PH3010 / MSci Skills Miniproject

Statistical Data Analysis

Week 2: Goodness-of-fit, LS fitting with correlated data,

Autumn term 2017

Glen D. Cowan
RHUL Physics

Outline – lecture 2

Today:

Goodness-of-fit

Fitting correlated data

More exercises

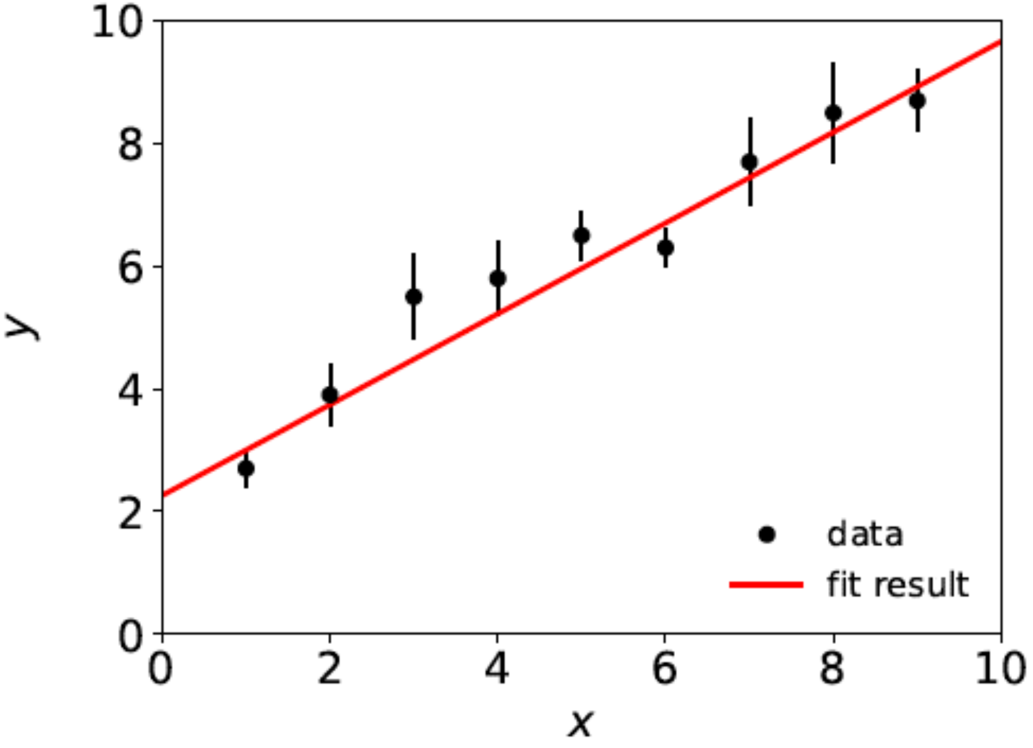
Discussion of project report

Next week:

