A Comparative Study of Data Transformation Techniques

Muhammad Ibrahim
Dept of Computer Science & Software Engineering, NED University of Engineering and Technology
APCIC, Pakistan Council of Scientific and Industrial Research (PCSIR)
Khi, PAK
email: *m.ibrahim2094@gmail.com*

***Astract* ── This study paper reviews about compartaive study of data transformation techniques used prior to applying a linear regression model. It also sheds light on using alternate generalized linear models that would allow for flexibilty in terms of getting normalized error terms. The paper covers six types of data transformation techniques – Log Transformation, Exponential, Reciprocal, Box Cox, SquareRoot and Arcsine.**

Keywords* ── *Transformation, Normal Distribution, Linear Model, Generalized Linear models

I. INTRODUCTION

In statistics, data transformation is the application of a deterministic mathematical function each point in a data set — that is, each data point z_i is replaced with the transformed value $y = f(z_i)$, where f is a function.[1] The purpose of transforming data is to make the data follow assumptions of statiscal inference or undergo a parametrical statistical test or fit over a model. In our case in this paper if a data does not fit over a linear regression model we have two basic choices, to let go of the linear regression model and adopt a more suitable model or perform transformation on the data so that the linear regression model works for that transformed data. Many variables donot meet assumptions of linear relationship, normality, independence and homoscheadsaticsicty of error terms [7]; when used on linear model or in statistical tests, these may give misleading results. That is why data transformation prior to appling our model is very important in this case. The unit, the experimental design and the possible robustness of F statistics to 'small deviations' to Normal are among the main indicators for the choice of the type of transformation.[5] Transformations work on trial and error basis, where mostly results are checked after applying transformations and compared. It is also important to note that while some assumptions hold prior to transformation, they may not always hold once transformation is applied.

II. SYMBOL LITERACY

The following symbols are covered in equations or experessions used in this paper.

Symbol	**Definition**
Y	Response/ Dependent Variable
x	Predictor/ Independent Variable
bo , b1	Regression Parameters
P	Proportion

III. LOG TRANSFORMATION

The most popular and widely used data transformation is the Log Transformation. It is suitable for the data where the variance is proportional to square of the mean or the coefficient of variation is constant or where effects are multiplicative. Log transforation reduces skew and makes relationships more clear between variables. Log transformation discussed here is applied to fit a linear regression model. There are a few conditions under which log transformation may be applied that is when to apply on predictor variable(s) or response or both. Note that in most cases of logarithmic transformations, 'log' and natural log 'ln' are generally going to be treated as the same. Logarithms work with data exhibiting exponential growth or power curves.

A) Log Transforming the predictor variable(s)

Predictor variables are transformed when there is a lack of finding linearity. Note that all other assumptions of independence and normality of variance of error terms must hold true under this condition. Non-linearity must be the only problem. The scatter plot for one predictor and the residual

against fitted plot for multiple predictors are checked to discover linearity problems. Taking log of predictor variable(s) in our regression function transforms data to map a linear relationship.

$$y = b_0 + b_1 \log(x)$$

B) Log Transforming the response variable

Transforming the *y* values should be considered when non-normality and/or unequal variances are the problems with the model. As an added bonus, the transformation on *y* may also help to "straighten out" a curved relationship.[2] In some cases the R squared of the transformed model might be lower, but we would still consider the fact that the untransformed model does not satisfy equal variance condition and hence would use the transformed model. Taking log of the dependent variable in our regression function transforms data that gives equal variance of error terms.

$$\log(y) = b_0 + b_1 x$$

C) Log Transforming both the response and predictor variables

Transforming both the values are considered when nothing seems right. The relationship is non-linear and error terms are also unequal. Transforming the x values is intended to primarily fix the problem of non-linearity. Transforming the y values would fix unequal variances and also contribute to make the relationship linear. However, if non-linearity exists, we cannot use the model to check if error variances are equal. Taking log transformation of only the x values will also not result in making the relationship linear rather both x and y values would have to be transformed together to give a linear relationship and equal error variance under the following regression function

$$\log(y) = b_0 + b_1 \log(x)$$

D) Logit Transformation

The logit transformation is used in logistic regression and for fitting linear models to categorical data (log-linear models). A logit function is defined as the log of the odds function.

$$\text{logit}(p) = \log(p / 1 - p)$$

where p is the probability of event occurring (range 0 – 1). In other words, the logit of probability p is the log of odds function.

