Problems With Transformed Data

A. Alsakka, Physics, Umeå University
abdullahalsakka@hacettepe.com

M. L. Smith, Retired
mlsmith55@gmail.com

UMEÅ UNIVERSITY

January 22, 2023

Contents

Introduction

Scientists often use graphs to represent observations and mathematics to explain correlations.

It is common to propose several mathematical descriptions then choose the best model using statistics as a guide.

One should select the proper metric for data analysis. The experiment of Galileo is a good example where free fall distance is correlated to time.

One would select meters and seconds for the Galileo experiment; miles and days would be inappropriate.

Introduction

Sometimes presentations are better received when a metric is transformed; to compress widely spread data (the dynamic range problem) for ease of presentation.

Metric transformation is usually a mistake when applied to analysis.

We explore some problems when inappropriate metrics are chosen. One example presents astronomical observations of distant supernovae.

Another example, presented in the **Appendix**, follows cancer cell replication over time.

We find that investigators select the proper metrics for one experiment but inappropriate metrics for the other.

Introduction

A common fault is transforming data by applying the log function, log(x), then using the "new" data. Some reasons why using log(data) often leads to incorrect results:

1. Data are measured using metrics (m, km, seconds, etc.), however $\log(x)$ has no metric.
2. Standard deviations, log(σ), are distorted often becoming worse the further the measurement is taken from the origin.
3. The best data pair, sometimes the origin (0,0) without significant error, cannot be used because $\log(0) = -\infty$. Exclusion of any data pair with insignificant error can never be justified.
4. Since the σ values are distorted, computerized goodness of fit estimates are improperly estimated; χ^2 and r^2 are often incorrect, complicating model discrimination.

Standard Deviations

The the standard deviation, σ, is an estimate of the uncertainty of the dependent measurement [3]. This usually means the correct value has a $\approx 67\%$ probability of residing between the upper and lower limits of σ, presuming a Gaussian distribution of measurables.

Consider a function of $f(x)$ where the x value has an estimated error of $\pm x_{\text{err}}$. Our error for the y values at any arbitrary x value is

$$y_{\text{err}} = \frac{f(x + x_{\text{err}}) - f(x - x_{\text{err}})}{2}$$

Therefore the standard deviation, σ, is an estimate of how much the y value might deviate from the mean, so $\sigma = y_{\text{err}}$. The dependent variable is usually chosen as the variable with greater σ and the $\sigma_{(x)}$ is often considered to be negligible to ease calculations.

Chi-squared test, χ^2

Pearson's χ^2 is a commonly used measure of a model compliance with the data and calculated as

$$\chi^2 = \sum \left(\frac{(\text{Observed value} - \text{Expected value})^2}{\text{Denom}} \right).$$

where the Denom is either the Expected value or σ^2. A smaller χ^2 usually indicates a better model [4].

Some investigators use the *reduced* χ^2, which is simply the "normalized" χ^2 as

$$\chi^2_{red} = \chi^2 / (N - P)$$

where N is the number of data pairs and P the parameter count. Models sporting many parameters will display slightly larger values of χ^2_{red}.

r-squared test, r^2

The r-squared, r^2, statistic is a commonly used measure of the correlation between the dependent and independent variables [5,6]. It is calculated as

$$r^2 = 1 - \frac{\sum(y_{\text{observed}} - y_{\text{expected}})^2}{\sum(y_{\text{observed}} - y_{\text{mean}})^2}.$$

Formally, r^2 only applies to linear models, for more general application the *adjusted* r^2 is used as

$$r^2_{adj} = 1 - \frac{(1 - r^2)(N - 1)}{(N - P - 1)}$$

where N is the number of data pairs and P the parameter count. There are many other statistical tests which are useful to estimate model worth, the F-test being very good.

Codes and data

The popular Python 3 programming language used here is available for free and one helpful website is

https://www.anaconda.com/products/distribution

We checked our routines with the Jupyter Notebook and Spyder IDE using both Apple® Macintosh and Windows® 11 on a PC laptop. The publicly available data are saved as .csv files for use with *scipy*, *pandas* and *numpy* Python libraries.

Where to find the data? Where to find the programs?

All programs and data are available in a GitHub repository -

https://github.com/MLSResearchGroup/About-Transforming-Data

Data

The data are from online sources and articles [1], [2].

For the first demonstration, we use observational data of supernovae type Ia (SNe Ia) emissions reported by the Riess team of astronomers [1]. We use the *mag* (or *distance mag*) and σ as given and "back-calculate" the supernovae distances, D_L, and σ values using a spreadsheet which is saved in the GitHub repository.

The second experiment (**Appendix**) are counts of cancer cells observed every 4 hours for 236 hours during unhindered replication [2]. Ten independent counts were made at each time and the mean and σ were calculated using a spreadsheet and saved in the .csv file.

Supernovae Type Ia Cosmology Test
background

Astrophysicists commonly correlate galactic distances with redshifts (*z*). Strong correlation is considered evidence supporting claims about Universe expansion rate, spacetime geometry, matter density, dark matter and dark energy.

Intergalactic distances are often measured from light intensity, using a metric termed *mag* or *distance mag*. This historic method follows the logarithmic scaling of luminosity as perceived by the human eye.

