American Journal Of Physics

Problems With Transformed Data

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

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

UMEÅ UNIVERSITY

November 21, 2022

Contents

Introduction

Scientists commonly use graphs to represent observations and data and mathematics to explain correlations. The experiment of Galileo is an example where free fall distance is correlated to time.

It is important to select the proper metric for data analysis. One would select meters and seconds for the Galileo experiment; miles and days would be inappropriate.

We explore some problems when inappropriate metrics are chosen. One example follows cancer cell replication over time, our second example presents astronomical observations of supernovae. We show that investigators select the proper metrics for one experiment but inappropriate metrics for the second.

Introduction

A common fault is choosing data without a metric or transforming data by applying the log function, log(x), into data without a metric. Some reasons why changing the *metric* to 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, σ, are compressed. This often becomes worse the further the measurement is taken from the origin.
3. The best data pair, sometimes the origin at 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. The σ values are distorted, goodness of fit estimates are improperly estimated; χ^2 and r^2 are incorrect complicating model discrimination.

Standard Deviation Problems

The σ, the standard deviation, is an estimate of the uncertainty of the dependent measurement. 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 "real" values.

For example, say we have 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

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

Remember, the standard deviation, σ, is an estimate of how much the y value might deviate from the mean. Our y_{error} is how far our highest or lowest y values might deviate from the "real" (mean) y value. So we write $\sigma = y_{\text{err}}$

Chi-squared test

χ^2 is a measure, a test, of the model fit to the observed data and χ^2 is calculated as

$$\chi^2 = \sum \left(\frac{(\text{Experimental value} - \text{Calculated value})^2}{(\sigma)^2} \right)$$

In general a smaller χ^2 indicates a better model. We use the *reduced* χ^2 in our tests, which is simply the normalized χ^2 calculated by

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

where N are the number of data pairs and P is the parameter count. Usually, the smaller the χ^2_{red}, the better our model. Note that models sporting many parameters (P) will display larger values of χ^2_{red}.

r-squared test

The r-squared, r^2, statistic is a measure of the correlation between the dependent and independent variables. 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 are the number of data pairs and P is the parameter count, which is 2 for our two examples.

There are many other statistical tests which are useful to estimate model worthiness, the F-test being very good.

Programs

We use the Python 3 programming language to write all codes. We checked these with the Jupyter Notebook and Spyder IDE using both Apple® Macintosh and Windows® 11 on a PC laptop. We save the publicly available data 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 saved in a GitHub repository for public use.

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

Data

Both sets of data are from online sources and articles [1], [2].

The first set are counts of cancer cells every 4 hours for 236 hours during replication [1]. Ten independent counts were made at each time and the mean and σ were calculated.

For the second set, we use data of supernovae type Ia (SNe Ia) emissions reported by the Riess team of astronomers [2]. We use the values of *mag* (*distance mag*) and σ as given and "back-calculate" the supernovae distances, D_L, and related σ values which are saved in the repository.

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 growth rate and t, the time (hours). To transform the dependent data we take the natural logarithm of equation (1) on both sides which yields

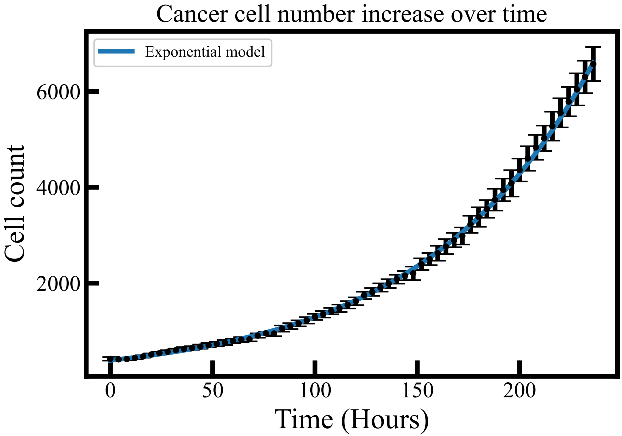
logarithmic model (Log)

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

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

Cancer Cell Replication

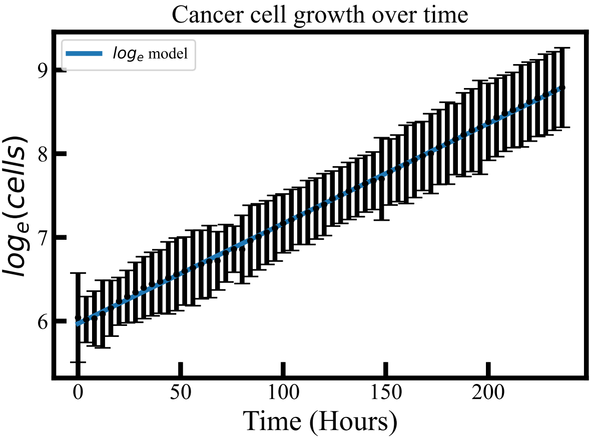
Exponential Model Plot



Note the increase of standard deviations with time.

