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?

\[ \text{Percent Difference} = \frac{(x_1 - x_2)}{\frac{1}{2}(x_1 + x_2)} \]
Comparison of precision or accuracy?

\[ \text{Precent Error} = \frac{x_{\text{measured}} - x_{\text{Known}}}{x_{\text{Known}}} \]
Multiple Measurements of the Same Quantity

Our statement of the best value and uncertainty is:

\[( \langle t \rangle \pm \sigma_t ) \text{ sec at the 68\% confidence level for N measurements} \]

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 how 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

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

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

It can be shown that (see Taylor Sec. 5.4) \(\sigma_t\) 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 \(<t> \pm \sigma_t\)
2. 95% of measurements of t will fall within \(<t> \pm 2\sigma_t\)
3. 98% of measurements of t will fall within \(<t> \pm 3\sigma_t\)
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 \(<t> \pm \sigma_t\)
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 $\langle t \rangle$.

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.
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**: \( (<t> \pm \sigma_t) \text{ sec} \)

II. **Relative (fractional) Error**: \(<t> \text{ sec} \pm (\sigma_t/<t>)\%\)

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 \( q_{\text{best}} = x_{\text{best}} \pm y_{\text{best}} \) the error is \( \delta q \approx \delta x + \delta y \)
   Follows from \( q_{\text{best}} \pm \delta q = (x_{\text{best}} \pm \delta x) \pm (y_{\text{best}} \pm \delta y) = (x_{\text{best}} \pm y_{\text{best}}) \pm (\delta x \pm \delta y) \)

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

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

4. Exponentiation (powers)  [Taylor, p. 56]
   For \( q_{\text{best}} = (x_{\text{best}})^n \) the error is \( (\delta q / q_{\text{best}}) \approx n (\delta x / x_{\text{best}}) \)
   Follows from 2 by setting \( \delta x / x_{\text{best}} = (\delta y / y_{\text{best}}) \)
General formula for uncertainty of a function of one variable

\[ \delta q = \left| \frac{\partial q}{\partial x} \right| \delta x \]  
[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

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

[Taylor, Eq. 3.47]

Worst case

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

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

[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

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

[Taylor, Eq. 3.48]

Best case

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).

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

Fixed, shifts peak of distribution

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

Best case

Error for a function of Two Variables: Addition in Quadrature

Fixed Distribution centered at X with width $\sigma_x$

Distribution centered at Y with width $\sigma_y$
Independent (Random) Uncertainties 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

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

- Distribution Function
- Independent Variable
- Center of Distribution (mean)
- Normalization Constant
- Width of Distribution (standard deviation)

Figure 5.10. Two normal, or Gauss, distributions.
Standard Deviation of Gaussian Distribution

\[
\text{Prob(\text{within } \sigma)} = \int_{X-\sigma}^{X+\sigma} G_{X,\sigma}(x) \, dx = \frac{1}{\sigma \sqrt{2\pi}} \int_{X-\sigma}^{X+\sigma} e^{-\frac{(x-X)^2}{2\sigma^2}} \, dx.
\]

Area under curve (probability that \(-\sigma < x < +\sigma\)) is 68%.

1% or \(~3\sigma\) “Highly Significant”
5% or \(~2\sigma\) “Significant”
1\sigma “Within errors”

5 ppm or \(~5\sigma\) “Valid for HEP”

See Sec. 10.6: Testing of Hypotheses

More complete Table in App. A and B

Figure 5.13. The probability \(\text{Prob(\text{within } t\sigma)}\) 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, \(\text{erf}(t)\).
Mean of Gaussian Distribution as "Best Estimate"

**Principle of Maximum Likelihood**

To find the most likely value of the mean (the best estimate of \( \bar{x} \)), 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

