Systematic Errors (2)
Working with Systematic Errors

Roger Barlow
Huddersfield University

Aachen Online Statistics School

16^{th} March 2023

Why do we quote systematic errors separately?

Results are always given like

In conclusion, we have measured $m = 12.1 \pm 0.3 \pm 0.4$, where the first error is statistical and the second is systematic

Or even '$\pm$ statistical, $\pm$systematic, $\pm$luminosity uncertainty, $\pm$theory uncertainty, $\pm$branching ratio uncertainty'

Why quote them separately?

Why not just 12.1 ± 0.5?

Minor reason - shows whether result is statistics limited
Major reason - to enable combination of this result with others that share a systematic uncertainty

Errors with Correlations

What is the error on $f(x, y)$?

For undergraduates

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2$$

For graduates

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\rho \left(\frac{\partial f}{\partial x}\right) \left(\frac{\partial f}{\partial y}\right) \sigma_x \sigma_y$$

If there are several functions and several variables this generalises to

$$\mathbf{V}_f = \tilde{\mathbf{G}} \mathbf{V_x} \mathbf{G} \tag{1}$$

where V_f and V_x are the covariance matrices and $G_{ij} = \frac{\partial f_j}{\partial x_i}$

Example - the straight line fit

Note: for compatibility with traditional usage, x is now called y

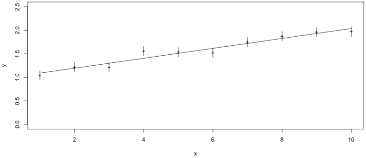
$$y = mx + c$$

$$f_1 \equiv m = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\sum (x_i - \overline{x}) y_i}{N(\overline{x^2} - \overline{x}^2)}$$

$$f_0 \equiv c = \overline{y} - m\overline{x} = \frac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} = \frac{\sum (\overline{x^2} - x_i \overline{x}) y_i}{N(\overline{x^2} - \overline{x}^2)}$$

$$\mathbf{V_y} = \sigma^2 \mathbf{I}$$

$$G_{i1} = \frac{x_i - \overline{x}}{N(\overline{x^2} - \overline{x}^2)} \qquad G_{i0} = \frac{\overline{x^2} - x_i \overline{x}}{N(\overline{x^2} - \overline{x}^2)}$$



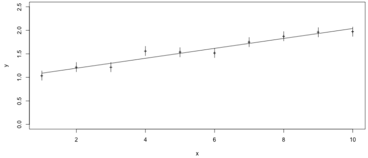
Equation 1 gives the usual errors, and also the correlation:

$$V_m = \frac{\sigma^2}{N(\overline{x^2} - \overline{x}^2)} \qquad V_c = \frac{\sigma^2 \overline{x^2}}{N(\overline{x^2} - \overline{x}^2)} \qquad Cov = -\frac{\overline{x}\sigma^2}{N(\overline{x^2} - \overline{x}^2)} \qquad \rho = -\frac{\overline{x}}{\sqrt{\overline{x^2}}}$$

in this example, $m = 0.105 \pm 0.011, c = 0.983 \pm 0.068, \rho = -0.886$

Even though the y_i are independent, m and c are correlated

Example - the straight line fit



Correlation $\rho = -\frac{\overline{x}}{\sqrt{\overline{x^2}}}$

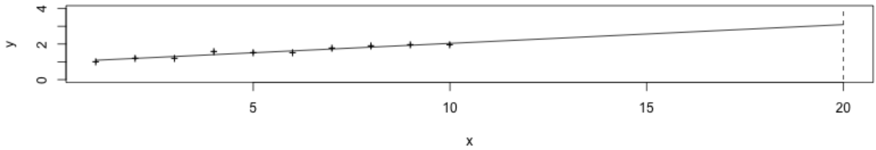
Fluctuations in measurement(s) affect slope and intercept in opposite directions.

Correlation vanishes if $\overline{x} = 0$. Or write $y = m(x - \overline{x}) + c'$

Re-parametrising to kill correlation is sometimes worth doing.

