9.7

Define a function

$$A(t) \equiv \frac{1}{N}\sum_{i=1}^{N}[(x_i - \bar{X}) + t\cdot(y_i - \bar{Y})]^2 \geq 0$$

$A(t) \geq 0$, since the function is a square of real numbers.
Using the substitutions

$$\sigma_x \equiv \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(x_i - \bar{X})^2 \qquad \text{Eq. (4.6)}$$

$$\sigma_{xy} \equiv \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y}) \qquad \text{Eq. (9.8)}$$

$$A(t) = {\sigma_x}^2 + 2t\sigma_{xy} + t^2{\sigma_y}^2 \geq 0$$

Now find *t* for which *A(t_{min})* is a minimum:

$$\partial A(t)/_{\partial t} = 0 = 2\sigma_{xy} + 2t_{min}\cdot{\sigma_y}^2 \;\; \Longrightarrow \;\; t_{min} = -\sigma_{xy}/{\sigma_y}^2$$

Then since for any t, $A(t) \geq 0$

$$A_{min}(t_{min}) = {\sigma_x}^2 + 2\sigma_{xy}\left(-\sigma_{xy}/{\sigma_y}^2\right) + \left(-\sigma_{xy}/{\sigma_y}^2\right)^2{\sigma_y}^2 \geq 0$$

$$= {\sigma_x}^2 - \left(2\sigma_{xy}/\sigma_y\right)^2 + \left(\sigma_{xy}/\sigma_y\right)^2 \geq 0$$

$$= \left(\sigma_x + \sigma_{xy}/\sigma_y\right)\left(\sigma_x - \sigma_{xy}/\sigma_y\right) \geq 0$$

Multiplying through by ${\sigma_y}^2 \geq 0$

$$= \left(\sigma_x\sigma_y + \sigma_{xy}\right)\left(\sigma_x\sigma_y - \sigma_{xy}\right) \geq 0$$

which is true if

$$\left({\sigma_x}^2{\sigma_y}^2 - {\sigma_{xy}}^2\right) \geq 0 \quad \Longrightarrow \quad {\sigma_x}^2{\sigma_y}^2 \geq {\sigma_{xy}}^2$$

Now, since by definition $\sigma_x > 0$ and $\sigma_y > 0$,

$$\sigma_x\sigma_y \geq |\sigma_{xy}|, \quad \text{QED}$$

Schwartz Inequality

Combining the Schwartz inequality

$$|\sigma_{xy}| \le \sigma_x \sigma_y$$

With the definition of the covariance

$$\sigma_q{}^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x{}^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y{}^2 + 2\left(\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right)\sigma_{xy}$$

yields

$$\sigma_q{}^2 \le \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x{}^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y{}^2 + 2\left|\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right|\sigma_x \sigma_y$$

Then completing the squares

$$\sigma_q{}^2 \le \left[\left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y\right]^2$$

And taking the square root of the equation, we finally have

$$\sigma_q \le \left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y$$

At last, the upper bound of errors is

$$\sigma_q = \left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y$$

And for independent and random variables

$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial x}\cdot\sigma_x\right)^2 + \left(\frac{\partial q}{\partial y}\cdot\sigma_y\right)^2}$$

Another Useful Relation

Taylor Problem 4.5

Show $\sum_{i=1}^{N}[(x_i - \bar{x})]^2 = \sum_{i=1}^{N} x_i{}^2 - \frac{1}{N}\left[\sum_{i=1}^{N} x_i\right]^2$

Given $\sum_{i=1}^{N}[(x_i - \bar{x})]^2 = \sum_{i=1}^{N}[x_i{}^2 - 2x_i\bar{x} + \bar{x}^2]$

