Statistics & Probability

What is error?

Two kinds: Random & Systematic

Random error is the scatter you get from repeated measurements

Systematic error is systematic: It happens to the same degree each time you make a measurement.

Multiple measurements will average out (reduce) random error, but won't affect systematic error.

If you do many measurements, the error in the avg. will be completely dominated by systematic error, but this is very difficult to estimate.

Be wary of error estimates - always look to see how it is calculated.
As an engineer, a detailed understanding of statistics and error is **critical**; you must master this material.

OK, now for some elementary statistics:

The most important **probability distribution** is the Gaussian or normal distribution:

\[ p(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\[ p(x) = \text{probability density} \]

= probability of measurement being in interval \([x, x+\Delta x]\) divided by \(\Delta x\)

We can define the cumulative probability:
\[ P(x) = \int_{-\infty}^{x} p(x') \, dx' \]

This is the probability that a meas.
is less than some value \( x \).

The normal distribution is characterized by 2 quantities:

\[ \mu = \text{mean of distribution} \]
\[ \sigma = \text{population standard deviation} \]

About 69% of observations lie within 1\( \sigma \) of \( \mu \), 95% within 2\( \sigma \), and 99% within 3\( \sigma \).

In working my statistics, we define an expectation value

\[ E(x) = \text{what you expect to get if you do a meas. many times and average it together!} \]
Mathematically:

\[ E(x) = \int_{-\infty}^{\infty} x \ p(x) \, dx \]

\[ = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx \]

\[ = \int_{-\infty}^{\infty} \frac{x-\mu}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx + \mu \int_{-\infty}^{\infty} p(x) \, dx \]

Let \( z = x - \mu \)

\[ E(x) = \int_{-\infty}^{\infty} \frac{z}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}} \, dz + \mu \int_{-\infty}^{\infty} p(x) \, dx \]

\[ \int_{-\infty}^{0} = 0 \quad \text{odd} \]

\[ \int_{0}^{\infty} = 1 \quad \text{even} \]

The integral of an odd function over an even domain is zero, so first integral vanishes!
Thus $E(X) = \mu$

What about the mean of $N$ observations?

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$\therefore E(\bar{X}) = E \left( \frac{1}{N} \sum_{i=1}^{N} X_i \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} E(X_i) = \frac{1}{N} \sum_{i=1}^{N} \mu = \mu$$

This may have been self-evident, but we can use the same approach for the variance:

What is $E((X-\mu)^2)$?

$$E((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 \rho(x) \, dx$$

$$= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2 \sigma^2}} \, dx$$

Let $Z = (X-\mu)$
\[ E((x - \mu)^2) = \int_{-\infty}^{\infty} \frac{e^{-\frac{z^2}{2\sigma^2}}}{\sqrt{2\pi}} \, dz \]

Integrate by parts,

\[ = -\frac{\sigma^2}{\sqrt{2\pi}} \left[ e^{-\frac{z^2}{2\sigma^2}} \right]_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}} \, dz \]

vanishes at \( \pm \infty \) = 1

\[ \therefore E((x - \mu)^2) = \sigma^2 \]

What about the variance of the mean of \( N \) measurements?

\[ E((\bar{x} - \mu)^2) \]

\[ = E\left( \left( \frac{1}{N} \sum_{i=1}^{N} x_i - \mu \right)^2 \right) \]

\[ = \frac{1}{N^2} E\left( \left( \sum_{i=1}^{N} x_i - N\mu \right)^2 \right) = \frac{1}{N^2} E\left( \left( \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu) \right)^2 \right) \]

\[ = \frac{1}{N^2} E\left( \sum_{i=1}^{N} (x_i - \mu)^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - \mu)(x_j - \mu) \right) \]
\[ = \frac{1}{N} \mathbb{E} \left( \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \right) \\
\quad + \mathbb{E} \left( \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} (x_i - \mu)(x_j - \mu) \right) \\
\quad = \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E} ( (x_i - \mu)^2 ) \\
\quad + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} \mathbb{E} ( (x_i - \mu)(x_j - \mu) ) \\
\]  

First expectation is just \( \sigma^2 \).

Second term is zero provided \( x_i \) and \( x_j \) are independent.

\[ \mathbb{E} (x_i - \mu)(x_j - \mu) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu)(x_j - \mu) \cdot p(x_i) p(x_j) \, dx_i \, dx_j \]

\[ \equiv \int_{-\infty}^{\infty} (x_i - \mu) \cdot p(x_i) \, dx_i \int_{-\infty}^{\infty} (x_j - \mu) \cdot p(x_j) = 0 \quad i \neq j \]
Thus:

\[ \sigma_{\bar{x}}^2 = \frac{\sigma^2}{N} \]

That is why multiple measurements reduce the random error!

Okay, how do we estimate \( \mu \) and \( \sigma \)?

In general we have a sample of \( N \) observations drawn from an underlying population of possible observations.

We define the sample mean

\[ \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \]

We have already shown that

\[ E(\bar{X}) = \mu \]

That is, \( \bar{X} \) is an unbiased estimate of \( \mu \).
Now for the variance:

We define the sample variance:

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

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

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

\[ = \frac{1}{N-1} \left\{ N\bar{x}^2 - 2N\bar{x} + N\bar{x}^2 \right\} \]

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

Note that to calculate this we do not have to store all the \( x_i \)'s!

We just need to keep \( N, \sum_{i=1}^{N} x_i \)

and \( \sum_{i=1}^{N} x_i^2 \Rightarrow \) that's how a calculator with only a few memories can calculate \( S_x^2 \).
Ok, how is $s_x^2$ related to $\sigma^2$?

