Error analysis in biology

Marek Gierliński
Division of Computational Biology

Hand-outs available at http://is.gd/statlec

Previously on errors…

How to make a good plot

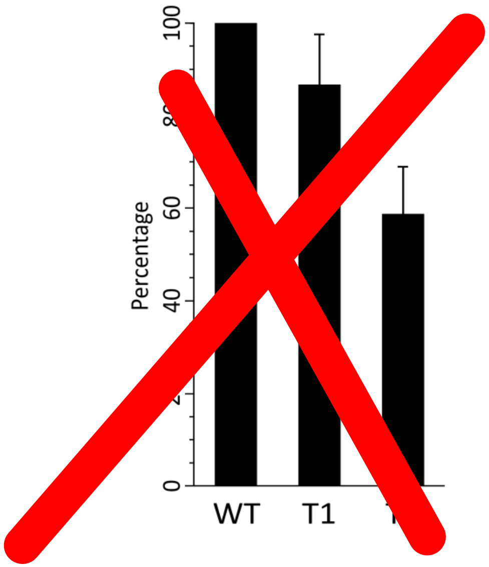
- Clarity of presentation!
- Good lines and symbols
- Clear labels
- Logarithmic plots

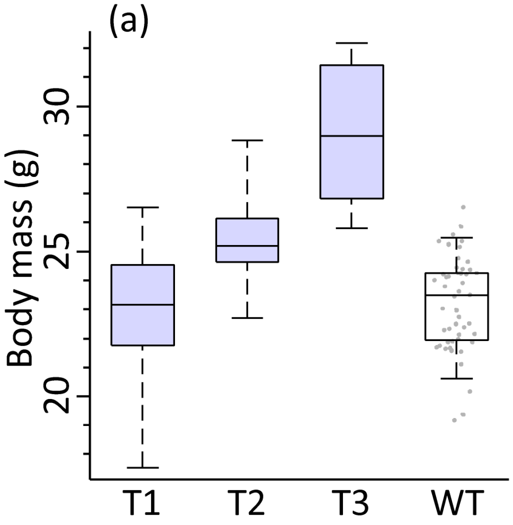
Box plots

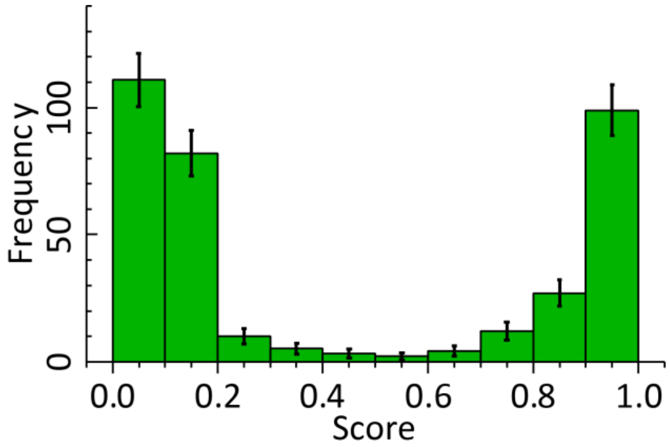
- Good alternative for SD
- Show distribution of data

Bar plots

- Only for additive quantities
- Baseline must be zero
- Never in log space
- Careful with continuous variable
- Always show upper and lower error bars!







Significant figures

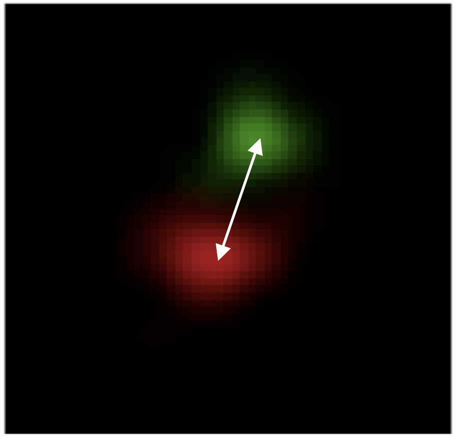
- Carry meaningful information
- Quote only significant figures
- Use error in the error

Number	1.23457456
Error	0.02377345

Example

- Measure positions of two fluorescent dots under a microscope (in μm)

	x	y	z
Dot 1	3.68	3.12	5.44
Dot 2	3.90	3.86	4.02



- Measurement errors for x-y and z direction:
  - □ $\Delta_{xy} = 120$ nm
  - □ $\Delta_z = 200$ nm

- Find distance between the dots

$$R = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} = 1.62\ \mu\text{m}$$

- What is the error of R?
- We need to **propagate** errors of x, y and z

7. Error propagation

"If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is"

"Young man, in mathematics you don't understand things; you just get used to them"

John von Neumann

Derivative

- Consider a function $y = f(x)$
- Derivative of f

$$\frac{df}{dx} \approx \frac{\text{change in } y}{\text{change in } x} = \frac{\Delta y}{\Delta x} \qquad \text{for small } \Delta x$$