The actual luminosity distance is based better on D_L, as parsec (pc), kiloparsec (Kpc) or megaparsec (Mpc) which are measures of large distances. The best data currently comes from measuring the luminosity of distant supernovae (SNe Ia) explosions and the associated galaxy redshift, z. These emissions are considered standard candles.

Supernovae Type Ia Cosmology Test

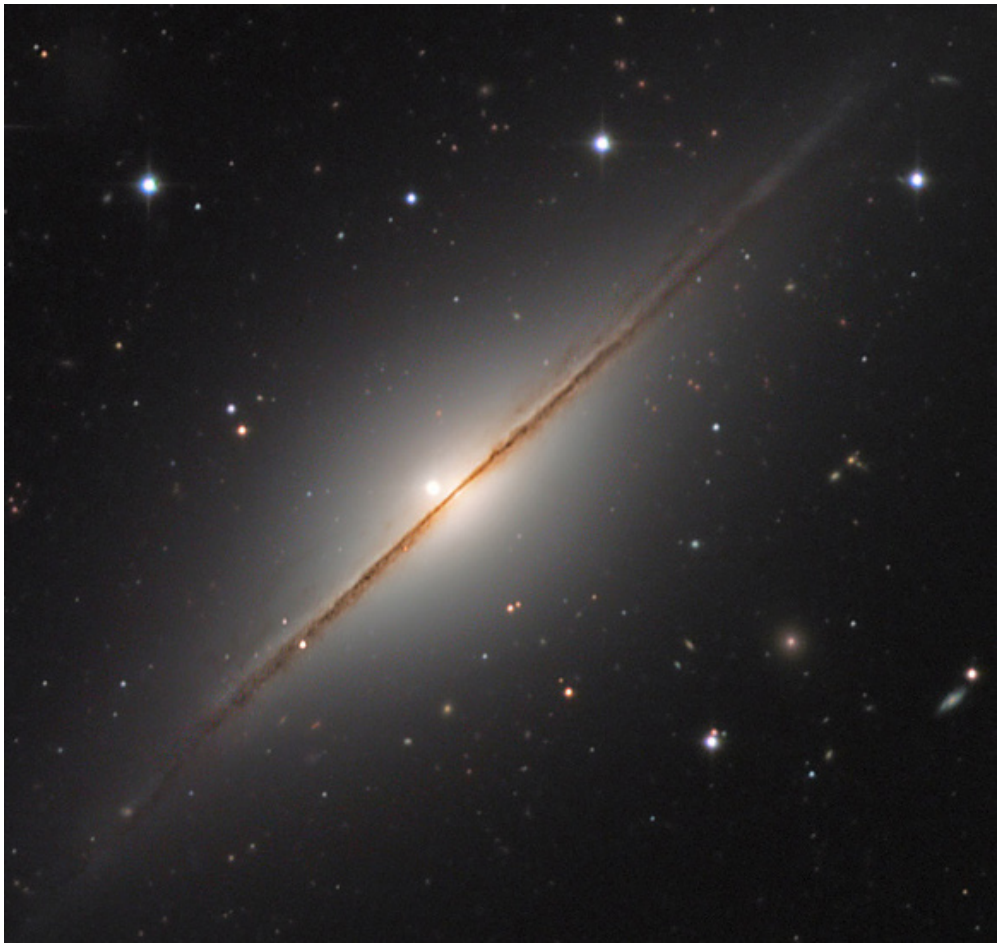


Figure: The supernova near the Little Sombrero galaxy. The bright light just to the left of the galactic center is a type Ia supernova and an example of light and possibly dust contamination. July 17, 2021; Image from NASA.

For the SNe Ia data the mag is a function of D_L as

$$mag = 5 \log_{10} D_L + 25$$

and the luminosity distance, D_L in Mpc units, is "back-calculated" from *mag* as

$$D_L = 10^{(mag-25)/5}.$$

Astrophysicists use the redshift, z, as a measure of galactic recession velocity, this is measured with little error. We calculate the relative recession velocity using

$$\xi = \frac{1}{1+z} = \frac{\nu}{\nu_0}$$

where ξ is the ratio of the observed, lower frequency, ν, over ν_0 - the frequency unmodified by the Doppler effect. The value ξ is often termed the expansion factor or sometimes a.

The Einstein-DeSitter (E-DS) model considers our Universe being primarily matter without spacetime or dark energy [7]. To find the local Hubble constant, H_0, with mag and z data we use

magE-DS model

$$mag = 5\log_{10}\left[\frac{c(1+z)}{H_0}\sinh\left(\frac{z}{1+z}\right)\right] + 25.$$

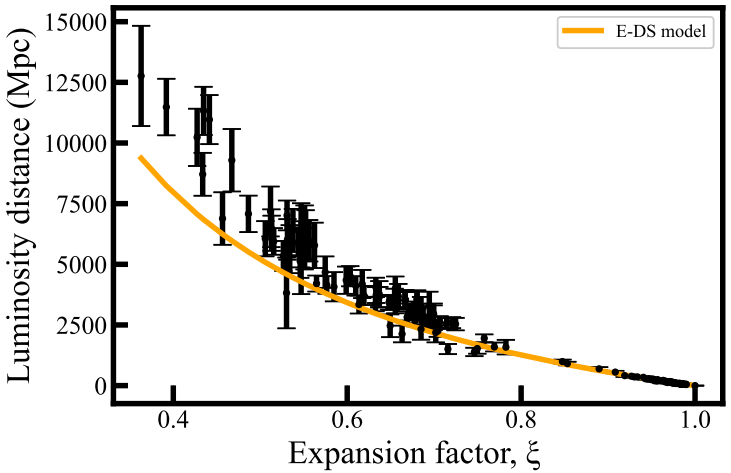
With the D_L and ξ data we use the simpler