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

The combined probability for the full data set is the product

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

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

**Best Estimate of X is from maximum probability or minimum summation**

\[
\text{Minimize } \sum_{i=1}^{N} \frac{(x_i-X)^2}{\sigma} \quad \text{Solve for derivative set to 0} \quad \sum_{i=1}^{N} (x_i - X) = 0 \quad \text{Best estimate of X} \quad X_{\text{best}} = \frac{\sum x_i}{N}
\]
Uncertainty of “Best Estimates” of Gaussian Distribution

**Principle of Maximum Likelihood**

To find the most likely value of the mean (the best estimate of \( \bar{x} \)), 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

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

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

**Best Estimate of \( X \) is from maximum probability or minimum summation**

\[
\text{Minimize } \sum_{i=1}^{N} \frac{(x_i-X)^2}{\sigma} \quad \text{Solve for derivative wrst } X \text{ set to 0} \quad \sum_{i=1}^{N} (x_i-X) = 0 \quad \text{Best estimate of } X \quad X_{best} = \frac{\sum x_i}{N}
\]

**Best Estimate of \( \sigma \) is from maximum probability or minimum summation**

\[
\text{Minimize } \sum_{i=1}^{N} \frac{(x_i-X)^2}{\sigma} \quad \text{Solve for derivative wrst } \sigma \text{ set to 0} \quad \text{See Prob. 5.26} \quad \text{Best estimate of } \sigma \quad \sigma_{best} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{(x_i-X)^2}{\sigma}}
\]
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

\[
Prob_x(x_1, x_2) = Prob_x(x_1) \, Prob_x(x_2)
\]

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

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

**Best Estimate of \(\chi\) is from maximum probability or minimum summation**

**Minimize Sum**

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

**Solve for derivative wrst \(\chi\) 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 \(\chi\)**

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

This leads to

\[
x_{W\_avg} = \frac{w_1 x_1 + w_2 x_2}{w_1 + w_2} = \frac{\sum w_i \, x_i}{\sum w_i} \text{ where } w_i = \frac{1}{(\sigma_i)^2}
\]

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

\[
\sigma_{weighted \, avg} = \frac{1}{\sum_i w_i}
\]
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).

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

\[ y(x) = A + B \times \]
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:

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

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. 8)

(see Taylor Ch. 9)
Intermidiate 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 simple trick for many models is to linearize the model in the independent variable.

• Refer to Baird Ch.5 and the associated homework problems.
“Old School” graph paper is still a useful tool, especially for reality checking during the experimental design process.
Intermediate Lab
PHYS 3870

Linear Regression

References: Taylor Ch. 8
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 $\sigma_y$
Question 1: What is the Best Linear Fit (A and B)?

For the linear model \( y = A + Bx \)

Intercept: \[ A = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{N \Sigma x^2 - (\Sigma x)^2} \]

Slope \[ B = \frac{N \Sigma xy - \Sigma x \Sigma y}{N \Sigma x^2 - (\Sigma x)^2} \]

\( \sigma_A = \sigma_y \frac{\Sigma x^2}{N \Sigma x^2 - (\Sigma x)^2} \)

\( \sigma_B = \sigma_y \frac{N}{N \Sigma x^2 - (\Sigma x)^2} \)

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

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

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

To find the best simultaneous values for \( A \) and \( B \), find the maximum Prob or minimum \( \chi^2 \)

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

<table>
<thead>
<tr>
<th>Minimize Sum</th>
<th>Solve for derivative wrst A and B set to 0</th>
<th>Best estimate of A and B</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \chi^2 \equiv \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i)]^2}{\sigma_y^2} )</td>
<td>( \frac{\partial \chi^2}{\partial A} \equiv -\frac{2}{\sigma_y^2} \sum_{i=1}^{N} [y_i - (A + Bx_i)] = 0 )</td>
<td>( AN + B \sum x_i = \sum y_i )</td>
</tr>
<tr>
<td></td>
<td>( \frac{\partial \chi^2}{\partial B} \equiv -\frac{2}{\sigma_y^2} \sum_{i=1}^{N} x_i [y_i - (A + Bx_i)] = 0 )</td>
<td>( A \sum x_i + A \sum x_i^2 = \sum x_i y_i )</td>
</tr>
</tbody>
</table>
**“Best Estimates” of Linear Fit**

