Internship project UMC Utrecht - Centre for Translational Immunology

Anaïs Fopma

Supervisor: Dr. Julia Drylewicz

Formatting: *Biostatistics* journal

Research master's programme: Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

Exploration of transformations for deuterium labelling data

Anaïs M. Fopma*

Methodology & Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht,the Netherlands.

a.m.fopma@uu.nl

Summary

Deuterium labelling is a technique commonly used in biomedical research to estimate the lifespan of cells in terms of, for example, their proliferation rate and their death rate. To transform the data, typically an arcsine squared transformation is used. However, there is no consensus on which transformation is the most appropriate for this type of data. This study compares different data transformations and investigates which are the most appropriate in terms of accuracy, precision, and certainty.

Key words: Deuterium labelling; Data transformations.

1. Introduction

Deuterium labelling is a technique commonly used in biomedical research to estimate the lifespan of cells in terms of, for example, their proliferation rate and their death rate. Estimating parameters like these has many applications in medicine. For example, deuterium labelling has been used in experiments investigating homeostasis of lymphocytes (Baliu-Piqué *and others* 2021), the self-renewal of T-cells in bone marrow and the periphery (Baliu-Piqué *and others* 2018), and

estimation of the average lifespan of leukocytes (Borghans *and others* 2018). Labelling with deuterium has many advantages, as it is a non-toxic and stable isotope and does not itself interfere with the kinetics of cells. The technique allows us to study these dynamics and the way in which they change or are disrupted by disease or interventions.

In deuterium-labelling studies, subjects are administered either deuterium-labelled glucose (D2-glucose) or deuterium-labelled water (D2O) for a given period of time (Borghans *and others* 2018). This period of time is the labelling phase of the experiment, and typically consists of a few hours or days in the case of D2-glucose, and several weeks in the case of D2O. The period after administration of D2-glucose or D2O is stopped, is called the de-labelling phase of the experiment. When cells divide, deuterium is incorporated into the newly synthesized DNA strands of the cells. As a consequence, the level of deuterium enrichment (which is the fraction of DNA that is labelled with deuterium) can be measured by means of gas chromatography and mass spectrometry (GC-MS) analysis of the DNA in the isolated cell population. Parameters such as production or death rates can then be derived from this information.

The administration of D2-glucose and D2O each require different approaches for estimation of the parameters. When D2O is administered, enrichment in the body fluids reaches its maximum and is washed out from the body at a rate much slower than D2-glucose (Borghans *and others* 2018). Therefore, DNA labelling will continue after the end of the labelling phase. For this reason, the availability of deuterium in the body fluids during the labelling and de-labelling phase of D2O experiments changes slowly and is typically described by a mathematical model fitted to the urine or plasma enrichment data. Glucose, however, turns over rapidly, and the ratio of D2-glucose over glucose in the serum can fluctuate depending on food consumption. Therefore, the availability of deuterium is often averaged when D2-glucose is used for labelling. This is done by estimating the area under the curve of deuterium enrichment in plasma (Macallan *and others* 2009). In this project, we are interested in fitting models to data obtained by D2O administration. With

these data, we can estimate the parameters using different approaches. Often least squares (LS) estimation or maximum likelihood (ML) estimation is used for this purpose. To transform the data, typically an arcsine squared transformation is used. However, there is no consensus on which transformation is the most appropriate for these type of data. Therefore, the research question to be answered is: What is the most appropriate transformation to apply to deuterium labelling data? This question is explored by comparing different data transformations with both the LS and the ML approach, and by investigating which transformations are the most appropriate in terms of accuracy, precision, and certainty.

2. Method

This study consists of simulation research and reports this according to the ADEMP structure proposed by Morris *and others* (2019), namely: Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures. The data simulated are those of a 9 week-labelling experiment, meaning that the end of deuterium administration is at $\tau = 63$. For simplicity it is assumed that *U(t)* is an on/off function which is equal to 1 between 0 and τ and equal to 0 otherwise. It is also assumed that the amplification factor $c = 1$. The data are based on the following measurement time points in number of days: 0, 7, 14, 28, 42, 63, 70, 84, 105, 126. The following model is considered:

For $t \leqslant \tau$, $L' = p - pL$.