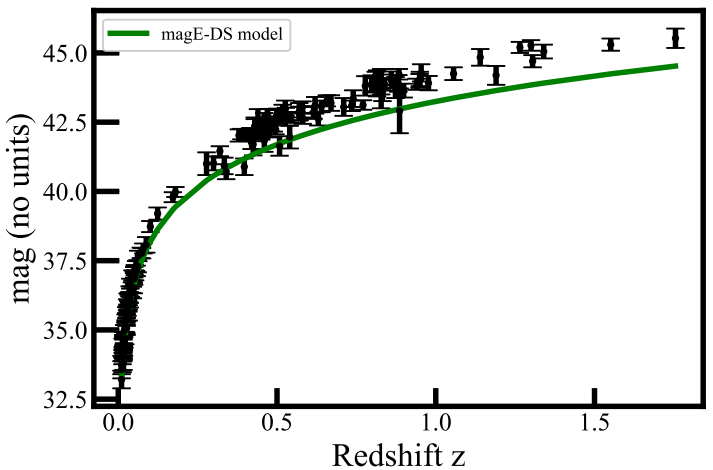
E-DS model

$$D_L = \frac{c}{H_0\xi}\sinh\left(1-\xi\right)$$

here c is lightspeed as km s^{-1} and H_0 as km s^{-1} Mpc^{-1}.

Supernovae Type Ia Cosmology Test
E-DS model, using "gold" data from Riess et al. 2004





The E-DS model was suggested before the tremendous size of our Universe was known. The best data available to Einstein was published by Hubble in 1929 [8].

Physicists now use *normalized* parameters to calculate properties of our Universe: for matter density we use Ω_m; Ω_k describes the portion which is not matter but spacetime; Ω_Λ to describe galactic motions and Universe expansion [9].

normalized Ω parameters

Normalized parameters means the parameters sum to 1; keep these relationships in mind

$$\Omega_m + \Omega_\Lambda = 1$$

for the ΛCDM model, the *standard model* of cosmology and

$$\Omega_m + \Omega_k = 1$$

for the *arctanh* models; the Universe without dark energy. Remember that $1 - \Omega_m = \Omega_\Lambda$ or $1 - \Omega_m = \Omega_k$.

We write the luminosity distance, D_L, and ξ in terms of the normalized parameters for the *standard model* as

ΛCDM model, distance vs. recession velocity

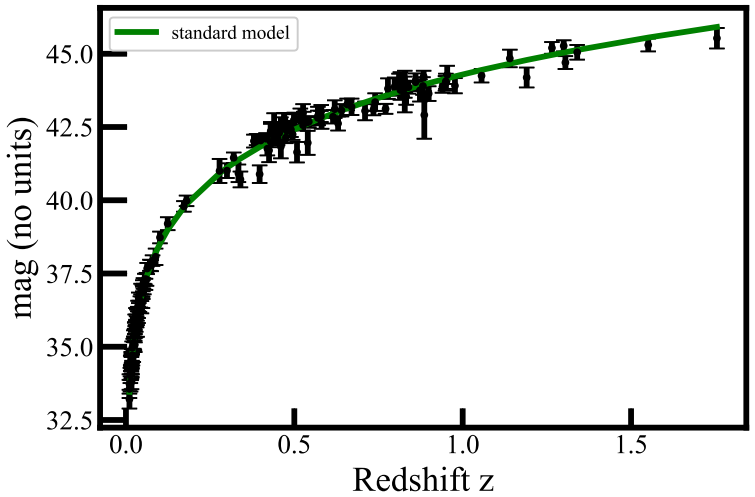
$$D_L = \frac{c}{H_0 \xi} \sinh \left\{ \int_{\xi_1}^{1} \frac{d\xi}{\xi \sqrt{\frac{\Omega_m}{\xi} + \Omega_\Lambda \xi^2}} \right\}$$

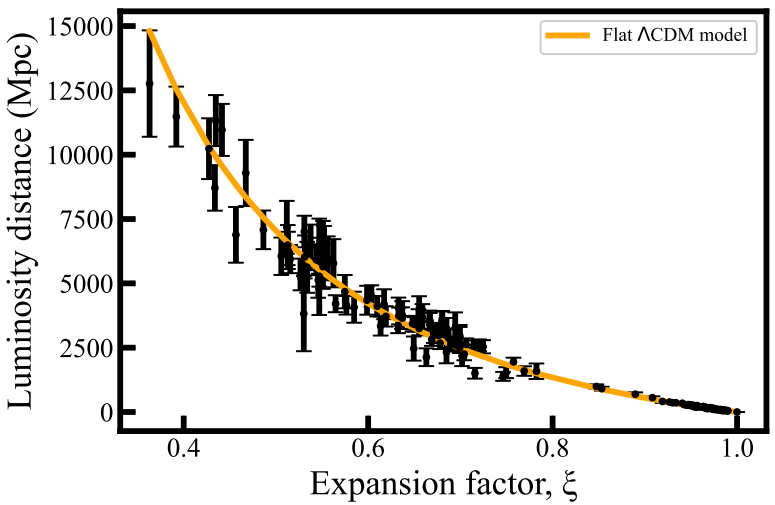
This integral cannot be calculated analytically, so we must use numerical integration methods in Python. We also write the *standard model* in terms of *mag* and redshift z as

The standard model of cosmology, mag vs. redshift [9,10]

$$mag = \frac{c(1+z)}{H_0} \sinh \left\{ \int_{0}^{z} \frac{dz}{\sqrt{(1+z)^2(1+\Omega_m z) - z(2+z)(\Omega_\Lambda)}} \right\}$$

Supernovae Type Ia Cosmology Test
The *standard* and ΛCDM models





Note the difference in size of σ values between plots.