- Derivative = slope

A few derivatives

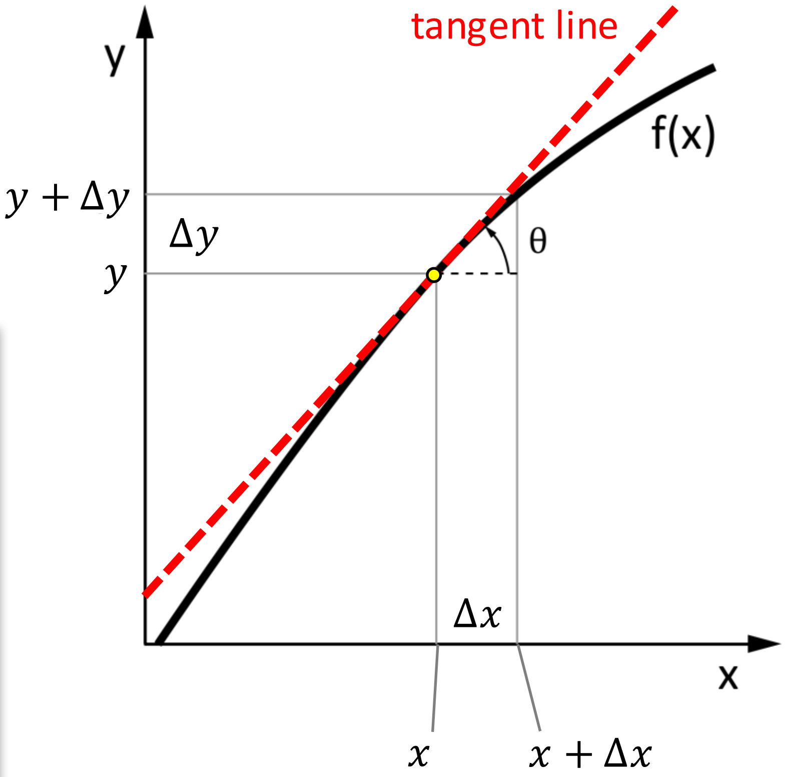
$$\frac{d}{dx} ax = a$$

$$\frac{d}{dx} x^r = r x^{r-1}$$

$$\frac{d}{dx} x^2 = 2x$$

$$\frac{d}{dx} \frac{1}{x} = -\frac{1}{x^2}$$

$$\frac{d}{dx} \log_a x = \frac{1}{x \ln a}$$



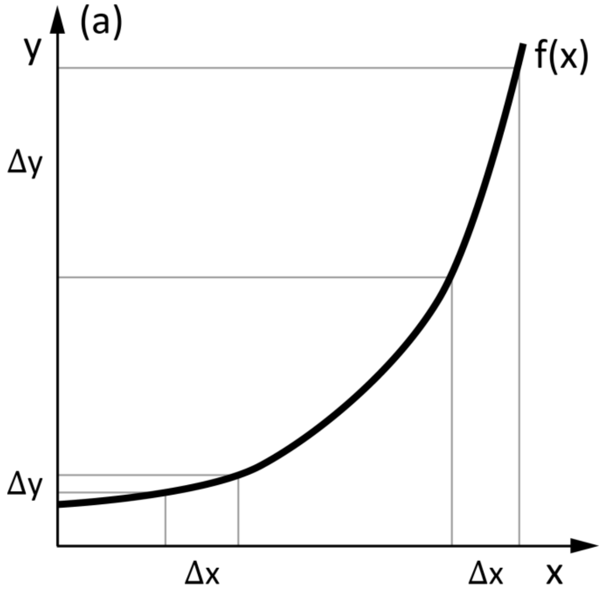
$$\boxed{\frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}}$$

Error propagation (single variable)

- Consider a quantity $x \pm \Delta x$

- Transform to a new variable $y = f(x)$
- For example
  - □ $y = ax$
  - □ $y = \log(x)$
  - □ $y = \sqrt{x}$

- Find error of y, Δy



Error propagation (single variable)

- Consider a quantity $x \pm \Delta x$

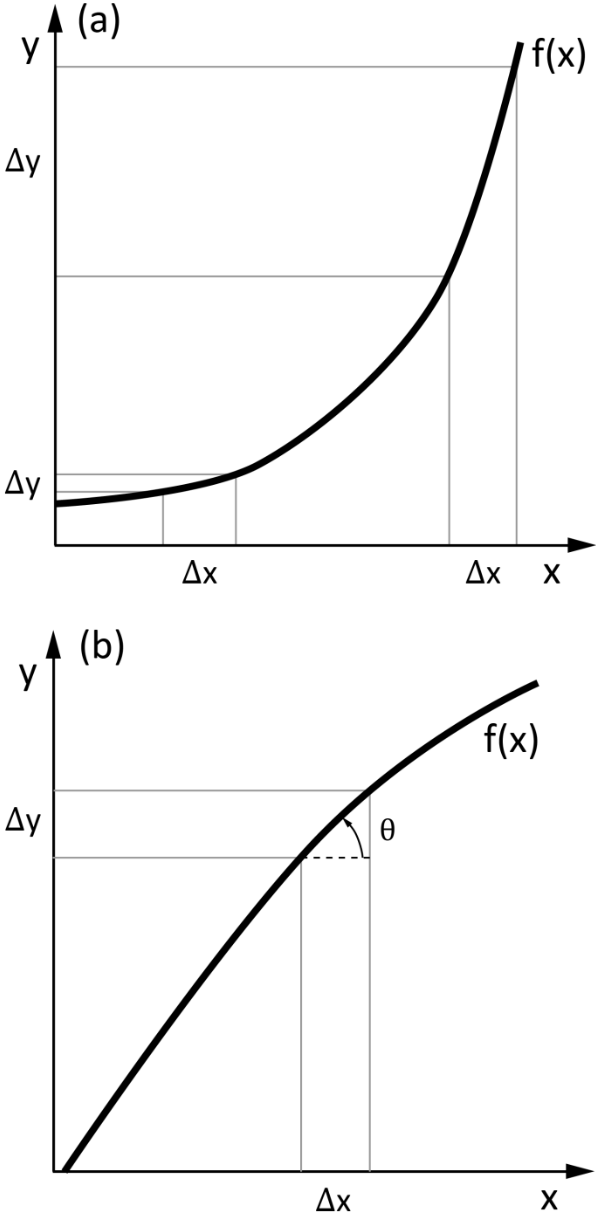
- Transform to a new variable $y = f(x)$
- Find error of y, Δy

- If errors are small

$$\frac{\Delta y}{\Delta x} \approx \frac{df}{dx}$$

- Hence

$$\Delta y \approx \left|\frac{df}{dx}\right| \Delta x \quad \text{or} \quad \Delta y^2 \approx \left(\frac{df}{dx}\right)^2 \Delta x^2$$



Error propagation: scaling

$x = 5 \pm 3$

$y = 10x$

$y = 50 \pm ?$

$y = f(x) = ax$

$$\Delta y = \left|\frac{df}{dx}\right| \Delta x = |a| \Delta x$$

$$\frac{d}{dx} ax = a$$

$\Delta y = |a| \Delta x$

$\Delta y = |10| \times 3 = 30$

$y = 50 \pm 30$

Error propagation: logarithm

$$x = 4.0 \pm 0.5$$

$$y = \log_2 x$$

$$y = 2 \pm ?$$

$$y = f(x) = \log_2 x$$

$$\Delta y = \left|\frac{df}{dx}\right| \Delta x = \left|\frac{1}{x \ln 2}\right| \Delta x$$

$$\Delta y \approx 1.44 \frac{\Delta x}{x}$$

$$\frac{d}{dx} \log_a x = \frac{1}{x \ln a}$$

$$\Delta y \approx 1.44 \frac{0.5}{4} = 0.18$$

$$y = 2.0 \pm 0.2$$

Error propagation: many variables

- Consider n **independent** (not correlated) variables $x_1, x_2, \ldots, x_n$
- Each of them with error $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$

- New variable $y = f(x_1, x_2, \ldots, x_n)$
- For example

$$y = x_1 + x_2$$

$$y = \frac{x_1}{x_2}$$

- It can be shown that

$$\mathit{\Delta} y^2 \approx \left(\frac{\partial f}{\partial x_1}\right)^2 \mathit{\Delta} x_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \mathit{\Delta} x_2^2 + \cdots + \left(\frac{\partial f}{\partial x_n}\right)^2 \mathit{\Delta} x_n^2$$

