Statistics & Probability

What is error?

Two kinds: Random & Systematic

Random error is the scatter you get from repeated measurements

Systematic error is <u>systematic</u>: it happens to the same degree <u>each time</u> you make a measurement.

Multiple measurements will average out (reduce) random error, but <u>won't</u> affect systematic error!

If you do many measurements, the error in the avg. will be <u>completely</u> dominated by systematic error, but this is <u>very</u> 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:

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

$p(x) \equiv$ probability density

$$= \frac{\text{probability of meas. being in interval } [x, x+dx]}{dx}$$

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:

$\mu \equiv$ mean of distribution

$\sigma \equiv$ population standard deviation

About 69% of observations lie within 1σ of μ, 95% within 2σ, and 99% within 3σ.

In working w/ statistics, we define an <u>expectation value</u>

$E(x) \equiv$ what you expect to get if you do a meas. many times and average it together!

Mathematically:

$$E(x) \equiv \int_{-\infty}^{\infty} x\, p(x)\, dx$$

(x: meas. quantity; $p(x)$: prob. density)

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

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

Let $z \equiv x - \mu$

$$\therefore E(x) = \int_{-\infty}^{\infty} \underbrace{\frac{z}{\sqrt{2\pi}\sigma^2}}_{\text{odd}} \underbrace{e^{-\frac{z^2}{2\sigma^2}}}_{\text{even}} dz + \mu \underbrace{\int_{-\infty}^{\infty} p(x)\, dx}_{=1}$$

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 <u>variance</u>:

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

$$E((x-\mu)^2) \equiv \int_{-\infty}^{\infty} (x-\mu)^2 \, p(x) \, dx$$

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

Let $z = (x-\mu)$

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

Integrate by parts:

$$= -\frac{\sigma^2 z}{\sqrt{2\pi}\,\sigma} e^{-\frac{z^2}{2\sigma^2}} \Big|_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{z^2}{2\sigma^2}} dz$$

$\underbrace{\qquad\qquad}$ vanishes at $\pm\infty$; $\underbrace{\qquad\qquad}$ $= 1$

$$\therefore E((x-\mu)^2) \equiv \sigma^2$$

What about the variance of the mean of N measurements?

$$E((\bar{x}-\mu)^2) \equiv \sigma_{\bar{x}}^2$$

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

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

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

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

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

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

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

First expectation is just σ^2

Second term is <u>zero</u> provided x_i & x_j are <u>independent</u>

$$E\left((x_i-\mu)(x_j-\mu)\right) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i-\mu)(x_j-\mu)\cdot P(x_i)\,P(x_j)\,dx_i\,dx_j$$

$$\equiv \int_{-\infty}^{\infty}(x_i-\mu)\,P(x_i)\,dx_i \int_{-\infty}^{\infty}(x_j-\mu)\,P(x_j) = 0 \quad i\neq j$$

Thus:

$$\sigma_{\bar{x}}^2 \equiv \frac{\sigma^2}{N}$$

That is why multiple measurements reduce the random error!

Ok, how do we <u>estimate</u> μ and σ?

In general we have a <u>sample</u> of N observations drawn from an underlying <u>population</u> of possible observations!

We define the <u>sample mean</u>

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

we have already shown that

$$E(\bar{x}) = \mu$$

That is, $\bar{x}$ is an <u>unbiased estimate</u> of μ.

Now for the variance:

We <u>define</u> the sample variance:

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

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

$$= \frac{1}{N-1} \left\{ \sum_{i=1}^{N} x_i^2 - 2\bar{x} \underbrace{\sum_{i=1}^{N} x_i}_{\equiv N\bar{x}} + N\bar{x}^2 \right\}$$

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

Note that to calculate this we <u>don't</u> 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 calc. S_x^2.

Ok, how is S_x^2 related to σ^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:

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

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

Thus:

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

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

$$= \frac{1}{N-1}\left\{N\sigma^2 - \frac{N}{N}\sigma^2\right\} = \sigma^2 !$$

So S_x^2 is an unbiased estimate for σ^2!

The $N-1$ part comes about because we are calculating the <u>mean</u> from the same sample that we are calculating the <u>variance</u>

We have reduced the number of <u>degrees of freedom</u> by one.

Propagation of Errors:

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

Each individual meas. has some error associated with it. How do these combine?

Let x_i, y_i be random variables w/ 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.

$$\therefore \bar{z} = \frac{1}{N} \sum_{i=1}^{N} z_i = c_1 \bar{x} + c_2 \bar{y}$$

What about the variance?

$$s_z^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \bar{z})^2$$

$$= \frac{1}{N-1} \left\{ \sum_{i=1}^{N} (z_i^2 - \bar{z}^2) \right\}$$

Substituting in:

$$z_i^2 = c_1^2 x_i^2 + c_2^2 y_i^2 + 2 c_1 c_2 x_i y_i$$

$$\bar{z}^2 = c_1^2 \bar{x}^2 + c_2^2 \bar{y}^2 + 2 c_1 c_2 \bar{x} \bar{y}$$

$$\therefore S_z^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} \Big(C_1^2 (x_i^2 - \bar{x}^2) + C_2^2 (y_i^2 - \bar{y}^2) + 2 C_1 C_2 (x_i y_i - \bar{x}\bar{y}) \Big)$$

$$= C_1^2 S_x^2 + C_2^2 S_y^2 + \frac{2 C_1 C_2}{N-1} \sum_{i=1}^{N} (x_i y_i - \bar{x}\bar{y})$$

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

$$S_{xy}^2 \equiv \frac{1}{N-1} \sum_{i=1}^{N} (x_i y_i - \bar{x}\bar{y})$$

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

This is the <u>covariance</u> of x and y. If they are <u>independent</u> then:

$$E(S_{xy}^2) = 0$$

So for addition:

$$S_z^2 = C_1^2 S_x^2 + C_2^2 S_y^2 + 2 C_1 C_2 S_{xy}^2$$

If we <u>subtract</u> two variables we get a similar result:

$$z_i = x_i - y_i$$

then $\bar{z} = \bar{x} - \bar{y}$; $s_z^2 = s_x^2 + s_y^2 - 2s_{xy}^2$

note $+$ sign!

