Linear regression and the normality assumption

A F Schmidt* [a] and Chris Finan [a]

a. Institute of Cardiovascular Science, Faculty of Population Health, University College London, London WC1E 6BT, United Kingdom.

***** Contact: 0044 (0)20 3549 5625

E-mail address: amand.schmidt@ucl.ac.uk (A.F.Schmidt)

Word count abstract: 210

Word count text: 2017

Number of references: 13

Number of tables: 0

Number of figures: 3

Abstract

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

Design Linear regression assumptions are illustrated using simulated data and an empirical example on the relation between time since type 2 diabetes diagnosis and glycated haemoglobin (HbA_{1c}). Simulation results were evaluated on coverage; e.g., the number of times the 95% confidence interval included the true slope coefficient.

Results While outcome transformations bias point estimates, violations of the normality assumption in linear regression analyses do not. Instead this 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 larger than 10) violations of this normality assumption 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.

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

Keywords Epidemiological methods; Bias; Linear regression; Assumptions

What is new?

- To ensure 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 transformation function). These transformations also change the target estimate (the estimand) and hence bias point estimates. Unless these transformations are distributive (in the mathematical sense) in nature inverse transforming model parameters does not necessarily decrease bias.
- Linear regression models with residuals deviating from the normal distribution often still produce valid results (without performing arbitrary outcome transformations), especially in large sample size settings (e.g., when there are 10 observations per parameter).
- Conversely, linear regression models with normally distributed residuals are not necessarily valid. Graphical tests are described to evaluate the following modelling assumptions on: the parametric model, absence of extreme observations, homoscedasticity and independency of errors.
- 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 will typically do more harm than good, a prime example of this is the pervasive use of, bias inducing, ‘arbitrarily’ outcome transformations.

Introduction

Linear regression models are often used to explore the relation between a continuous outcome and independent variables; note that binary outcomes may also be used [1,2]. To fulfil "the" normality assumption researchers frequently perform arbitrary outcome transformation. For example, using information on more than 100,000 subjects Tyrrel *et al* 2016[3] explored the relation between height and deprivation using a rank-based inverse normal transformation, or Eppinga *et al* 2017[4] who explored the effect 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 focussing on the normality assumption, more consideration should be given to the detection of 1) trends between the residuals and the independent variables, 2) multivariable outlying outcome or predictor values, and 3) 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 HbA_{1c} (outcome) is considered [5].

Bias due to outcome transformations

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

$$y = \beta_0 + \beta_1 x + \epsilon \ \ [eq\ 1].$$

Here y is the continuous outcome variable (e.g., HbA_{1c}) x an independent variable (e.g., years since T2DM diagnosis), parameter β_0 the $\bar{y}$ value when $x = 0$ (e.g., the intercept term representing the *average* HbA_{1c} at time of diagnosis), and ϵ the errors which are 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 $\hat{\epsilon}$ (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) average HbA_{1c}.

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 the residuals better approximate a normal distribution often results in a biased estimate of β_1. To see this let's define $g(\cdot)$ as an arbitrary function used to transform the outcome resulting in an effect estimate $\beta_{1,t} = g(y_x) + g(y_{x+1})$, 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 $g(x) = x + c$ (with c a constant), hence $\hat{\beta}_{1,t}$ is a biased estimate of β_1 in the sense that $\bar{\beta}_{1,t} \neq \beta_1$.

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" $\beta_{1,t} = g(y_x) + g(y_{x+1}) = g(y_x + y_{x+1})$; where we assume $g(x) \neq x + c$ 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" effect estimate $g^{-1}(\beta_{1,t})$ will not equal β_1. Take for example a logarithmic transformation $\log_{10} 10 + \log_{10} 100 \neq \log_{10}(10 + 100)$ or the square root transformation $\sqrt{10} + \sqrt{100} \neq \sqrt{10 + 100}$.

Readers should note that this bias pertains only to *arbitrary* transformation where the original measurement scale has clinical relevance (and is not normally represented 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 in so far one is interested in interpreting $\hat{\beta}_1$, if for example 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, as stated before, in using linear regression models we assume researchers are interesting in *estimating* the magnitude of an association. If, instead, a researcher is interested in testing a (null-) hypothesis non-parametric methods will often be more appropriate.

