CAUTIONARY NOTE ON THE TWO-STEP TRANSFORMATION TO NORMALITY

Accepted for publication in *Journal of Information Systems*

DOI: 10.2308/isys-52255

Mikko Rönkkö
University of Jyväskylä
Department of Computer Science and Information Systems
P.O. Box 35
FI-40014 Jyväskylä FINLAND
phone: + 358 40 805 3177
email: mikko.ronkko@jyu.fi

Miguel I. Aguirre-Urreta
Florida International University
College of Business
11200 S.W. 8th St., RB258A
Miami, FL 33199, USA
email: miguel.aguirreurreta@fiu.edu

CAUTIONARY NOTE ON THE TWO-STEP TRANSFORMATION TO NORMALITY

ABSTRACT

Templeton and Burney (2017) proposed a two-step normality transformation as a remedy for non-normally distributed data, which is commonly found in AIS research. We argue that, rather than transforming the data towards normality, researchers should first seek to analyze and understand the sources of non-normality. Using simulated datasets, we demonstrate three sources of non-normality and their consequences for regression estimation. We then demonstrate that the two-step transformation cannot solve any of these problems and that each source of non-normality can be handled with alternative, existing techniques. We further present two empirical examples to demonstrate these issues with real datasets.

Keywords: normal distribution, regression analysis, two-step transformation

INTRODUCTION

In a recent publication, Templeton and Burney (2017) proposed the use of a two-step normality transformation, first introduced to information systems (IS) research by Templeton (2011), as a remedy for non-normally distributed data, which are common in accounting information systems (AIS) research. The technique belongs to the class of rank-based inverse normal transformations, whose first introduction to the literature is often attributed to Fisher and Yates (1938). A number of alternative techniques were also presented in the 1950's and 1960's (Bliss, 1967; Blom, 1958; Tukey, 1962; Van der Waerden, 1952). This type of transformations are commonly called *normal scores* in statistical software such as SPSS[1]. In the variant proposed by Templeton and Burney (2017), the data are first transformed into percentile ranks and these percentile ranks are then converted to normally distributed variables by applying the inverse of the cumulative normal probability distribution. Unfortunately, the technique has several problems that go unaddressed by Templeton and Burney (2017) and which we discuss in more detail below.

In their article, Templeton and Burney (2017) focused exclusively on comparing the outcomes of performing analyses on original and transformed variables, and whether the results using transformed data would be more preferable for researchers that seek to publish their work. For example, the two-step transformation may increase the correlation between two variables, and this is seen as a positive outcome by the authors. More precisely, Templeton and Burney (2017) concluded that "Generally, researchers can expect the Two-Step to increase effect sizes and correspondingly, probabilities of significant findings" (p. 17). We see this rationale for the

[1] In SPSS normal scores can be calculated by choosing Transform > Rank Cases and choosing "Normal scores" from "Rank Types".

justification of the two-step approach as problematic. That a technique produces estimates that are larger or stronger than would otherwise be the case does not necessarily imply that the estimates are more accurate, as it may be possible that the estimates are simply biased. The same logic applies to the frequency of significant results.

Consider the following analogy. If you plan to go to a beach on the weekend, you want the weather forecast to predict a sunny weekend. However, we do not judge a forecaster by how frequently she forecasts nice weather, but rather by how accurate her predictions of the weather are, which we can only observe after we know how the weather turned out to be. If we valued forecasters only by the frequency with which they forecasted nice weather, an ideal forecaster would be one that would simply say that it will be always sunny. Yet, a forecaster that did this would be completely useless because such predictions are not informative of the upcoming weather. Similarly, the quality of our research tools should not be judged on their ability to produce results that are seen as desirable by researchers, but rather based on whether the tools produce results that can be expected to be correct in some well-defined sense. For example, when our technique produces a statistically significant estimate, we want to be confident that there actually is an effect in the population of interest. In other words, we want our analysis tools to both correctly indicate when an effect exists (low Type II / false negative error rate) and when it does not (low Type I / false positive error rate). As an extreme example, consider the fact that the bootstrap-based confidence intervals in commonly used PLS software can be configured to always indicate a statistically significant effect (Rönkkö, McIntosh, & Antonakis, 2015)[2].