Error propagation: sum or difference

$x_1 = 8 \pm 3$

$x_2 = 10 \pm 4$

$x_1 + x_2 = 18 \pm ?$

$y = f(x_1, x_2) = x_1 + x_2$

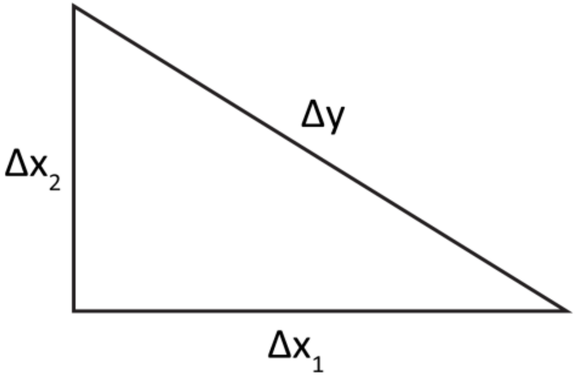
$$\Delta y^2 = \left(\frac{\partial f}{\partial x_1}\right)^2 \Delta x_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \Delta x_2^2$$

$$\Delta y^2 = \Delta x_1^2 + \Delta x_2^2$$

$$\frac{\partial f}{\partial x_1} = 1$$

$$\frac{\partial f}{\partial x_2} = 1$$

Geometrical interpretation



$\Delta y^2 = 3^2 + 4^2 = 25$

$\Delta y = 5$

$x_1 + x_2 = 18 \pm 5$

Error propagation: ratio or product

$$x_1 = 25.0 \pm 2.5$$

$$x_2 = 10 \pm 1$$

$$\frac{x_1}{x_2} = 2.5 \pm ?$$

$$y = f(x_1, x_2) = \frac{x_1}{x_2}$$

$$\Delta y^2 = \left(\frac{\partial f}{\partial x_1}\right)^2 \Delta x_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \Delta x_2^2$$

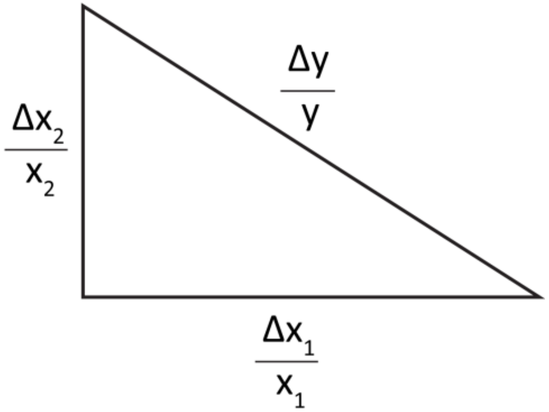
$$\Delta y^2 = y^2 \left[\left(\frac{\Delta x_1}{x_1}\right)^2 + \left(\frac{\Delta x_2}{x_2}\right)^2\right]$$

$$\left(\frac{\Delta y}{y}\right)^2 = \left(\frac{\Delta x_1}{x_1}\right)^2 + \left(\frac{\Delta x_2}{x_2}\right)^2$$

$$\frac{\partial f}{\partial x_1} = \frac{1}{x_2}$$

$$\frac{\partial f}{\partial x_2} = -\frac{x_1}{x_2^2}$$

Geometrical interpretation



$$\left(\frac{\Delta y}{y}\right)^2 = \left(\frac{2.5}{25}\right)^2 + \left(\frac{1}{10}\right)^2 = 0.02$$

$$\frac{\Delta y}{y} \approx 0.14, \qquad \Delta y \approx 0.35$$

$$\frac{x_1}{x_2} = 2.5 \pm 0.4$$

Fluorescent dots

- Two dots: (x_1, y_1, z_1) and (x_2, y_2, z_2)

$$R = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

- Propagated error of $R(x_1, y_1, z_1, x_2, y_2, z_2)$:

$$\Delta R^2 = \left(\frac{\partial R}{\partial x_1}\right)^2 \Delta_{xy}^2 + \left(\frac{\partial R}{\partial y_1}\right)^2 \Delta_{xy}^2 + \left(\frac{\partial R}{\partial z_1}\right)^2 \Delta_z^2$$

$$+ \left(\frac{\partial R}{\partial x_2}\right)^2 \Delta_{xy}^2 + \left(\frac{\partial R}{\partial y_2}\right)^2 \Delta_{xy}^2 + \left(\frac{\partial R}{\partial z_2}\right)^2 \Delta_z^2$$

- Homework: do the calculations and confirm that

$$\Delta R = \frac{\sqrt{2}}{R} \sqrt{[(x_1 - x_2)^2 + (y_1 - y_2)^2]\Delta_{xy}^2 + (z_1 - z_2)^2 \Delta_z^2}$$

	x	y	z
Dot 1	3.68	3.12	5.44
Dot 2	3.90	3.86	4.02

$\Delta_{xy} = 120$ nm
$\Delta_z = 200$ nm

$R = 1.6 \pm 0.3\ \mu\text{m}$

When error propagation is not necessary

- Experiment to measure IC_{50} of a drug
- Logarithmic version: $pIC_{50} = -\log\frac{IC_{50}}{1\ \mathrm{M}}$

	R1	R2	R3	R4	R5
IC_{50} (nM)	25	85	43	108	12
pIC_{50}	7.6	7.2	7.4	6.7	7.9

- Two ways of finding mean pIC_{50} and its error
  - □ propagate from IC_{50}
  - □ direct calculation

$M = 54.6\ \mathrm{nM}$

$SE = 18.2\ \mathrm{nM}$

$$pIC_{50} = -\log\frac{54.6\ \mathrm{nM}}{1\ \mathrm{M}} = 7.26$$

error propagation $\Delta pIC_{50} = 0.43\frac{SE}{M} = 0.14$

$$\boxed{pIC_{50} = 7.3 \pm 0.1}$$

$M_p = 7.39$

$SE_p = 0.17$

calculating directly from logarithmic data:

$$\boxed{pIC_{50} = 7.4 \pm 0.2}$$

- Results are not identical
- Log of the mean is not mean of the logs!

- Errors are large (~30%), so error propagation formula does not work well
- Do not use error propagation if you can calculate errors directly from replicated data

Error propagation summary

- When a quantity is transformed, its error must be propagated

- Single variable

$$y = f(x)$$

$$\Delta y \approx \left|\frac{df}{dx}\right| \Delta x$$

- Multiple variables

$$y = f(x_1, x_2, \ldots, x_n)$$

$$\mathit{\Delta} y^2 \approx \left(\frac{\partial f}{\partial x_1}\right)^2 \mathit{\Delta} x_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \mathit{\Delta} x_2^2 + \cdots + \left(\frac{\partial f}{\partial x_n}\right)^2 \mathit{\Delta} x_n^2$$

Function	Error
$y = ax$	$\Delta y = a\Delta x$
$y = ax^b$	$\dfrac{\Delta y}{y} = b\dfrac{\Delta x}{x}$
$y = a\log_b cx$	$\Delta y = \dfrac{a}{\ln b}\dfrac{\Delta x}{x}$
$y = ae^{bx}$	$\dfrac{\Delta y}{y} = b\Delta x$
$y = 10^{ax}$	$\dfrac{\Delta y}{y} = a\ln(10)\Delta x$
$y = ax_1 \pm bx_2$	$\Delta y = \sqrt{a^2\Delta x_1^2 + b^2\Delta x_2^2}$
$y = x_1 x_2, \qquad y = \dfrac{x_1}{x_2}$	$\dfrac{\Delta y}{y} = \sqrt{\left(\dfrac{\Delta x_1}{x_1}\right)^2 + \left(\dfrac{\Delta x_2}{x_2}\right)^2}$