We need the following identities:

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

(this is because $\sum_{i=1}^{N} (x_i - \bar{x}) = 0$ by the definition of $\bar{x}$)

Also, we had:

$$E((\bar{x} - \mu)^2) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$$

and

$$E((x_i - \mu)^2) = \sigma^2$$

Thus:

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

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

So \( s^2 \) is an unbiased estimate for \( \sigma^2 \).

The \( N-1 \) part comes about because we are calculating the mean from the same sample that we are calculating the variance from. We have reduced the number of degrees of freedom by one.

---

**Propagation of Errors:**

Usually when doing experimental calculations, you must combine several measurements to get the final result.

Each individual measurement has some error associated with it. How do these combine?
Let \( x_i, y_i \) be random variables with mean \( \bar{x}, \bar{y} \) and st. dev. \( s_x \) & \( s_y \).

Let \( Z_i = C_1 x_i + C_2 y_i \)

where \( C_1 \) & \( C_2 \) are csts.

\[
\overline{Z} = \frac{1}{N} \sum_{i=1}^{N} Z_i = C_1 \overline{x} + C_2 \overline{y}
\]

What about the variance?

\[
\sigma_Z^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Z_i - \overline{Z})^2
\]

\[
= \frac{1}{N-1} \left\{ \sum_{i=1}^{N} (Z_i^2 - \overline{Z}^2) \right\}
\]

Substituting in:

\[
Z_i^2 = C_1^2 x_i^2 + C_2^2 y_i^2 + 2 C_1 C_2 x_i y_i
\]

\[
\overline{Z}^2 = C_1^2 \overline{x}^2 + C_2^2 \overline{y}^2 + 2 C_1 C_2 \overline{x} \overline{y}
\]
\[ s_e^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left( e_i^2 (x_i - \bar{x})^2 + c_2^2 (y_i - \bar{y})^2 \right) \]

\[ + 2 c_1 c_2 (x_i y_i - \bar{x} \bar{y}) \]

\[ = c_1^2 s_x^2 + c_2^2 s_y^2 + \frac{2 c_1 c_2}{N-1} \sum_{i=1}^{N} (x_i y_i - \bar{x} \bar{y}) \]

Let's look at this last term. We define:

\[ s_{xy}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i y_i - \bar{x} \bar{y}) \]

\[ = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \]

This is the covariance of \( x \) and \( y \). If they are independent, then:

\[ E(s_{xy}^2) = 0 \]

So for addition:

\[ s_e^2 = c_1^2 s_x^2 + c_2^2 s_y^2 + \frac{2 c_1 c_2}{N-1} s_{xy}^2 \]
If we subtract two variables we get a similar result:

\[ z_i = x_i - y_i \]

then \[ \bar{z} = \bar{x} - \bar{y} \]

\[ s^2_z = s^2_x + s^2_y - 2s_{xy} \]

Note the sign!

A positive covariance (note: \( s_{xy} \) may be + or -) reduces the error in the difference between variables.

OK, how about multiplication and division?

Let \[ z_i = \frac{x_i}{y_i} \]

where \( x_i \) and \( y_i \) are normally distributed.

Note that \( z \) is not normally distributed! It only approaches this if \( \frac{s_x}{x} \) and \( \frac{s_y}{y} \ll 1 \)!
Let's look at this more closely.

Let \( x_i = \bar{x} + x'_i \)

\[ \text{mean deviation} \]

and \( y_i = \bar{y} + y'_i \)

Suppose \( \frac{sx}{x} \ll 1 \) and \( \frac{sy}{y} \ll 1 \)

This means that \( x'_i \) and \( y'_i \) are small:

\[ z_i = \frac{x'_i}{y'_i} = \frac{\bar{x}_i + x'_i}{\bar{y}_i + y'_i} = \frac{x}{y} \left( 1 + \frac{x'_i}{x} \right) \left( 1 + \frac{y'_i}{y} \right) \]

\[ \approx \frac{x}{y} \left( 1 + \frac{x'_i}{x} \right) \left( 1 - \frac{y'_i}{y} + O\left( \frac{(y'_i)^2}{y^2} \right) \right) \]

\[ \approx \frac{x}{y} \left( 1 + \frac{x'_i}{x} - \frac{y'_i}{y} + O\left( \frac{y'_i}{x} + \frac{(y'_i)^2}{y^2} \right) \right) \]

We ignore the higher order terms!
\[ \bar{z} = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{y_i} \approx \frac{1}{y} \sum_{i=1}^{N} \left( 1 + \frac{x_i}{x} + \frac{y_i}{y} \right) \]

\[ \therefore \bar{z} \approx \frac{x}{y} \]

Now for \( S_{z^{2}} \):

\[ S_{z^{2}} \approx \left( \frac{\bar{x}^{2}}{x^{2}} \right) \left( \frac{S_{x}^{2}}{x^{2}} + \frac{S_{y}^{2}}{y^{2}} - 2 \frac{S_{xy}}{xy} \right) \]

\[ \approx \frac{\bar{x}^{2}}{x^{2}} \]

Thus:

\[ \frac{S_{z^{2}}}{\bar{z}^{2}} = \frac{S_{x}^{2}}{x^{2}} + \frac{S_{y}^{2}}{y^{2}} - 2 \frac{S_{xy}}{xy} \]

so for division you add the fractional or relative variance!
For multiplication you get the same:

\[ z_i = x_i y_i \]
\[ z \approx x \cdot y \quad \text{provided} \quad \frac{s_x}{x}, \frac{s_y}{y} \ll 1 \]

and

\[ \frac{s_{z_i}^2}{s_{x_i}^2} \approx \frac{s_x^2}{x^2} + \frac{s_y^2}{y^2} + 2 \frac{s_{xy}}{x \cdot y} \quad \uparrow \]

note (+) Sign

What about a more general functional relationship?