A positive covariance (note: s_{xy}^2 may be $+$ or $-$) <u>reduces</u> the error in the difference between variables.

OK, how about multiplication and division?

Let $z_i = x_i / y_i$

where x_i and y_i are normally distributed.

Note that z is <u>not</u> normally distributed! It only approaches this <u>if</u> $s_x/\bar{x}$ and $s_y/\bar{y} \ll 1$!

Let's look at this more closely.

Let $x_i = \bar{x} + x_i'$

($\bar{x}$: mean, x_i': deviation)

and $y_i = \bar{y} + y_i'$

Suppose $\frac{s_x}{\bar{x}} << 1$ and $\frac{s_y}{\bar{y}} << 1$

This means that x_i' and y_i' are small:

$$z_i = \frac{x_i}{y_i} = \frac{\bar{x} + x_i'}{\bar{y} + y_i'} = \frac{\bar{x}}{\bar{y}} \frac{\left(1 + \frac{x_i'}{\bar{x}}\right)}{\left(1 + \frac{y_i'}{\bar{y}}\right)}$$

$$\approx \frac{\bar{x}}{\bar{y}} \left(1 + \frac{x_i'}{\bar{x}}\right)\left(1 - \frac{y_i'}{\bar{y}} + O\left(\left(\frac{y_i'}{\bar{y}}\right)^2\right)\right)$$

$$\approx \frac{\bar{x}}{\bar{y}} \left(1 + \frac{x_i'}{\bar{x}} - \frac{y_i'}{\bar{y}} + O\left(\frac{x_i' y_i'}{\bar{x}\bar{y}}, \left(\frac{y_i'}{\bar{y}}\right)^2\right)\right)$$

We <u>ignore</u> the higher order terms!

$$\bar{z} \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{x_i}{y_i} \approx \frac{\bar{x}}{\bar{y}}\frac{1}{N}\sum_{i=1}^{N}\left(1+\frac{x_i'}{\bar{x}}+\frac{y_i'}{\bar{y}}\right)$$

Sum to zero

$$\therefore \bar{z} \approx \frac{\bar{x}}{\bar{y}}$$

Now for s_z^2:

$$s_z^2 \approx \left(\frac{\bar{x}^2}{\bar{y}^2}\right)\left(\frac{s_x^2}{\bar{x}^2}+\frac{s_y^2}{\bar{y}^2}-\frac{2\,s_{xy}^2}{\bar{x}\,\bar{y}}\right)$$

$\approx \bar{z}^2$

Thus:

$$\frac{s_z^2}{\bar{z}^2} \approx \frac{s_x^2}{\bar{x}^2}+\frac{s_y^2}{\bar{y}^2}-2\frac{s_{xy}^2}{\bar{x}\,\bar{y}}$$

So for division you add the <u>fractional</u> or <u>relative</u> variance!

For multiplication you get the same:

$$z_i \equiv x_i y_i$$

$$\bar{z} \approx \bar{x}\,\bar{y} \quad \underline{\text{provided}} \quad \frac{s_x}{\bar{x}}, \frac{s_y}{\bar{y}} \ll 1$$

and

$$\frac{s_z^2}{\bar{z}^2} \approx \frac{s_x^2}{\bar{x}^2} + \frac{s_y^2}{\bar{y}^2} + 2\frac{s_{xy}^2}{\bar{x}\,\bar{y}}$$

↑ note (+) sign

What about a more general functional relationship?

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

Let $x_i = \bar{x} + x_i'$, $y_i = \bar{y} + y_i'$

We shall <u>expand</u> f in a 2-D

Taylor series (Note: the T.S. can be generalized to n-dimensions)

So:

$$f(x_i, y_i) = f(\bar{x}, \bar{y}) + x_i' \left.\frac{\partial f}{\partial x}\right|_{\bar{x},\bar{y}} + y_i' \left.\frac{\partial f}{\partial y}\right|_{\bar{x},\bar{y}} + O(x_i'^2, y_i'^2, x_i' y_i')$$

(Ignore 2nd order terms!)

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

$$\approx f(\bar{x}, \bar{y}) + \left.\frac{\partial f}{\partial x}\right|_{\bar{x},\bar{y}} \frac{1}{N}\sum_{i=1}^{N} x_i' + \left.\frac{\partial f}{\partial x}\right|_{\bar{x},\bar{y}} \frac{1}{N}\sum_{i=1}^{N} y_i'$$

sum to zero!

$$= f(\bar{x}, \bar{y})$$

The <u>error</u> is $O(x_i'^2, y_i'^2, x_i' y_i' \cdot \nabla\nabla f)$

↑ dyadic

Similarly, we get for the variance:

$$s_z^2 \cong \left(\frac{\partial f}{\partial x}\Big|_{\bar{x},\bar{y}}\right)^2 s_x^2 + \left(\frac{\partial f}{\partial y}\Big|_{\bar{x},\bar{y}}\right)^2 s_y^2 + 2\left(\frac{\partial f}{\partial x}\Big|_{\bar{x},\bar{y}}\right)\left(\frac{\partial f}{\partial y}\Big|_{\bar{x},\bar{y}}\right) s_{xy}^2$$

again providing that the higher order terms can be neglected.