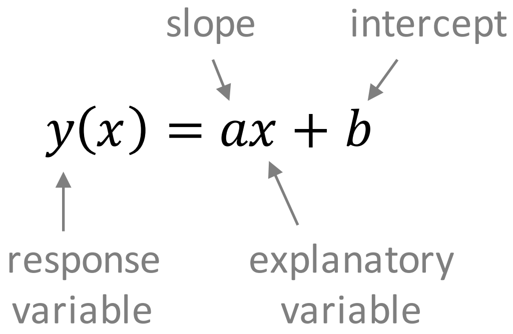
8. Simple linear regression

"It is proven that the celebration of birthdays is healthy. Statistics show that those people who celebrate the most birthdays become the oldest"

S. den Hartog

Linear regression

- Sample linear regression:



slope

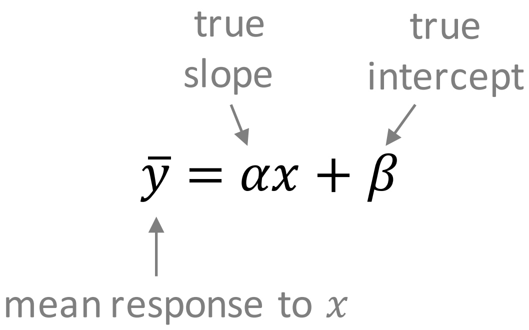
intercept

$y(x) = ax + b$

response variable

explanatory variable

- Population true regression

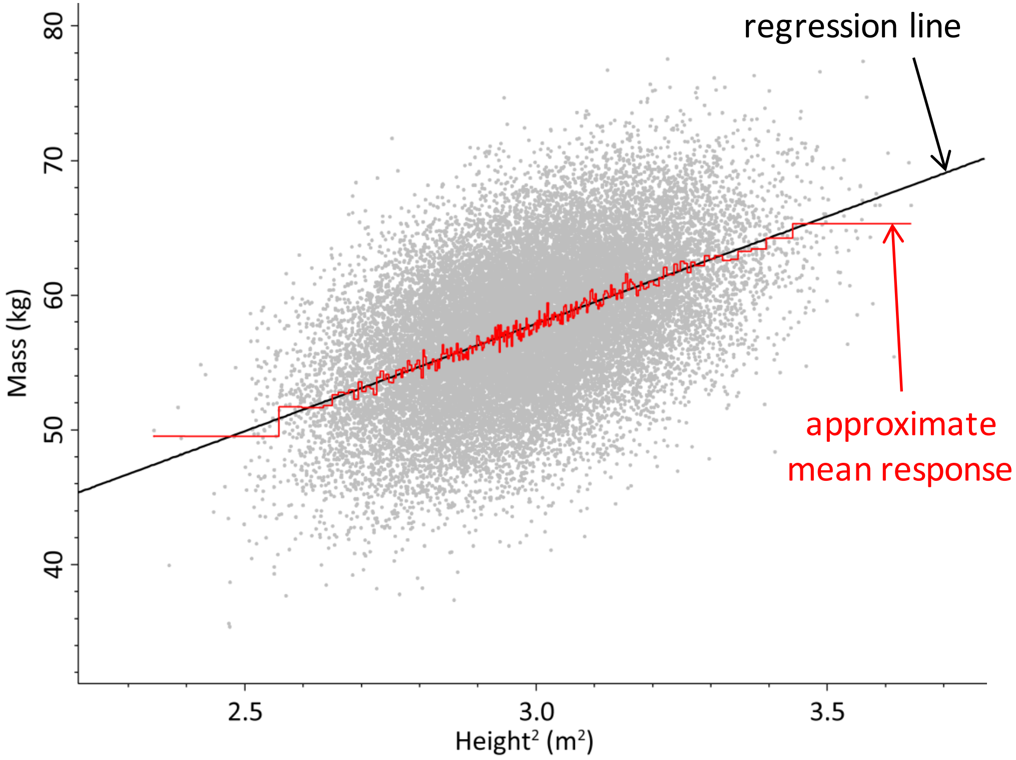


true slope

true intercept

$\bar{y} = \alpha x + \beta$

mean response to x



regression line

approximate mean response

Data from Hong Kong Growth Survey (25,000 adolescent youths). Body mass (m) is plotted against squared height (h^2)

Simple linear fit

- Consider a sample (x_i, y_i), $i = 1, \dots, n$

- Actual response:

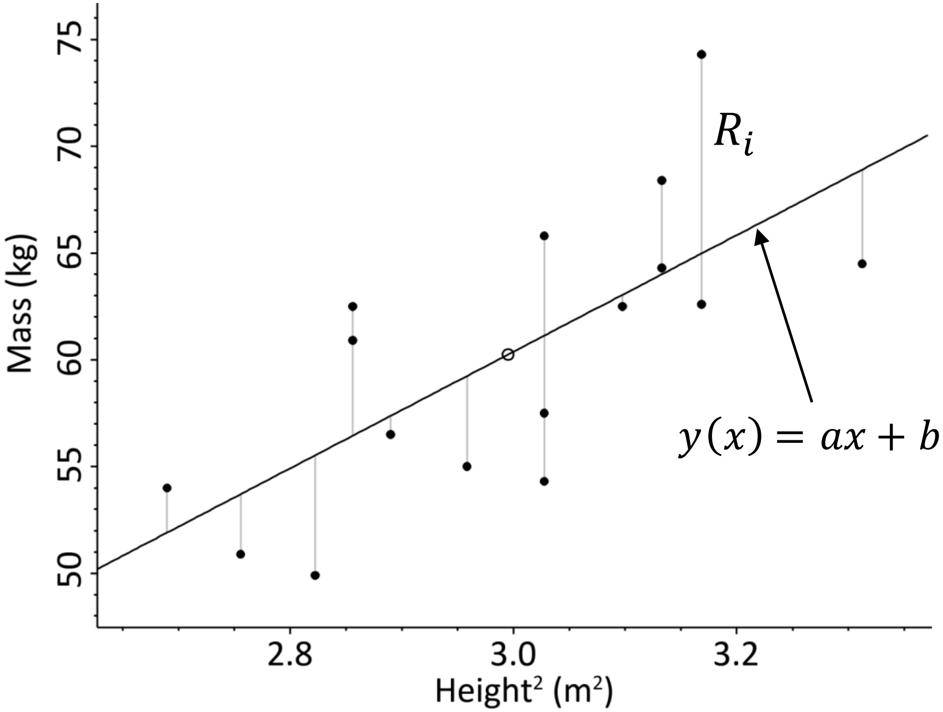
$$y_i = ax_i + b + R_i$$

- $R_i = y_i - y(x_i)$ are **residuals**

- For best-fitting model minimise

$$Q = \sum_{i=1}^{n} R_i^2 = \min$$

- Minimum: $\frac{\partial Q}{\partial a} = 0$ and $\frac{\partial Q}{\partial b} = 0$