Cancer Cell Replication

Log Model Plot



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

Cancer Cell Replication

Results

Estimates of initial cell count, hourly growth rate, r^2_{adj} and χ^2_{red} from the fitted models. For the Log model we use the $log_e(mean\ count)$ and $log_e(\sigma)$ values for evaluation.

Model	N_0 (initial cell count)	β (hour^{-1}) growth rate	r^2_{adj}	χ^2_{red}
Exp	389.6± 2.3	0.012$\pm 4E^{-5}$	0.9995	0.17
Log	391.5± 1.0	0.012$\pm 5E^{-5}$	0.9992	0.0038

Since the r^2_{adj} values approaches 1 and χ^2_{red} is small for both calculations there is no advantage to transform the data. The obvious distortion of the σ values, which is reflected in the very small χ^2_{red}, is a warning that data transformation is unnecessary and yields invalid statistical results.

Supernovae Type Ia
background

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

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

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

Supernovae Type Ia

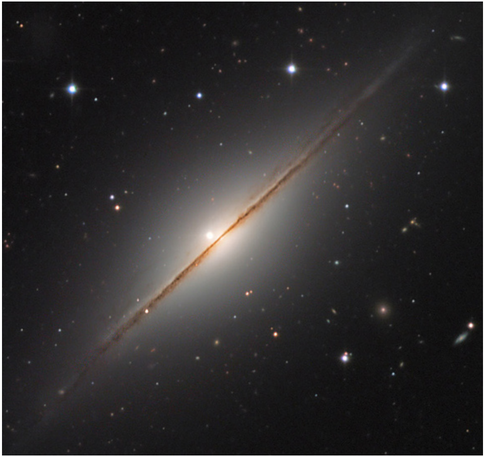


Figure: The supernova near the Little Sombrero galaxy. The bright light just to the left of the galactic center is the type Ia supernova as recorded July 17, 2021. Image from NASA.

Supernovae Type Ia

Calculating astronomical distances and velocities

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, to measure galaxy recession velocity, which is easy to measure 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 frequency, ν, over ν_0 - the frequency unmodified by the Doppler effect. The value ξ is often termed the expansion factor or sometimes a.

Supernovae Type Ia
The Einstein-DeSitter model

The Einstein-DeSitter (E-DS) model considers our Universe being primarily matter without spacetime or dark energy [7]. For evaluation with the mag and z data we use

magE-DS model

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

For use with the D_L and ξ data we can use the simpler

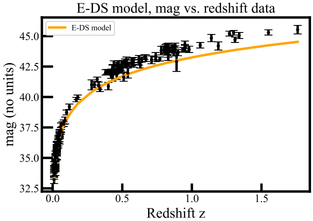
E-DS model

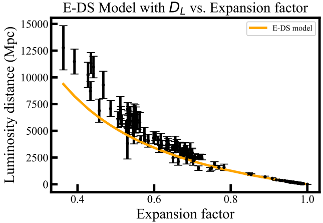
$$D_L = \frac{c}{H_0\xi}\sinh\left(1-\xi\right)$$

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

Supernovae Type Ia

Einstein-DeSitter model, using "Gold" data





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 [9].

Supernovae Type Ia

Normalized parameters

Physicists now use *normalized* parameters to calculate properties of our Universe. For matter density we use Ω_m, the Ω_k to describes the portion which is not matter but spacetime and Ω_Λ describes galactic motions and expansion of the Universe [8].

normalized 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* model (**see appendix**) ; the Universe without dark energy. Remember that $1 - \Omega_m = \Omega_\Lambda$ and $1 - \Omega_m = \Omega_k$.