[2] This can be achieved by using the individual sign-change correction and interpreting empirical confidence intervals as statistical tests by checking if zero is included in the interval. While there are debates on the merits of PLS (Rigdon, 2016; e.g. Rönkkö, McIntosh, & Antonakis, 2015; Rönkkö, McIntosh, Antonakis, & Edwards, 2016), the purpose of this example is not to get into that debate. In fact, also some proponents of the PLS technique caution against the use of the

Clearly, like a forecaster that will always say that the weather will be sunny regardless of what the actual weather looks like, a technique that will always indicate the presence of an effect regardless of whether the effect exists in the population is not a useful technique for testing the existence of the effect.

Unfortunately, whether results are correct is in most cases impossible to assess with empirical datasets because we do not know the true value of the effect under examination. For example, if the value of the correlation between two untransformed variables is .3 and that between the two-step transformed variables is .5, we are not able to state which of the two results is more accurate unless we know the true (population) value of the correlation. Judging which of the two estimates is more accurate is only possible if we know the true effect, to which we can then compare our estimates.

Because the true effects in populations from which empirical datasets are drawn are rarely, if ever, known, comparing statistical techniques can generally only be achieved through the use of simulated data, where the characteristics of interest are both known and under the control of the researcher (Bandalos, Hancock, & Mueller, 2006). If a technique can be demonstrated to work perfectly under conditions that are fully under researcher's control, then we can conclude that the technique may also work sufficiently well under non-ideal real-world scenarios. However, if we cannot get a technique to work even under ideal conditions, the claim that it would work in real world conditions that may be non-ideal would be implausible. For this

---

individual sign-change correction (Henseler, Hubona, & Ray, 2016). Instead, we bring up this example because it was the only example of a technique that we could come up with that a) has been used in published research, b) can be explained in a way that sounds reasonable, and c) always produces the result that a researcher wants to see, but d) is clearly not a valid approach for drawing inferences from the data. Note also that the sign-change correction is by no means specific to PLS estimates, and could be applied to any set of bootstrap replications regardless of the analysis technique and would produce the same 100% false positive rate.

reason, empirical examples have their place in methodological research, but their role is limited to illustrating analysis techniques, not to assessing their properties (Boomsma, 2013; Goodhue, Lewis, & Thompson, 2012).

The issue of non-normality is a complex one and unfortunately there are often no simple and general solutions to complex problems, as indicated by the extensive research on transformations in the structural equation modeling literature (e.g., Liu, Chen, Lu, & Song, 2015; Montfort, Mooijaart, & Meijerink, 2009; Mooijaart, 1993; Yuan, Chan, & Bentler, 2000). Furthermore, it is important to consider why a focal variable may not be normally distributed, because different causes of non-normality require different solutions. In this commentary, we address a number of issues with the two-step approach proposed by Templeton and Burney (2017) using simple regression analysis applied to simulated datasets with known properties. Our simulated demonstrations show that the two-step procedure is not an effective solution to any of the possible sources of non-normality in the data; moreover, its use introduces new problems for both statistical inference and interpretation of the results. Our findings are consistent with prior research questioning the usefulness of rank-based inverse normal transformations (Beasley, Erickson, & Allison, 2009).

SOURCES AND CONSEQUENCES OF NON-NORMALITY

Non-normal data may be problematic for a given statistical analysis, but whether this is the case depends on both the purpose of the analysis and the source of non-normality. The assumptions of statistical procedures are generally not about sample data, but rather about the distribution of the variables in the population from which data is assumed to originate[3].

[3] Strictly speaking, sample data can never be normally distributed because normal distribution is a continuous probability distribution where observations can take infinitely many different values, but a sample is a finite set of observations.

Moreover, the normality assumption pertains neither to observed (fixed) exogenous variables nor to observed endogenous variables, but rather to the unobserved exogenous part of the model (exogenous latent variables, including error terms). For example, ordinary least squares (OLS) regression analysis does not assume that the observed variables are normally distributed in the sample, but rather that the error term (a latent variable) is normally distributed in the population from which the data were sampled. Applying the two-step transformation therefore addresses an incorrect problem (Beasley et al., 2009) and its application can do more harm than good.

We demonstrate the problems with the two-step approach with the following simple regression model:

$$y = \beta_0 + \beta_1 x + u, \quad (1)$$

where y is the observed dependent variable, x is observed independent variable, and u is the unobserved (i.e., latent) error term. For simplicity, we set the intercept $\beta_0 = 0$ and the regression coefficient $\beta_1 = 1$. The dependent variable y can be non-normal under three scenarios (or a combination of them), each of which implies a different solution:

1. x is non-normal
2. u is is non-normal
3. the relationship between x and y is non-linear.