The *arctanh* model describes a universe containing matter with non-Euclidean spacetime but without the cosmological constant (no dark energy) [7]. The relationship takes the form

arctanh model, distance vs. recession velocity

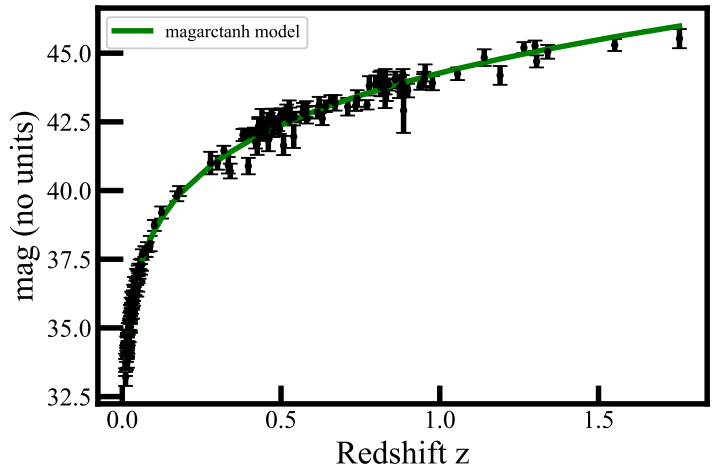
$$D_L = \frac{c}{H_0 \xi \sqrt{|\Omega_k|}} \sinh\left\{ \sqrt{|\Omega_k|} \int_{\xi_1}^{1} \frac{d\xi}{\xi \sqrt{\frac{\Omega_m}{\xi} + \Omega_k}} \right\}$$

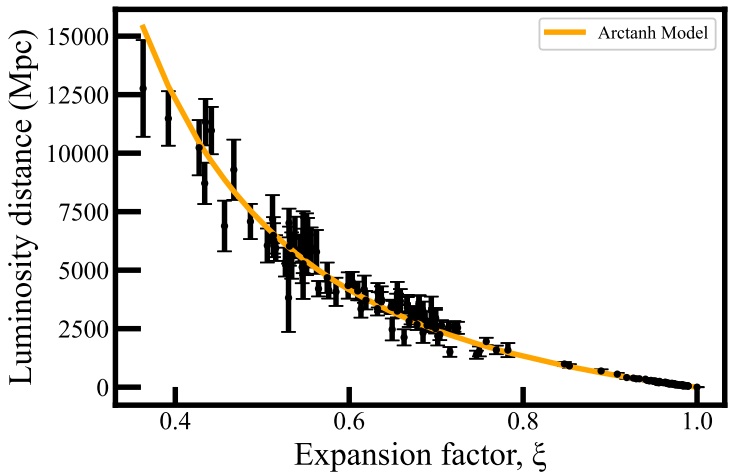
and this can be integrated to become

$$D_L = \frac{c}{H_0 \xi \sqrt{|\Omega_k|}} \sinh\left\{ 2\left(\operatorname{arctanh}(\sqrt{|\Omega_k|}) - \operatorname{arctanh}\left(\frac{\sqrt{|\Omega_k|}}{\sqrt{\frac{\Omega_m}{\xi} + \Omega_k}} \right) \right) \right\}$$

Supernovae Type Ia Cosmology Test

Arctanh model, *mag vs z* and D_L vs ξ plots





These figures are similar in appearance to the *standard* and ΛCDM model figures though the models are quite different.

Table: Results from the Einstein-DeSitter, *standard* and *arctanh* models using the gold SNe Ia data with curve_fit regression routine. All calculations use relative weights of σ values.

model	H_o (km s^{-1} Mpc^{-1})	Ω_m with Ω_k or (Ω_Λ)	r^2_{adj}
E-DS	57.4±0.8	-	0.887
*mag*E-DS	54.8±0.8	-	0.988
ΛCDM	64.8±0.9	(0.48±0.05)	0.968
standard	63.8±0.8	(0.47±0.05)	0.995
3P*standard*	65.4±0.9	0.32±2.26	0.970
arctanh	63.9±0.8	0.008±0.075	0.965
*mag*arctanh	62.7±0.8	0.001±0.075	0.994

The 3P*standard* model includes all Ω_m, Ω_Λ and Ω_k, parameters. Note the standard deviation of Ω_m for the 3P*standard* model is very large due to the use of 3 parameters (H_0, Ω_m, Ω_Λ).