IV. BOX COX TRANSFORMATION

A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape. Box-Cox transformations are a family of power transformations on *Y* such that $Y = Y^{\lambda}$, where λ is a parameter to be determined using the data.[2] The Box Cox procedure uses the method of maximum likelihood to estimate 'λ' as well as other regression parameters. The exponent lambda (λ) varies from -5 to 5. All values of λ are considered and the optimal value for the data is selected; The optimal value is the one which results in the best approximation of a normal distribution curve of the error terms.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

For negative values,

$$y(\boldsymbol{\lambda}) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \text{if } \lambda_1 \neq 0; \\ \log(y+\lambda_2), & \text{if } \lambda_1 = 0. \end{cases}$$

Note that if λ is taken zero then actually a natural log transformation on the response variable is done as seen previously.

V. EXPONENTIAL TRANSFORMATION

The power transformation has its independent variable in its base, whereas an exponential transformation takes its independent variable in its exponent.

VI. RECIPROCAL TRANSFORMATION

Reciprocal transformation is taken for data expressing right skewness; it converts it to a normal distribution which can be assumed by any statistical methods. Reciprocal transformation maps non-zero values of x to 1/x (or -1/x for negative values) under the regression function such that

$$Y = b_0 + b_1 (1/x)$$

The reciprocal of a ratio may often be interpreted as easily as the ratio itself: e.g. population density (people per unit area) becomes area per person. The reciprocal reverses order among values of the same sign: largest becomes smallest, etc. The negative reciprocal preserves order among values of the same sign.

VII. SQUARE ROOT TRANSFORMATION

Square root transformations are applied on count data or small whole numbers and for other measures where group means are correlated with within group variances. The square root must be considered when variance is proportional to the mean. [6] This can also be used for percentage data where range is between 0% to 30% or 70% to 100%.[4] Count data is often encountered as Poisson Distribution and to stay within the linear framework square root transformation is applied. The other alternate is to a generalized linear model: Poisson Regression.

VIII. ARCSINE TRANSFORMATION

Arcsine or Angular transformation is generally used when data is in the form of percentages or proportions. In theory and in practice such data is not normally distributed, usually binomial. Arcsine does not transform values greater than 1, so even for percentages first they are converted to proportions. Note that percentages used in transformation must be of some count data. It gives results as nominal variables which should not be treated as measurement variables. The result is given in radians that range from $-\pi / 2$ to $\pi / 2$.

$$Y = \text{arcsine } \sqrt{p} = \sin{-1} \sqrt{p}$$

Square root and arcsine transformations are extensively used in biological research.

IX. CONCLUSION

Many researchers do not recommend data transformations arguing it causes problems in inferences and mischaracterizes data sets, which can hinder interpretation. There are other researchers who consider data transformation necessary to meet the assumptions of parametric models. But majority agrees to see data transformation as to improve experimental accuracy.

Most often after applying some mathematical operation and obtaining result you need to back transform your results for them to be interpretable. For ex. for log, raise 10 to the power of the number; for square root, square power the variable. It is important to note that the severity of consequences is related to the severity of violation. Considering the extent of violations reflects directly from how the model is intended to be used. If model is to be used for predictive purposes, then predictive intervals are very sensitive to deviations from normality. Whereas the same deviations from normality are forgiving in hypothesis test, confidence interval and in case of finding relationship between predictor and response. Nevertheless, all the tests and intervals are very sensitive to independence and moderately sensitive to departures from equal variances.

X. ACKNOWLEDGEMENT

This research review paper was written under Department of Applied Physics, Computer, and Instrumentation Centre (APCIC) at Pakistan Council for Scientific and Industrial Research (PCSIR Labs) as part of the data analytics internship program.

XI. REFERENCES

[1]. Handbook of Medical Statistics, Ji Qian Fang, Sun Yat-Sen University, China.

[2]. Stat 501, Eberly College of Science, Penn State University

[3]. www.statisticshowto.datasciencecentral.com

[4]. Rajender Parsad, I.A.S.R.I, Library Avenue, New Delhi, “Transformation of Data

[5]. João Paulo Ribeiro-Oliveira*, Denise Garcia de Santana, Vanderley José Pereira and Carlos Machado dos Santos Instituto de Ciências Agrárias, Universidade Federal de Uberlândia, Avenida João Naves de Ávila - Data transformation: an underestimated tool by inappropriate use.

[6]. Bartlett, M. S. (1936). The square root transformation in analysis of variance. Journal of Royal Statistical Society, 3(1), 68-78.

[7]. Kutner Nachtsheim Neter Li, Applied Linear Statistical Models 5th Edition, 127 -137.