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

"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 unnecessarily to get estimates with the BLUE property. However, in small–sample size settings, 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 study by Aitken [6] and chapter 2 of Faraway [7].

To empirically assess the relevance of the normality assumption, we performed an illustrative simulation using four scenarios with a single independent variable and an error distribution, following either: 1) the standard normal

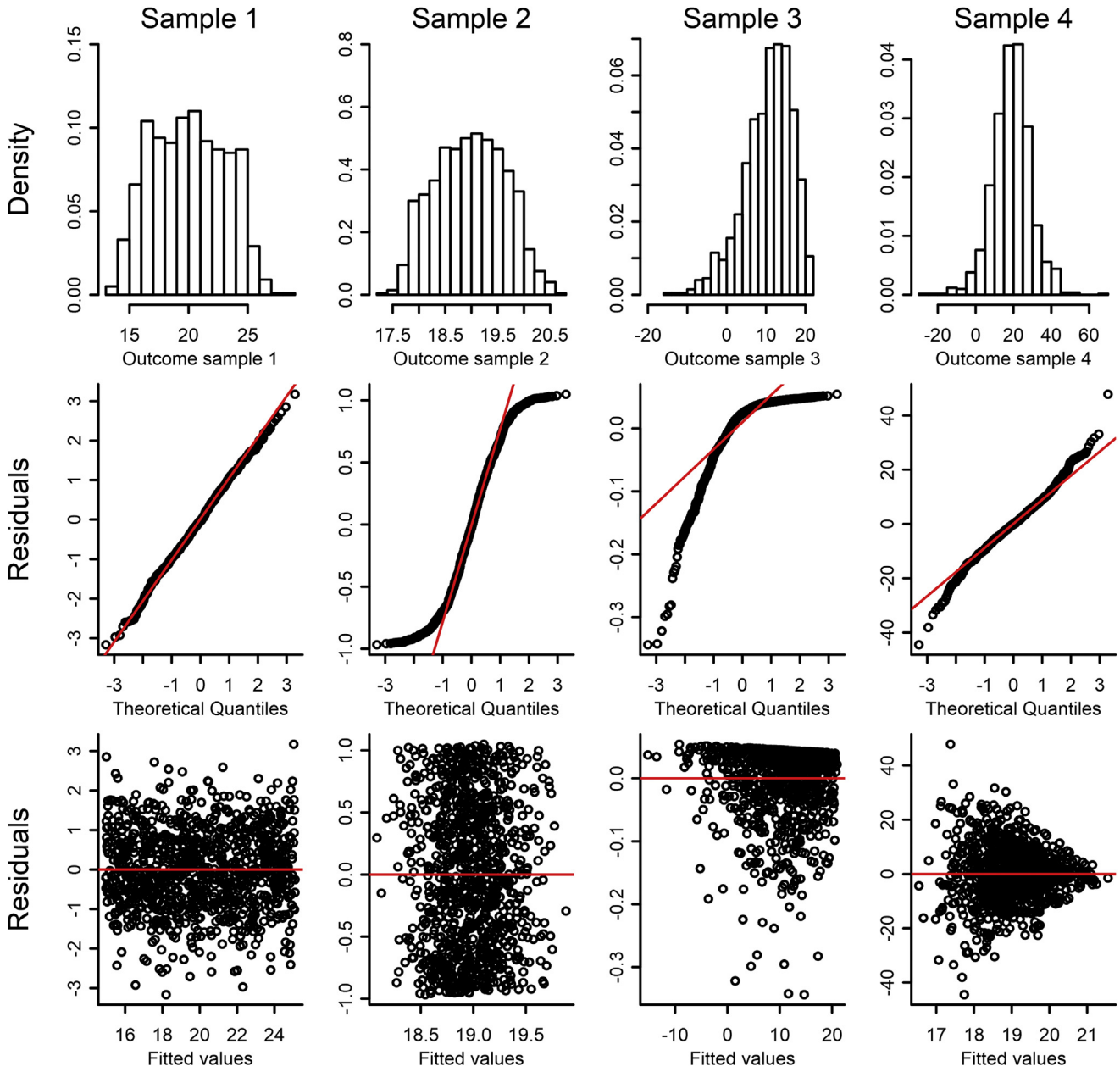

Fig. 1. Graphical tests to explore normality of the outcome (row 1), normality of the residuals (row 2), and potential trends between the residuals and the fitted values (row 3) for four different linear regression scenarios. N.b. The columns represent a 1,000 subjects sampled from four scenarios: normally 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 compares the observed model residuals to the expected residuals from the standard normal distribution with the diagonal line indicating perfect fit. The bottom panel compares 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 $N(10,3^2)$, and on $U(-50,50)$ in scenario 1, and the square of $N(10,3^2)$ in scenario 3.

distribution, 2) a uniform distribution, 3) a beta distribution, and 4) a normal distribution where the errors depend on x (i.e., heteroscedasticity). Fig. 1 depicts a sample of 1,000 subjects from each of the four scenarios, the top row shows the outcome distribution, the middle figures compares the residuals to the standard normal quantiles, 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 observed in scenario 4; note fitted values are defined as $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ or informally, outcome = fitted values + 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 $(\overline{\beta}_1 = \beta_1)$, we evaluate performance on the number of times the 95% confidence interval included β_1 (i.e., coverage). Fig. 2 shows that despite the errors not following a normal distribution in scenarios 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 lower than the nominal 0.95 level. Moreover, as the sample size increased, coverage does not improve.