Random selection of 16 points from the Hong Kong Growth Survey

Simple linear fit: the solution

- Minimise the sum of squared residuals and find:

$$a = \frac{S_{xy}}{S_{xx}}$$

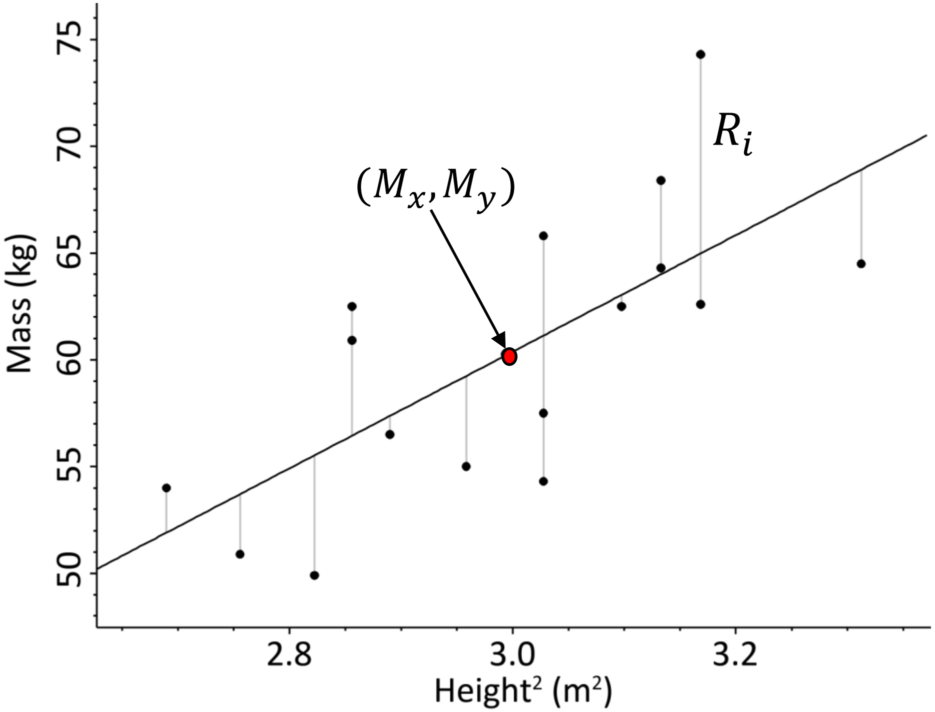
$$b = M_y - aM_x$$

- These are the estimators of true unknown slope, α, and intercept, β

$$M_x = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad M_y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - M_x)^2 \qquad S_{yy} = \sum_{i=1}^{n}\left(y_i - M_y\right)^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - M_x)(y_i - M_y)$$



The best-fitting line always passes through data centroid, (M_x, M_y)

Uncertainties of a and b

- We have raw data $(x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_n)$
- From this we find a and b

Idea

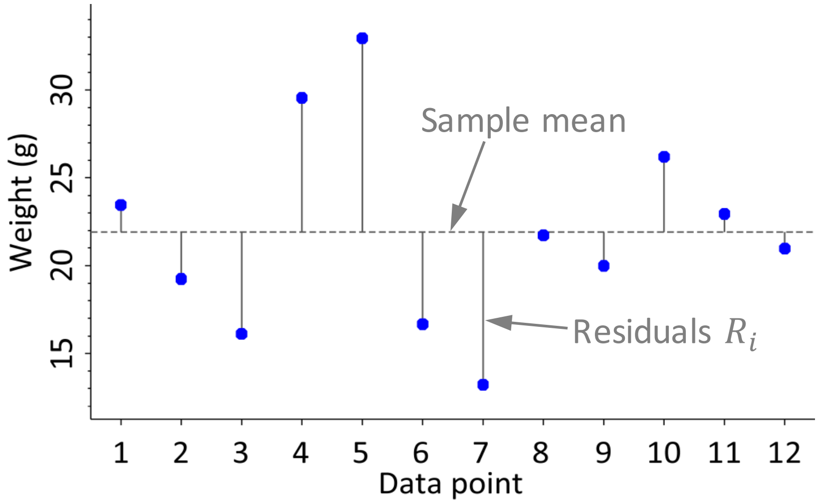
- Use scatter of data around the regression line
- Try finding uncertainties of y_i
- Propagate them into a and b

Analogy to standard deviation

- Sample y_i is scattered around the mean, M
- Standard deviation is calculated from squared residuals

$$R_i = y_i - M$$

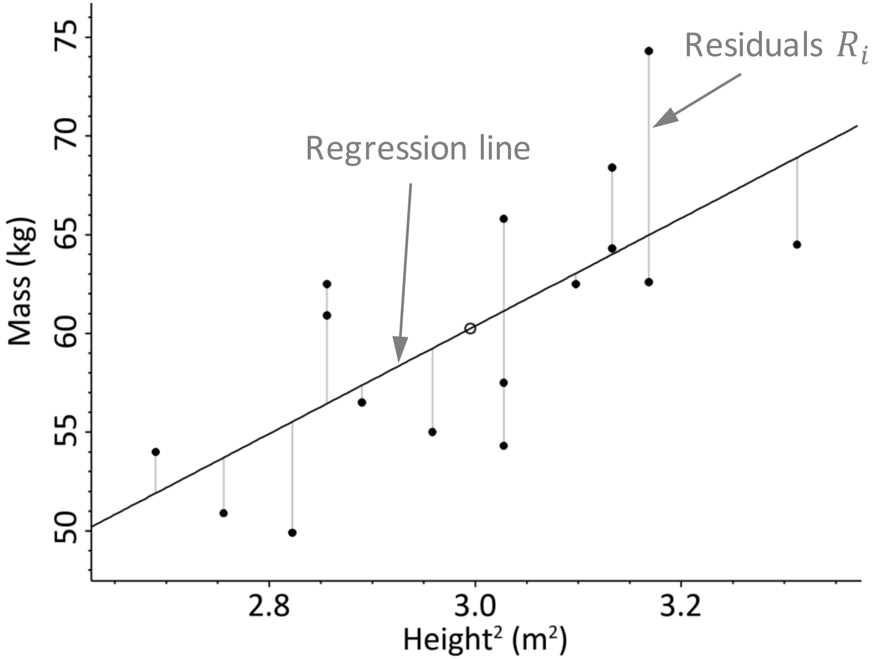
$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} R_i^2}$$



- Sample (x_i, y_i) is scattered around the regression line, $y(x)$
- Standard deviation is calculated from squared residuals

$$R_i = y_i - y(x_i)$$

$$SD_R = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} R_i^2} = \sqrt{\frac{S_{yy} - aS_{xy}}{n-2}}$$