For $t > \tau$, $L' = -pL$.

Noise was added to each simulated data set by drawing and adding 10 values from a normal distribution with a mean of 0 and a standard deviation of $p/10$ (i.e. 0.001 in the case of $p = 0.01$ and 0.00005 in the case of $p = 0.0005$).

2.1 *Aims*

The aim of the current project was to compare different transformations on deuterium labelling data, in order to identify which transformation is the most appropriate.

2.2 *Data-generating mechanisms*

We considered a full-factorial design for the simulation study with two variables. The first was the value of the production rate p. This is taken to be either 0.01 or 0.0005. The second was the transformation that was applied to the data. The data could be left as is, without transformation, or transformed with an arcsine square root transformation, a square root transformation, a log transformation, a quadratic root transformation, or an arcsine quadratic root transformation. This resulted in a total of twelve different data-generating mechanisms. The analyses were conducted using R version 4.1.1 (2021-08-10).R Core Team (2021) The input seed was 987.

2.3 *Estimands*

Our estimand was the production rate p.

2.4 *Methods*

Each simulated data set analysed with two methods:

1. Least squares estimation. This approach finds the best fit for the data by minimizing the sum of the squared residuals, that is:

$$SSR = \sum_{i=1}^{n} (y_i - f(p_i))^2,$$

   where y_i is the i^{th} observed value in the simulation, and p_i is the i^{th} estimated value.

2. Maximum likelihood estimation. This approach finds the best fit for the data by maxi-

mizing the likelihood function *L(p)* so that the data is most probable under the assumed model. Maximizing *L(p)* is equivalent to maximizing *log* L(p). *log L(p)* is defined as the log likelihood function, that is:

$$l(p) = logL(p) = L(p) = log\prod_{i=1}^{n} f(X_i|p) = \sum_{i=1}^{n} logf(X_i|p),$$

where f is

$$2\pi(exp(\sigma)^2)^{\frac{-n}{2}})exp\left(\frac{-SSR}{2(exp(\sigma)^2)}\right).$$

2.5 *Performance measures*

We assessed bias, given by:

$$bias = p - E(\hat{p}),$$

the mean squared error (MSE), given by:

$$MSE = (\frac{1}{n})\sum_{i=1}^{n}(\hat{p} - p)^2,$$

the relative standard error (RSE), given by:

$$RSE = \frac{\sigma}{\sqrt{n}},$$

and the confidence interval (CI) coverage, defined as the proportion of simulations in which the CI contains the true value of p:

$$coverage = P(CI_{lower} \leqslant \hat{p} \leqslant CI_{upper}).$$

For the calculation of the CI coverage, a bootstrap is necessary. The number of iterations for this bootstrap was chosen to be 200. Morris *and others* (2019) recommend to calculate the number of simulations necessary based on the Monte Carlo standard error. However, this was not feasible in the current project due to a lack of computational power and limited time. Therefore, the number of simulations for each data-generating mechanism for each method was set to 100.

3. Results

The estimates of the performance measures obtained with the LS approach are displayed in Table 1 and those obtained with the ML approach are displayed in Table 2. A visualisation of these estimates is given in Figure 1. We can derive several conclusions from the results.

First, we can take a look at each of the performance measures when $p = 0.01$. The bias is the lowest for ML when a square root transformation is used, and for LS when a quadratic root transformation is used. The MSE is the lowest for ML when a square root transformation is used, and for LS when a log transformation is used. The RSE is lowest for ML when an arcsine quadratic root transformation is applied, and for LS when a log transformation is applied. The coverage is the highest when a log transformation is used, both for ML and LS.

Second, we can see that the results slightly differ when $p = 0.0005$. Both bias and MSE are lowest for ML with an arcsine quadratic root transformation, and for LS with a log transformation. RSE is the lowest for ML in the case of an arcsine quadratic root transformation, and for LS in the case of a log transformation. Finally, coverage is the highest at 0.81 for ML with either no transformation or a quadratic root transformation. Again, it is the highest for LS when a log transformation is used.