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

Only for add^n & subtr. is it <u>exact</u> because $\nabla\nabla f \equiv 0$ for this case and the higher order terms <u>vanish</u>.

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

Let $\underset{\sim}{x} = x_1, x_2, \ldots, x_n$

$z = f(\underset{\sim}{x})$

We need to find the variance of each variable x_1, x_2, etc. and the covariance of all pairs $S^2_{x_1 x_2}$, etc.

We can define the matrix of variance & covariance

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

↑ number of elements contributing to each x_i, x_j

Now if $\frac{V_{ii}}{\bar{x}_i^2} << 1$ for all i

(this corresponds to a small error in each variable, e.g. $\frac{S_x^2}{\bar{x}^2} << 1$, $\frac{S_y^2}{\bar{y}^2} << 1$, etc.)

And if we define:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}, \cdots \frac{\partial f}{\partial x_n}\right)$$

Then we get the result:

$$s_z^2 = (\underset{\sim}{\nabla} f)\, \underset{\approx}{V}\, (\underset{\sim}{\nabla} f)^T$$

Which gives us an easy way of calculating the variance in $z = f(\underset{\sim}{x})$!

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 $f(\underset{\sim}{x}) = x_1 x_2$

Let $x_1 \equiv x$, $x_2 \equiv y$

$$\therefore \quad V_{ij} = \frac{N}{N-1}\left(\overline{x_i x_j} - \bar{x}_i \bar{x}_j\right)$$

$$\text{so} \quad V_{xx} = \frac{N}{N-1}\left(\overline{xx} - \bar{x}\,\bar{x}\right)$$

$$= \frac{1}{N-1}\sum^{N}\left(x^2 - \bar{x}^2\right) = s_x^2$$

$$V_{xy} = \frac{N}{N-1}\left(\overline{xy} - \bar{x}\,\bar{y}\right) = \frac{1}{N-1}\sum^{N}(xy - \bar{x}\bar{y})$$
$$\equiv s_{xy}^2$$

$$\underset{\approx}{V} = \begin{pmatrix} s_x^2 & s_{xy}^2 \\ s_{xy}^2 & s_y^2 \end{pmatrix}$$

$$\underset{\sim}{\nabla} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \equiv (y, x)$$

$$\underset{\sim}{\nabla} f \;\underset{\approx}{V}\; (\underset{\sim}{\nabla} f)^T = (y \;\; x)\;\underset{\approx}{V}\begin{pmatrix} y \\ x \end{pmatrix}$$

$$= \left(y s_x^2 + x s_{xy}^2,\; y s_{xy}^2 + x s_y^2\right)\begin{pmatrix} y \\ x \end{pmatrix}$$

$$= y^2 s_x^2 + xy s_{xy}^2 + xy s_{xy}^2 + x^2 s_y^2$$

$$= y^2 s_x^2 + x^2 s_y^2 + 2xy\, s_{xy}^2$$

So

$$\frac{s_z^2}{(xy)^2} = \frac{s_x^2}{x^2} + \frac{s_y^2}{y^2} + 2\,\frac{s_{xy}^2}{xy}$$

Finally, we examine vector functions of vector variables:

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

$$\underset{\sim}{\mu}_z \equiv E(\underset{\sim}{z}), \quad \underset{\sim}{\mu}_x = E(\underset{\sim}{x})$$

and the matrix of covariance for each:

$$\underset{\approx}{\Sigma}_x^2 \equiv E\left[(\underset{\sim}{x} - \underset{\sim}{\mu}_x)(\underset{\sim}{x} - \underset{\sim}{\mu}_x)^T\right]$$

$$\underset{\approx}{\Sigma}_z^2 = E\left[(\underset{\sim}{z} - \underset{\sim}{\mu}_z)(\underset{\sim}{z} - \underset{\sim}{\mu}_z)^T\right]$$

we have taken $\underset{\sim}{z}$ and $\underset{\sim}{x}$ to be column vectors.

Suppose we know $\underset{\sim}{x}$ and $\underset{\approx}{\Sigma}_x^2$. We wish to calculate $\underset{\approx}{\Sigma}_z^2$.

We do a multivariable Taylor series expansion of $\underset{\sim}{z}$ about $\underset{\sim}{\mu}_x$:

$$\underset{\sim}{z} = \underset{\sim}{f}(\underset{\sim}{x}) = \underset{\sim}{f}(\underset{\sim}{\mu}_x) + \underset{\sim}{\nabla}\underset{\sim}{f}\Big|_{\underset{\sim}{\mu}_x} \cdot (\underset{\sim}{x} - \underset{\sim}{\mu}_x) + \frac{1}{2}\underset{\sim}{\nabla}\underset{\sim}{\nabla}\underset{\sim}{f} : (\underset{\sim}{x} - \underset{\sim}{\mu}_x)(\underset{\sim}{x} - \underset{\sim}{\mu}_x)^T + \ldots$$

If the deviation between $\underset{\sim}{x}$ and $\underset{\sim}{\mu}_x$ is small (e.g. that $\underset{\sim}{\Sigma}_x^2$ is small), or $\underset{\sim}{f}(\underset{\sim}{x})$ is linear in $\underset{\sim}{x}$, then we can ignore the quadratic or higher order terms in the Taylor series!