Predicting our present & future
The *arctanh* model success and *magarctanh* failure

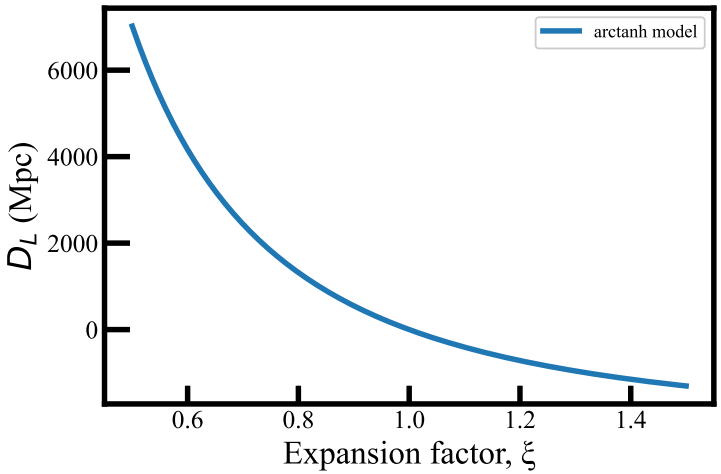
A strength of science is the ability to predict the future and describe the present. Most interesting, calculation of the origin at $z = 0$, our present situation, is impossible with the *mag*arctanh and *standard* models

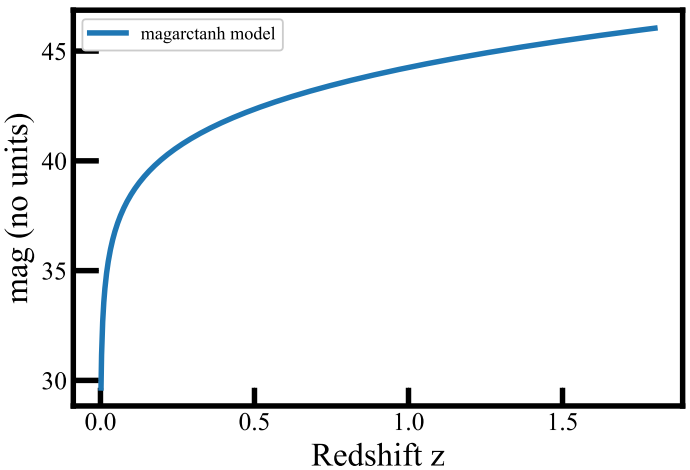
- the log of our position at the origin 0 is $-\infty$
- but our position and recession velocity are known with 100% certainty

Our present and future conditions are presented in the next two figures.

Also - a plot of the expansion factor, ξ, *vs.* redshift *z* shows that the redshift does not correlate in a linear manner with astronomical recession velocity.