Suppose \( z_i = f(x_i, y_i) \) where \( f \) is some nasty function.

Let \( x_i = x + x_i' \), \( y_i = y + y_i' \)

We shall expand \( f \) in a 2-D
Taylor series (Note: the T.S. can be generalized to $n$-dimensions)

So:

$$f(x_i, y_i) = f(x, y) + x_i \frac{\partial f}{\partial x}\bigg|_{x, y} + y_i \frac{\partial f}{\partial y}\bigg|_{x, y} + O(x_i^2, y_i^2, x_i y_i)$$

(Ignore second order terms!)

\[\therefore \bar{z} = \frac{1}{N} \sum_{i=1}^{N} f(x_i, y_i)\]

\[\approx f(x, y) + \frac{\partial f}{\partial x}\bigg|_{x, y} \frac{1}{N} \sum_{i=1}^{N} x_i + \frac{\partial f}{\partial y}\bigg|_{x, y} \frac{1}{N} \sum_{i=1}^{N} y_i\]

\[= f(x, y)\]

The error is $O(x_i^2, y_i^2, x_i y_i \cdot \text{mf})$
Similarly, we get for the variance:

$$s_2^2 = \left( \frac{\partial f}{\partial x} \bigg|_{\bar{x}, \bar{y}} \right)^2 s_x^2 + \left( \frac{\partial f}{\partial y} \bigg|_{\bar{x}, \bar{y}} \right)^2 s_y^2$$

$$+ 2 \left( \frac{\partial f}{\partial x} \bigg|_{\bar{x}, \bar{y}} \right) \left( \frac{\partial f}{\partial y} \bigg|_{\bar{x}, \bar{y}} \right) s_{xy}$$

again providing that the higher order terms can be neglected.

This formula reduces to the earlier ones for addition, substr., mult., & div.

Only for add & substr. is it exact because $\nabla \nabla f = 0$ for this case and the higher order terms vanish.

Let's generalize this result to an $n$-dimensional ($n$-variable) system

Let $\mathbf{x} = x_1, x_2, \ldots, x_n$
\[ z = f(x) \]

we need to find the variance of each variable \( x_1, x_2, \text{ etc.} \) and the covariance of all pairs \( S_{x_1x_2}, \text{ etc.} \).

We can define the matrix of variance & covariance

\[
V_{ij} = \frac{N}{N-1} \left( \bar{x}_i \bar{x}_j - \bar{x}_i \bar{x}_j \right)
\]

\( \uparrow \)

number of elements contributing to each \( x_i, x_j \)

Now if \( \frac{V_{ii}}{x_i^2} << 1 \) for all \( i \)

(this corresponds to a small error in each variable, e.g. \( \frac{s^2}{x_i^2} << 1 \), \( \frac{s^2}{y^2} << 1 \), etc.)

And if we define:

\[ \nabla f = \left( \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}, \ldots, \frac{\partial f}{\partial x_n} \right) \]
Then we get the result:

$$s_x^2 = \left( \Delta f \right) \left( \Delta f \right)^T$$

Which gives us an easy way of calculating the variance in $z = f(x)$!

Note: in this derivation the subscripts $i$ and $j$ denote different variables entirely, not different measurements contributing to the same average.

Let's look at the example $f(x) = x_1 x_2$

Let $x_1 = x$, $x_2 = y$

$$V_{ij} = \frac{N}{N-1} \left( \bar{x}_i \bar{x}_j - \bar{x}_i \bar{x}_j \right)$$

so

$$V_{xx} = \frac{N}{N-1} \left( \bar{x} \bar{x} - \bar{x} \bar{x} \right)$$

$$= \frac{1}{N-1} \sum (x^2 - \bar{x}^2) = s_x^2$$
\[ v_{xy} = \frac{N}{N-1} \left( \bar{x} \bar{y} - \bar{x} \bar{y} \right) = \frac{1}{N-1} \sum (x \bar{y} - \bar{x} \bar{y}) \equiv s_{xy}^2 \]

\[ \nabla = \begin{pmatrix} s_x^2 & s_{xy}^2 \\ s_{xy}^2 & s_y^2 \end{pmatrix} \]

\[ \nabla f = \left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) \equiv (y, x) \]

\[ \nabla f \nabla (\nabla f)^T = (y, x) \nabla \begin{pmatrix} y \\ x \end{pmatrix} \]

\[ = (y s_x^2 + x s_{xy}^2, y s_{xy}^2 + x s_y^2) \begin{pmatrix} y \\ x \end{pmatrix} \]

\[ = y^2 s_x^2 + xy s_{xy}^2 + xy s_{xy}^2 + x^2 s_y^2 \]

\[ = y^2 s_x^2 + x^2 s_y^2 + 2xy s_{xy}^2 \]

So

\[ \frac{s_{xy}^2}{(xy)^2} = s_x^2 \frac{s_y^2}{y^2} + 2 \frac{s_{xy}^2}{xy} \]
Finally, we examine vector functions of vector variables:

Let \( \mathbf{z} = f(\mathbf{x}) \) be a vector function of \( \mathbf{x} \). We define:

\[
\mu_\mathbf{z} \equiv E(\mathbf{z}), \quad \mu_\mathbf{x} = E(\mathbf{x})
\]

and the matrix of covariance for each:

\[
\Sigma^2_\mathbf{z} = E\left[ (\mathbf{z} - \mu_\mathbf{z})(\mathbf{z} - \mu_\mathbf{z})^T \right]
\]

\[
\Sigma^2_\mathbf{x} = E\left[ (\mathbf{z} - \mu_\mathbf{z})(\mathbf{z} - \mu_\mathbf{z})^T \right]
\]

we have taken \( \mathbf{z} \) and \( \mathbf{x} \) to be column vectors.