Thus:

$$\underset{\sim}{z} \cong \underset{\sim}{f}(\underset{\sim}{\mu}_x) + \underset{\sim}{\nabla}\underset{\sim}{f}\Big|_{\underset{\sim}{\mu}_x} \cdot (\underset{\sim}{x} - \underset{\sim}{\mu}_x)$$

and thus:

$$\underset{\sim}{\mu}_z \equiv E(\underset{\sim}{z}) \cong \underset{\sim}{f}(\underset{\sim}{\mu}_x) + \underset{\sim}{\nabla}\underset{\sim}{f}\Big|_{\underset{\sim}{\mu}_x} \cdot E(\underset{\sim}{x} - \underset{\sim}{\mu}_x)$$

but $E(\underset{\sim}{x} - \underset{\sim}{\mu}_x) \equiv 0$ by definition!

Thus $\underset{\sim}{\mu}_z \cong \underset{\sim}{f}(\underset{\sim}{\mu}_x)$ + quadratic terms!

Putting this together, we get:

$$\underset{\sim}{z} - \underset{\sim}{\mu}_z \approx \underset{\sim}{\nabla} f \Big|_{\underset{\sim}{\mu}_x} (\underset{\sim}{x} - \underset{\sim}{\mu}_x)$$

So:

$$\underset{\approx}{\Sigma}_z^2 \equiv E\left((\underset{\sim}{z} - \underset{\sim}{\mu}_z)(\underset{\sim}{z} - \underset{\sim}{\mu}_z)^T \right)$$

$$\approx E\left[(\underset{\sim}{\nabla} f)\ (\underset{\sim}{x} - \underset{\sim}{\mu}_x)(\underset{\sim}{x} - \underset{\sim}{\mu}_x)^T (\underset{\sim}{\nabla} f)^T \right]$$

$$= (\underset{\sim}{\nabla} f)\ E\left((\underset{\sim}{x} - \underset{\sim}{\mu}_x)(\underset{\sim}{x} - \underset{\sim}{\mu}_x)^T \right) (\underset{\sim}{\nabla} f)^T$$

$$= (\underset{\sim}{\nabla} f)\ \underset{\approx}{\Sigma}_x^2\ (\underset{\sim}{\nabla} f)^T$$

<u>provided</u> that $\frac{1}{2} \underset{\sim}{\nabla}\underset{\sim}{\nabla} f : \underset{\approx}{\Sigma}_x^2 \ll \underset{\sim}{f}$

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 <u>one</u> answer.

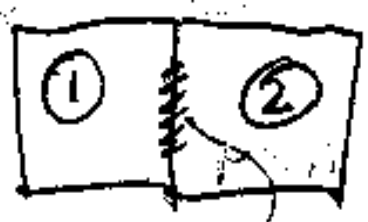
Often in data analysis you want to fit a curve to <u>many</u> data points

How do we do this? Use (usually) <u>linear regression</u>

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

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

Suppose we have a membrane between two reservoirs



→ 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:

$$V \frac{\partial C}{\partial t} = -Ah(C_1 - C_{eq})$$

where $C\big|_{t=0} = C_0$

We can solve this equation:

$$C = C_{eq} + (C_0 - C_{eq})\, e^{-\left(\frac{htA}{V}\right)}$$

We wish to determine h by measuring C as a function of time

In order for us 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 eqn:

$$(C - C_{eq}) = (C_0 - C_{eq})\, e^{-\frac{hA}{V}t}$$

$$\ln(C - C_{eq}) = \ln(C_0 - C_{eq}) - \frac{hA}{V}t$$

So if we plot $\ln(C - C_{eq})$ vs $\frac{At}{V}$

we get a line w/ intercept of $\ln(C_0 - C_{eq})$ and a slope of h!

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:

$$y \equiv \ln(C - C_{eq}), \quad a \equiv \frac{Ah}{V},$$

$$b \equiv \ln(C_0 - C_{eq})$$

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 pt. and model in the y direction.

We pick a & b so that this is a minimum

We let:

$$\left.\begin{aligned} \frac{\partial \text{sum}}{\partial a} &= 0 \\ \frac{\partial \text{sum}}{\partial b} &= 0 \end{aligned}\right\} \text{2 eqns for } a, b!$$

$$\frac{\partial \text{sum}}{\partial a} = \sum_{i=1}^{N} -2(y_i - (at_i + b))t_i$$

$$= \sum_{i=1}^{N} -2y_i t_i + 2a\sum_{i=1}^{N} t_i^2 + 2b\sum_{i=1}^{N} t_i = 0$$

and:

$$\frac{\partial \text{sum}}{\partial b} \equiv \sum_{i=1}^{N} -2(y_i - (at_i + b))$$

$$= \sum_{i=1}^{N} -2y_i + 2a\sum_{i=1}^{N} t_i + 2Nb = 0$$

This is a pair of <u>linear</u> equations for a & b!

We have the solution:

$$a\overline{t^2} + b\bar{t} = \overline{yt} \qquad (\text{let } \bar{x} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i)$$

$$a\bar{t} + b = \bar{y}$$

$$\therefore a = \frac{\overline{yt} - \bar{y}\bar{t}}{\overline{t^2} - \bar{t}^2}$$

and $b = \bar{y} - a\bar{t}$

which can be used to get the 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:

$$y_1 = a t_1 + b$$
$$y_2 = a t_2 + b$$
$$\vdots$$
$$y_N = a t_N + b$$

Let's write this in <u>matrix form</u>:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \approx \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_N & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

we shall use $\underset{\sim}{x}$ to be the vector containing the modelling parameters:

$$x_1 \equiv a \quad x_2 \equiv b$$

If we let the matrix of modelling functions be $\underset{\approx}{A}$ (e.g. $A_{11} = t_1$, etc.)

then we want to satisfy:

$$\underset{\sim}{y} \simeq \underset{\approx}{A}\,\underset{\sim}{x}$$

Actually, we want to minimize the residual:

$$\underset{\sim}{r} = \underset{\sim}{y} - \underset{\approx}{A}\,\underset{\sim}{x}$$

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

$\|\underset{\sim}{r}\|_2 \equiv \|\underset{\sim}{y} - \underset{\approx}{A}\,\underset{\sim}{x}\|_2$ is as small as possible

↑ Euclidean norm

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

$$\underset{\sim}{r}^T\underset{\sim}{r} \equiv (\underset{\sim}{y} - \underset{\approx}{A}\,\underset{\sim}{x})^T(\underset{\sim}{y} - \underset{\approx}{A}\,\underset{\sim}{x})$$

is a minimum.