The first scenario is not problematic because OLS regression makes no assumptions about the distribution of the independent variables, the second scenario is a violation of one of the standard regression assumptions but the consequences of failing this assumption are typically not serious for applied research, and the third scenario is a serious problem that needs to be addressed by adjusting the model. We now consider each of these three cases in detail and assess the performance of the two-step transformation under each condition.

Non-normal Independent Variable

In our first scenario, the unobserved error u has standard normal distribution and x is non-normal, distributed as a chi-square with one degree of freedom (i.e., the square of a standard normal distribution). Here the OLS assumption of a normally distributed error term u holds in the population, but y has a mixture distribution, which is severely non-normal. However, because all OLS assumptions hold its application is not problematic. To demonstrate this, we generated a large sample of 1000 observations from this known model[4], and estimated an OLS regression with the (1) original data, (2) data where the two-step transformation was applied to y, and (3) data where the two-step transformation was applied to both x and y. The results are shown in Table 1 below.

----- Insert Table 1 about here -----

The results show that, with the original data, the OLS estimator can accurately recover the population parameters, whereas applying the two-step transformation produces estimates of the relationship and variance explained that are substantially biased, and negatively so. The error is even more marked when the two-step transformation is applied to both the independent and the dependent variables. In contrast to the results presented by Templeton and Burney (2017), whose correlation estimates using the two-step transformation were larger than correlations obtained from the untransformed data, this simple example demonstrates the effect can also be the opposite. We explain the mechanism leading to the bias later in this research. Moreover, as noted

[4] Our empirical demonstrations were implemented in R (R Core Team, 2016). The analysis file is included as Online Supplement A. While statistical techniques are commonly evaluated with Monte Carlo simulations where repeated samples are drawn from the same population model, we opted for one large sample for simplicity of presentation. To verify that these results are not idiosyncratic to our sample, we include a proper Monte Carlo simulation as Online Supplement B.

above, this type of comparison – between the true value of a relationship and the different estimates obtained from different analytical approaches – is not possible without recourse to the use of controlled, simulated data, where the value of the relevant parameter is known.

Non-normal Error Term

In the second scenario, the independent variable x is normally distributed, but the error term u follows a chi-square distribution with one degree of freedom centered at zero in the population. As before, we estimate three different models, one without any transformation, then applying the two-step transformation to the dependent variable y only followed by its application on both the independent variable x and dependent variable y. The results are shown in Table 2 below.

----- Insert Table 2 about here -----

Again, OLS without any transformation produces an accurate estimate of the population regression coefficient but applying the two-step transformation to the data produces substantially biased estimates and R^2 values, but in the opposite direction from the first example. In the current scenario, the transformation of the independent variable had little influence on the results because x was already normal in the population and the sample size was large.

We now explain the source of the bias and why the direction of the bias differs between the two scenarios. The two-step approach can be thought of as a non-parametric transformation that maps the original observations to the transformed values:

$$y_{two-step} = y + shift \quad (2)$$

Figure 1 below shows the mapping from original y values to transformed values (first plot) and the amount by which each observation is shifted sideways (second plot) as a function of the

original value for our latest example ($shift$). The dashed line depicts identity mapping that does not transform the values in any way.

----- Insert Figure 1 about here -----

The figure shows that, in this scenario, both small and very large values of the original y variable are decreased (shifted left) and values closer to the mean of the original variable are relatively less affected. What this means is that, in this scenario, the transformation makes the positive tail of the distribution shorter and the negative tail longer. The original data were substantially right skewed because the chi-square distribution is bounded at zero and has a long positive tail, and transforming the data in this way reduces the skewness. To understand the source of the bias, we need to first understand how $shift$ is related to x and u. In linear models, such as linear regression, the information provided by the covariance matrix is sufficient for estimating the model. The covariances for our last example are presented in Table 3 below. The table shows that $shift$ is negatively correlated with u, but positively correlated with x. This is natural because extreme positive values that are shifted the most are mostly due to extreme values of u, which followed the chi square distribution having a long positive tail. The positive correlation between x and $shift$ is due to shifting observations in the negative tail of y to the left. How these dependencies between x, u, and $shift$ influence regression estimates can be understood from two different perspectives.