Plotting our present & future
The *arctanh* and *magarctanh* models

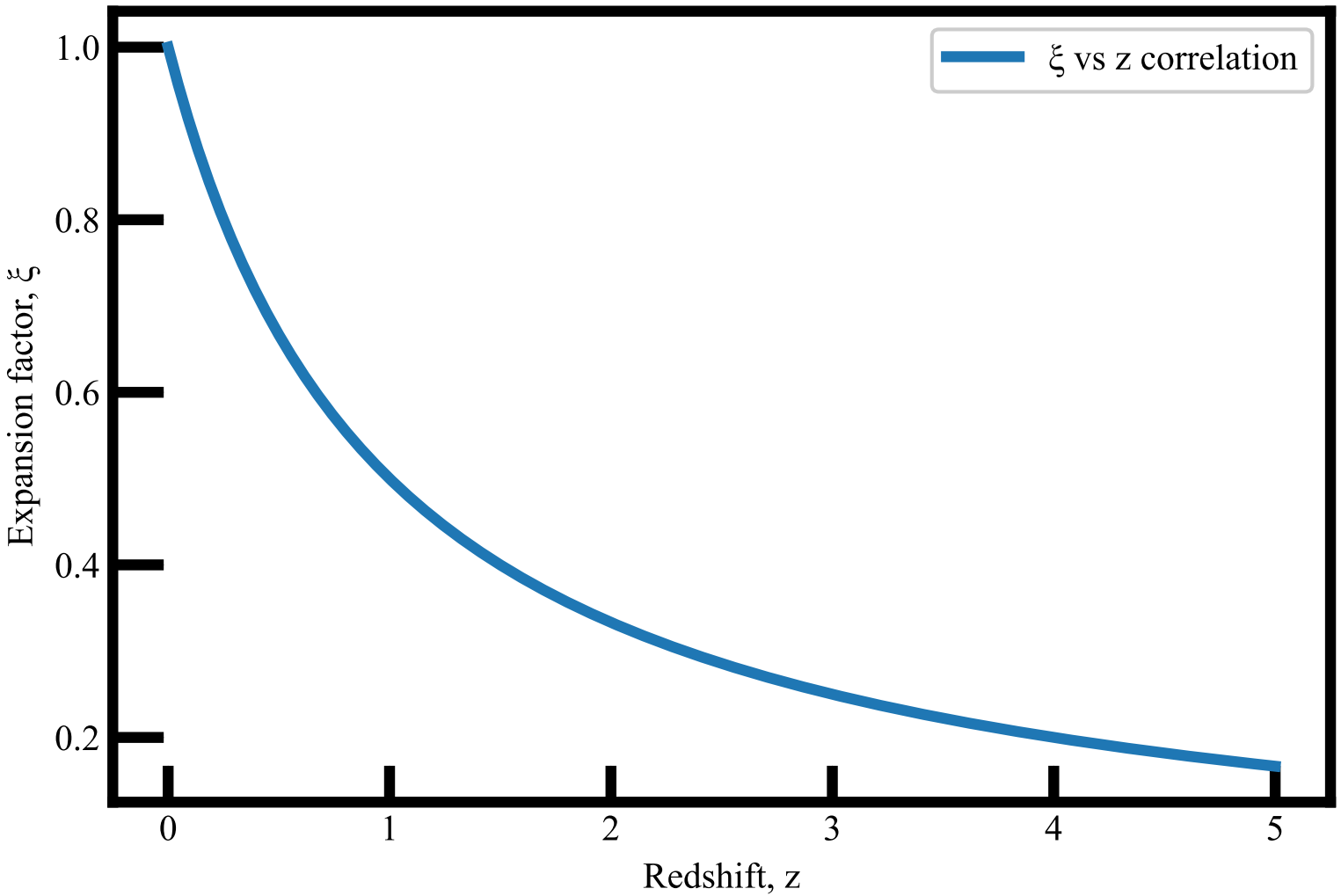




The calculated curve of the *mag*arctanh model, bottom plot, approaches 0,0 but Python returns an error message for a value of *mag* at z=0.

Plotting transformed metric

The ξ vs z relationship



This trace of the expansion factor, ξ, vs redshift, z, displays a nonlinear relationship.

Supernovae Type Ia Cosmology Test
Results

- The E-DS model functions well for nearby SNe but fails at larger distances*
- The Hubble constant from the E-DS calculations is ≈55-58 km s^{-1} Mpc^{-1}, much lower than recent estimates of ≈62-74 [11]
- The values for Ω_m from the *standard* and ΛCDM models are significantly higher than the *arctanh* models
- The Ω_k is calculated to be very large (≈1) for the *arctanh* model but presumed to be ≈0 for the *standard* model
- Because σ values are smaller after log transformation, those values of r^2_{adj} should be considered doctored
- Parameter values and statistics differ between metrics and the statistics are distorted by the log(data) transformation [12,13]

*The statistics calculated using these data do not necessarily support a preference for the *standard* model over the *arctanh* model but only over the E-DS model

Conclusions

We demonstrate that data transformation shall be avoided

- Transforming data just to visualize a straight line may lead investigators towards incorrect results
- The σ values associated with SNe Ia *mag* values [1, 10] suffer artificial contraction
- Discarding the origin when employing the log(data) transformation cannot be justified
- Calculated parameters and standard deviations should be similar using either metric values or transformed values but are not
- The example following cancer cell counts during replication showed that transforming good data, especially the σ values, can lead to faulty statistics and is a waste of time (**Appendix**)

Bibliography

1. A.G. Riess, et al., Type Ia Supernova Discoveries at z> 1 from the Hubble Space Telescope: Evidence for Past Deceleration and Constraints on Dark Energy Evolution. *Astrophys. J.* **607**, 665 (2004).
2. K.E. Johnson, et al., Cancer cell population growth kinetics at low densities deviate from the exponential growth model and suggest an allee effect. *PLoS Biology* **17**, e3000399 (2019).
3. M. Hargrave, Standard Deviation Formula and Uses vs. Variance. *Investopedia* (2022, July 6). https://www.investopedia.com/terms/s/standarddeviation.asp
4. A. Hayes, Chi-Square (χ^2) Statistic. *Investopedia* (2022, October 23). https://www.investopedia.com/terms/c/chi-square-statistic.asp
5. J. Fernando, R-Squared Formula, Regression, and Interpretations. *Investopedia*(2021, September 12). https://www.investopedia.com/terms/r/r-squared.asp
6. S. Gupta, R-Squared: Formula Explanation - Analytics Vidhya. *Medium* (2021, December 28). https://medium.com/analytics-vidhya/r-squared-formula-explanation-6dc0096ce3ba

Bibliography

7 A.M. Oztas, et al., Spacetime Curvature is Important for Cosmology Constrained with Supernova Emissions. *Int. J. Theor. Phys.* **47**, 2474 (2008). DOI 10.1007/s10773-008-9680-7

8 E. Hubbel, A relation between distance and radial velocity among extra-galactic nebulae. *Proc. Nat. Acad. Sci. USA* **15**, 168. DOI 10.1073/pnas.15.3.168

9 S.M. Carroll, et al., The Cosmological Constant. *Ann. Rev. Astron. Astrophys.* **30**, 499 (1992). DOI 10.1146/annurev.aa.30.090192.002435

10 A.G. Riess, et al, Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *Astronom. J.* **116**, 1009 (1998). DOI 10.1086/300499

11 W.L. Freedman, et al., Calibration of the Tip of the Red Giant Branch. *Astrophys. J.* **891**, 57. DOI 10.3847/1538-4357/ab7339

12 M.L. Smith & A.M. Oztas, Log-transformation problems: astrophysics examples. (2017) https://www.youtube.com/watch?v=Y1nEQmg2yJA&feature=youtu.be

13 M.L. Smith & A.M. Oztas, Lawrence Krause Do Your Homework. (2017) https://www.youtube.com/watch?v=IN0jqhqjW3Y&feature=youtu.be

Notes on terms
Some terms applied to spacetime geometry

The following terms are used interchangeably by many physicists though the exact meanings are not identical

- elliptical (concave) geometry and closed universe
- flat geometry and Euclidean and quasi-Euclidean geometry
- hyperbolic (convex) geometry and open universe*
- *hyperbolic geometry is impossible in a universe with matter - only possible in a universe with anti-gravity