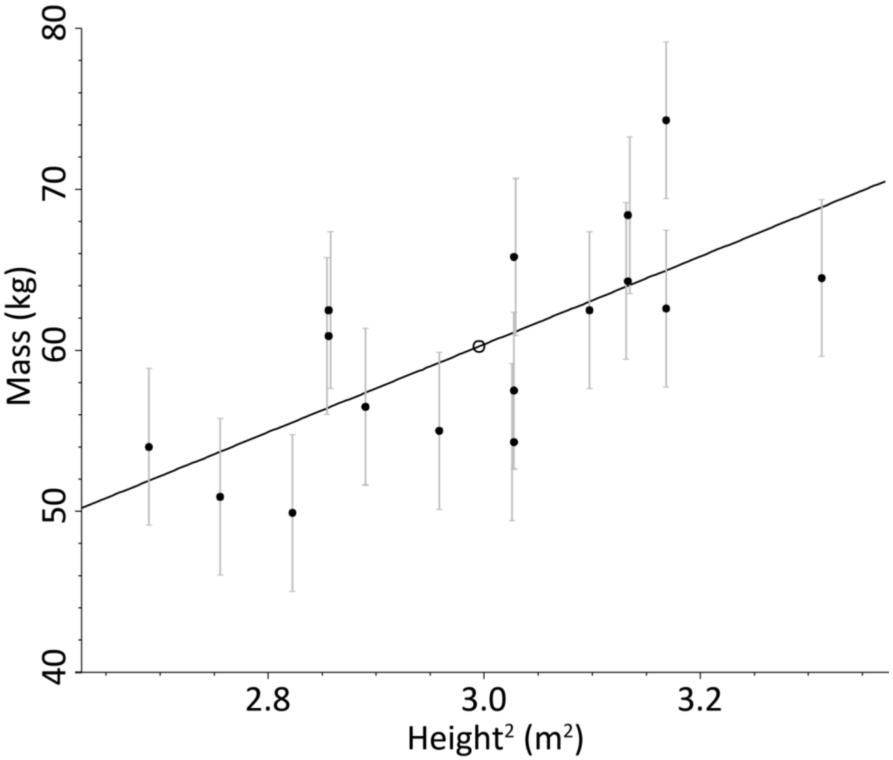
Confidence intervals on fit parameters

- We can quantify scatter around the regression line by SD_R
- Assume the scatter is due to uncertainty (noise/variability) in y

- Let's use SD_R as a common uncertainty of y_i

$$\Delta y_i = SD_R$$

- This is our **guessed** error of y_i
- It estimates a **typical** error in response



Confidence intervals on fit parameters

- Fit parameters depend on data points

$$a = a(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n)$$

$$b = b(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n)$$

$$\Delta x_i = 0 \qquad \Delta y_i = SD_R$$

$$a = \frac{S_{xy}}{S_{xx}}$$

$$b = M_y - aM_x$$

- Hence, we can propagate estimated errors $\Delta y_i = SD_R$:

$$\Delta a^2 = \sum_{i=1}^{n} \left(\frac{\partial a}{\partial y_i}\right)^2 \Delta y_i^2 \qquad \Delta b^2 = \sum_{i=1}^{n} \left(\frac{\partial b}{\partial y_i}\right)^2 \Delta y_i^2$$

- After some rather straightforward calculations we get

$$\Delta a^2 = \frac{SD_R^2}{S_{xx}} \qquad \Delta b^2 = SD_R^2 \left[\frac{1}{n} + \frac{M_x^2}{S_{xx}}\right]$$

Confidence intervals on fit parameters

- Δa and Δb represent the width of their sampling distributions

- Hence, they are standard errors

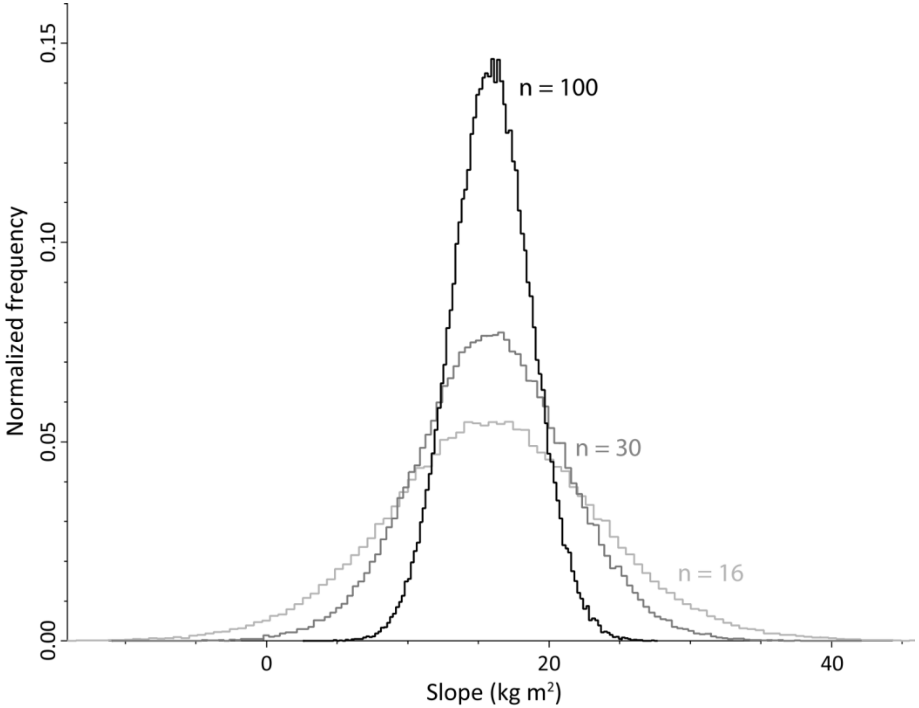
$$SE_a = \frac{SD_R}{\sqrt{S_{xx}}} \qquad SE_b = SD_R\sqrt{\frac{1}{n} + \frac{M_x^2}{S_{xx}}}$$

- We can find confidence intervals

$$a - \mathrm{t}^* SE_a \le \alpha \le a + \mathrm{t}^* SE_a$$

$$b - \mathrm{t}^* SE_b \le \beta \le b + \mathrm{t}^* SE_b$$

- where t^* is the critical value from t-distribution with $n-2$ degrees of freedom