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 $\frac{p}{n} > 10$, but realize that this may likely 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 “best linear unbiased estimators” (BLUE). Linear regression estimates are BLUE when the errors have mean zero, are uncorrelated and have equal variance across different values of the independent variables (i.e., homoscedasticity)[6]. The normality assumption is thus not necessary to get estimates with the BLUE property. However, in small sample size settings (relative to p) the standard error estimates may be biased (and hence confidence intervals and p-values as well) when the errors do not follow a normal distribution. For formal proofs of the BLUE characteristics please see the historically relevant Aitken, 1936 [6], and chapter 2 of Faraway, 2015 [7]

To empirically assess the relevance of the normality assumption we performed an illustrative simulation using 4 scenarios with a single independent variable and an error distribution, following either: 1) the standard normal distribution, 2) a uniform distribution, 3) a beta distribution, 4) a normal distribution where the errors depend on x (i.e., heteroscedasticity). Figure 1 depicts a sample of 1,000 subjects from each of the 4 scenarios, the top row shows the outcome distribution, the middle figures depicts quantile-quantile (QQ) plots exploring how well the model residuals follow the normal distribution (diagonal line of perfect fit); showing clear deviations in scenarios 2 and 3. With the bottom row revealing a trend between the residuals and the fitted values; with a clear relationship being observed in scenario 4; note the fitted values are defined by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, or informally, $\text{outcome} = \text{fitted values} + \text{residuals}$.

Based on these scenarios $3, 10, 100, 1000, 10\,000$ and $100\,000$ subjects were sampled (repeated 10 000 times) and the linear model of equation 1 was fitted to the data. Given that in these settings point estimates will be unbiased on average $(\bar{\beta}_1 = \beta_1)$, we evaluated performance on the number of times the 95% confidence interval included β_1 (i.e., **coverage**). Figure 2 shows that despite the errors not following a normal distribution, in scenario 2-3 coverage is ~0.95 in larger sample sizes. However, in scenario 4 despite the residuals more closely following a normal distribution coverage in large sample sizes is consistently lower than the nominal 0.95 level. Moreover, as the sample size increased coverage did not improve.

Model diagnostics

As the above illustrates, linear models without normally distributed residuals may nevertheless produce valid results, especially given sufficient sample size. Conversely, the following modelling assumptions are sample size invariant and should be carefully checked regardless of

the size of the collected data: miss-specification of the parametric model, presence of extreme observations, homoscedasticity and independency of errors.

An example of model **miss-specification** would be if the linear model of equation 1 was used, when in reality the association was curved. To detect such a model miss-specification one can compare the residuals to the fitted values, for example figure 3 shows the residuals plotted against the fitted values from the model association time since T2DM diagnosis to HbA_{1c} level. The slope becomes negative at about 9.5 years since diagnosis. A different example of miss-specification would be if unknown to the analysist the association differed between males and females (interaction). While interaction or non-linearity are often cited forms of model miss-specification, as we discuss next, other assumption violations *may* be indicative of miss-specification as well.

In (multivariable) linear regression an **outlier** is defined as an observed outcome value y_i that is far away from the predicted outcome value $\hat{y}_i$. Outliers can influence model parameters, and are therefore important to detect, for example by comparing the fitted values to the Studentized residuals (see Appendix page 16). Similar to outliers, unusual x values may be over-influential as well. Such observations are said to have high **leverage** and can be detected using the leverage statistic (as shown in the Appendix page 18). Removal of observations with high leverage and/or outlying outcome values may seem like a logical decision, however applying this as a general rule will often severely bias a model. Outlying values may of course indicate errors, however these errors may pertain to the model not necessarily to the data. Similarly, observations with high leverage may point to data issues, however it may also be indicative of interesting subgroups.

Correlated errors often arise in time series, for example when modelling the association between mortality and temperature the previous day(s) temperature is influential as well. More generally, correlated errors occur when clustering in the data is ignored. As a hypothetical example, subjects in our HbA_{1c} dataset may have been related, if ignored such clustering will artificially decrease the standard errors and may even bias point estimates. **Heteroscedastic** occurs when the variance of the residuals depends of the predicted value (see Figure 1: row 4, column 1). Similar, to the omission of a cluster indicator, heteroscedasticity may be indicative of an omitted interaction term affecting the variance instead of the mean. Given that interactions are scale dependent [8] arbitrary outcome transformation are often applied here as well, however, as discussed this may bias results. Instead, in the presence of heteroscedasticity or correlated errors, a relatively straightforward solution is to replace the erroneously attenuated standard errors by larger heteroscedastic robust standard errors [9] (see Appendix).

As an example, in the Appendix we have applied the above discussed modelling diagnostics on the HbA_{1c} data. Based on these steps we come to the conclusion that conditional on the covariates, age, marital status, and body mass index (BMI), time since T2DM diagnosis has a non-linear relation with HbA_{1c}; where its level initially increases, only to decrease around 9.5 years after T2DM diagnosis.