This is exactly equivalent to what we did before!

Let's multiply this out:

$$r^2 \equiv \underset{\sim}{y}^T \underset{\sim}{y} - \underset{\sim}{y}^T \underset{\approx}{A} \underset{\sim}{x} - \underset{\sim}{x}^T \underset{\approx}{A}^T \underset{\sim}{y} + \underset{\sim}{x}^T \underset{\approx}{A}^T \underset{\approx}{A} \underset{\sim}{x}$$

We take the gradient of this w.r.t. $\underset{\sim}{x}$ and set it equal to zero!

$$\underset{\sim}{\nabla}_{\underset{\sim}{x}} r^2 = 0 = -2 \underset{\approx}{A}^T \underset{\sim}{y} + 2 \underset{\approx}{A}^T \underset{\approx}{A} \underset{\sim}{x} = 0$$

Thus we have the problem:

$$\underset{\approx}{A}^T \underset{\approx}{A} \underset{\sim}{x} = \underset{\approx}{A}^T \underset{\sim}{y} \quad !$$

This gives us a set of $\underline{\underline{n}}$ eq'ns for the n unknowns in $\underset{\sim}{x}$!

These are known as the normal equations.

Linear Regression Error

Let's apply all this back to linear regression. Suppose we have some data and we are trying to fit it, and calc. the uncertainty in the modelling parameters.

We measure g by tossing a lead weight in the air and meas. its pos'n as a fⁿ (t)

We get:

t_i (sec)	h_i (m)
0	0.3
1	14.8
2	20.7
3	15.9
4	2.0

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:

$$h = (1)\,h_0 + (t)\,V_0 + \left(-\tfrac{1}{2}t^2\right) g$$

$h_0 \uparrow x_1$, $V_0 \uparrow x_2$, $g \uparrow x_3$

So:

$$\underset{\approx}{A}\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -.5 \\ 1 & 2 & -2 \\ 1 & 3 & -9/2 \\ 1 & 4 & -8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \underset{\sim}{h}\begin{pmatrix} 0.3 \\ 14.8 \\ 20.7 \\ 15.9 \\ 2.0 \end{pmatrix}$$

Using the normal eq'ns we get:

$$\underset{\approx}{A}^T \underset{\approx}{A}\, \underset{\sim}{x} = \underset{\approx}{A}^T \underset{\sim}{h}$$

$$\underset{\sim}{x} = \left[\underset{\approx}{A}^T \underset{\approx}{A}\right]^{-1} \underset{\approx}{A}^T \underset{\sim}{h}$$

Or numerically:

$$\underset{\sim}{X} = \begin{pmatrix} .886 & .257 & -.086 & -.143 & .086 \\ -.771 & .186 & .571 & .386 & -.371 \\ -.286 & .143 & .286 & .143 & -.286 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \end{pmatrix}$$

This yields:

$x_1 = 0.197$, $x_2 = 19.7$ $x_3 = 9.64$

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

To do this we look at the deviation:

$$S_h^2 \approx \frac{1}{N-3} \sum_{i=1}^{N} \left(h_i - h(t_i)\right)^2$$

↑ because 3 adjustable parameters det. from data

↑ using fitted parameters

The quantity $h_i - h(t_i)$ is just the residual;

$$\underset{\sim}{r} = \underset{\sim}{h} - \underset{\approx}{A}\,\underset{\sim}{x} \quad !$$

$$\therefore\ s_h^2 = \frac{1}{N-3}\,\|\underset{\sim}{r}\|_2^2 = 0.110$$

↳ you should really have more pts for accuracy!

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

We had:

$$\underset{\sim}{x} = \left\{\left[\underset{\approx}{A}^T\underset{\approx}{A}\right]^{-1}\underset{\approx}{A}^T\right\}\underset{\sim}{h}$$

$$\equiv \underset{\approx}{K}$$

Thus $x_i \equiv \sum_{j=1}^{N} K_{ij}\,h_j$

i'th row of $\underset{\approx}{K}$!

We can thus use the formula for a linear combination of random variables!

$$S_{x_i}^2 = \sum_{j=1}^{N} K_{ij}^2 S_{h_i}^2 = 0.0315$$

these are all the same in this problem!

Thus $g = 9.64 \pm 0.18$

to the 69% confidence level (1σ)

Often we want to develop correlations for their predictive value; for example what is the expected ht. of the lead wt. at some time t?

$$h_m = X_1 + X_2 t - \tfrac{1}{2} t^2 X_3 \; !$$

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:

$$S_{h_m}^2 = S_{x_1}^2 + (t)^2 S_{x_2}^2 + \left(-\tfrac{1}{2}t^2\right)^2 S_{x_3}^2$$

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:

$$S_{h_m}^2 = S_{x_1}^2 + (t) S_{x_2}^2 + \left(-\tfrac{1}{2}t^2\right)^2 S_{x_3}^2$$
$$+ 2(1)(t) S_{x_1 x_2}^2 + 2(1)\left(-\tfrac{1}{2}t^2\right) S_{x_1 x_3}^2$$
$$+ 2(t)\left(-\tfrac{1}{2}t^2\right) S_{x_2 x_3}^2$$

How do we calculate the covariance?