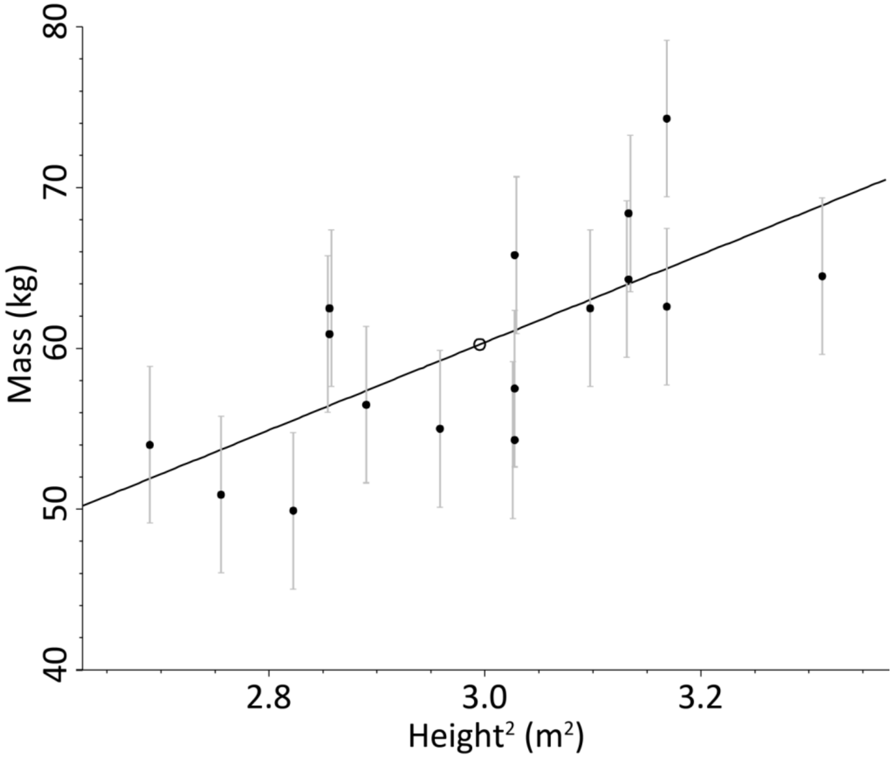
Sampling distribution of the slope. 100,000 samples of size n were randomly drawn from the Hong Kong Survey, and fitted with a straight line.

Example, Hong Kong sample

- Calculate
  $M_x = 2.995 \text{ m}^2$
  $M_y = 60.24 \text{ kg}$
  $S_{xx} = 0.4439 \text{ m}^4$
  $S_{yy} = 663.4 \text{ kg}^2$
  $S_{xy} = 12.12 \text{ kg m}^2$
- Find slope and intercept
  $a = 27.30 \text{ kg m}^{-2}$
  $b = -21.53 \text{ kg}$
- Find their standard errors
  $SD_R = 4.873 \text{ kg}$
  $SE_a = 7.314 \text{ kg m}^{-2}$
  $SE_b = 21.94 \text{ kg}$
- Critical $t^* = 2.145$ for $n - 2 = 14$ d.o.f.
- Finally, the 95% confidence intervals
  $a = 27 \pm 16 \text{ kg m}^{-2}$
  $b = -22 \pm 47 \text{ kg}$

	1	2	3	4	5	6	7	8
h (m)	1.66	1.70	1.64	1.74	1.72	1.82	1.78	1.74
m (kg)	50.9	56.5	54.0	57.5	55.0	64.5	62.6	54.3

	9	10	11	12	13	14	15	16
h (m)	1.68	1.76	1.69	1.74	1.77	1.69	1.78	1.77
m (kg)	49.9	62.5	62.5	65.8	68.4	60.9	74.3	64.3



Linear fit prediction errors

- Linear fit gives prediction for every x:

$$y(x) = ax + b$$

- Can we find uncertainty of $y(x)$?
- y is a function of a and b, so we can propagate errors:

$$SE_y^2 = \left(\frac{\partial y}{\partial a}\right)^2 SE_a^2 + \left(\frac{\partial y}{\partial b}\right)^2 SE_b^2$$

This is wrong!

Linear fit prediction errors

- Linear fit gives prediction for every x:

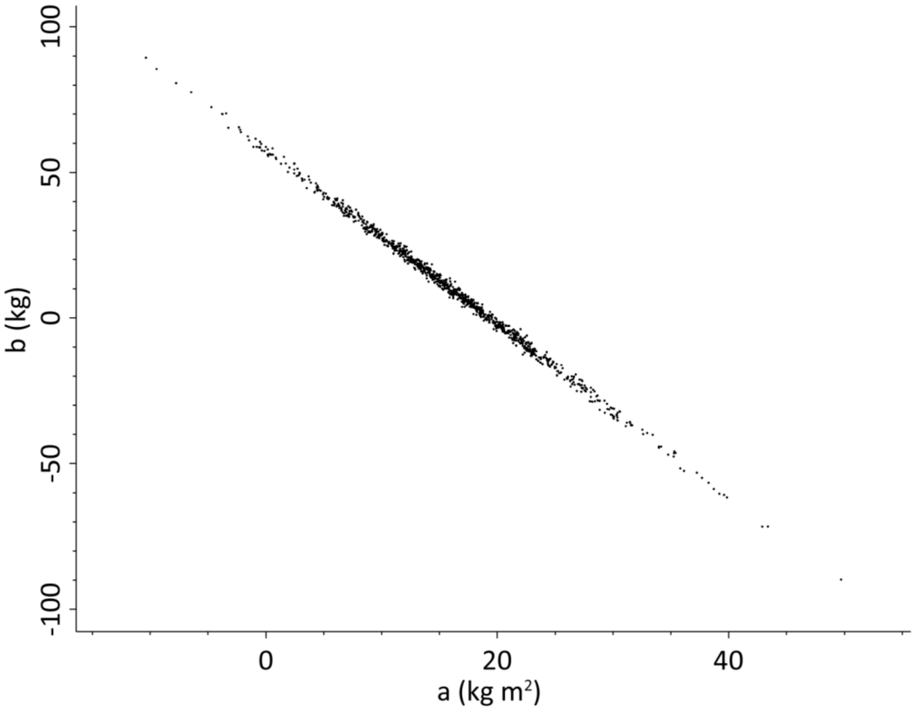
$$y(x) = ax + b$$

- Can we find uncertainty of $y(x)$?
- y is a function of a and b, so we can propagate errors
- Keep in mind that a and b are strongly correlated

$$SE_y^2 = \left(\frac{\partial y}{\partial a}\right)^2 SE_a^2 + 2\frac{\partial y}{\partial a}\frac{\partial y}{\partial b}\mathrm{Cov}(a,b) + \left(\frac{\partial y}{\partial b}\right)^2 SE_b^2$$

- After some derivations

$$SE_y = SD_R\sqrt{\frac{1}{n} - \frac{(x - M_x)^2}{S_{xx}}}$$



Correlation between fit parameters. 1000 samples of size $n = 16$ were drawn from the Growth Survey, fitted with the linear regression, parameters a and b found.