Introduction to machine learning

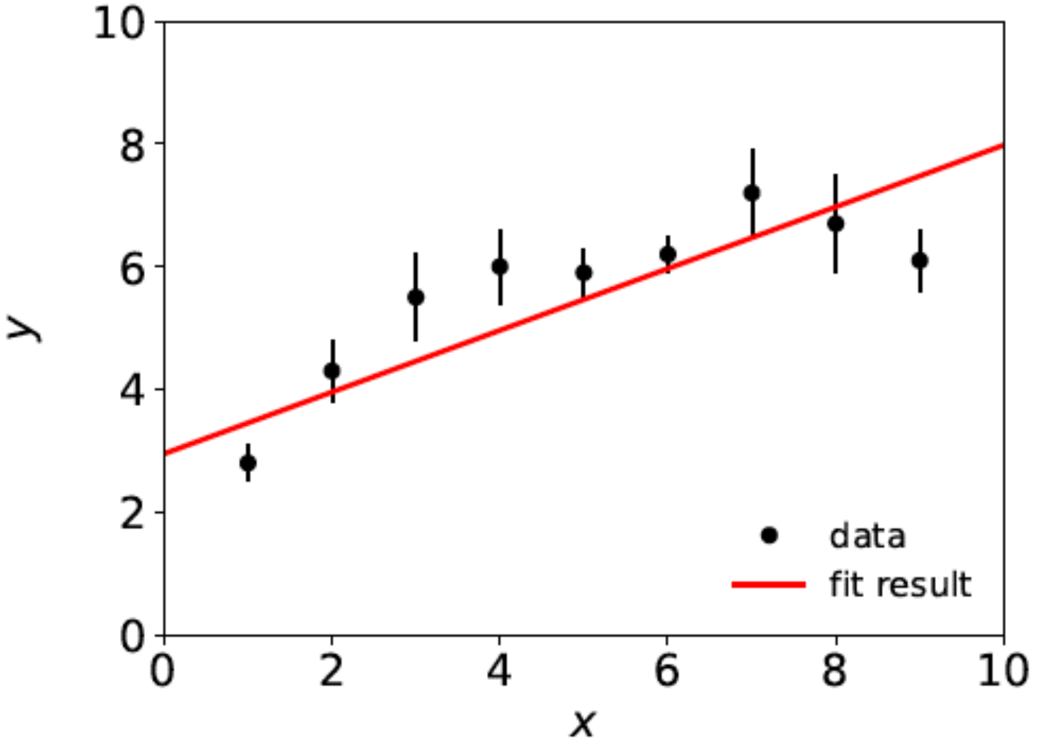
A “good” fit

Last week we fitted data that were reasonably well described by a straight line:



A “bad” fit

But what if a straight-line fit looks like this:



Maybe here we should fit a higher-order polynomial?

Goodness-of-fit: the questions

How do we quantify the level of agreement between the data and the hypothesized form of the fit function?

How do we decide whether to try a different fit function?

⚠ Note first the following common misunderstanding:

If the fit is “bad”, you may expect large statistical errors for the fitted parameters. This is not the case.

The statistical errors say how much the parameter estimates will fluctuate under repetition of the experiment, under assumption of the hypothesized fit function. This is not the same as the degree to which the function is able to describe the data.

If the hypothesized $f(x; \boldsymbol{\theta})$ is not correct, the fitted parameters will have some systematic uncertainty – a more complex question that we will not take up here.

Quantifying goodness-of-fit

We can quantify the goodness-of-fit directly from the value of $\chi^2(\boldsymbol{\theta})$ evaluated at its minimum, i.e., at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$:

$$\chi^2_{\min} = \sum_{i=1}^{N} \frac{(y_i - f(x_i; \hat{\boldsymbol{\theta}}))^2}{\sigma_i^2}$$

If the fitted function is in good agreement with the data, then the numerator of each term in the sum should be small.

If $f(x_i; \hat{\boldsymbol{\theta}})$ were equal to the true mean of y_i, then we would expect the residual $y_i - f(x_i; \hat{\boldsymbol{\theta}})$ to have an rms value of σ_i. Each term in the sum would contribute 1, and we'd have $\chi^2_{\min} = N$.

This is not quite true: if we have fitted m parameters and the hypothesized function is correct, the *expected* value of $\chi^2_{\min}$ is $N - m$ (called the *number of degrees of freedom* of the fit.)

Distribution of $\chi^2_{\min}$

$\chi^2_{\min}$ is a function of the data so is itself a random variable.

If the hypothesized fit function is correct and the data are Gaussian distributed, one can show $\chi^2_{\min}$ follows a chi-square distribution with $n = N - m$ degrees of freedom (here let $\chi^2_{\min} = z$):

