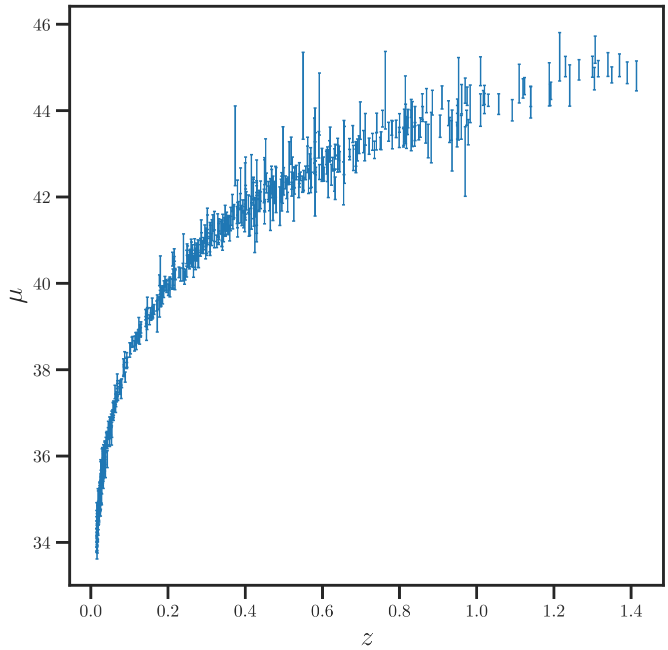
Data Science Lab Astrophysics 2022: Lab Worksheet 4: Linear Regression

Florian List*

March 31, 2022



In this week's lab, we will be looking at **supernova** measurements from the Supernova Cosmology Project (`https://supernova.lbl.gov/Union/`), specifically the "Union2.1" dataset (see `https://supernova.lbl.gov/Union/figures/SCP_Union_Astroph_25_04.pdf` for the accompanying paper). Type Ia supernovae are so-called **standard candles** in cosmology: since the peak luminosity of every supernova is (roughly) the same, their visual magnitude as seen from Earth can be used to determine the distance to their host galaxy.

*florian.list@univie.ac.at

Using this data, you will estimate the Hubble constant and the age of our Universe. Notice that in practice, there are some further intricacies involved (such as K-corrections, Malmquist bias, Galactic extinction, etc.), which we will ignore in this lab.

Getting the data

Download the dataset containing 580 supernovae measurements from the Moodle page of the course (`mu_vs_z.txt`). You can load the file in PYTHON using PANDAS with

```
import pandas as pd
filename = "mu_vs_z.txt"
data_full = pd.read_csv(filename, delim_whitespace=True)
```

The data contains 5 columns:

- *Supernova*: Supernova name
- *Redshift*: Redshift of the supernova's host galaxy
- *Modulus*: Distance modulus μ, related to distance d [pc] via $d = 10^{\mu/5+1}$
- *Error*: Uncertainty of distance modulus measurement ($1\sigma_\mu$)
- *LMGProb*: Probability that the supernova was hosted by a low-mass galaxy

Note: In this lab, we will always treat the redshift z (and quantities that are analytically derived from the redshift such as the recessional velocity) as the *independent* variable (on the "x"-axis) that we assume to know perfectly without any uncertainties, and the distance (modulus) as the *dependent* variable (on the "y"-axis) with associated uncertainties.

Exploring and manipulating the data

1. Open the supernova data, inspect the columns of the PANDAS dataframe, and plot the distance modulus μ as a function of redshift z. Also plot the uncertainties of the distance moduli σ_μ.

   For now, we will work only with nearby supernovae at redshifts $z \leq 0.1$.

2. Apply a redshift cut to the supernova dataset and select those with $z \leq 0.1$. Convert the distance modulus μ to a distance d in Megaparsec (Mpc) and make a plot of distance d vs. redshift z.

   For small distances d, the redshift z is directly proportional to the recessional velocity v, related via the formula

   $$z = \frac{v}{c}, \tag{1}$$

   where c is the speed of light.

3. Use this formula to convert the redshifts to recessional velocities, and plot the distance d as a function of recessional velocity v.

Determining the Hubble constant from nearby supernovae

Hubble's famous law (also known as the Hubble–Lemaître law) states that the recessional velocity v with which far-away objects like other galaxies move away from us due to the expansion of the Universe is proportional to the distance of the object d, i.e.

$$v = H_0 \, d, \tag{2}$$

where H_0 [km/s/Mpc] is the **Hubble constant** that describes the (current) speed of the expansion of the Universe. The reciprocal of the Hubble constant, which has units of time, is known as the **Hubble time** $t_H = 1/H_0$. The Hubble time would be the current age of the Universe if its expansion had been linear throughout cosmic history (which is in fact not quite the case).