Example - the straight line fit
Continued

Extrapolation of a straight line - what is y at $x = 20$?



$y = 0.983 + 20 \times 0.105$
Error from $\sqrt{0.068^2 + 20^2 \times 0.011^2} = 0.23$ Wrong
Correct Error from
$\sqrt{0.068^2 + 20^2 \times 0.011^2 - 2 \times 0.886 \times 20 \times 0.068 \times 0.011} = 0.16$

Building a correlation matrix

or covariance matrix, or variance matrix...

Matrix element $V_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$

Given correlated x_1 and x_2, model as $x_1 = y_1 + z, x_2 = y_2 + z$, where y_1, y_2, z independent with errors σ_1, σ_2, S.

$V_{11} = \langle (y_1 + z)(y_1 + z) \rangle - \langle (y_1 + z) \rangle^2 = \sigma_1^2 + S^2$.
V_{22} similar
$V_{12} = V_{21} = \langle (y_1 + z)(y_2 + z) \rangle - \langle (y_1 + z) \rangle \langle (y_2 + z) \rangle = S^2$

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + S^2 & S^2 \\ S^2 & \sigma_2^2 + S^2 \end{pmatrix}$$

For more variables, build up larger matrix where off-diagonal elements come from shared features, on-diagonal gives total variance.

Building a correlation matrix

continued

Suppose experiment A measures y_1 and y_2 with shared systematic uncertainty S_A, and experiment B measures y_3 and y_4 with shared S_B

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + S_A^2 & S_A^2 & 0 & 0 \\ S_A^2 & \sigma_2^2 + S_A^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 + S_B^2 & S_B^2 \\ 0 & 0 & S_B^2 & \sigma_4^2 + S_B^2 \end{pmatrix}$$

Similar for (more common) shared multiplicative uncertainty - (e.g. efficiency, luminosity, normalisation...)

$y_1 \pm \sigma_1 \pm S_1$ and $y_2 \pm \sigma_2 \pm S_2$ with $S_1 = \xi y_1, S_2 = \xi y_2$

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 + S_1^2 & S_1 S_2 \\ S_1 S_2 & \sigma_2^2 + S_2^2 \end{pmatrix}$$

PDG, HFLAV and similar groups do this on an industrial scale

Using the matrix

Independent measurements

Maximum Likelihood $\rightarrow$ Least Squares $\rightarrow$ minimise $\chi^2 = \sum_i \left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2$

What if the y_i are not independent but correlated with non-diagonal covariance matrix V_y?
Change to some $\mathbf{y}' = \mathbf{R}\mathbf{y}$ with rotation matrix $\mathbf{R}$ such that all $Cov(y'_i, y'_j) = 0$

V' diagonal by construction. $\mathbf{V}'^{-1} = \begin{pmatrix} 1/\sigma_1'^2 & 0 & 0 & ... \\ 0 & 1/\sigma_2'^2 & 0 & ... \\ 0 & 0 & 1/\sigma_3'^2 & ... \\ ... & & & \end{pmatrix}$

$\mathbf{y}' = \mathbf{R}\mathbf{y}$ so $\mathbf{V}' = [\tilde{R} V^{-1} R]^{-1}$ and
$\chi^2 = (\tilde{\mathbf{y}}' - \tilde{\mathbf{f}}')\mathbf{V}'^{-\mathbf{1}}(\mathbf{y}' - \mathbf{f}') = (\tilde{\mathbf{y}} - \tilde{\mathbf{f}})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{f})$
Forget about the primed system and get $\chi^2 = (\tilde{\mathbf{y}} - \tilde{\mathbf{f}})\mathbf{V}^{-1}(\mathbf{y} - \mathbf{f})$

How does this all link to the Hessian matrix? (1)

$$\frac{\partial^2 \ln L}{\partial a_i \partial a_j}$$

$\hat{a}_1$ and $\hat{a}_2$ are functions of the data: maximise
$\ln L(a_1, a_2) = \sum_i \ln P(x_i; a_1, a_2)$