Best Estimates of A and B are from maximum probability 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

\[
\begin{bmatrix} N \sum x_i \sum x_i^2 \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}
\]

This is a standard eigenvalue problem

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

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 \( y = A + B \cdot x \)

Intercept:

\[
A = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{N \Sigma x^2 - (\Sigma x)^2}
\]

\[
\sigma_A = \sigma_y \frac{\Sigma x^2}{N \Sigma x^2 - (\Sigma x)^2}
\]

Slope

\[
B = \frac{N \Sigma xy - \Sigma x \Sigma xy}{N \Sigma x^2 - (\Sigma x)^2}
\]

\[
\sigma_B = \sigma_y \frac{N}{N \Sigma x^2 - (\Sigma x)^2}
\]

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 root-mean-square deviation of the measured to predicted value of y

For the linear model \( y = A + B \cdot x \)

**Intercept:**

\[
A = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{N \Sigma x^2 - (\Sigma x)^2}
\]

\[
\sigma_A = \sigma_y \frac{\Sigma x^2}{N \Sigma x^2 - (\Sigma x)^2}
\]

**Slope**

\[
B = \frac{N \Sigma xy - \Sigma x \Sigma xy}{N \Sigma x^2 - (\Sigma x)^2}
\]

\[
\sigma_B = \sigma_y \frac{N}{N \Sigma x^2 - (\Sigma x)^2}
\]

**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 root-mean-square 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

\[
B = \frac{\sum xy}{\sum x^2} \quad \text{with} \quad \sigma_B = \frac{\sigma_y}{\sum x^2}
\]

Slope

uncertainty of measurements in y

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

where
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:
\[ A = \frac{\sum \omega x^2 \sum wy - \sum \omega x \sum wy}{N \sum \omega x^2 - (\sum \omega x)^2} \]
\[ \sigma_A = \sigma_y \frac{\sum \omega x^2}{N \sum \omega x^2 - (\sum \omega x)^2} \]

Slope
\[ B = \frac{N \sum \omega xy - \sum \omega x \sum wy}{N \sum \omega x^2 - (\sum \omega x)^2} \]
\[ \sigma_B = \frac{\sum \omega x}{N \sum \omega x^2 - (\sum \omega x)^2} \]
Least Squares Fits to Other Curves

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

This leads to a standard 3x3 eigenvalue problem,

Which can easily be generalized to any order polynomial.

Extend the linear solution to include on more term, for a second order polynomial.
a) Exponential function
   (1) Useful for exponential models
   (2) “linearized” approach
       (a) recall semilog paper – a great way to quickly test model
       (b) recall linearization
           (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 linearization
      (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
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
         \[ A \sin(w \cdot t + b) = A \sin(wt) + B \cos(wt) \]
   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 \( \omega \) 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.

<table>
<thead>
<tr>
<th>“x”: Time ( t )</th>
<th>-4</th>
<th>-2</th>
<th>0</th>
<th>2</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>“y”: Position ( y )</td>
<td>3</td>
<td>-16</td>
<td>6</td>
<td>9</td>
<td>-8</td>
</tr>
</tbody>
</table>

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?
Problem 8.24

Enter the data:

Number of data points: \( N := 5 \quad n := 0..(N - 1) \)

\[
\begin{array}{|c|c|}
\hline
 t_n & y_n \\
\hline
-4\text{-sec} & 3\text{-cm} \\
-2\text{-sec} & -16\text{-cm} \\
0\text{-sec} & 6\text{-cm} \\
2\text{-sec} & 9\text{-cm} \\
4\text{-sec} & -8\text{-cm} \\
\hline
\end{array}
\]

\( \omega := 1\text{-rad}\cdot\text{sec}^{-1} \)