1. **Least-square estimate** Given the distances d to the supernovae and the recessional velocities v of their host galaxies (as determined from their redshifts), find the least-square estimate for the Hubble time t_H (for now ignoring the uncertainties in d). Recall that we are treating the recessional velocities v as the "x"-variable and the (uncertain) distances as the "y"-variable. Then, convert this estimate for t_H to an estimate for H_0.

   Now, let us include our knowledge about different uncertainties for different supernovae into the fit. Since we only have uncertainties for the distance moduli μ, we first need to convert the uncertainties in μ into uncertainties in d. In order to do so, recall the uncertainty propagation formula

$$\sigma[d(\mu)] = \left|d'(\mu)\right| \sigma_\mu, \tag{3}$$

   where $d'(\mu)$ denotes the derivative of $d(\mu)$ and σ_μ is the uncertainty of the distance modulus (different for each data point).

2. Use the above formula to obtain uncertainties $\sigma_d := \sigma[d(\mu)]$ for the distances $d = d(\mu)$.

3. **Maximum likelihood estimate** Now, assuming Gaussianity of the errors σ_d, determine the maximum likelihood estimate for t_H. Again, convert this estimate to an estimate for H_0.

4. Plot again the distances d as a function of the recessional velocity v and add the least-square estimate and the maximum likelihood estimate to the plot. How do they compare?

Considering all supernovae and determining the age of the Universe

Now, we will look again at *all* the supernovae in the data (not only those with $z \leq 0.1$). For the remainder of this lab, we will take the redshift z to be the independent variable (on the "x"-axis), rather than the recessional velocity.

1. **Polynomial regression** As a first "cosmology-agnostic" approach, fit a polynomial of degree $r = 4$ (so there are $K = 5$ model parameters including the constant intercept) to the $\mu = \mu(z)$ distance moduli by computing the maximum likelihood estimate for $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_4)^T \in \mathbb{R}^5$, taking into account the uncertainties σ_μ for μ. Plot the polynomial with the maximum likelihood parameters over the data. Is this polynomial a good model for the data? Do you expect this model to generalize to new supernova data at higher redshifts $z > 1.5$?

2. **"Cosmological" regression** Now, let us consider a regression model that is motivated by cosmology. From theory, we know that the relationship between the (luminosity) distance d and the redshift z should be given by

$$d(z; t_H) = (1 + z)\, c\, t_H \int_0^z \frac{dz}{E(z)}, \tag{4}$$

where

$$E(z) = \sqrt{\Omega_\Lambda + \Omega_m (1 + z)^3}. \tag{5}$$

For the current dark energy and matter density parameters Ω_Λ and Ω_m, respectively, we take the parameters determined by the Planck Collaboration $\Omega_\Lambda = 0.685$ and $\Omega_m = 0.315$.

Note that while $d(z; t_H)$ is nonlinear in z, it is *linear* in our free parameter t_H, so we can determine the maximum likelihood estimate for t_H using what we have learned in the lecture. The steps to do this are:

(a) Convert the uncertainties for μ into uncertainties for d again, now for all the supernovae (using Eq. (3)).

(b) Build the feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N \times 1}$ using Eq. (4). Recall the definition of the feature matrix (which only has a single column now because we only have a single parameter t_H): the feature matrix should be such that $\mathbf{d} = \boldsymbol{\Phi}\, t_H + \boldsymbol{\varepsilon}$, where $\mathbf{d} \in \mathbb{R}^N$ contains the N distance measurements and $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ is a noise vector.

(c) Build the precision matrix $\mathbf{C}^{-1}$ containing the inverse of the uncertainty variances for d on the diagonal.

(d) Find the maximum likelihood estimate for t_H.

Overplot the maximum likelihood fit onto the data. Is this a good model for the data? What is the value of t_H (and consequently of H_0) that you obtained?

Hint: for the integral in Eq. (4), you can use the function `scipy.integrate.quad`.

From the Hubble time t_H (together with the cosmological parameters Ω_Λ and Ω_m), we can compute the age of the Universe $t = t(z)$ at redshift z via the formula

$$t(z) = t_H \int_z^\infty \frac{dz}{(1+z)\,E(z)}, \tag{6}$$

where $E(z)$ is defined as above.

3. **Age of the Universe** Using the maximum likelihood estimate for t_H obtained in the previous part, compute the current age of our Universe (at redshift $z = 0$).

   Hint 1: The function `scipy.integrate.quad` also accepts `numpy.infty` as an integration boundary.

   Hint 2: make sure to be careful with the units, especially for the Hubble time.