$$f_{\chi^2}(z;n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$$

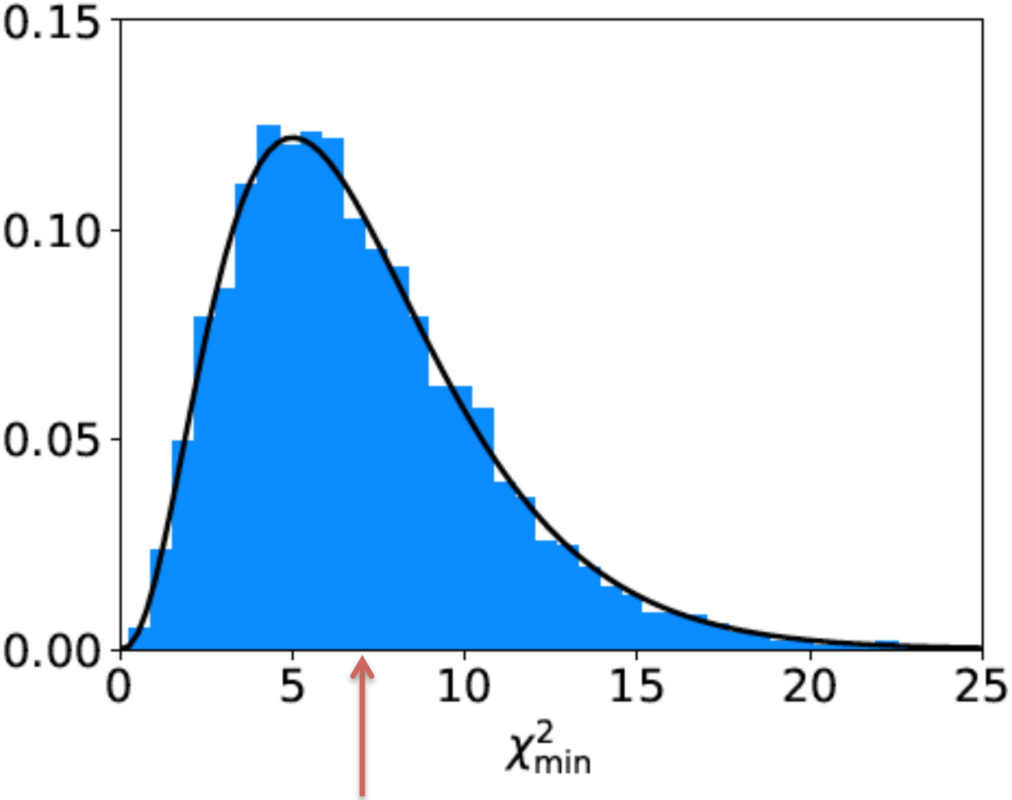
Mean and standard deviation of the chi-square distribution:

$$\begin{aligned} \langle z \rangle &= n \\ \sigma_z &= \sqrt{2n} \end{aligned}$$

If the hypothesized fit function is not correct, then one would obtain a distribution of $\chi^2_{\min}$ shifted to higher values.

Distribution of $\chi^2_{\min}$ from straight-line fit

$\chi^2_{\min}$ from straight-line fit with $N = 9$ data points, $m = 2$ fitted parameters.



Curve: chi-square pdf for $n = 7$ degrees of freedom.

Histogram: values of $\chi^2_{\min}$ from straight-line fit from repeated Monte Carlo simulation of the experiment.

mean $= N - m = 7$

How to interpret χ^2_{min}

A simple way to assess the goodness-of-fit is simply to compare χ^2_{min} to the number of degrees of freedom, $n_{\mathrm{dof}} = N - m$.

$\chi^2_{\mathrm{min}} \sim n_{\mathrm{dof}}$ → fit is "good"

$\chi^2_{\mathrm{min}} \gg n_{\mathrm{dof}}$ → fit is "bad"

$\chi^2_{\mathrm{min}} \ll n_{\mathrm{dof}}$ → fit is better than what one would expect given fluctuations that should be present in the data.

Often this is done using the ratio $\chi^2_{\mathrm{min}}/n_{\mathrm{dof}}$, i.e. fit is good if the "chi-square per degree of freedom" comes out not much greater than 1.

Often report as, e.g., $\chi^2_{\mathrm{min}}/n_{\mathrm{dof}} = 8.2/7$. It is best to communicate both χ^2_{min} and n_{dof}, not just their ratio.

p-value from $\chi^2_{\min}$

Another way to assess the goodness-of-fit is to give the probability, assuming the fit function is correct, to obtain a $\chi^2_{\min}$ value as high as the one we got or higher:

$$p = \int_{\chi^2_{\min}}^{\infty} f_{\chi^2}(z;n)\,dz$$

This is an example of what is called a *p*-value of the hypothesis (here the hypothesized form of the fit function).

