PHAS0007 Reading Week Task 2019: Using least-squares fitting to calculate the Hubble constant

Revision date: October 31, 2019

Dr Louise Dash (louise.dash@ucl.ac.uk), Dr Pablo Lemos

1. Introduction

For this assignment, you are going to calculate unweighted and weighted linear least-squares fits using several different methods:

- From first principles;
- Using the numpy `polyfit()` function to fit to an order-1 polynomial;
- Using the scipy `curve_fit()` function to fit to the function $y = mx + c$.

You will be using astronomical data, prepared by Dr Pablo Lemos (Demonstrator in the Tuesday morning team, and Postdoctoral Researcher in the UCL Astrophysics group), from observations of type Ia Supernovae to calculate a value for the Hubble constant.

2. Background information

In 1927, Georges Lemaître discovered a proportionality law between the distances to astronomical bodies and their recession velocity Lemaître (1927). This relation was then corrected by Edwin Hubble (Hubble, 1929), and was henceforth known (somewhat unfairly) as **Hubble's law**:

$$v = H_0 D, \tag{1}$$

where v is the radial velocity (positive for recessing bodies), and D is the distance to the body. Distances in cosmology are not straightforward to measure (if you are interested, see Hogg (1999)), and so the uncertainties in these distances can be large. The proportionality constant is known as **Hubble constant** H_0. For historical reasons, it is usually measured in kilometers per second per megaparsec (a parsec is a distance unit frequently used in astronomy, 1pc $\approx 3.0857 \times 10^{16}$m).

The numerical value of the Hubble constant has been a source of controversy in the scientific community for nearly a century. The first attempts to estimate its value differed between claims of $H_0 \sim 50$ km s^{-1}Mpc^{-1} and $H_0 \sim 100$ km s^{-1}Mpc^{-1}.

Currently, we have two main methods to estimate this constant: We can take measurements of the *Cosmic Microwave Background* (CMB), a radiation that was emited after the Big Bang, and use those measurements and theoretical models to estimate the Hubble constant. Alternatively, we can follow Lemaître's procedure, and measure distances and recession velocities of type Ia Supernovae (SNe), and use **linear fitting** to calculate the Hubble constant through Eq. 1. In this exercise, you will follow the latter method.

2.1 Data

The data you will use comes from the Joint Light Curve Analysis (JLA) of the SDSS-II and SNLS supernova samples Betoule et al. (2014). The data has been processed specifically for the purpose of this exercise. Your columns will be velocities (in km/s) and distances (in Mpc). You can then use the inverse of Eq. 1 to calculate the Hubble constant.

3. Weighted least-squares fit

The fit we did in session 5 assumed that each of the data points were equally valid.

Frequently though, we find that experimental data do not have equal uncertainties, and some data points may have significantly larger uncertainties than others, as is the case with the astronomical data you will use for this assignment.

We can take the relative sizes of the error bars into account by weighting them proportionally. The weight we will give each data point is

$$w_i = \frac{1}{(\Delta y_i)^2}. \tag{2}$$

If these weights are included in the least-squares derivation, we end up with new expressions for the slope m:

$$m = \frac{\sum_i w_i \sum_i w_i x_i y_i - \sum_i w_i x_i \sum_i w_i y_i}{\sum_i w_i \sum_i w_i x_i^2 - \left(\sum_i w_i x_i\right)^2} \tag{3}$$

$$= \frac{\sum_i w_i \sum_i w_i x_i y_i - \sum_i w_i x_i \sum_i w_i y_i}{\delta}, \tag{4}$$

and the intercept c:

$$c = \frac{\sum_i w_i x_i^2 \sum_i w_i y_i - \sum_i w_i x_i \sum_i w_i x_i y_i}{\delta}; \tag{5}$$

where we have used an abbreviation for the denominator

$$\delta = \sum_i w_i \sum_i w_i x_i^2 - \left(\sum_i w_i x_i \right)^2 . \tag{6}$$

The uncertainties in these quantities are given by

$$\Delta m = \sqrt{\frac{\sum_i w_i}{\delta}} \tag{7}$$

and

$$\Delta c = \sqrt{\frac{\sum_i x_i^2 w_i}{\delta}}. \tag{8}$$

You can find more background to these equations in Chapter 6, particularly section 6.3, of Hughes and Hase (2010) and in the PHAS0007 Data Analysis and Statistics booklet (Llorente Garcia et al., 2019)

4. Task instructions

For this assignment you will need to submit a fully self-contained Jupyter notebook, with code cells, text cells and results, laid out in a clear and logical order.