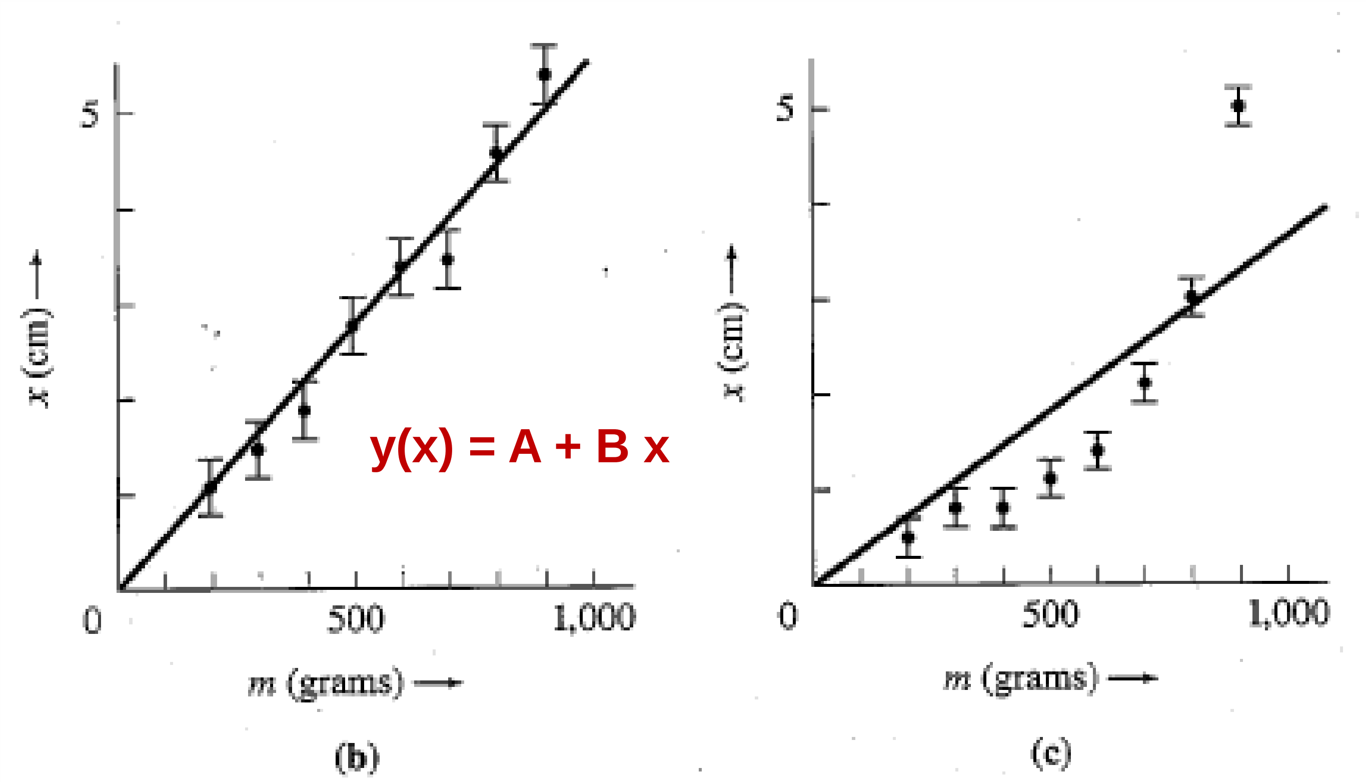
$$= \sum_{i=1}^{N}[x_i{}^2] - 2\bar{x}\sum_{i=1}^{N}[x_i] + \bar{x}^2\sum_{i=1}^{N}[i]$$

$$= \sum_{i=1}^{N}[x_i{}^2] - 2\bar{x}(N\bar{x}) + \bar{x}^2(N)$$

$$= \sum_{i=1}^{N}[x_i{}^2] - N\bar{x}^2$$

$$= \sum_{i=1}^{N}[x_i{}^2] - N\left[\sum_{i=1}^{N}[x_i]\right]^2, \text{ QED}$$

Question 2: Is it Linear?



Coefficient of Linear Regression: $r \equiv \frac{\sum[(x-\bar{\bar{x}})(y-\bar{\bar{y}})]}{\sqrt{\sum(x-\bar{\bar{x}})^2 \sum(y-\bar{\bar{y}})^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Consider the limiting cases for:

- r=0 (no correlation) [for any x, the sum over y-Y yields zero]
- r=±1 (perfect correlation). [Substitute yi-Y=B(xi-X) to get r=B/|B|=±1]

Tabulated Correlation Coefficient

Table C. The percentage probability $Prob_N(|r| \geq r_o)$ that N measurements of two uncorrelated variables give a correlation coefficient with $|r| \geq r_o$, as a function of N and r_o. (Blanks indicate probabilities less than 0.05%.)