Suppose we know \( \mu_\mathbf{x} \) and \( \Sigma^2_\mathbf{x} \). We wish to calculate \( \Sigma^2_\mathbf{z} \).

We do a multivariable Taylor series expansion of \( \mathbf{z} \) about \( \mu_\mathbf{x} \):
\[ z = f(x) = f(\mu_x) + \nabla f(\mu_x) \cdot (x - \mu_x) \]

\[ + \frac{1}{2} \nabla^2 f(\mu_x)(x - \mu_x)^T + \ldots \]

If the deviation between \( x \) and \( \mu_x \) is small (e.g. that \( \nabla^2 f(\mu_x) \) is small), or if \( f(x) \) is linear in \( x \), then we can ignore the quadratic or higher order terms in the Taylor series.

Thus:

\[ z = f(\mu_x) + \nabla f(\mu_x) \cdot (x - \mu_x) \]

and thus:

\[ E(z) = E(f(\mu_x)) + \nabla f(\mu_x) \cdot E(x - \mu_x) \]

but \( E(x - \mu_x) = 0 \) by definition.

Thus, \( \mu_z \approx f(\mu_x) + \text{quadratic terms} \).
Putting this together, we get:

\[ \frac{\partial^2 f}{\partial x^2} \approx \nabla^2 f \left|_{x=x_0} \right. \]

So:

\[ \sum_{x} \nabla^2 f \approx E \left( (x - \mu_x)(x - \mu_x)^T \right) \]

\[ = E \left[ (\nabla f) (x - \mu_x)(x - \mu_x)^T (\nabla f)^T \right] \]

\[ = (\nabla f) \left( E((x - \mu_x)(x - \mu_x)^T) \right)(\nabla f)^T \]

\[ = (\nabla f) \sum_{x} \nabla^2 f \]

provided that \( \frac{1}{\sum_{x} \nabla^2 f} \ll f \)

so that we can neglect the quadratic term in the Taylor series expansion.

We will use this formulation again in calculating linear regression error!
Linear Regression

In HW problem, wrote routine to fit a parabola to 3 points. In this case there was only one answer.

Often in data analysis you want to fit a curve to many data points. How do we do this? Use (usually) linear regression.

This does not mean that we fit data with a line, rather that we use an algorithm which reduces the problem to a set of linear equations!

Let’s look at a simple problem: measure mass transfer coefficients.

Suppose we have a membrane between two reservoirs

\[ \text{membrane area} = A \]
We start with some concentration in reservoir 1 of \( C_0 \) and we assume that reservoir 2 is maintained at a constant \( C_{eq} \). If the volume of reservoir 1 is \( V \), then:

\[
V \frac{dC}{dt} = -A_v h(C_t - C_{eq})
\]

where \( C_t \big|_{t=0} = C_0 \)

We can solve this equation:

\[
C = C_{eq} + (C_0 - C_{eq}) e^{-\left(\frac{h t A}{V}\right)}
\]

We wish to determine \( h \) by measuring \( C \) as a function of time.

In order for use to use linear regression, the model must be linear in the modelling parameter.

In this case, the modelling parameter is \( h \). We can rewrite the 

\[(c - c_{eq}) = (c_0 - c_{eq}) e^{-\frac{bA}{V}t} \]

\[\ln(e - c_{eq}) = \ln(c_0 - c_{eq}) - \frac{bA}{V}t\]

So if we plot \(\ln(c - c_{eq})\) vs \(\frac{A}{V}\), we get a line with intercept of \(\ln(c_0 - c_{eq})\) and a slope of \(b\)!

How do we get the best fit line?

We look at the deviation of the points from the line, and try to minimize this in some way.

Let's take:

\[y = \ln(c - c_{eq}), \quad a = \frac{A}{V}, \quad b = \ln(c_0 - c_{eq})\]

We want to fit a series of points \((t_i, y_i)\) by the model \(y = at + b\)
We shall use the method of least squares. We form the sum:
\[
\text{sum} = \sum_{i=1}^{N} (y_i - (at_i + b))^2
\]
which is the sum of the squared distance between data and model in the \( y \) direction.

We pick \( a \) & \( b \) so that this is a minimum.

We let:
\[
\frac{\partial \text{sum}}{\partial a} = 0 \quad \text{\{2 equations for } a, b\}
\]
\[
\frac{\partial \text{sum}}{\partial b} = 0
\]
\[
\frac{\partial \text{sum}}{\partial a} = \sum_{i=1}^{N} -2(y_i - (at_i + b)) t_i
\]
\[
= \sum_{i=1}^{N} -2y_i t_i + 2a \sum_{i=1}^{N} t_i^2 + 2b \sum_{i=1}^{N} t_i = 0
\]
and:
\[
\frac{\partial \text{sum}}{\partial b} = \sum_{i=1}^{N} -2(y_i - (at_i + b)) \\
= \sum_{i=1}^{N} -2y_i + 2a \sum_{i=1}^{N} t_i + 2Nb = 0
\]

This is a pair of \textit{linear} equations for \(a\) & \(b\)!

We have the solution:
\[
\begin{align*}
\overline{a}t^2 + b \overline{t} &= \overline{yt} \quad (\text{let } \overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i) \\
\overline{a} \overline{t} + b &= \overline{y} \\
\therefore a = \frac{\overline{yt} - \overline{y} \overline{t}}{\overline{t}^2 - \overline{t}^2}
\end{align*}
\]

and \(b = \overline{y} - a \overline{t}\)

which can be used to get the \textit{mass transfer coefficient}!
This is not the only way to approach the problem. Let's look at it as a system of equations!