Read this document through in its entirety before you start.

Your notebook should answer the questions below in the given order. Use the question numbers as numbers for section headers in your notebook.

You will also need to have access to the data file and the supplementary Jupyter notebook, both of which are available on Moodle. The supplementary notebook contains the equations from this script already typeset for you, and the code snippet you will need for question **3)**),

You will be dealing with lots of different variables in this assignment . Try to come up with a coherent variable naming scheme—this will make life easier for both you and the marker.

4.1 What you need to do:

1) **a)** Use `np.loadtxt()` to load the data from the data file into appropriately named arrays. Your code should import the data from the same directory as the notebook. Do not rename the data file or make any changes to the data file itself.

b) Plot the data, using y-errorbars, on a suitably labelled plot.

c) Adapt your code from your session 5 submission as appropriate, implementing the feedback you were given in session, to calculate an unweighted fit to the data. Your code should output the gradient, intercept, and uncertainties at full precision.

d) Implement equations 2–8 given above for the weighted least squares fit, and output the fitting parameters, again at full precision.

2) Download the "Script for linear fitting using Python" from the PHAS0007 Data Analysis and Statistics Moodle. This notebook contains code that uses Numpy's `polyfit()` function to fit data to an order-1 polynomial, i.e. a straight line. Paste code from this notebook, adapting it as necessary, to calculate:

a) an unweighted least-squares fit of the data;

b) a weighted least-squares fit of the data.

Again, output these results at full precision. If you have done this correctly, the results from sections **1)** and **2)** should agree almost exactly.

3) Paste the code snippet from the supplementary Jupyter notebook (on Moodle) into your notebook. This code will use the `curve_fit()` function from scipy to calculate quantities from which you can obtain the fit parameters.

a) Read the the function documentation to find out how `curve_fit` works, and what the inputs and outputs represent. Explain this briefly in a text cell.

Add code as necessary to extract the fit parameters from the `curve_fit` outputs. You may also find it helpful to look carefully at Dr Llorente Garcia's fitting code to see how to obtain the errors from the covariance matrix.

Again, output the results at full precision for both the weighted and unweighted fits. The results from all three methods should agree almost exactly.

b) Find information (you can use textbooks or internet sources) about the covariance matrix,or matrix of covariance. In your own words, and citing any sources you referred to, explain in a paragraph what the covariance matrix represents. You do *not* need to go into any mathematical detail for this, just a general conceptual explanation at a level that could easily be understood by an average first-year student.

4) You should now have six sets of results: for the unweighted and weighted fits using each of the three methods. The results from all three methods should agree almost exactly.

a) Use a single code cell to output all of these results (gradient, intercept, and their respective uncertainties), still at full precision, using appropriate text strings. To make your output clear and nicely formatted, it will be useful to know that

- `\t` within a `print()` will act as a tab;
- `\n` within a `print()` will give a line break; and
- an empty `print()` will give a blank line.

b) Choose the results from one method only. Use a text cell to explain which of the three methods you personally have chosen, and why. There's no "right"

answer here, as the results from each method should be the same—just explain your personal preference.

c) Using these results, create a new plot that includes the data with errorbars and both the unweighted and weighted fitted lines.

d) Output your fitting parameters at the precision appropriate for quoting final results. You will need to put some thought into the appropriate precision for the intercept, and justify your decision in a text cell.

e) For both the weighted and unweighted fits, calculate the residual at each data point. Plot the residuals on a new plot, and comment in a text cell on what you notice. In particular, to what extent do you think a straight line fit is appropriate for this data?

5) Use your results together with Eq. 1 to calculate a value for the Hubble constant H_0, for both the weighted and unweighted fits, not forgetting to propagate the errors. Use a paragraph in a text cell to discuss whether the unweighted or weighted fit is more appropriate given your results above.

6) Thus comment on the significance of your results for the Universe.

---

5. General information

5.1 How to reference sources

You should reference sources using the Harvard referencing style, i.e. by using the "Author, Date" referencing style you see in this document. You can find more information about how to do this for various types of source in the UCL library guide to referencing: http://www.ucl.ac.uk/library/ docs/guides/references-plagiarism.