r_o ⟵ r value

N data points

N	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
3	100	94	87	81	74	67	59	51	41	29	0
4	100	90	80	70	60	50	40	30	20	10	0
5	100	87	75	62	50	39	28	19	10	3.7	0
6	100	85	70	56	43	31	21	12	5.6	1.4	0
7	100	83	67	51	37	25	15	8.0	3.1	0.6	0
8	100	81	63	47	33	21	12	5.3	1.7	0.2	0
9	100	80	61	43	29	17	8.8	3.6	1.0	0.1	0
10	100	78	58	40	25	14	6.7	2.4	0.5		0
11	100	77	56	37	22	12	5.1	1.6	0.3		0
12	100	76	53	34	20	9.8	3.9	1.1	0.2		0
13	100	75	51	32	18	8.2	3.0	0.8	0.1		0
14	100	73	49	30	16	6.9	2.3	0.5	0.1		0
15	100	72	47	28	14	5.8	1.8	0.4			0
16	100	71	46	26	12	4.9	1.4	0.3			
17	100	70	44	24	11	4.1	1.1	0.2			
18	100	69	43	23	10	3.5	0.8	0.1			
19	100	68	41	21	9.0	2.9	0.7	0.1			
20	100	67	40	20	8.1	2.5	0.5	0.1			
25	100	63	34	15	4.8	1.1	0.2				
30	100	60	29	11	2.9	0.5					
35	100	57	25	8.0	1.7	0.2					
40	100	54	22	6.0	1.1	0.1					
45	100	51	19	4.5	0.6						0
	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	
50	100	73	49	30	16	8.0	3.4	1.3	0.4	0.1	
60	100	70	45	25	13	5.4	2.0	0.6	0.2		
70	100	68	41	22	9.7	3.7	1.2	0.3	0.1		
80	100	66	38	18	7.5	2.5	0.7	0.1			
90	100	64	35	16	5.9	1.7	0.4	0.1			
100	100	62	32	14	4.6	1.2	0.2				

Consider the limiting cases for:

- r=0 (no correlation)
- r=±1 (perfect correlation).

To gauge the confidence imparted by intermediate r values consult the table in Appendix C.

The values in Table C were calculated from the integral

$$Prob_N(|r| \geq |r_o|) = \frac{2\Gamma[(N-1)/2]}{\sqrt{\pi}\,\Gamma[(N-2)/2]} \int_{|r_o|}^{1} (1 - r^2)^{(N-4)/2}\, dr.$$

See, for example, E. M. Pugh and G. H. Winslow, *The Analysis of Physical Measurements* (Addison-Wesley, 1966), Section 12-8.

Probability that analysis of N=70 data points with a correlation coefficient of r=0.5 is **not** modeled well by a linear relationship is 3.7%.
Therefore, it is very probably that y is linearly related to x.

If

$Prob_N(|r|>r_o)<32\%$ ➔ it is probably that y is linearly related to x
$Prob_N(|r|>r_o)<5\%$ ➔ it is very probably that y is linearly related to x
$Prob_N(|r|>r_o)<1\%$ ➔ it is highly probably that y is linearly related to x

Uncertainties in Slope and Intercept

Taylor:

For the linear model y = A + B x

Intercept: $A = \frac{\sum x^2 \sum y - \sum x \sum xy}{N\sum x^2 - (\sum x)^2}$ $\sigma_A = \sigma_y \frac{\sum x^2}{N\sum x^2 - (\sum x)^2}$ **(Prob (8.16)**

Slope $B = \frac{N\sum xy - \sum x \sum xy}{N\sum x^2 - (\sum x)^2}$ $\sigma_B = \sigma_y \frac{N}{N\sum x^2 - (\sum x)^2}$

where $\sigma_y = \sqrt{\frac{1}{N-2}\sum[y_i - (A + Bx_i)]^2}$

Relation to R^2 value:

$$\sigma_A = \sigma_B \sqrt{\frac{1}{N}\sum x^2}$$

$$\sigma_B = B\sqrt{\frac{1}{N-2}[(1/R^2) - 1]}$$

$$r \equiv \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$