Journal of Clinical Epidemiology 98 (2018) 146–151

Journal of Clinical Epidemiology

COMMENTARY

Linear regression and the normality assumption

Amand F. Schmidt[a,b,c,*], Chris Finan[a]

[a]Faculty of Population Health, Institute of Cardiovascular Science, University College London, London WC1E 6BT, United Kingdom
[b]Groningen Research Institute of Pharmacy, University of Groningen, Groningen, The Netherlands
[c]Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Utrecht, The Netherlands

Accepted 12 December 2017; Published online 16 December 2017

Abstract

Objectives: Researchers often perform arbitrary outcome transformations to fulfill the normality assumption of a linear regression model. This commentary explains and illustrates that in large data settings, such transformations are often unnecessary, and worse may bias model estimates.

Study Design and Setting: Linear regression assumptions are illustrated using simulated data and an empirical example on the relation between time since type 2 diabetes diagnosis and glycated hemoglobin levels. Simulation results were evaluated on coverage; i.e., the number of times the 95% confidence interval included the true slope coefficient.

Results: Although outcome transformations bias point estimates, violations of the normality assumption in linear regression analyses do not. The normality assumption is necessary to unbiasedly estimate standard errors, and hence confidence intervals and *P*-values. However, in large sample sizes (e.g., where the number of observations per variable is >10) violations of this normality assumption often do not noticeably impact results. Contrary to this, assumptions on, the parametric model, absence of extreme observations, homoscedasticity, and independency of the errors, remain influential even in large sample size settings.

Conclusion: Given that modern healthcare research typically includes thousands of subjects focusing on the normality assumption is often unnecessary, does not guarantee valid results, and worse may bias estimates due to the practice of outcome transformations.

Keywords: Epidemiological methods; Bias; Linear regression; Modeling assumptions; Statistical inference; Big data

1. Introduction

Linear regression models are often used to explore the relation between a continuous outcome and independent variables; note however that binary outcomes may also be used [1,2]. To fulfill "the" normality assumption, researchers frequently perform arbitrary outcome transformation. For example, using information on more than 100,000 subjects, Tyrrell et al. 2016 [3] explored the relationship between height and deprivation using a rank-based inverse normal transformation and Eppinga et al. 2017 [4] who explored the effects of metformin on the square root of 233 metabolites.

In this paper, we argue that outcome transformations change the target estimate and hence bias results. Second, the relevance of the normality assumption is challenged; namely, that non-normally distributed residuals do not impact bias, nor do they (markedly) impact tests in large sample sizes. Instead of focusing on the normality assumption, more consideration should be given to the detection of trends between the residuals and the independent variables; multivariable outlying outcome or predictor values; and general errors in the parametric model. Unlike violations of the normality assumption, these issues impact results irrespective of sample size. As an illustrative example, the association between years since type 2 diabetes mellitus (T2DM) diagnosis and glycated hemoglobin (HbA_{1c}) levels is considered [5].

2. Bias due to outcome transformations

First, let us define a linear model and which part of the model the normality assumption pertains to:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad [1]$$

Conflict of interest statement: The authors of this paper do not have a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the paper.

Funding: A.F.S. is funded by University College London (UCL) Hospitals National Institute for Health Research Biomedical Research Center and is an UCL Springboard Population Health Sciences Fellow. The funders did not in any way influence this manuscript.

* Corresponding author. Tel.: 0044 (0)20 3549 5625.
E-mail address: amand.schmidt@ucl.ac.uk (A.F. Schmidt).

https://doi.org/10.1016/j.jclinepi.2017.12.006

What is new?

Key findings

- To ensure that the residuals from a linear regression model follow a normal distribution, researchers often perform arbitrary outcome transformations (here arbitrary should be interpreted as using an unspecified function). These transformations also change the target estimate (the estimand) and hence bias point estimates. Unless these transformations are distributive (in the mathematical sense), inverse-transforming model parameters does not necessarily decrease bias.

What this adds to what was known?