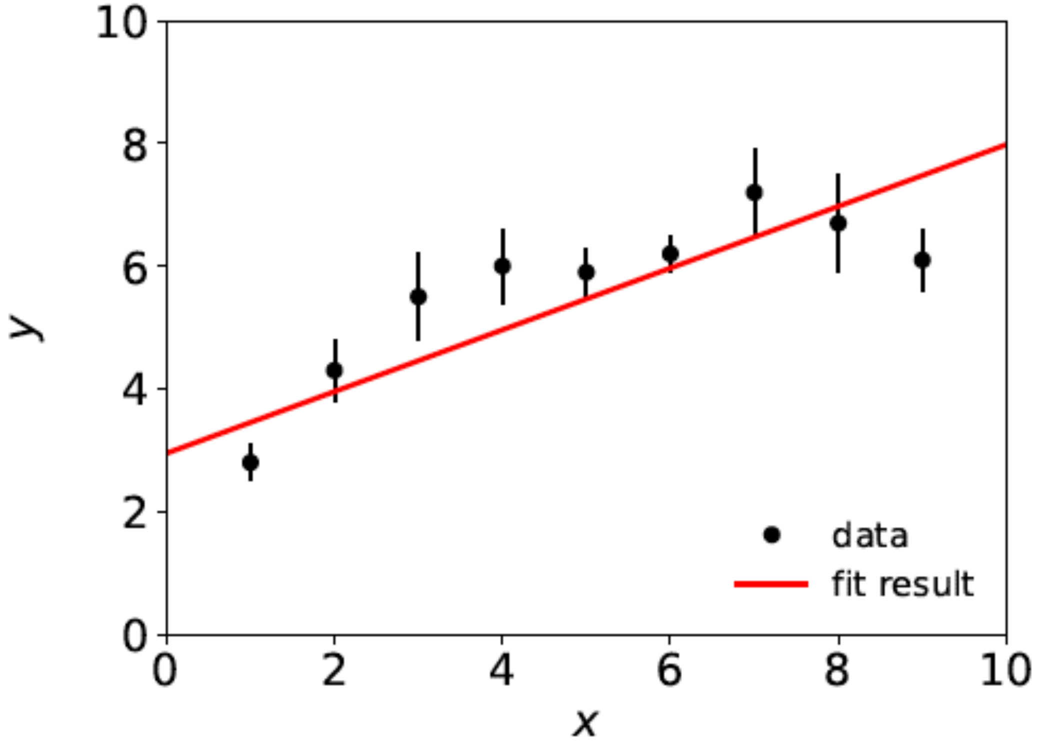
p-value is not the same as the probability that the hypothesis is true!

Nevertheless, a small *p*-value indicates that the hypothesis is disfavoured.

Compute using: `scipy.stats.chi2.sf`

$\chi^2_{\min}$ from the “bad” fit

Straight-line fit with $N = 9$ data points, $m = 2$ fitted parameters.



$\chi^2_{\min} = 20.9$ for $n_{\mathrm{dof}} = 7$

$\chi^2_{\min} / n_{\mathrm{dof}} = 3.7$

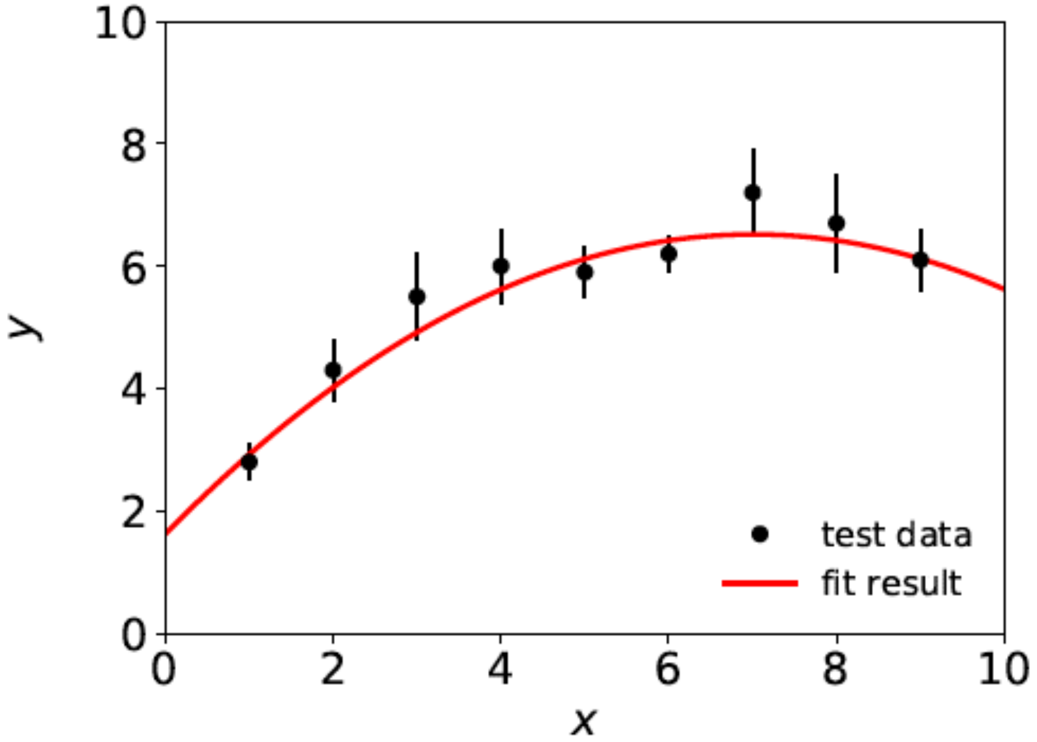
p-value = 0.0039

So is the straight-line hypothesis correct? It could be, but if so we would expect a $\chi^2_{\min}$ as high as observed or higher only 4 times out of a thousand.