Discussion and recommendations

In this brief outline of much larger theoretical works [6,10] we show that given sufficient sample size, linear regression models without normally distributed errors are valid. Despite this well-known characteristic, arbitrarily outcome transformations are often applied in an attempt to force the residuals to follow a normal distribution. As discussed such transformation frequently bias slope coefficients (as well as standard errors) and should be discouraged. What constitutes large sample size obviously differs between analyses, before we mentioned a ratio of 10

observations per parameter, however lower values have been found sufficient as well [11]. Conversely, larger values (e.g., 50) may be necessary when variables are correlated or variable distributions result in localized (multivariate) sparse data settings. As such in no way should this manuscript be misconstrued into arguing that linear regression should always be used, and especially not without critical reflection of modelling assumptions. Instead we simply wish to make the point that the linear model often performs adequately, even when some assumptions are violated. This robust behaviour of linear regression can be extended in many ways, for example generalized least square can be used in the presence of correlated errors, weighted least squares in the presence of heteroscedasticity, or RIDGE and LASSO regression in the presence of sparse data (e.g., $\frac{n}{p} \leq 1$). All these methods are in essence still linear models making a thorough understanding of the underlying modelling assumptions, as presented here, crucial.

Ideally, model decisions should be based on prior, topic specific, knowledge. If such external information is absent graphical tests (as presented here) should be used to detect grossly wrong assumption, not to optimize fit, which likely biases results far beyond any assumption violation[12,13].

In conclusion, in large sample size settings linear regression models are fairly robust to violations of the normality assumption and hence arbitrary - bias inducing - outcome transformations are usually unnecessary. Instead, researchers should focus on detection of model miss-specifications such as outlying values, high leverage, heteroscedasticity, correlated errors, non-linearity, and interactions which may bias results irrespective of sample size.

Conflict of interest statement

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

Author contribution

AFS and CF contributed to the idea, design, and analyses of the study and drafted the manuscript.

Guarantor

AFS had full access to all of the data and takes responsibility for the integrity of the data presented.

Funding

AFS is funded by UCLH NIHR Biomedical Research Centre and is a UCL Springboard Population Health Sciences Fellow. The funders did not in any way influence this manuscript.

References

[1] Schmidt AF, Groenwold RHH, Knol MJ, Hoes AW, Nielen M, Roes KCB, et al. Exploring interaction effects in small samples increases rates of false-positive and false-negative findings: Results from a systematic review and simulation study. J Clin Epidemiol 2014;67:821–9.

[2] Austin PC, Laupacis A. A Tutorial on Methods to Estimating Clinically and Policy-Meaningful Measures of Treatment Effects in Prospective Observational Studies: A Review. Int J Biostat 2011;7:1–32.

[3] Tyrrell J, Jones SE, Beaumont R, Astley CM, Lovell R, Yaghootkar H, et al. Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank. BMJ 2016;352:i582.

[4] Eppinga RN, Kofink D, Dullaart RPF, Dalmeijer GW, Lipsic E, Van Veldhuisen DJ, et al. Effect of Metformin on Metabolites and Relation with Myocardial Infarct Size and Left Ventricular Ejection Fraction after Myocardial Infarction. Circ Cardiovasc Genet 2017;10.

[5] Shu PS, Chan YM, Huang SL. Higher body mass index and lower intake of dairy products predict poor glycaemic control among Type 2 Diabetes patients in Malaysia. PLoS One 2017;12.

[6] Aitken AC. IV.—On Least Squares and Linear Combination of Observations. Proc R Soc Edinburgh 1936;55:42–8.

[7] Faraway JJ. Linear Models with R. 2015.

[8] Schmidt AF, Klungel OH, Nielen M, de Boer A, Groenwold RHH, Hoes AW. Tailoring treatments using treatment effect modification. Pharmacoepidemiol Drug Saf 2016;25:355–62.

[9] Zeileis A. Object-oriented Computation of Sandwich Estimators. J Stat Softw 2006;16:1–16.

[10] White H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test

for Heteroskedasticity. Econometrica n.d.;48:817–38.