4. Discussion

Of all 16 cases in which the performance of the methods was assessed, a square root transformation seems to perform best in 2 cases, a quadratic root transformation in 2 cases as well, an arcsine quadratic root transformation in 4 cases, and a log transformation in 8 cases. Overall, these 4 transformations seem to be the most appropriate transformations for the simulated data. The square root transformation leads to fairly precise and accurate results. The quadratic root transformation leads to reasonable results across all performance measures. Even though the log transformation performs best in most cases, it does not yield accurate, precise, or certain results

when ML estimation is used. Overall, the recommendation based on this small simulation study is to use an arcsine quadratic root transformation for ML estimation, and a log transformation for LS estimation.

It is important to bear in mind that a low number of bootstraps and simulations was used, and that problems with convergence were encountered while conducting the study. These problems were mainly solved by using different initial values, but in some cases, convergence may still not have been reached. An additional limitation is that ten time points were used, therefore we do not know if the results are also valid when fewer time points are available. Besides this, a homogeneous population was assumed, and the results may differ in the case of a heterogeneous population. A final important limitation is that the noise added to the simulated data was not varied. The noise can have a large impact on the results, and it would perhaps be even more influential in the case of low enrichment.

For these reasons, it is important that in potential future research, a larger number of simulations is used in order to be able to draw more valid conclusions on which transformations are most suitable. Additionally, it would be interesting to use fewer time points, different noise in the simulated data, and to investigate the results for a heterogeneous population as well.

5. Documentation

The documentation of this project, including the R code and instructions to reproduce the results, is available in the research repository at `https://github.com/anaisfopma/Research-Repository`.

References

Baliu-Piqué, Mariona, van Hoeven, Vera, Drylewicz, Julia, van der Wagen, Lotte E., Janssen, Anke, Otto, Sigrid A., van Zelm, Menno C., de Boer, Rob J., Kuball, Jürgen and Borghans, Jose AM. (2021). Cell-density independent increased lymphocyte production and loss rates post-autologous HSCT. *Elife* **10**, e59775. Publisher: eLife Sciences Publications Limited.

Baliu-Piqué, Mariona, Verheij, Myrddin W., Drylewicz, Julia, Ravesloot, Lars, De Boer, Rob J., Koets, Ad, Tesselaar, Kiki and Borghans, José AM. (2018). Short lifespans of memory T-cells in bone marrow, blood, and lymph nodes suggest that T-cell memory is maintained by continuous self-renewal of recirculating cells. *Frontiers in immunology* **9**, 2054. Publisher: Frontiers.

Borghans, José AM, Tesselaar, Kiki and de Boer, Rob J. (2018). Current best estimates for the average lifespans of mouse and human leukocytes: Reviewing two decades of deuterium-labeling experiments. *Immunological reviews* **285**(1), 233–248. Publisher: Wiley Online Library.

Macallan, Derek C., Asquith, Becca, Zhang, Yan, De Lara, Catherine, Ghattas, Hala, Defoiche, Julien and Beverley, Peter CL. (2009). Measurement of proliferation and disappearance of rapid turnover cell populations in human studies using deuterium-labeled glucose. *Nature protocols* **4**(9), 1313–1327. Publisher: Nature Publishing Group.

Morris, Tim P., White, Ian R. and Crowther, Michael J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine* **38**(11), 2074–2102. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8086.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Table 1. *Estimates of performance for the least squares approach.*

p	Transformation	Bias	MSE	RSE	CI Coverage
0.01	None	0.002306934	0.000532194	0.04749049	0.37
	asin(sqrt(x))	-0.000173604	3.01E-06	0.01836813	0.71
	sqrt(x)	-0.000304078	9.25E-06	0.01420013	0.8
	log(x)	3.09E-05	9.55E-08	0.001947107	0.95
	x$\hat{0}$.25	-8.22E-05	6.76E-07	0.006052567	0.91
	asin(x$\hat{0}$.25)	-0.000590515	3.49E-05	0.01567873	0.83
0.0005	None	-0.000314476	9.89E-06	0.1033912	0.76
	asin(sqrt(x))	-5.63E-05	3.17E-07	0.05137726	0.82
	sqrt(x)	-0.000115016	1.32E-06	0.04011087	0.69
	log(x)	7.39E-08	5.46E-13	0.000215442	0.99
	x$\hat{0}$.25	-4.75E-05	2.25E-07	0.03391412	0.88
	asin(x$\hat{0}$.25)	-5.19E-05	2.70E-07	0.04213078	0.88