A better fit

If we decide the agreement between data and hypothesis is not good enough (exact threshold is a subjective choice), we can try a different model, e.g., a 2nd order polynomial:

$$f(x;\boldsymbol{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2$$



$\chi^2_{\min} = 3.5$ for $n_{\text{dof}} = 6$

$\chi^2_{\min} / n_{\text{dof}} = 0.58$

p-value = 0.75

Least squares with correlated measurements

Up to now we have assumed that the measurements $y_1,\ldots,y_N$ are all independent.

This means that if one value fluctuates, say, high, then this has no influence on whether one of the others will fluctuate high or low.

But there could be cases where the y_i are correlated, i.e., they have nonzero covariances

$$\mathrm{cov}[y_i, y_j] = V_{ij} = \rho_{ij}\sigma_i\sigma_j$$

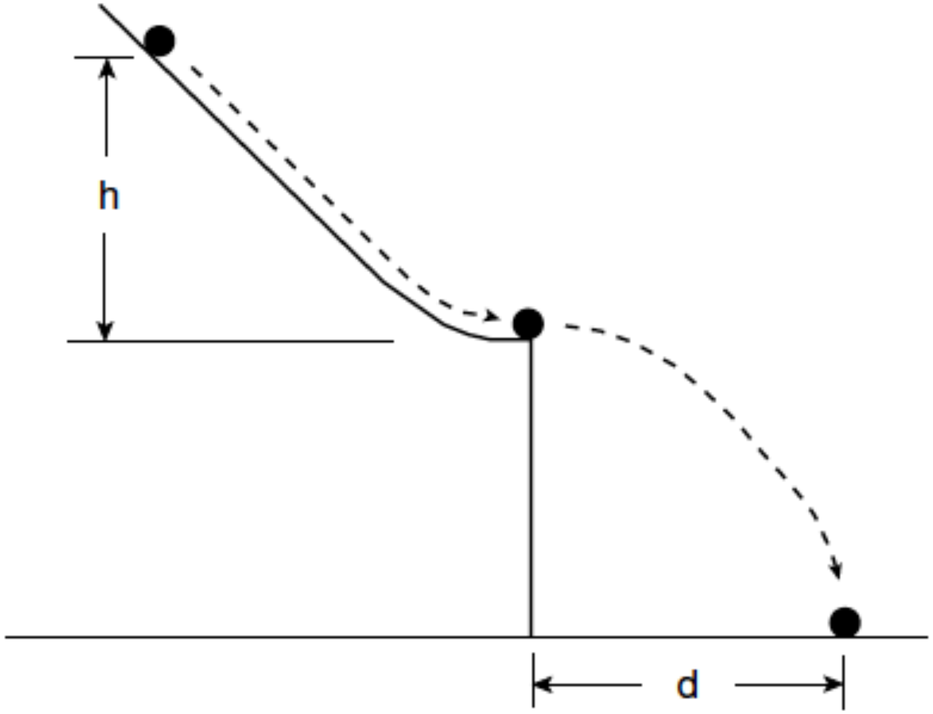
In this case, the formula for $\chi^2(\boldsymbol{\theta})$ becomes

$$\chi^2(\boldsymbol{\theta}) = \sum_{i,j=1}^{N} (y_i - f(x_i;\boldsymbol{\theta}))V_{ij}^{-1}(y_j - f(x_j;\boldsymbol{\theta}))$$

where V^{-1} is the inverse of the covariance matrix of the data V.

Goodness-of-fit with Galileo's data

Last week we used data from Galileo...