Cancer Cell Replication
Theory

Cancer cell replication is a good example of exponential growth. We model the cell count increase over time as

Exponential model (Exp)

$$N(t) = N_0 e^{\beta t} \tag{1}$$

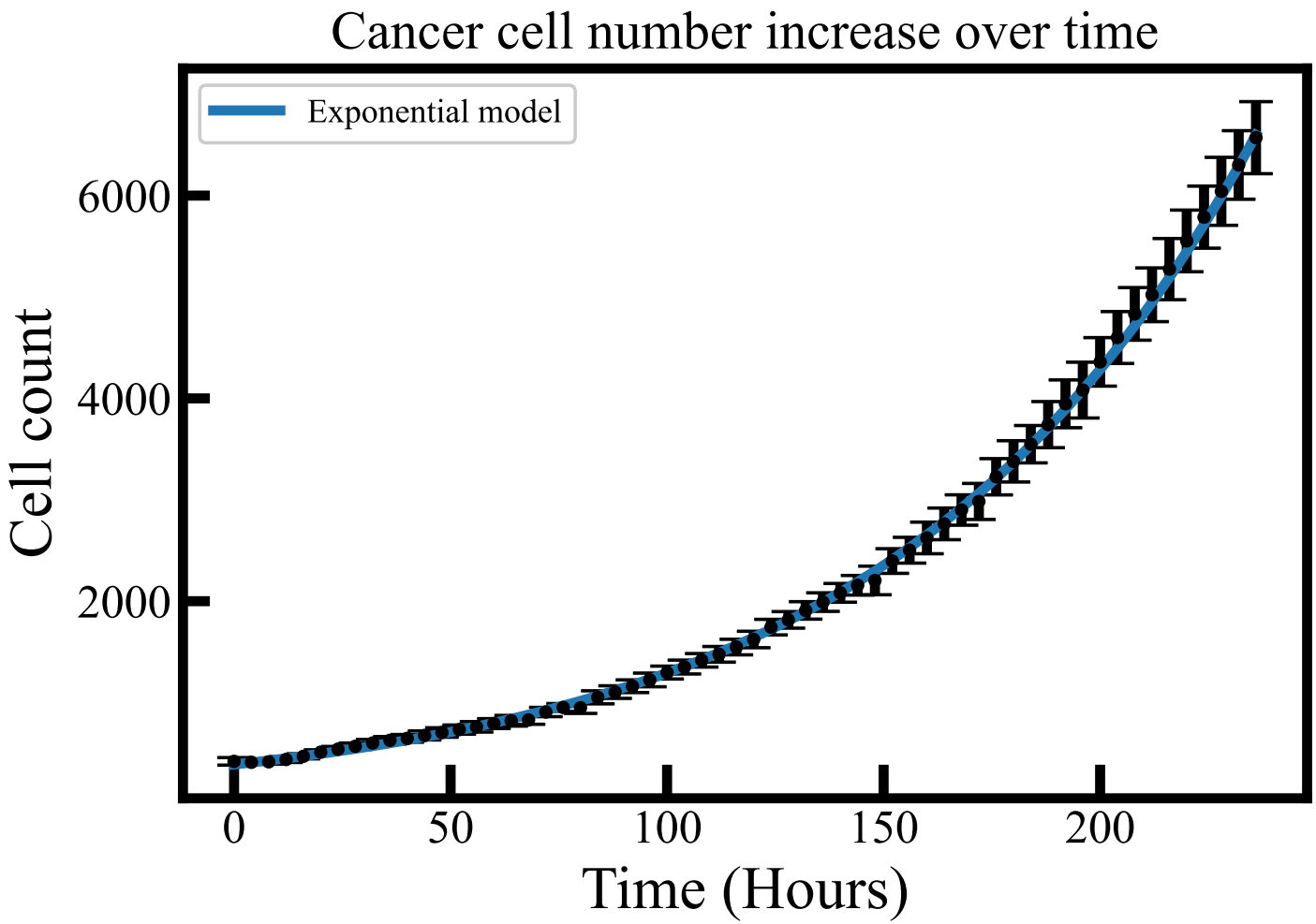
Here N_0 is the initial cell number, β the replication rate and t, the time (hours). To transform the dependent data we take the natural logarithm of equation (1) on both sides yielding

logarithmic model (Log)