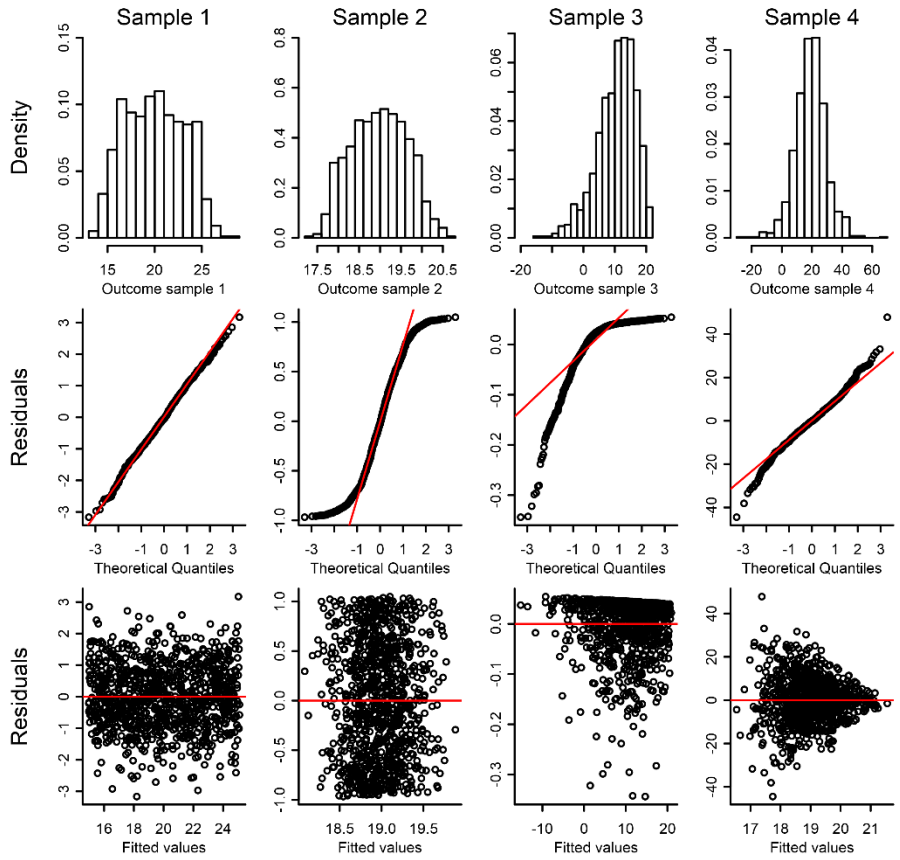
[11] Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. J Clin Epidemiol 2015;68:627–36.

[12] James G, Witten D, Hastie T, Tibishirani R. An Introduction to Statistical Learning. 2013.

[13] Chatfield C. Model Uncertainty, data mining and statistical inference. J R Stat Soc A 1995;158:419–66.

Figure captions

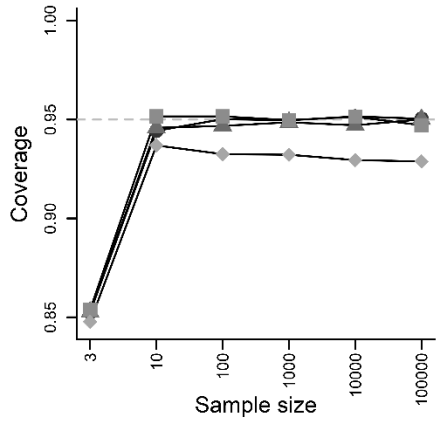
Figure 1 Graphically exploring the normality of the outcome (row 1), normality of the residuals (row 2), and potential trends between the residuals and the fitted values (row 3) for 4 different linear regression scenarios.



N.b. The columns represent a 1,000 subjects sampled from 4 scenarios: normally distributed errors $\varepsilon \sim N(0,1)$ (column 1), uniformly distributed errors$\varepsilon \sim U(-1,1)$ (column 2), skewed beta distributed errors $\varepsilon \sim B(10,0.05)$ (column 3), and heteroscedastic but normally distributed errors $\varepsilon \sim x_i N(0,1)$ (column 4). Top row contains histograms of the outcome. The middle row contains QQ plots comparing the observed model residuals to the expected residuals from the normal distribution with the red diagonal line indicating perfect fit. The bottom panel compared the residuals to the fitted values In all scenarios the outcome was generated based on $y_i = 20 +$

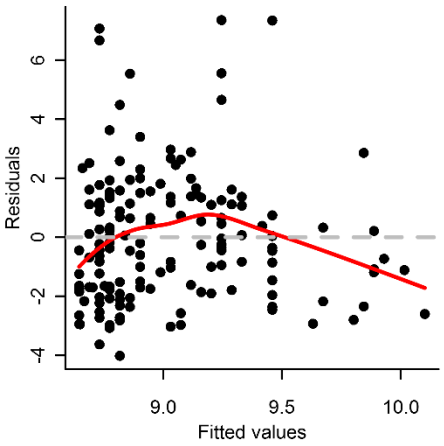
$\beta_1 x_i + \varepsilon$. In scenarios 2 and 4 x_i was (arbitrarily) generated based on $N(10,3)$, $U(-50,50)$ in scenario 1, and the square of $N(10,3)$ in scenario 3

Figure 2 The impact of sample size on coverage of linear regression model parameters with differently distributed errors.



n.b. results from scenario 1-3 are depicted by a circle, a triangle or a square, respectively. Scenario 4, where the normally distributed errors depend on the predictor variable, is depicted by a diamond.

Figure 3 A residual plot of the linear regression model regressing HbA_{1c} on years since type 2 diabetes diagnosis.



N.b. the red curve represents a LOESS (a generalization of the locally weighted scatterplot smoother) curve.