We are trying to satisfy:

\[ y_1 = \alpha \cdot t_1 + b \]
\[ y_2 = \alpha \cdot t_2 + b \]
\[ \vdots \]
\[ y_N = \alpha \cdot t_N + b \]

Let's write this in matrix form:

\[
\begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_N
\end{pmatrix} =
\begin{pmatrix}
  t_1 \\
  t_2 \\
  \vdots \\
  t_N
\end{pmatrix}
\begin{pmatrix}
  \alpha \\
  b
\end{pmatrix}
\]

We shall use \( x \) to be the vector containing the modelling parameters:

\[ x_1 = \alpha \quad x_2 = b \]
If we let the matrix of modelling functions be $A$ (e.g. $A_{11} = t_1$, etc.) then we want to satisfy:

$$\tilde{y} = \tilde{A}\tilde{x}$$

Actually, we want to minimize the residual:

$$\tilde{r} = \tilde{y} - \tilde{A}\tilde{x}$$

Or, for least squares, adjust $\tilde{x}$ s.t.:

$$\|\tilde{r}\|_2 = \|\tilde{y} - \tilde{A}\tilde{x}\|_2$$

is as small as possible.

We want to pick $\tilde{x}$ s.t.

$$\tilde{r}^T\tilde{r} = (\tilde{y} - \tilde{A}\tilde{x})^T(\tilde{y} - \tilde{A}\tilde{x})$$

is a minimum.
This is exactly equivalent to what we did before!

Let’s multiply this out:

\[ n^2 = y^T y - y^T A x - x^T A^T y + x^T A^T A x \]

We take the gradient of this w.r.t. \( \bar{x} \) and set it equal to zero!

\[ \nabla_{\bar{x}} n^2 = 0 = -2 A^T y + 2 A^T A \bar{x} = 0 \]

Thus we have the problem:

\[ A^T A \bar{x} = A^T y \]

This gives us a set of \( n \) eq’ns for the \( n \) unknowns in \( \bar{x} \)!

These are known as the normal equations.
Let's apply all this back to linear regression. Suppose we have some data and we are trying to fit it, and so calculate the uncertainty in the modelling parameters.

We measure $g$ by tossing a lead weight in the air and measuring its position as a function of time.

We get:

<table>
<thead>
<tr>
<th>$t$ (sec)</th>
<th>$h$ (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>14.8</td>
</tr>
<tr>
<td>2</td>
<td>20.7</td>
</tr>
<tr>
<td>3</td>
<td>15.9</td>
</tr>
<tr>
<td>4</td>
<td>2.0</td>
</tr>
</tbody>
</table>

We fit this to the equation:

$$h = h_0 + v_0 t - \frac{1}{2} g t^2$$
We want to both calculate $g$, and determine the uncertainty in $g$.

First let's get $g$. We have the problem:

$$h = (1) h_0 + (t) V_0 + (-\frac{1}{2} t^2) g$$

So:

$$\begin{pmatrix}
1 & 0 & 0 \\
1 & 1 & -0.5 \\
1 & 2 & -2 \\
1 & 3 & -3/2 \\
1 & 4 & -8 \\
\end{pmatrix} \begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
\end{pmatrix} = \begin{pmatrix}
0.3 \\
14.8 \\
20.7 \\
15.9 \\
2.0 \\
\end{pmatrix}$$

Using the normal eq's we get:

$$A^T A \hat{x} = A^T \hat{b}$$

$$x = [A^T A]^{-1} A^T \hat{b}$$
or numerically:

\[ X = \begin{pmatrix} 0.886 & 0.257 & -0.086 & -1.43 & 0.086 \\ -0.771 & 0.186 & 0.571 & 0.376 & -3.71 \\ -0.286 & 0.143 & 0.286 & -0.143 & -2.186 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \end{pmatrix} \]

This yields:

\[ x_1 = 0.197, \quad x_2 = 19.7, \quad x_3 = 9.64 \]

Now we want to examine the error in the position measurements \( h_i \).

To do this, we look at the deviation:

\[ S^2_h = \frac{1}{N-3} \sum_{i=1}^{N} (h_i - h(t_i))^2 \]

using fitted parameters because 3 adjustable parameters det. from data.
The quantity \( h_i - h(t_i) \) is just the residual:
\[
\tilde{\nu} = h - Ax
\]

\[
\therefore \hat{s}_h^2 = \frac{1}{N-3} \| \tilde{\nu} \|_2^2 = 0.110
\]

\( \Rightarrow \) you should really have more pts for accuracy!

Oh, how do we use this to get the error in \( g(x_3) \)?

We had:

\[
\mathbf{x} = \left[ \mathbf{A}^T \mathbf{A} \right]^{-1} \mathbf{A}^T h
\]

\[
\uparrow
\]

\[
\Rightarrow \mathbf{K}^2
\]

Thus \( x_i \equiv \sum_{j=1}^{N} K_{ij} h_j \)

\[
\text{\text{ith row of } K^2!}
\]
We can thus use the formula for a linear combination of random variables:

\[ s_{x_i}^2 = \sum_{j=1}^{N} k_{ij}^2 s_{h_j}^2 \]

\[ \uparrow \]

these are all the same in this problem!

Thus \( g = 9.64 \pm 0.18 \)

to the 69\% confidence level (1\sigma)

Often we want to develop correlations for their predictive values, for example, what is the expected height of the lead weight at some time \( t \)?

\[ h_m = x_1 + x_2 t - \frac{1}{2} t^2 x_3 \]

But what is the error? We have a linear combination of \( x_1, x_2, x_3 \) each of which have some error.
You might expect that:
\[ s_{h_{\text{mu}}}^2 = s_{x_1}^2 + (t)^2 s_{x_2}^2 + (-\frac{1}{2} t^2)^2 s_{x_3}^2 \]

But this is incorrect!!