Linear fit prediction errors

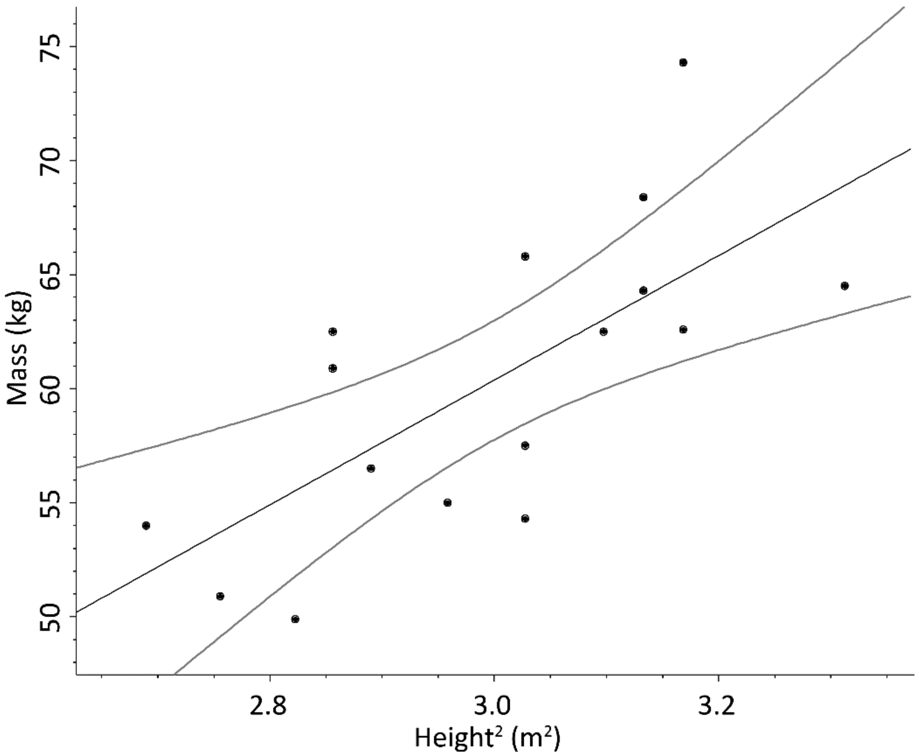
- Standard error of $y(x)$ is

$$SE_y(x) = SD_R\sqrt{\frac{1}{n} - \frac{(x - M_x)^2}{S_{xx}}}$$

- From this we can find confidence intervals

$$y(x) - t^* SE_y \leq \bar{y}(x) \leq y(x) + t^* SE_y$$

- t^* is a critical value from t-distribution with $n - 2$ degrees of freedom
- We can find these errors at any x



Linear regression summary

- Simple linear regression $y(x) = ax + b$
- Best-fitting parameters

$$a = \frac{S_{xy}}{S_{xx}} \qquad b = M_y - aM_x$$

- Assumption: x is measured accurately, scatter is caused by error in y
- Guessed common error of y, $\Delta y_i = SD_R$
- Standard error of the slope and intercept can be propagated from SD_R

$$SE_a = \frac{SD_R}{\sqrt{S_{xx}}} \qquad SE_b = SD_R\sqrt{\frac{1}{n} + \frac{M_x^2}{S_{xx}}}$$

- Confidence intervals are t^*SE for $n - 2$ degrees of freedom
- Standard error of y can be propagated from a and b

$$SE_y = SD_R\sqrt{\frac{1}{n} - \frac{(x - M_x)^2}{S_{xx}}}$$

- Confidence interval is t^*SE_y for $n - 2$ degrees of freedom

UNIVERSITY OF DUNDEE

THE BARTON GROUP

01000100010000010101010001000001

The Data Analysis Group

Hand-outs available at http://is.gd/statlec

Centre for Gene Regulation & Expression

wellcome trust

General curve fitting

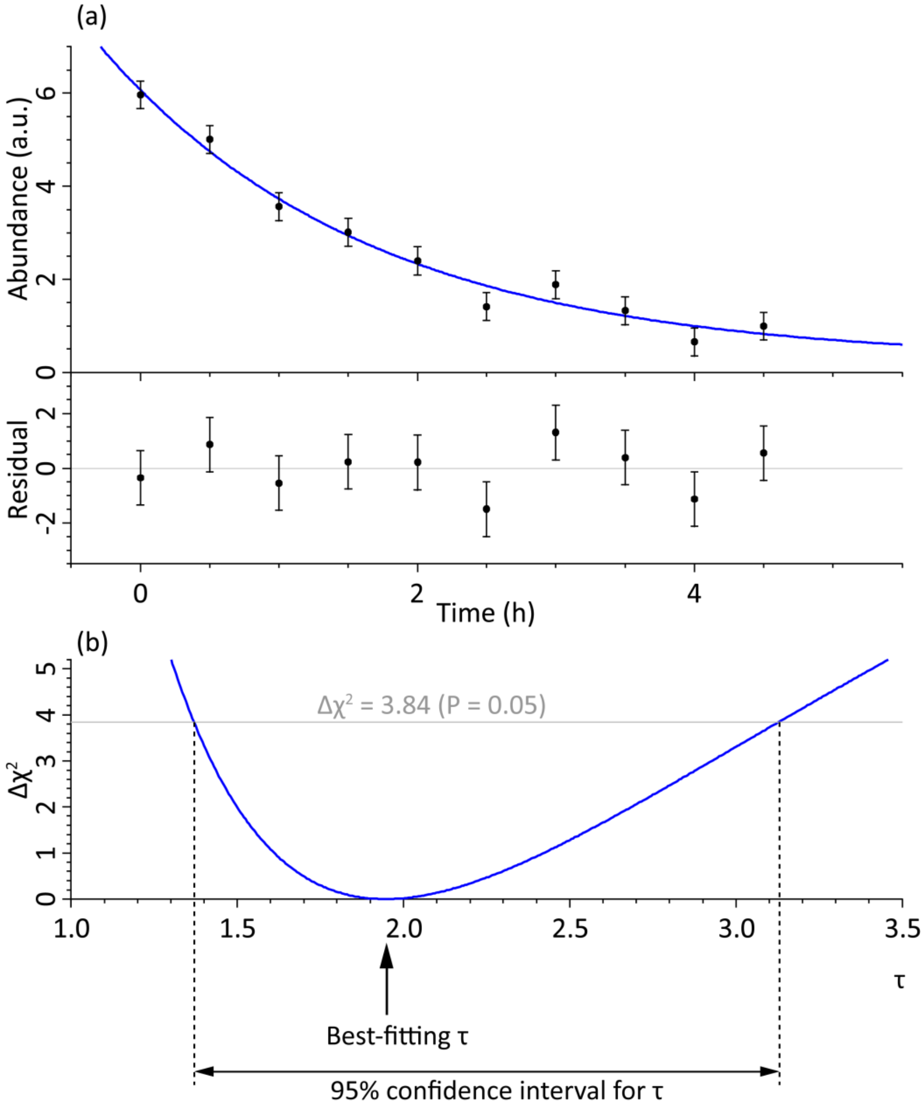
- We want to fit a model to *x*-*y* data
- Example: exponential decay
- Fit data points $(t_i, y_i, \Delta y_i)$ with a function

$$y(t) = A_0 + Ae^{-t/\tau}$$

- Find best-fitting time scale, τ, and find its error (95% CI)
- Minimize goodness of the fit:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - y(t_i)}{\Delta y_i}\right]^2$$

- From tables of χ^2 distribution we can find 95% probability corresponds to $\chi^2 = 3.84$
- Move the fitted parameter, τ, left and right from the best-fitting value until χ^2 increases by 3.84
- We find $\tau = 2^{+1.1}_{-0.6}$


(a)
Abundance (a.u.)
Residual
Time (h)
(b)
Δχ²
Δχ² = 3.84 (P = 0.05)
τ
Best-fitting τ
95% confidence interval for τ