4. Model diagnostics

As shown linear models without normally distributed residuals may nevertheless produce valid results, especially given sufficient sample size. Conversely, the following modeling assumptions are sample size invariant and should be carefully checked regardless of the size of the collected data: correct specification of the parametric model, abscence 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 miss-specification, one can compare the residuals to the fitted values. For example, Fig. 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 9.2 years since diagnosis. A different example of miss-specification would be if unknown to the analyst the association differed between males and females (interaction). While interaction or nonlinearity are often cited forms of model miss-specification, as we discuss further, 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 $\widehat{y}_i$. 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 logical; however, applying this as a general rule could severely bias a model. Outlying values may of course indicate errors;

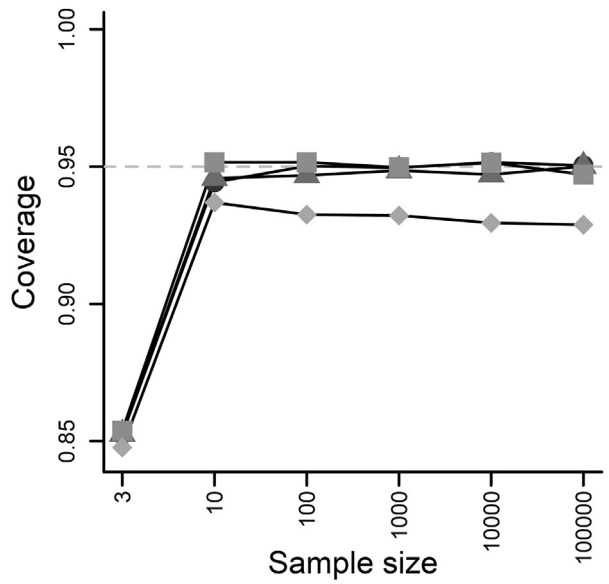

Fig. 2. The impact of sample size on coverage of linear regression model parameters with differently distributed errors. N.b. results from scenario 1 to 3 are depicted by a circle, a triangle, or a square, respectively. Scenario 4 is depicted by a diamond.

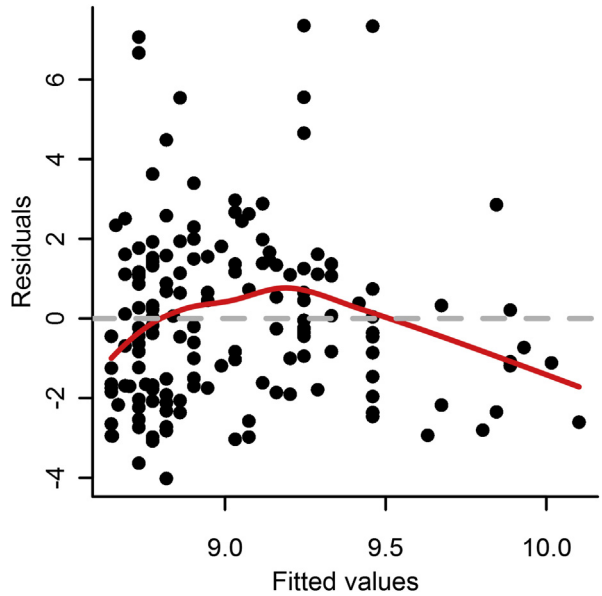

Fig. 3. A residual plot of a linear regression model regressing HbA_{1c} on years since type 2 diabetes diagnosis. N.b. the solid line represents a LOESS (a generalization of the locally weighted scatterplot smoother) curve. HbA_{1c}, glycated hemoglobin.

however, these errors may pertain to the model but not necessarily to the data. Similarly, observations with high leverage may point to data issues; however, they may also be indicative of interesting subgroups.

Correlated errors often arise in time series, for example, when modeling the association between mortality and temperature, the previous day's temperature is influential as well. More generally, correlated errors occur when clusters in the data are ignored. As a hypothetical example, subjects in our HbA_{1c} data set may have been related, if ignored, such clustering will artificially decrease the standard errors. Heteroscedasticity occurs when the variance of the residuals depends on the predicted values (see Fig. 1: row 3, column 4). 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], outcome transformation are often applied here as well. Instead of relying on outcome transformations 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 modeling 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 time since T2DM diagnosis has a nonlinear relation with HbA_{1c}; where, its level initially increases, only to decrease around 9.2 years after T2DM diagnosis.

5. 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 typically 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 on modeling assumptions. Instead, we wish to make the point that the linear model often performs adequately even when some assumptions are violated. This robust behavior 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 "least absolute shrinkage and selection operator" (lasso) regressions in the presence of sparse data (e.g., $n/p \leq 1$). All these methods are in essence still linear models, making a thorough understanding of the underlying modeling 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 may bias results far beyond any violated assumption [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, nonlinearity, and interactions, which may bias results irrespective of sample size.

Acknowledgments

Author contribution: A.F.S. and C.F. contributed to the idea, design, and analyses of the study and drafted the manuscript. Guarantor: A.F.S. had full access to all of the data and takes responsibility for the integrity of the data presented.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2017.12.006.

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. pii: e001564.

[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:e0172231.

[6] Aitken AC. IV.—on least squares and linear combination of observations. Proc R Soc Edinb 1936;55:42−8.

[7] Faraway JJ. Linear Models with R. Florida: Chapman & Hall/CRC; 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 HA. Heteroskedasticity-consistent covariance matrix estimator and a direct test for Heteroskedasticity. Econometrica 1980;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. New York: Springer; 2013.
[13] Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc A 1995;158:419−66.