The problem is that \( x_1, x_2, x_3 \) are not independent! Their covariance is non-zero!

Thus we should have:
\[ s_{h_{\text{mu}}}^2 = s_{x_1}^2 + (t) s_{x_2}^2 + (-\frac{1}{2} t^2)^2 s_{x_3}^2 \]
\[ + 2 \left( 1 \right) (t) s_{x_1 x_2} + 2 \left( 1 \right) (-\frac{1}{2} t^2) s_{x_1 x_3} \]
\[ + 2 \left( t \right) (-\frac{1}{2} t^2) s_{x_2 x_3} \]

How do we calculate the covariance?

Recall:
\[ x_i = \sum_{j=1}^{N} K_{ij} h_j \]
where \( K = [A^T A]^{-1} A^T \)
Thus we let:

\[ x_1 = \sum_{j=1}^{N} x_j h_j \]

and

\[ x_2 = \sum_{j=1}^{N} \beta_j h_j \]

where \( \alpha \) is the 1st row of \( \mathbf{R}^2 \)

\[ \mathbf{R}^2 \]

where \( \mathbf{R} \) is the 2nd row of \( \mathbf{R} \)

Now we need to determine the covariance

\[ \sigma_{x_1x_2}^2 = E \left( (x_1 - \mu_{x_1})(x_2 - \mu_{x_2}) \right) \]

value for \( x_1 \) & \( x_2 \) expected for an infinite sample

\[ \mu_{x_1} = \sum_{j=1}^{N} x_j h_j \]

from infinite sample

\[ \mu_{x_2} = \sum_{j=1}^{N} \beta_j h_j \]
So:

$$\bar{r}_{x_{1}x_{2}} = E\left(\left(\sum_{j=1}^{N} x_{j} (h_{j} - \mu_{h_{j}})\right)\left(\sum_{j=1}^{N} \beta_{j} (h_{j} - \mu_{h_{j}})\right)\right)$$

Now if all of the $h_{j}$ are independent then

$$E\left(\left(h_{j} - \mu_{h_{j}}\right)\left(h_{i} - \mu_{h_{i}}\right)\right) = 0$$

for $i \neq j$

$$\therefore \bar{r}_{x_{1}x_{2}} = E\left(\sum_{j=1}^{N} x_{j} \beta_{j} (h_{j} - \mu_{h_{j}})^{2}\right)$$

$$= \sum_{j=1}^{N} \alpha_{j} \beta_{j} \bar{r}_{h_{j}}$$

provided all the $h_{j}$ have the same uncertainty

$$\therefore \bar{r}_{x_{1}x_{2}} = x^{T} \beta \bar{r}_{h}$$

So for this problem if we let

$$\bar{y} = 3 \times 2 \text{ row of } \bar{x}$$
\[ x_1 = \alpha_1 \beta, \quad x_2 = \beta_2 \beta, \quad x_3 = \gamma \beta \]

\[ s_h^2 = \frac{1}{n-3} \sum \varepsilon^2, \quad \varepsilon = y - \hat{\Psi} \]

\[ s_{x_1}^2 = (\alpha_1 \alpha_1^T) s_h^2 \]

\[ s_{x_2}^2 = (\beta_2 \beta_2^T) s_h^2 \quad \text{etc.} \]

\[ h_m = (1) x_1 + (t) x_2 + (\frac{1}{2} t^2) x_3 \]

\[ \sigma_h^{-2} = (1) \sigma_{x_1}^{-2} + (t) \sigma_{x_2}^{-2} + \left( \frac{1}{2} t^2 \right) \sigma_{x_3}^{-2} \]

\[ + 2 (1) (t) \alpha_1 \beta_2^T \sigma_h^{-2} + 2 (t) \left( -\frac{1}{2} t^2 \right) \alpha_2 \beta_2 \sigma_h^{-2} \]

\[ + 2 (t) \left( -\frac{1}{2} t^2 \right) \alpha_2 \beta_2 \sigma_h^{-2} \]

\[ = \left\| \begin{bmatrix} (1) \alpha_1 + (t) \beta_2 + \left( -\frac{1}{2} t^2 \right) \gamma \end{bmatrix} \right\|_2^2 \sigma_h^{-2} \]
If we represent the vector of modelling functions as \( \bar{\alpha}(t) \), we can collapse this:

\[
\bar{\alpha}(t) = (1, t, -\frac{1}{2} t^2)
\]

Then

\[
\bar{\alpha}_w = \frac{1}{\| \bar{\alpha} \|^2} \bar{\alpha}
\]

Let's generalize the regression analysis we developed above.

We have \( n \) observations \( b_i \) which we wish to fit with \( m \) parameters and modelling functions:

\[
b_i \approx \sum_{j=1}^{m} x_j \phi_j(t_i)
\]

or:

\[
b \approx Ax
\]

where \( A_{ij} = \phi_j(t_i) \)
We define the residual:

\[ e = b - A \hat{x} \]

We define the best fit modelling parameters \( \hat{x} \) by:

\[ \min_x \| e \|_2^2 = \min_x (e^T e) \]

\( \hat{x} \) has the solution:

\[ A^T A \hat{x} = A^T b \]

or

\[ \hat{x} = (A^T A)^{-1} A^T b = \hat{X} b \]

Note that \( A \) is an \( nxm \) matrix, \( b \) is a \( nx1 \) array, \( \hat{x} \) is a \( mx1 \) array and \( \hat{X} \) is a \( m \times n \) matrix

If we recall that:

\[ \sigma_{\hat{x}}^2 = E \{ (z - \mu_{\hat{x}})^2 \} \]

where \( \mu_{\hat{x}} = E (z) \)
Then the matrix of covariance of the regression parameters is:

$$\Sigma_{xx} = E \left\{ (x - \mu_x)(x - \mu_x)^T \right\}$$

Substituting $x = \tilde{x} b$ and $\mu_x = \tilde{x} \mu_b$ we get:

$$\Sigma_{xx} \approx \tilde{x} \ E \left\{ (b - \mu_b)(b - \mu_b)^T \right\} \tilde{x}^T$$

$$= \tilde{x} \Sigma_{bb} \tilde{x}^T$$

where $\Sigma_{bb}$ is the matrix of covariance of the observations $b$.