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

\[
\begin{bmatrix}
\sum_{n} (y_n \cdot f(t_n)) \\
\sum_{n} (y_n \cdot g(t_n))
\end{bmatrix} =
\begin{bmatrix}
\sum_{n} (f(t_n))^2 & \sum_{n} (f(t_n) \cdot g(t_n)) \\
\sum_{n} (f(t_n) \cdot g(t_n)) & \sum_{i} (g(t_n))^2
\end{bmatrix}
\begin{bmatrix}
A \\
B
\end{bmatrix}
\]

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

\[
\begin{pmatrix}
E_1 \\
E_2
\end{pmatrix} =
\begin{pmatrix}
M11 \cdot A + M12 \cdot B \\
M21 \cdot A + M22 \cdot B
\end{pmatrix}
\]
Problem 8.24

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

\[
\begin{pmatrix}
E_1 \\
E_2
\end{pmatrix} = \begin{pmatrix}
M_{11} \cdot A + M_{12} \cdot B \\
M_{21} \cdot A + M_{22} \cdot B
\end{pmatrix}
\]

This can be solved symbolically to find A and B:

Given

\[M_{11} \cdot A + M_{12} \cdot B = E_1\]

\[M_{21} \cdot A + M_{22} \cdot B = E_2\]

Find \((A, B)\) →

\[
\begin{pmatrix}
\frac{E_1 \cdot M_{22} - E_2 \cdot M_{12}}{M_{11} \cdot M_{22} - M_{12} \cdot M_{21}} \\
\frac{M_{11} \cdot M_{22} - M_{12} \cdot M_{21}}{M_{11} \cdot M_{22} - M_{12} \cdot M_{21}}
\end{pmatrix}
\]

That is, for A

\[A = \frac{(M_{22} \cdot E_1 - E_2 \cdot M_{12})}{(M_{11} \cdot M_{22} - M_{21} \cdot M_{12})}\]
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:

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

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

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

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:

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

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

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

which yields a value for \( B \) of

\[
B = 11.095 \text{ cm}
\]
Intermediate Lab
PHYS 3870

Correlations
Uncertainties 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

\[
\delta q(x, y, z, \ldots) \approx \left| \frac{\partial q}{\partial x} \right| \delta x + \left| \frac{\partial q}{\partial y} \right| \delta y + \left| \frac{\partial q}{\partial z} \right| \delta z + \ldots
\]

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

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

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

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

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 \((x_i,y_i)\),

\[
q_i = q(x_i, y_i)
\]

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

We then find the simple result for the mean of \( q \)

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

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

\[
\sigma_{xy} \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \to 0
\]
**Schwartz Inequality**

Show that

\[ |\sigma_{xy}| \leq \sigma_x \sigma_y \]

See problem 9.7

Define a function

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

\[ A(t) \geq 0, \text{ since the function is a square of real numbers.} \]

Using the substitutions

\[
\sigma_x \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})^2 \quad \text{Eq. (4.6)}
\]

\[
\sigma_{xy} \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y}) \quad \text{Eq. (9.8)}
\]

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

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

\[
\frac{\partial A(t)}{\partial t} = 0 = 2\sigma_{xy} + 2t_{\text{min}} \cdot \sigma_y^2 \quad \Rightarrow \quad t_{\text{min}} = -\frac{\sigma_{xy}}{\sigma_y^2}
\]

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

\[
A_{\text{min}}(t_{\text{min}}) = \sigma_x^2 + 2\sigma_{xy} \left(-\frac{\sigma_{xy}}{\sigma_y^2}\right) + \left(-\frac{\sigma_{xy}}{\sigma_y^2}\right)^2 \sigma_y^2 \geq 0
\]

\[
= \sigma_x^2 - \left(2\frac{\sigma_{xy}}{\sigma_y}\right)^2 + \left(\frac{\sigma_{xy}}{\sigma_y}\right)^2 \geq 0
\]