That means $\frac{\partial \ln L}{\partial a_i}|_{a=\hat{a}} = 0 \qquad \forall i$

Expanding this to first order about a^{true}, as

$$\frac{\partial \ln L}{\partial a_1}|_{a=a^{true}} + \frac{\partial^2 \ln L}{\partial a_1^2}(\hat{a}_1 - a_1^{true}) + \frac{\partial^2 \ln L}{\partial a_1 \partial a_2}(\hat{a}_2 - a_2^{true}) = 0$$

$$\frac{\partial \ln L}{\partial a_2}|_{a=a^{true}} + \frac{\partial^2 \ln L}{\partial a_1 \partial a_2}(\hat{a}_1 - a_1^{true}) + \frac{\partial^2 \ln L}{\partial^2 a_2}(\hat{a}_2 - a_2^{true}) = 0$$

So $\mathbf{H}(\hat{\mathbf{a}} - \mathbf{a}^{\mathbf{true}}) = -\frac{\partial \ln L}{\partial \mathbf{a}}|_{a=a^{true}}$ and $\hat{\mathbf{a}} - \mathbf{a}^{\mathbf{true}} = -\mathbf{H}^{-1}\frac{\partial \ln L}{\partial \mathbf{a}}|_{a=a^{true}}$

Now apply Equation 1 with $\mathbf{G} = \mathbf{H}^{-1}$

How does this all link to the Hessian matrix? (2)

We need to know the variance matrix $\mathbf{V}$ of the gradients $\frac{\partial \ln L}{\partial a_i}|_{a=a^{true}}$

This is $\left\langle \frac{\partial \ln L}{\partial a_i} \frac{\partial \ln L}{\partial a_j} \right\rangle - \left\langle \frac{\partial \ln L}{\partial a_i} \right\rangle \left\langle \frac{\partial \ln L}{\partial a_j} \right\rangle$. evaluated at $\mathbf{a} = \mathbf{a^{true}}$

Unitarity says $\int ... \int L dx_1\, dx_2 ... dx_N = 1$, and differentiating wrt any a_i must give zero, so

$\int ... \int \frac{\partial L}{\partial a_i} dx_1\, dx_2 ... dx_N = \int ... \int L \frac{\partial \ln L}{\partial a_i} dx_1\, dx_2 ... dx_N = \left\langle \frac{\partial \ln L}{\partial a_i} \right\rangle = 0$

Differentiating again, and using the $\frac{\partial \ln L}{\partial a} = \frac{1}{L} \frac{\partial L}{\partial a}$ switch, gives

$\left\langle \frac{\partial \ln L}{\partial a_j} \frac{\partial \ln L}{\partial a_k} \right\rangle = - \left\langle \frac{\partial^2 \ln L}{\partial a_j \partial a_k} \right\rangle$

Now we approximate the expectation values by actual values we see and get $\mathbf{V} = -\mathbf{H}$

and Equation 1 gives $\mathbf{V_{\hat{a}}} = -\mathbf{H^{-1}}$

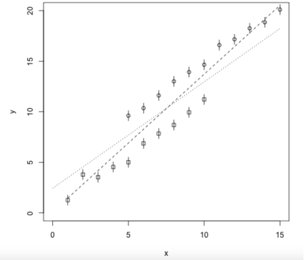
Averaging
BLUE

Given several (correlated) results y_i, how do you average them?
Best Linear Unbiased Estimator (L Lyons et al, NIM **A270** 110 (1988))
Minimise $\chi^2 = \sum_{i,j}(y_i - \hat{y})V_{ij}^{-1}(y_j - \hat{y})$
$\hat{y}\sum_{i,j} V_{ij}^{-1} = \sum_{i,j} V_{ij}^{-1} y_j$
Write as $\hat{y} = \sum_i w_i y_i$ with $w_i = \frac{\sum_j V_{ij}^{-1}}{\sum_{i,j} V_{ij}^{-1}}$
Error on $\hat{y}$ given by $\sqrt{\tilde{\mathbf{w}}\mathbf{V}\mathbf{w}}$
Notice that $\sum_i w_i = 1$ which is intuitive
Notice that some w_i may be negative (if correlations are large) which is counterintuitve

This assumes the elements of **V** are known exactly. If not, care needed.