- Linear regression models with residuals deviating from a normal distribution often still produce valid results (without performing arbitrary outcome transformations), especially in large sample size settings.
- Conversely, linear regression models with normally distributed residuals are not necessarily valid. Graphical tests are described to evaluate the following assumptions: the appropriateness of the parametric model, absence of extreme observations, homoscedasticity, and independency of errors.

What is the implication and what should change now?

- Linear regression models are often robust to assumption violations, and as such logical starting points for many analyses. In the absence of clear prior knowledge, analysts should perform model diagnoses with the intent to detect gross assumption violations, not to optimize fit. Basing model assumption solely on the data under consideration may do more harm than good. A prime example of this is the pervasive use of bias-inducing outcome transformations.

Here, y is the (continuous) outcome variable (e.g., HbA_{1c}), x is an independent variable (e.g., years since T2DM diagnosis), parameter β_0 is the $\bar{y}$ value when $x = 0$ (e.g., the intercept term representing the average HbA_{1c} at time of diagnosis), and ε represents the errors which is also the only part assumed to follow a normal distribution. Often one is interested in estimating β_1 (e.g., the slope), in this example, the amount HbA_{1c} changes each year, and the residuals $\hat{\varepsilon}$ (the observed errors) are a nuisance parameter of little interest. Note that $\hat{\beta}$ notation represents an estimate of a population quantity such as β, and similarly, $\bar{y}$ represents an estimate of the population mean HbA_{1c} concentration.

Throughout this manuscript, it is assumed that y is measured on a scale of clinical interest, for example HbA_{1c} as a percentage, or lipids in mmol/L or mg/dL. In these cases, transforming the outcome to ensure that the residuals better approximate a normal distribution often results in a biased estimate of β_1. To see this let us define $g(\cdot)$ as an arbitrary function used to transform the outcome resulting in an effect estimate $\beta_{1,t} = g(y_{x+1}) - g(y_x)$, with $x + 1$ indicating a unit increase from x to $x + 1$ and index t for "transformed". Clearly $\beta_{1,t}$ cannot equal β_1 unless the transformation pertains simple addition $g(y) = y + c$ (with c a constant), hence $\hat{\beta}_{1,t}$ is a biased estimate of β_1 in the sense that $\overline{\beta}_{1,t} \neq \beta_1$.

Often one tries to reverse such transformations by applying $g^{-1}(\cdot)$ on $\beta_{1,t}$. Such back transformations can only equal β_1 when the function $g(\cdot)$ is "distributive" $\beta_{1,t} = g(y_{x+1}) - g(y_x) = g(y_{x+1} - y_x)$; where we assume $g(y) \neq y + c$ in which case $\beta_{1,t} = \beta_1$. However, functions most often used for outcome transformations do not have this distributive property, and hence the "back-transformed" $g^{-1}(\beta_{1,t})$ will not equal β_1. To provide a numerical example let's look to the logarithmic transformation $\log_{10}10 - \log_{10}100 \neq \log_{10}(10 - 100)$, and the square root transformation $\sqrt{10} - \sqrt{100} \neq \sqrt{10 - 100}$.

Readers should note that this bias pertains only to transformation where the original measurement scale has clinical relevance (and is not regularly presented on the transformed scale), and not to the general use of the logarithmic scale (or any other mathematical functions) as an outcome. For example, the acidity of a solution is typically indicated by the pH (potential of hydrogen), which is best understood on the logarithmic scale. Similarly, this type of bias is only relevant if one is interested in interpreting $\hat{\beta}_1$. For example, if one is concerned with prognostication, outcome transformations are less of an issue. Furthermore, hypothesis tests from linear regression models using arbitrary-transformed outcomes are still valid. However, when using linear regression models, we assume researchers are interested in estimating the magnitude of an association. If, instead, a researcher is only interested in testing a (null-) hypothesis, nonparametric methods will often be more appropriate.

3. The normality assumption in large sample size settings

We define large sample size as a setting where the n observations are larger than the number of p parameters one is interested in estimating. As a pragmatic indication, we use $n/p > 10$, but realize that this will differ from application to application.

To discuss the relevance of the normality assumption, we look to the Gauss–Markov theorem [6], which states that the ideal linear regression estimates are both unbiased and have the least amount of variance, a property called the