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

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

\[
\left(\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2 \right) \geq 0 \quad \Rightarrow \quad \sigma_x^2 \sigma_y^2 \geq \sigma_{xy}^2
\]

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

\[ |\sigma_{xy}| \leq \sigma_x \sigma_y \]

With the definition of the covariance

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

yields

Then completing the squares

\[ \sigma_q^2 \leq \left(\left|\frac{\partial q}{\partial x}\right| \sigma_x + \left|\frac{\partial q}{\partial y}\right| \sigma_y\right)^2 \]

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

\[ \sigma_q \leq \left|\frac{\partial q}{\partial x}\right| \sigma_x + \left|\frac{\partial q}{\partial y}\right| \sigma_y \]

At last, the upper bound of errors is

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

And for independent and random variables
Another Useful Relation

Taylor Problem 4.5

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

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

$= \sum_{i=1}^{N} [x_i^2] - 2\bar{x} \sum_{i=1}^{N} [x_i] + \bar{x}^2 \sum_{i=1}^{N} [i]$  
$= \sum_{i=1}^{N} [x_i^2] - 2\bar{x}(N\bar{x}) + \bar{x}^2(N)$  
$= \sum_{i=1}^{N} [x_i^2] - N\bar{x}^2$  
$= \sum_{i=1}^{N} [x_i^2] - N[\sum_{i=1}^{N} [x_i]]^2$, QED
Question 2: Is it Linear?

Coefficient of Linear Regression:

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

Consider the limiting cases for:

- \( r=0 \) (no correlation) \[ \text{[for any } x, \text{ the sum over } y-Y \text{ yields zero]} \]
- \( r=\pm 1 \) (perfect correlation). \[ \text{[Substitute } y_i-Y=B(x_i-X) \text{ to get } r=B/|B|=\pm 1] \]
Tabulated Correlation Coefficient

Consider the limiting cases for:
- \( r=0 \) (no correlation)
- \( r=\pm 1 \) (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

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

For a correlation analysis of \( N=70 \) data points with a correlation coefficient of \( r=0.5 \), the probability that it 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 \( x \)

---

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

<table>
<thead>
<tr>
<th>( N )</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>100</td>
<td>94</td>
<td>87</td>
<td>81</td>
<td>74</td>
<td>67</td>
<td>59</td>
<td>51</td>
<td>41</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td>60</td>
<td>50</td>
<td>40</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>87</td>
<td>75</td>
<td>62</td>
<td>50</td>
<td>39</td>
<td>28</td>
<td>19</td>
<td>10</td>
<td>3.7</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>85</td>
<td>70</td>
<td>56</td>
<td>43</td>
<td>31</td>
<td>21</td>
<td>12</td>
<td>5.6</td>
<td>1.4</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>83</td>
<td>67</td>
<td>51</td>
<td>37</td>
<td>25</td>
<td>15</td>
<td>8.0</td>
<td>3.1</td>
<td>9.6</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>81</td>
<td>63</td>
<td>47</td>
<td>33</td>
<td>21</td>
<td>12</td>
<td>5.3</td>
<td>1.7</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>80</td>
<td>61</td>
<td>43</td>
<td>29</td>
<td>17</td>
<td>8.8</td>
<td>3.6</td>
<td>1.0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>78</td>
<td>58</td>
<td>40</td>
<td>25</td>
<td>14</td>
<td>6.7</td>
<td>2.4</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

N data points

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 \( x \).
Uncertainties in Slope and Intercept

Taylor:

For the linear model \( y = A + B x \)

\[
A = \frac{\sum x^2 \sum y - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}
\]

\[
\sigma_A = \sigma_y \frac{\sum x^2}{N \sum x^2 - (\sum x)^2}
\]

(Prob (8.16))

\[
B = \frac{N \sum xy - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}
\]

\[
\sigma_B = \sigma_y \frac{N}{N \sum x^2 - (\sum x)^2}
\]

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

Relation to \( R^2 \) value:

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

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

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