You can find an example of how to include a bibliography in a Jupyter Notebook in the supplementary notebook for the task (on Moodle).

Remember that when you paste or otherwise reuse code from another source, you *must* make sure you clearly state the original source in the code comments, and summarise any changes that you have made.

5.2 Typesetting equations

You will need to include equations in your Jupyter notebook text cells. The supplementary notebook contains the equations from the script already in LaTeX format that you may copy and paste from—include this in your references. For the typesetting of any additional equations, you may find one of the following online LaTeX equation editors useful (you can pick the maths symbols you want from a palette and then copy and paste the LaTeX code into your notebook).

- http://www.sciweavers.org/free-online-latex-equation-editor

- http://www.codecogs.com/latex/eqneditor.php

6. Assessment

Your work will be graded using a rubric markscheme based on the following assessment criteria:

- Whether the values you have calculated are correct (40% of available marks);
- The quality of your plot, code comments, and coding style (35% of available marks).
- The quality and coherence of the text commentary (including referencing), discussion, and the conclusions you have drawn from your results. (25% of available marks).

The grading will be taking into account that you have significantly more time for this task than the in-session assignments, and you should bear this in mind, particularly when writing your text commentary. Do not restrict your text cells to merely answering the specific questions given in the task—make sure your text commentary results in a completely self-contained document.

Remember that this is a formal assignment—your document should be clearly laid out in sections, with complete, grammatically correct sentences and paragraphs rather than bullet point lists.

6.1 Anonymity

The grading of this assignment is anonymised. **Do not include your name in the filename or anywhere in the assignment itself.** Instead, please use your student ID (the 8-digit number on your ID card, for most of you this will start with 19) as the filename. Make sure you type the number correctly. If you have an e-sticker from Student Disability Services, please copy and paste the wording (but not your name!) into a clearly labelled text cell at the top of your assignment.

If you need to include a reference to one of your own submissions in your notebook, you should replace your name with your student number in the reference.

References

Betoule, M., Kessler, R., Guy, J., Mosher, J., Hardin, D., Biswas, R., Astier, P., El-Hage, P., Konig, M., Kuhlmann, S., Marriner, J., Pain, R., Regnault, N., Balland, C., Bassett, B. A., Brown, P. J., Campbell, H., Carlberg, R. G., Cellier-Holzem, F., Cinabro, D., Conley, A., D'Andrea, C. B., DePoy, D. L., Doi, M., Ellis, R. S., Fabbro, S., Filippenko, A. V., Foley, R. J., Frieman, J. A., Fouchez, D., Galbany, L.,

Goobar, A., Gupta, R. R., Hill, G. J., Hlozek, R., Hogan, C. J., Hook, I. M., Howell, D. A., Jha, S. W., Le Guillou, L., Leloudas, G., Lidman, C., Marshall, J. L., Möller, A., Mourão, A. M., Neveu, J., Nichol, R., Olmstead, M. D., Palanque-Delabrouille, N., Perlmutter, S., Prieto, J. L., Pritchet, C. J., Richmond, M., Riess, A. G., Ruhlmann-Kleider, V., Sako, M., Schahmaneche, K., Schneider, D. P., Smith, M., Sollerman, J., Sullivan, M., Walton, N. A., and Wheeler, C. J. (2014). Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples. *Astronomy and Astrophysics*, 568:A22. https://doi.org/10.1051/0004-6361/201423413.

Hogg, D. W. (1999). Distance measures in cosmology. *arXiv e-prints*, pages astro–ph/9905116. https://arxiv.org/abs/astro-ph/9905116.

Hubble, E. (1929). A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Contributions from the Mount Wilson Observatory*, 3:23–28.

Hughes, I. and Hase, T. (2010). *Measurements and their uncertainties. A practical guide to modern error analysis*. Oxford University Press. http://UCL.eblib.com/patron/FullRecord.aspx?p=584562.

Lemaître, G. (1927). Un Univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques. *Annales de la Société Scientifique de Bruxelles*, 47:49–59.

Llorente Garcia, I., Jones, P., and Skipper, N. (2019). Data analysis and statistics booklet, PHAS0007. UCL Moodle. https://moodle.ucl.ac.uk/course/view.php?id=16941&section=1.