$$\ln[N(t)] = \ln N_0 + \beta t \tag{2}$$

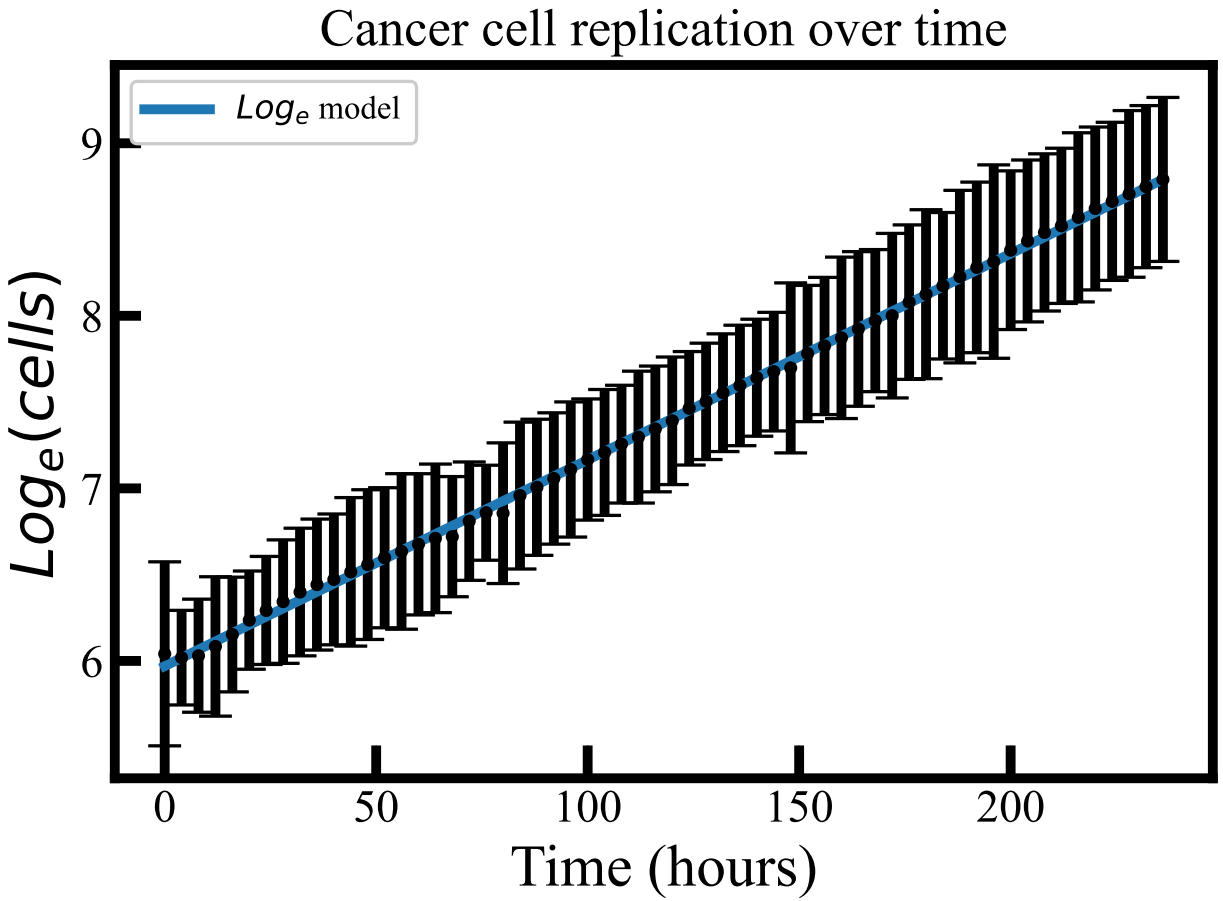
our logarithmic model is an example of coordinate transformation; the y-axis data are transformed into log_e*(data)*.

Cancer Cell Replication
Exponential Model Plot



Note the steady increase of standard deviation with time.

Cancer Cell Replication
Log Model Plot



The semi-log plot example. Note the smaller increase of $Log_e(\sigma)$ with time.

Cancer Cell Replication
Results

Table: Results from the Exp and Log models of cancer cell replication data using curve_fit regression routine. Both calculations use relative weights of σ values.

Model	N_0 (initial cell count)	β (hour^{-1}) growth rate	r^2_{adj}	χ^2
Exp	389.6$\pm$2.3	0.012$\pm 4E^{-5}$	0.9995	10.1
Log_e	391.4$\pm$1.0	0.012$\pm 5E^{-5}$	0.9992	0.22

- Both r^2_{adj} values are $\approx$1 but χ^2 values are dissimilar; these two regressions should be comparable but are not
- Distortion of the $Log_e(\sigma)$ values is reflected in the very small χ^2; warning that these results are suspect
- The value for N_0 are smaller than the cell count at $t = 0$, indicating some cells died during experiment initiation

Questions

Some questions may be asked

- Derive the well known equation for standard deviation

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)}{N}}$$

  where μ is the mean, x_i the observed values and N the number of observations, from the equation on page 6.
- Why does the Log model presents a slightly better estimate of initial cell count (420) than the Exp model?
- Does this justify using a Log model, or is it just happenstance?
- Evaluate the local Hubble law $\frac{c}{\xi} = H_0 D_L$ beginning with the E-DS or arctanh model?
- Which universe model do you think is closer to reality? Why?

Questions

Some more questions

- The expansion factor, $\xi = 1$, is our relative recession velocity and the position of our earth, $D_L = 0$, a certainty. Now consider our situation when the recession velocity $z \to 0$ and *mag*$= -\infty$.
  - Is it justified to discard the position of the earth, the only point with full certainty, when using the *mag* models?
  - Is the redshift, *z*, a useful transform of ξ? (**appendix**)
- The σ values associated with the *mag* data are relatively smaller than the σ values associated with D_L. Do we always want smaller errors, explain?
- When Einstein derived his cosmological model (E-DS), it was thought the universe was only the Milky Way and all stars are motionless.
  - Do you think this is reflected in the E-DS model?
  - Historical question - Why did Einstein and Stephen Hawking consider our Universe matter dominated?
- Difficult question - edit either the arctanh or magarctanh model to evaluate the data with one more free parameter - the Ω_k. The code will now have 3 parameters. If successful note the standard deviations.