Table 2. *Estimates of performance for the maximum likelihood approach.*

p	Transformation	Bias	MSE	RSE	CI Coverage
0.01	None	0.000367366	1.34958e-05	0.02008715	0.79
	asin(sqrt(x))	3.48E-05	1.208162e-07	0.01112538	0.83
	sqrt(x)	2.92E-05	8.535863e-08	0.007077125	0.87
	log(x)	0.003689039	0.001360901	0.03756869	0.98
	x$\hat{0}$.25	-0.0001148471	1.318986e-06	0.01166569	0.81
	asin(x$\hat{0}$.25)	-0.0001330969	1.771479e-06	0.006101403	0.92
0.0005	None	-0.0001271722	1.617276e-06	0.07902729	0.81
	asin(sqrt(x))	-2.161282e-05	4.67114e-08	0.02617524	0.74
	sqrt(x)	-3.930839e-05	1.54515e-07	0.03971332	0.73
	log(x)	-0.000969564	9.400543e-05	0.198259	0.46
	x$\hat{0}$.25	-4.763224e-05	2.26883e-07	0.03803349	0.81
	asin(x$\hat{0}$.25)	-4.845429e-06	2.347818e-09	0.00967009	0.80

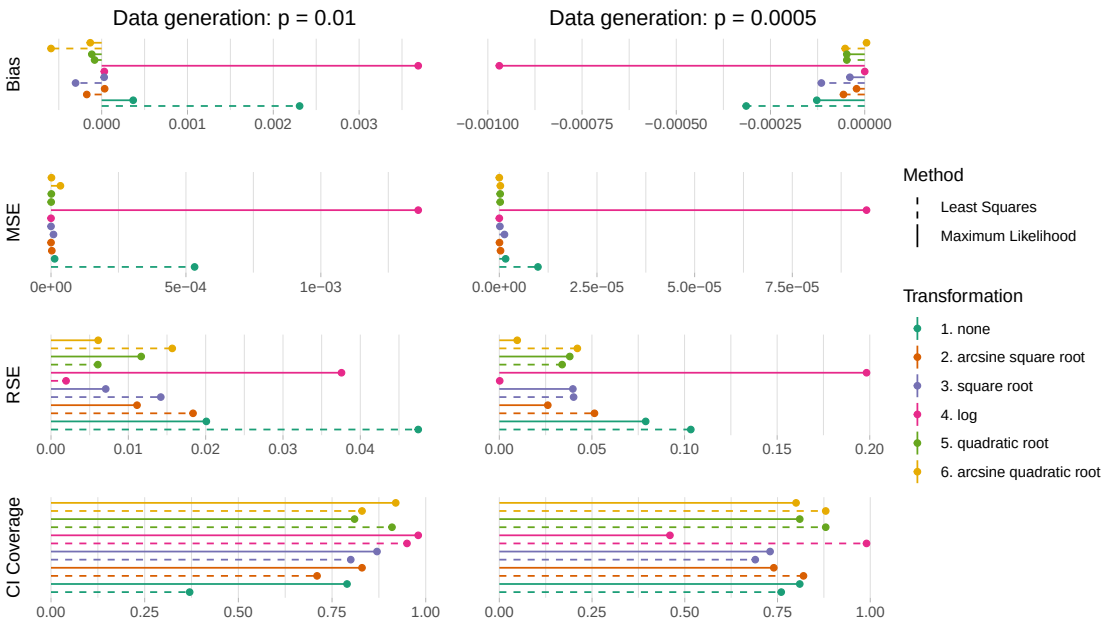

Fig. 1. Plot of the performance measures of interest.