Supernovae Type Ia

Standard model of cosmology

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

standard model, distance vs. recession velocity

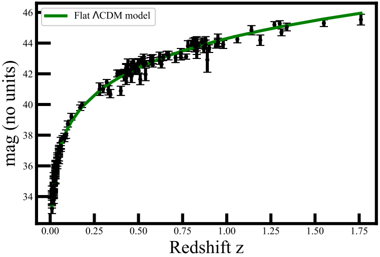
$$D_L = \frac{c}{H_0 \xi} \sinh \left\{ \int_{\xi_1}^{1} \frac{d\xi}{\xi \sqrt{\frac{\Omega_m}{\xi} + \Omega_\Lambda \xi^2}} \right\}$$

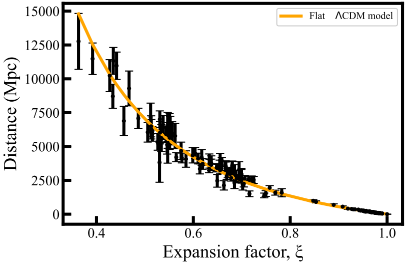
This integral cannot be calculated analytically, so we must use numerical integration methods in Python. We also write the *standard model* in terms of *mag* and redshift z [2] as

The standard model, mag vs. redshift

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

The *standard* and ΛCDM models





Note the difference of the σ sizes between plots.

Supernovae Type Ia
Results

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	Ω_m	r^2_{adj}	χ^2_{red}
	(Mpc km^{-1}s^{-1})	Ω_k or (Ω_Λ)		
E-DS	57.5±0.8	-	0.888	2.90
*mag*E-DS	55.4±0.9	-	0.984	4.17
The *standard* model	65.0±1.1	(0.43±0.6)	0.991	2.08
arctanh	63.9±0.8	0.01±0.08	0.965	1.29
*mag*arctanh	63.4±1.0	0.001±0.1	0.991	2.16

The H_o of the E-DS models, are significantly lower than the *standard* and *arctanh* models (**see appendix**). The r^2_{adj} and χ^2_{red} differ between the *mag* vs. *z* and D_L vs. ξ calculations.

Supernovae Type Ia

Results

- The E-DS model appears to model nearby distances well but fails at larger distances
- The Hubble constant from the two E-DS calculations is $\approx$55-58 Mpc km^{-1}s^{-1}, much lower than current estimates of $\approx$67-74 [10]
- Because σ values are smaller after log transformation, the values of r^2_{adj} and χ^2_{red} are doctored
- The values for Ω_m from the *standard* and ΛCDM models are vastly different than the *arctanh* models
- The Ω_k is calculated to be very large ($\approx$1) for the *arctanh* model but presumed to be $\approx$0 for the *standard* model
- Parameter values and statistics should not differ between metrics but are nearly always distorted by log(data) transformation

Conclusions

We demonstrate that transforming data should be avoided

- The example using cancer cell counts during replication showed that transforming good data, especially the σ values, can lead to faulty statistics
- Transforming data to visualize a straight line relationship may lead to incorrect results
- The σ values, as reported for SNe Ia distances [2], 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 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

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 5.
- How can we find the experimental values of β and N_0 from the semi-log model?
  - Discuss why we might expect the semi-log model to present a better estimate of the initial cell population (420) than the exponential model.
  - Does this justify using a semi-log model, or is it just happenstance?
- Derive the Hubble law $\xi = H_0 D_L$ from the E-DS model
- Which universe model do you think is closer to reality and why?

Questions

Some more questions

- The expansion factor, $\xi = 1$, is our relative recession velocity and the position of our earth, $D_L = 0$, is 100% certain. Now consider the situation of the earth when recession velocity $z \to 0$ and *mag* approaches $-\infty$.
  - Is it justified to discard the position of the earth, the only point with full certainty, when using these models?
  - Is the redshift, *z*, a useful transformation of ξ?
- 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?
  - Why did he consider our Universe matter dominated ?

Bibliography

1 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).

2 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).