Recall:

$$x_i \equiv \sum_{j=1}^{N} K_{ij} h_j$$

where $\underset{\approx}{K} \equiv \left[\underset{\approx}{A}^T \underset{\approx}{A}\right]^{-1} \underset{\approx}{A}^T$

Thus we let:

$$x_1 = \sum_{j=1}^{N} \alpha_j h_j$$ where α is 1st row of $\underset{\approx}{K}$

and

$$x_2 = \sum_{j=1}^{N} \beta_j h_j$$ where $\underline{\beta}$ is the 2nd row of $\underset{\approx}{K}$

Now we need to determine the covariance

$$\sigma^2_{x_1 x_2} = E\left((x_1 - \mu_{x_1})(x_2 - \mu_{x_2})\right)$$

↑ value for x_1 & x_2 expected for an infinite sample

$$\mu_{x_1} = \sum_{j=1}^{N} \alpha_j \mu_{h_j}$$

↑ from infinite sample

$$\mu_{x_2} = \sum_{j=1}^{N} \beta_j \mu_{h_j}$$

So:

$$\sigma^2_{x_1 x_2} \equiv E\left(\left(\sum_{j=1}^{N} \alpha_j (h_j - \mu_{h_j})\right)\left(\sum_{j=1}^{N} \beta_j (h_j - \mu_{h_j})\right)\right)$$

Now if all of the h_j are <u>independent</u> then

$$E\left((h_j - \mu_{h_j})(h_i - \mu_{h_i})\right) = \underline{\underline{0}}$$

for $i \neq j$

$$\therefore \sigma^2_{x_1 x_2} = E\left(\sum_{j=1}^{N} \alpha_j \beta_j (h_j - \mu_{h_j})^2\right)$$

$$= \sum_{j=1}^{N} \alpha_j \beta_j \sigma_h^2$$

↑ provided all the h_j have the same uncertainty

$$\therefore \sigma^2_{x_1 x_2} = \underset{\sim}{\alpha}\, \underset{\sim}{\beta}^T \sigma_h^2$$

So for this problem if we let

$$\underset{\sim}{\alpha} \equiv 3^{rd} \text{ row of } \underset{\approx}{K}$$

$$x_1 = \underset{\sim}{\alpha}\, \underset{\sim}{h} \;,\; x_2 = \underset{\sim}{\beta}\, \underset{\sim}{h} \;,\; x_3 = \underset{\sim}{\gamma}\, \underset{\sim}{h}$$

$$s_h^2 = \frac{1}{N-3} \| \underset{\sim}{r} \|_2^2 \;,\; \underset{\sim}{r} = \underset{\sim}{h} - \underset{\approx}{A}\, \underset{\sim}{x}$$

$$s_{x_1}^2 = \left(\underset{\sim}{\alpha}\, \underset{\sim}{\alpha}^T\right) s_h^2$$

$$s_{x_2}^2 = \left(\underset{\sim}{\beta}\, \underset{\sim}{\beta}^T\right) s_h^2 \quad \text{etc.}$$

and:

$$h_m = (1)\, x_1 + (t)\, x_2 + \left(\tfrac{1}{2}t^2\right) x_3$$

$$\sigma_{h_m}^2 = (1)^2 \sigma_{x_1}^2 + (t)^2 \sigma_{x_2}^2 + \left(\tfrac{1}{2}t^2\right)^2 \sigma_{x_3}^2$$
$$+ 2(1)(t)\, \underset{\sim}{\alpha}\, \underset{\sim}{\beta}^T \sigma_h^2 + 2(1)\left(\tfrac{1}{2}t^2\right) \underset{\sim}{\alpha}\, \underset{\sim}{\gamma}^T \sigma_h^2$$
$$+ 2(t)\left(\tfrac{1}{2}t^2\right) \underset{\sim}{\beta}\, \underset{\sim}{\gamma}^T \sigma_h^2$$
$$= \left\| \left[(1)\, \underset{\sim}{\alpha} + (t)\, \underset{\sim}{\beta} + \left(\tfrac{1}{2}t^2\right) \underset{\sim}{\gamma} \right] \right\|_2^2 \sigma_h^2$$

If we represent the vector of modelling functions as $\underset{\sim}{a}(t)$, we can collapse this:

$$\underset{\sim}{a}(t) \equiv \left(1, t, -\tfrac{1}{2}t^2\right)$$

Then

$$\sigma_{h_m}^2 = \| \underset{\sim}{a} \underset{\approx}{K} \|_2^2 \, \sigma_h^2$$

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:

$$\underset{\sim}{b} \approx \underset{\approx}{A} \underset{\sim}{x}$$

where $A_{ij} = \phi_j(t_i)$

We define the residual:

$$\underset{\sim}{r} \equiv \underset{\sim}{b} - \underset{\approx}{A}\,\underset{\sim}{x}$$

We define the best fit modelling parameters $\underset{\sim}{x}$ by:

$$\min_{\underset{\sim}{x}} \| \underset{\sim}{r} \|_2^2 \equiv \min_{\underset{\sim}{x}} \left(\underset{\sim}{r}^T \underset{\sim}{r} \right)$$

$\underset{\sim}{x}$ has the solution:

$$\underset{\approx}{A}^T \underset{\approx}{A}\, \underset{\sim}{x} = \underset{\approx}{A}^T \underset{\sim}{b}$$

or

$$\underset{\sim}{x} = \left[\underset{\approx}{A}^T \underset{\approx}{A} \right]^{-1} \underset{\approx}{A}^T \underset{\sim}{b} \equiv \underset{\approx}{K}\, \underset{\sim}{b}$$