If the observations are independent then $\Sigma_{bb}$ will be a diagonal matrix.

If all of the observations have the same variance, then:

$$\Sigma_{xx} \approx \frac{1}{n} \Sigma_{bb}^2$$
and we get
\[ \sum_{X}^2 = \Sigma_{X} K K^T \]

We may estimate \( \Sigma_{X} \) from:
\[ \Sigma_{X} \approx s_{\tilde{x}}^2 = \frac{1}{n-m} \| \tilde{A} \tilde{x} - \tilde{b} \|_2^2 \]
\[ = \frac{1}{n-m} \tilde{c}^T \tilde{c} \]

We may calculate the error in model predictions similarly. Suppose we wish to determine the model value at \( K \) times \( \tilde{x}_{\text{new}} \). Thus:
\[ \tilde{b}^{(m)} = A^{(m)} \tilde{x} \]

where \( A^{(m)} = \phi_{\text{new}} (t_{\text{new}}^{(m)}) \)

Note that \( A^{(m)} \) is a \( K \times M \) matrix.

What is the error in \( \tilde{b}^{(m)} \)? We have the matrix of covariance:
\[
\sum_b^2(m) \equiv E \sum_i \left( \frac{1}{n} - \mu_b(m) \right) \left( \frac{1}{n} - \mu_b(m) \right)^T \sum_x \frac{1}{n} \sum_i A(m)^T
\]
\[
= A(m)^T E \sum_i (x - \mu x)(x - \mu x)^T \sum_x \frac{1}{n} \sum_i A(m)^T
\]
\[
= A(m)^T \sum_x \frac{1}{n} \sum_i A(m)^T
\]

where \( \sum_x \frac{1}{n} \sum_i A(m)^T \) is the matrix of covariance determined before! In general, only the diagonal elements of \( \sum_b^2(m) \) are of interest:
\[
\sum_b^2(m) = \text{diag} \left\{ \frac{1}{n} \sum_i \sum_x \frac{1}{n} \sum_i A(m)^T \right\}
\]

which yields the error in the model predictions.

As a final note on linear regression, let us examine the case where the variance of \( \sum_b^2 \) is not constant!

In this case we wish to weight the points in our regression analysis differently.
This is weighted linear regression!

Suppose we have the diagonal matrix \( \Sigma_b \). The weighted problem is just:

\[
\min_x (A x - b)^T \Sigma_b^{-1} (A x - b)
\]

\[
= \min_x x^T \Sigma_b^{-1} x
\]

where \( \Sigma_b^{-1} \) is a diagonal matrix whose elements are just \( \frac{1}{\sigma_i^2} \).

Thus we weight each point by the inverse of the variance of that point. This makes sense. Suppose the reason why \( \hat{b}_i \) varied was that \( m_i \) points were averaged together to get \( b_i \).

Then we have \( \sigma_{b_i}^2 = \frac{\sigma_b^2}{m_i} \) and thus the above regression yields:
\[
\frac{1}{\sigma^2} x^T \begin{pmatrix}
\sigma^2 & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \sigma^2
\end{pmatrix} x 
\]

or the \(i^{th}\) point is counted \(m_i\) times. The result for \(x\) would be the same (on average) as if you had used all of the points individually!

OK, what is the formula for \(x\)?

Just as before, we take the gradient:

\[
\nabla_x \left\{ \nabla^T (\Sigma_x^{-1}) \nabla \right\} = 0
\]

which yields:

\[
A^T (\Sigma_x^{-1}) A x = A^T (\Sigma_x^{-1}) b
\]

or

\[
\tilde{x} = \Sigma_x^{-1} \Sigma_x b
\]

with error:

\[
\Sigma_x = \Sigma_x \Sigma_x \Sigma_x^T
\]
Thus far we have considered error propagation/estimation techniques which rely on the regression formulas specific to linear regression. There are other numerical ways to estimate the error in the parameters, however, which apply to both linear and non-linear regression!

One approach is undersampling.

Suppose we have a series of observations \( b_i (t_i) \). We wish to fit this with a model:

\[
b_i = f(x, t_i)
\]

where \( f \) may be linear or non-linear in \( x \). If it's linear, we just use linear regression. If it's non-linear, we still do regression, only now it's non-linear regression.

We want to estimate the error
and covariance in \( \Sigma \), e.g., we wish:
\[
\Sigma \approx \mathbb{E} \left\{ (x_i - \mu_x) (x_j - \mu_x)^T \right\}
\]

If the number of pts \( N \) is fairly large, we can do this with undersampling.

Suppose we take every \( m \)th point. This gives us a set of \( N/m \) points from which we can calculate \( \Sigma \). If we do this for the \( m \) independent sets of \( N/m \) points, we get \( m \) estimates for \( \Sigma \).

We can then get a sample covariance matrix:
\[
\Sigma^2 \approx \frac{1}{m-1} \left\{ x_i x_j - \bar{x}_i \bar{x}_j \right\}
\]

Note that this did not require knowing the error in \( \frac{1}{\Delta} \). It also
works equally well for linear and non-linear regression.

Its drawbacks are that it requires a pretty big $N$, and it requires multiple solutions for $x$. That can take awhile for non-linear problems.