3 A. Hayes, Chi-Square (χ^2) Statistic. *Investopedia* (2022, October 23). https://www.investopedia.com/terms/c/chi-square-statistic.asp

4 J. Fernando, R-Squared Formula, Regression, and Interpretations. *Investopedia*(2021, September 12). https://www.investopedia.com/terms/r/r-squared.asp

5 M. Hargrave, Standard Deviation Formula and Uses vs. Variance. *Investopedia* (2022, July 6). https://www.investopedia.com/terms/s/standarddeviation.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 S.M. Carroll, et al., The Cosmological Constant. *Ann. Rev. Astron. Astrophys.* **30**, 499 (1992). DOI 10.1146/annurev.aa.30.090192.002435

9 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

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

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

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

Supernovae Type Ia

Arctanh model of cosmology

The *arctanh* model describes a universe containing matter with geometry bent in space and time but without the cosmological constant (no dark energy). The relationship takes the form

arctanh model, distance vs. recession velocity

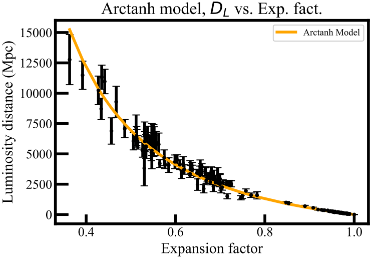
$$D_L = \frac{c}{H_0\xi\sqrt{|\Omega_k|}}\sinh\left\{\sqrt{|\Omega_k|}\int_{\xi_1}^{1}\frac{d\xi}{\xi\sqrt{\frac{\Omega_m}{\xi}+\Omega_k}}\right\}$$

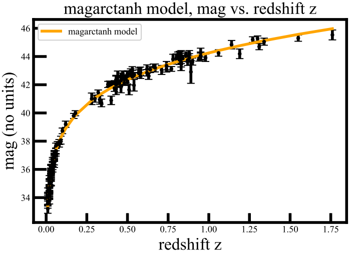
and this can be integrated to become

$$D_L = \frac{c}{H_0\xi\sqrt{|\Omega_k|}}\sinh\Bigg\{2\Bigg(\mathrm{arctanh}(\sqrt{|\Omega_k|}) - \mathrm{arctanh}\Big(\frac{\sqrt{|\Omega_k|}}{\sqrt{\frac{\Omega_m}{\xi}+\Omega_k}}\Big)\Bigg)\Bigg\}$$

Supernovae Type Ia

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





These figures are similar in appearance to the *standard* and ΛCDM model figures.

Predicting our present & future

The *arctanh* model success and *magarctanh* failure

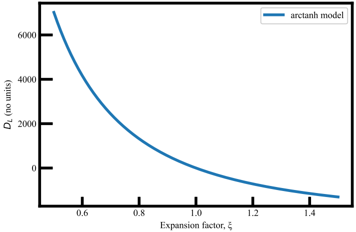
A strength of science is the ability to predict the future and describe the present. Calculations of the origin at $z = 0$ are impossible for the *mag*, log(data), models

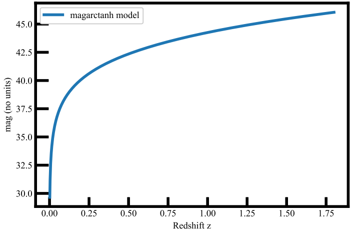
- The log of our position of 0 is $-\infty$
- but our position is known with 100% certainty

Our present and future conditions are presented in the next two figures. The failure of the *magarctanh* model is also observed for the *standard* model.

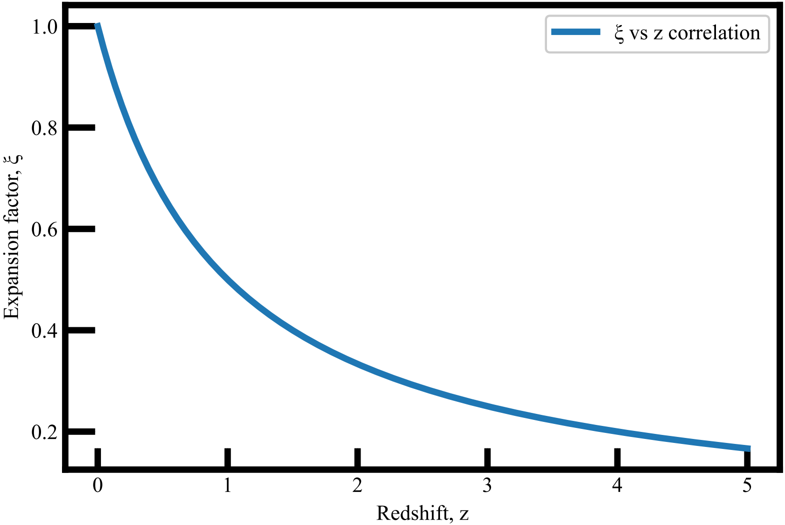
Plotting our present & future

The *arctanh* and *magarctanh* models





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



Trace of the expansion factor, ξ, vs redshift, z, relation.

Notes on terms

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

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