Note that $\underset{\approx}{A}$ is an $n \times m$ matrix, $\underset{\sim}{b}$ is a $n \times 1$ array, $\underset{\sim}{x}$ is a $m \times 1$ array and $\underset{\approx}{K}$ is a $m \times n$ matrix

If we recall that:

$$\sigma_z^2 = E\left\{ (z - \mu_z)^2 \right\}$$

where $\mu_z = E(z)$

Then the matrix of covariance of the regression parameters is:

$$\underset{\approx}{\Sigma}_x^2 \equiv E\left\{(\underset{\sim}{x} - \underset{\sim}{\mu}_x)(\underset{\sim}{x} - \underset{\sim}{\mu}_x)^T\right\}$$

Substituting $\underset{\sim}{x} = \underset{\approx}{K}\,\underset{\sim}{b}$ and $\underset{\sim}{\mu}_x = \underset{\approx}{K}\,\underset{\sim}{\mu}_b$ we get:

$$\underset{\approx}{\Sigma}_x^2 \equiv \underset{\approx}{K}\, E\left\{(\underset{\sim}{b} - \underset{\sim}{\mu}_b)(\underset{\sim}{b} - \underset{\sim}{\mu}_b)^T\right\} \underset{\approx}{K}^T$$

$$= \underset{\approx}{K}\, \underset{\approx}{\Sigma}_b^2\, \underset{\approx}{K}^T$$

where $\underset{\approx}{\Sigma}_b^2$ is the matrix of covariance of the observations $\underset{\sim}{b}$.

If the observations are <u>independent</u> then $\underset{\approx}{\Sigma}_b^2$ will be a diagonal matrix.

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

$$\underset{\approx}{\Sigma}_b^2 = \underset{\approx}{I}\,\sigma_b^2$$

and we get

$$\underset{\approx}{\Sigma}_x^2 = \sigma_b^2 \underset{\approx}{K}\underset{\approx}{K}^T$$

We may estimate σ_b^2 from:

$$\sigma_b^2 \approx s_b^2 \equiv \frac{1}{n-m} \| \underset{\approx}{A}\underset{\sim}{x} - \underset{\sim}{b} \|_2^2$$

$$= \frac{1}{n-m} \underset{\sim}{r}^T \underset{\sim}{r}$$

We may calculate the error in model predictions similarly. Suppose we wish to determine the model value at K times $\underset{\sim}{t}^{(m)}$. Thus:

$$\underset{\sim}{b}^{(m)} = \underset{\approx}{A}^{(m)} \underset{\sim}{x}$$

where $A_{ij}^{(m)} = \phi_j(t_i^{(m)})$

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

What is the error in $\underset{\sim}{b}^{(m)}$? We have the matrix of covariance:

$$\underset{\approx}{\Sigma}_{b^{(m)}}^2 \equiv E\left\{ (\underset{\sim}{b}^{(m)} - \underset{\sim}{\mu}_{b^{(m)}})(\underset{\sim}{b}^{(m)} - \underset{\sim}{\mu}_{b^{(m)}})^T \right\}$$

$$= \underset{\approx}{A}^{(m)} E\left\{ (\underset{\sim}{x} - \underset{\sim}{\mu}_x)(\underset{\sim}{x} - \underset{\sim}{\mu}_x)^T \right\} \underset{\approx}{A}^{(m)T}$$

$$= \underset{\approx}{A}^{(m)} \underset{\approx}{\Sigma}_x^2 \underset{\approx}{A}^{(m)T}$$

where $\underset{\approx}{\Sigma}_x^2$ is the matrix of covariance determined before! In general, only the diagonal elements of $\underset{\approx}{\Sigma}_{b^{(m)}}$ are of interest:

$$\underset{\sim}{\sigma}_{b^{(m)}}^2 = \text{Diag}\left\{ \underset{\approx}{\Sigma}_{b^{(m)}}^2 \right\}$$

which yields the error in the model predictions.

As a final note on linear regression, let us examine the case where the variance of $\underset{\sim}{\sigma}_b^2$ is <u>not</u> constant!

In this case we wish to weight the points in our regression analysis differently.

This is <u>weighted</u> linear regression!

Suppose we have the <u>diagonal</u> matrix $\underset{\approx}{\Sigma}_b^2$. The weighted problem is just:

$$\min_{\underset{\sim}{x}} (\underset{\approx}{A}\underset{\sim}{x} - \underset{\sim}{b})^T (\underset{\approx}{\Sigma}_b^2)^{-1} (\underset{\approx}{A}\underset{\sim}{x} - \underset{\sim}{b})$$

$$= \min_{\underset{\sim}{x}} \underset{\sim}{r}^T (\underset{\approx}{\Sigma}_b^2)^{-1} \underset{\sim}{r}$$

where $(\underset{\approx}{\Sigma}_b^2)^{-1}$ is a diagonal matrix whose elements are just $\frac{1}{\sigma_{b_i}^2}$.

Thus we weight each point by the inverse of the variance of that point. This makes sense. Suppose the reason why σ_{b_i} varied was that m_i points were <u>averaged</u> together to get b_i.

Then we have $\sigma_{b_i}^2 = \frac{\sigma_b^2}{m_i}$ and

thus the above regression yields:

$$\frac{1}{\sigma_b^2} \min_{\underset{\sim}{x}} \underset{\sim}{r}^T \begin{pmatrix} m_1 & & 0 \\ & m_2 & \\ 0 & & \ddots \\ & & & m_n \end{pmatrix} \underset{\sim}{r} ,$$

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

Ok, what is the formula for $\underset{\sim}{x}$?