Equivalent alternative for additive systematics

Fit parameters using several datasets each with some systematic additive uncertainty S_j



Method 1 For $j = 1...n$ experiments, construct large covariance matrix $\mathbf{V}$ with S_j^2 off-diagonal elements and minimise χ^2
Method 2 introduce explicit offsets.
$y'_{ij} = y_{ij} + \xi_j$ for value i of experiment j. ξ_j Gaussian with mean 0, sd S_j, included in χ^2
Fit the ξ_i together with the parameter(s) of interest. Variance matrix larger but now diagonal.

Which method should you use?

Method 2
Downside: n more parameters to fit
Upside (1): avoids matrix inversion
Upside (2): extracts the factors which can be useful to check behaviour

Which method should you use?

Method 2

Downside: n more parameters to fit

Upside (1): avoids matrix inversion

Upside (2): extracts the factors which can be useful to check behaviour

These two methods are actually (surprisingly!) equivalent

R.B. *Combining experiments with systematic errors.* NIM **A987** 164864 (2021)

Also Method 2 with multiplicative errors applied to prediction avoids 'D'Agostini bias' (G. D'Agostini NIM **A346** 306 (1994))

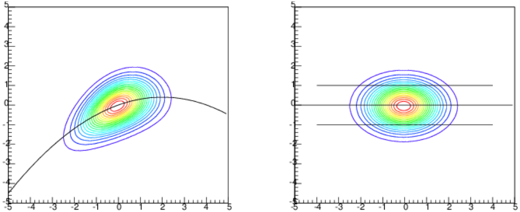
Adjust parameter(s) a to minimise $\chi^2 = (\tilde{\mathbf{y}} - \tilde{\mathbf{f}}(x;a))\mathbf{V}^{-1}(\mathbf{y} - \mathbf{f}(x;a))$

Bias possible if $\mathbf{V}$ includes normalising systematic errors:

$S_i = fy_i$ so increasing value increases error and lowers χ^2

Indicates separate fit to systematic factors is preferable in some cases

Profile Likelihood - motivation (not very rigorous)



You have a 2D likelihood plot with axes a_1 and a_2. You are interested in a_1 but not in a_2 ('Nuisance parameter')
Different values of a_2 give different results (central and errors) for a_1
Suppose it is possible to transform to $a_2'(a_1, a_2)$ so L factorises, like the one on the right. $L(a_1, a_2') = L_1(a_1)L_2(a_2')$
Whatever the value of a_2', get same result for a_1
So can present this result for a_1, independent of anything about a_2'.
Path of central a_2' value as fn of a_1, is peak - path is same in both plots
So no need to factorise explicitly: plot $L(a_1, \hat{\hat{a}}_2)$ as fn of a_1 and read off 1D values.
$\hat{\hat{a}}_2(a_1)$ is the value of a_2 which maximises $\ln L$ for this a_1

Nuisance Parameters 2

Marginalised likelihoods

Instead of profiling, just integrate over a_2.
Can be very helpful alternative, specially with many nuisance parameters
But be aware - this is strictly Bayesian

Frequentists are not allowed to integrate likelihoods wrt the parameter

$\int P(x;a)\,dx$ is fine, but $\int P(x;a)\,da$ is off limits

Reparametrising a_2 (or choosing a different prior) will give different values for a_1. With a bit of luck, even radical changes in the prior for a_2 will not effect the frequentist result for a_1.
But don't just leave it to luck. Check and make sure.

Conclusions

Systematic errors can readily be handled - with the help of the correlation matrix and other techniques