A similar approach is the bootstrap, which works pretty well with moderate sized data sets.

Suppose we have, say, 20 observations $b_i(t_i)$. These observations are a subset of all possible observations. We can produce an approximation to this infinite set by periodically replicating our original data set!

The idea is to then take samples of 20 observations from this replicated set and calculate $x$. We then compute $\sum x$ as before!
This technique is appealing, but only works if \( N \) is fairly large.

This is because there is a non-zero probability of getting the same point \( b_i(t_i) \) \( N \) times out of our replicated set, which leads to an ill-posed regression problem.

Fortunately, this probability goes as

\[
\sim \left( \frac{1}{N} \right)^{N-1}
\]

which gets pretty small fairly fast!

The regression covariance is just:

\[
\Sigma_{xixj}^2 = \frac{p}{p-1} \left\{ \overline{x_i x_j} - \overline{x_i} \overline{x_j} \right\}^2
\]

where we are solving the regression problem \( p \) times.
As a final note on statistics, we have assumed throughout that the deviation between the observations and the model is random! This is, in general, **Not True**!

Thus, you tend to underestimate your error by ignoring the non-zero covariance in your data.

It is important to always plot your residuals to see if there is a systematic deviation.

A systematic deviation means that something is missing from your model and/or there is a systematic error in your data.

Never forget that for large N, parameter error is dominated by systematic error!!
**Non-linear Regression**

So far we’ve focussed an **linear regression**; problems where the model is **linear** in the unknown modelling parameters. This is convenient, but not really necessary! Suppose we have the non-linear model:

$$b = f(x, t)$$

where $x$ is an array of modelling functions. We may still define the residual between observed and fitted values:

$$r_i = b_i - f(x, t_i)$$

\[ resid \uparrow \]
\[ meas \uparrow \]
\[ model \]
Thus we can form the sum of squares:

$$F(x) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

The best fitted values for $x$ are obtained via finding the minimum of $F(x)$.

Be wary: Non-linear optimization problems may have local minima which can trap the solver! Also, make sure you have set up the parameters so that $F(x)$ is well conditioned. The dependence on each parameter should be of similar magnitude. Let's look at an example:

Arrhenius Kinetics

rate $\sim K_0 e^{-E/RT}$
In senior lab you will measure the rate of oxidation of methane as a function of temperature, and use it to try to get the activation energy $E$ for the catalyst. If we could measure this rate, holding everything but $T$ constant, we could use linear regression. Unfortunately, the rate also depends on concentration, and that depends (experimentally) on $T$ as well!

The experiment is written up in Chem Eng. Education, 36(1) p. 34-40. Suppose we feed a reactor (well-mixed) a
concentration of methane $C_w$ at
flow rate $q_w$. We measure some
outlet concentration $C_r$. The
subscript "r" is because the reactor
is hot: due to expansion, the
actual concentration is $C/C_r = T_r/T$
from the ideal gas law - where $T_r$
is some ref. temp ($^\circ$C).
The system is further complicated
because the reaction is fractional
order, e.g.
\[ \text{rate} \sim C^n \]
where $n \neq 1$ (not first order)
we need to figure out $n$ too!
From a mass balance, we get the rate of \( r \times n \) per gram of catalyst:

\[
\frac{2r}{m} (C_{r0} - C_r) = (C_r T_r)^n K_0 e^{-\frac{E}{RT}}
\]

where \( m \) is the mass of catalyst. We can also look at the conversion ratio given by:

\[ X = 1 - \frac{C_r}{C_{r0}} \]

which yields (after rearrangement):

\[
\frac{X}{(1-X)^n} = \frac{m}{2r} C_{or}^{n-1} (\frac{T_r}{T})^n K_0 e^{-\frac{E}{RT}}
\]

If you were to fix \( T \) and plot \( X \) vs. \( C_{or} \), you find it decreases with \( C_{or} \), thus \( n < 1 \) for this system.
Ok, suppose we vary $C_{wo}$ and $T$, and measure $C_w$. How do we get the unknowns $n$, $K_0$, and $E$?

The classic approach is to do it in two steps:

1) Fix $T$ and plot

$$\ln \left\{ \frac{2 \nu}{m} (C_{wo} - C_w)^2 \right\} \text{ vs. } \ln \{C_w\}$$

You should get a straight line with slope $n$!

2) With this in hand, you can then do an Arrhenius Plot of:

$$\ln \left\{ \frac{2 \nu}{m} \frac{(C_{wo} - C_w)}{C_w^n} \left( \frac{T_r}{T} \right)^{-n} \right\} = \ln K_0 - \frac{E}{R} \frac{1}{T}$$
Thus, the \textit{slope} \( (v.s. T) \) is just \( \frac{\Delta}{R} \) and the intercept is \( \ln k_0 \).

This works fine if the data is good. Unfortunately, this is not usually the case. The problem is that \( \Delta \) is a much stronger function of \( T \) than \( C_{\alpha} \), and thus you get large errors in the first step \( (n) \), leading to large errors in the second step \( (E & K_0) \).

The fitting parameters are also strongly biased by errors at low conversions.

We can avoid this by using non-linear regression! We simply define the deviation from the model \( A \):
\[ \Delta = C_r - C_{r_0} + \frac{m}{Z_r} C_r K_0 e^{-\frac{E}{RT}} \left( \frac{1}{T} \right)^n \]

and minimize the objective function:

\[ F(K_0, E, n) = \sum \Delta_i^2 \]

over all the experiments! The sensitivity of the fitting parameters to error can then be easily calculated using the non-linear error propagation/gradient method, accounting for the error in both \( C_r \) and \( T \). Note that you should work with \( n, \ln K_0, \) and \( \frac{E}{RT} \) as fitting parameters so that the optimization problem is well-conditioned!