Just as before we take the gradient:

$$\nabla_{\underset{\sim}{x}} \left\{ \underset{\sim}{r}^T \left(\underset{\approx}{\Sigma}_b^{-1} \right) \underset{\sim}{r} \right\} = 0$$

which yields:

$$\underset{\approx}{A}^T \left(\underset{\approx}{\Sigma}_b^2 \right)^{-1} \underset{\approx}{A} \underset{\sim}{x} = \underset{\approx}{A}^T \left(\underset{\approx}{\Sigma}_b^2 \right)^{-1} \underset{\sim}{b}$$

or $\underset{\sim}{x} = \underset{\approx}{K}_w \underset{\sim}{b} \equiv \left\{ \underset{\approx}{A}^T \left(\underset{\approx}{\Sigma}_b^2 \right)^{-1} \underset{\approx}{A} \right\}^{-1} \underset{\approx}{A}^T \left(\underset{\approx}{\Sigma}_b^2 \right)^{-1} \underset{\sim}{b}$

with error: $\underset{\approx}{\Sigma}_x^2 = \underset{\approx}{K}_w \underset{\approx}{\Sigma}_b^2 \underset{\approx}{K}_w^T$

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:

$$b_i \approx f(\underset{\sim}{x}, t_i)$$

where f may be linear or non-linear in $\underset{\sim}{x}$. 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 $\underset{\sim}{x}$, e.g. we wish:

$$\underset{\sim}{\underset{\sim}{\Sigma}}^2_{\underset{\sim}{x}} \equiv E\left\{(\underset{\sim}{x} - \underset{\sim}{\mu}_x)(\underset{\sim}{x} - \underset{\sim}{\mu}_x)^T\right\}$$

If the # of pts N is fairly large, we can do this with <u>undersampling</u>

Suppose we take every m^{th} point. This gives us a set of N/m points from which we can calculate $\underset{\sim}{x}$. If we do this for the m indep. sets of N/m points, we get m estimates for $\underset{\sim}{x}$!

We can then get a sample covariance matrix:

$$\underset{\sim}{\underset{\sim}{\Sigma}}^2_{\bar{x}_i \bar{x}_j} \approx \frac{1}{m-1}\left\{\overline{x_i x_j} - \bar{x}_i \bar{x}_j\right\}$$

Note that this did <u>not</u> require knowing the error in $\underset{\sim}{x}$! 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 $\underset{\sim}{x}$. 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 <u>approximation</u> to this infinite set by periodically <u>replicating</u> our original data set!

The idea is to then take samples of 20 observations from this replicated set and calculate $\underset{\sim}{x}$. We then compute $\underset{\approx}{\Sigma}^2_x$ 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:

$$\underset{\sim}{\overset{z}{\sum}}_{x_i x_j} = \frac{p}{p-1}\left\{\overline{x_i x_j} - \bar{x}_i \bar{x}_j\right\}$$

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 on *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:

$$\underset{\sim}{b} = f(\underset{\sim}{x}, t)$$

where $\underset{\sim}{x}$ is an array of modelling functions. We may still define the *residual* between observed and fitted values:

$$r_i = b_i - f(\underset{\sim}{x}, t_i)$$

r_i: residual, b_i: meas, $f(\underset{\sim}{x}, t_i)$: model

Thus we can form the sum of squares:

$$F(\underset{\sim}{x}) = \sum_{i=1}^{N} (r_i)^2$$

The best fitted values for $\underset{\sim}{x}$ are obtained via finding the minimum of $F(\underset{\sim}{x})$!

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 $F(\underset{\sim}{x})$ is well conditioned. The dependence on each parameter should be of similar magnitude. Let's look at an example:

Arrhenius Kinetics

$$\text{rate} \sim K_0 e^{-E/RT}$$

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 exp't is written up in Chem Eng. Education, 36 (1) p. 34-40. Suppose we feed a reactor (well-mixed) a

concentration of methane C_{r_0} at flow rate Q_r. We measure some outlet concentration C_r. The subscript "r" is because the reactor is hot: due to expansion, the actual concentration is $\frac{C}{C_r} = \frac{T_r}{T}$ from the ideal gas law - where T_r is some ref. temp (°K).

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 rxn per gram of catalyst:

$$\frac{\varrho_r}{m}(C_{r0} - C_r) = \left(C_r \frac{T_r}{T}\right)^n K_0 e^{-\frac{E}{RT}}$$

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

$$X = 1 - \frac{C_r}{C_{r_0}}$$

which yields (after rearrangement):

$$\frac{X}{(1-X)^n} = \frac{m}{\varrho_r} C_{0r}^{n-1} \left(\frac{T_r}{T}\right)^n K_0 e^{-\frac{E}{RT}}$$

If you were to fix T and plot X vs. C_{0r}, you find it <u>decreases</u> w/ C_{0r}, thus $n < 1$ for this system.

Ok, suppose we vary C_{w0} 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

$$\ln\left\{\frac{q_w}{m}(C_{ow} - C_w)\right\} \text{ vs. } \ln\{C_w\}$$

You should get a straight line with slope n!

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

$$\ln\left\{\frac{q_w}{m}\frac{(C_{ow} - C_w)}{C_w^n}\left(\frac{T_w}{T}\right)^{-n}\right\} = \ln k_0 - \frac{E}{R}\frac{1}{T}$$

Thus, the slope (vs. $\frac{1}{T}$) is just E/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 C_P is a *much* stronger function of T than C_{P_0}, 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 Δ:

$$\Delta = C_r - C_{r_0} + \frac{m}{\rho_r} C_r^n K_0 e^{-\frac{E}{RT}} \left(\frac{T_w}{T}\right)^n$$

and minimize the objective function:

$$F(K_0, E, n) \equiv \sum \Delta_i^2$$

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 $\frac{E}{RT_w}$ as fitting parameters so that the optimization problem is well-conditioned!