h	d
1000	1500
828	1340
800	1328
600	1172
300	800

Goodness-of-fit with Galileo's data

...to fit several hypotheses for the functional relation between the initial height *h* and flight distance *d*:

$$d = \alpha h$$

$$d = \alpha h + \beta h^2$$

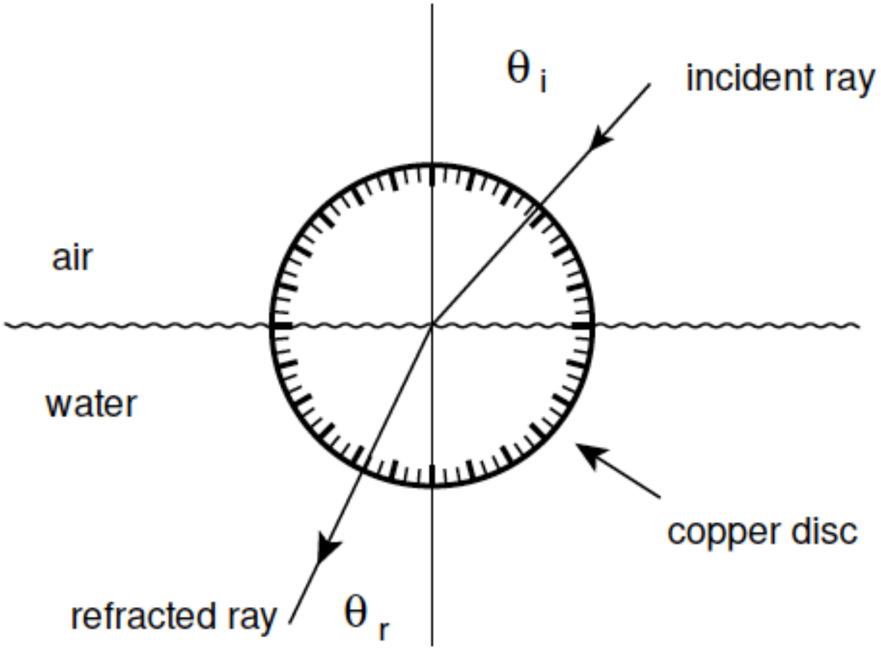
$$d = \alpha h^\beta$$

So now we can put ourselves in Galileo's position and, without knowledge of Newton's laws, find which hypotheses are compatible with or disfavoured by the data (find $\chi^2_{\rm min} / n_{\rm dof}$ and *p*-value).

You can also use your knowledge of Newtonian mechanics to work out the predicted law and compare to what you find empirically.

Exercise 3: refraction data from Ptolemy

Astronomer Claudius Ptolemy obtained data on refraction of light by water in around 140 A.D.:



Angles of incidence and refraction (degrees)

θ_i	θ_r
10	8
20	$15\frac{1}{2}$
30	$22\frac{1}{2}$
40	29
50	35
60	$40\frac{1}{2}$
70	$45\frac{1}{2}$
80	50

Suppose the angle of incidence is set with negligible error, and the measured angle of refraction has a standard deviation of ½°.

Laws of refraction

A commonly used law of refraction was

$$\theta_{\mathrm{r}} = \alpha\theta_{\mathrm{i}} \ ,$$

although it is reported that Ptolemy preferred

$$\theta_{\mathrm{r}} = \alpha\theta_{\mathrm{i}} \ - \ \beta\theta_{\mathrm{i}}^2 \ .$$

The law of refraction discovered by Ibn Sahl in 984 (and rediscovered by Snell in 1621) is

$$\theta_{\mathrm{r}} = \sin^{-1}\left(\frac{\sin\theta_{\mathrm{i}}}{r}\right) .$$

where $r = n_{\mathrm{r}}/n_{\mathrm{i}}$ is the ratio of indices of refraction of the two media.

Analysis of refraction data

Fit the parameters and find their statistical errors and (where relevant) covariance matrix.

Assess goodness-of-fit of each hypothesis ($\chi^2_{\rm min}/n_{\rm dof}$ and p-value).

What can you conclude?

(See The Feynman Lectures on Physics, Vol I., Addison-Wesley, 1963, Section 26-2, **`www.feynmanlectures.caltech.edu`** .)

Discussion of project report

Exercise 4 in the Least Squares script is optional.

There will be some exercises next week on Machine Learning.

Guidelines for report writing are on the Moodle page (section Reports 2017-2018), with LaTeX template, etc.

You should submit the electronic version of your report to the Turnitin Repository on the Moodle page Topic 6: Report Repository for Statistical Analysis Miniproject.

Deadline is 17 Nov 2017 at 10:00 a.m. (extended one week).

You should also submit 2 bound copies of the report to the departmental office (same deadline).

Do not put off writing to the last minute (start ~now).

Discussion of project report (2)

Your report should have

- A short introduction
- A section for each exercise
- Brief conclusions
- Bibliography
- Appendices (including all code you've written)

Use the relevant tools in LaTeX for the components of the report (sections, figures, bibliography, etc.).

Maximum word count is (including captions but not including appendices) is 3000.