----- Insert Table 3 about here -----

The first way to understand the source of bias is by focusing on how the regression coefficients are calculated. The regression coefficient of $y_{two-step}$ on x is given by

$$\frac{cov(x, y_{two-step})}{var(x)} \tag{3}$$

Which can be rewritten as

$$\frac{cov(x, y + shift)}{var(x)} \tag{4}$$

$$\frac{cov(x, y) + cov(x, shift)}{var(x)}$$

$$\frac{cov(x, y)}{var(x)} + \frac{cov(x, shift)}{var(x)},$$

which shows that the regression coefficient after the two-step transformation equals the sum of the original regression coefficient (given by $\frac{cov(x,y)}{var(x)}$) and the covariance between x and $shift$, divided by variance of x. Thus, the direction of the bias depends on the sign of the covariance. In the second example this sign was negative because x had a long positive tail but no negative tail. The first example presented the opposite scenario where u followed a chi-square distribution, leading to a positive covariance between x and $shift$ and therefore positive bias.

The second way to understand the direction of bias is to consider the example as an instance of omitted variable bias in regression. To understand why this is the case, we substitute the regression equation of y into the definition of $y_{two-step}$

$$y_{two-step} = y + shift \tag{5}$$

$$y_{two-step} = \beta_0 + \beta_1 x + shift + u$$

Regressing $y_{two-step}$ on x is essentially a misspecified model because $shift$ is not included in the estimated regression equation. The direction of bias follows directly from how regression behaves when variables are omitted from the estimated model (Wooldridge, 2009, pp. 89–93).

Model misspecification

The third possible source of non-normality arises when the relationship between x and y is non-linear, in which case a linear regression model is misspecified. Assume that the population model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u, \quad (6)$$

where both x and u are normally distributed and, for simplicity, all regression coefficients are set to 1. We fit six regression models to the data. Following the two earlier examples, the first three models are linear models fitted to original data, data where the two-step transformation is applied first to the dependent variable only, and finally to both x and y variables. The remaining three models are fitted to the same data but include a quadratic term (u-shape) and are thus correctly specified. The results are shown in Table 4 below.

----- Insert Table 4 about here -----

All misspecified models produced severely biased results. Fortunately, in all scenarios, the residuals vs. fitted plot that is commonly used for model diagnostics (e.g., Cohen, Cohen, West, & Aiken, 2003, Chapter 4; Draper & Smith, 1998, Chapter 2; Fox, 2015, Chapter 6), shown in Figure 2, reveal that the linear model is misspecified regardless of whether the two-step transformation was applied (first row of plots). Of the correctly specified models that contained the quadratic term, the one estimated with non-transformed data accurately reproduced the population parameters whereas applying the two-step transformation lead to substantially biased estimates and R^2. Moreover, while the diagnostic plot for the correctly specified model estimated with original data (first plot on the second row) demonstrates no problems, the plots generated using the two-step transformed data (last two plots on the second row) still indicated model misspecification thus, leading to the incorrect conclusion that the quadratic model was inappropriate for the data.

----- Insert Figure 2 about here -----

The three examples discussed here, representing different sources of non-normal data, provide a clear picture: when the dependent variable is non-normal, manipulating the data so that

it is closer to normality in the sample produces biased estimates. If the source of non-normality is a non-normal independent variable or error term, non-normality has virtually no impact on the accuracy of the estimates. On the other hand, in the scenario where non-normality is due to model misspecification, the most appropriate action is not to transform the dependent variable toward normality, but to diagnose the model and include nonlinear terms to make the model correctly specified (that is, where the structure of the relationships included in the model matches that in the population from which the data were sampled). Indeed, some textbooks on regression explicitly present transformations as a way to achieve linearity instead of normality (Chatterjee & Hadi, 2012, Chapter 6; Greene, 2012, Chapters 6–7; Kennedy, 2008, Chapter 6.3).

DOES LACK OF NORMALITY REALLY MATTER?

Given that none of the scenarios presented in the previous section provided evidence of major problems in the analysis of non-normal data except when this was due to model misspecification, one could question whether the consequences of non-normality have been exaggerated. To answer the question, we now focus on the assumptions and properties of the OLS estimator of linear regression models.

The Normality Assumption in OLS Regression

Neither the proof that OLS is the best linear unbiased estimator (the Gauss-Markov theorem) nor the unbiasedness of the standard errors requires any normality assumptions (Wooldridge, 2009, Chapter 3). However, to derive the exact sampling distribution of the estimates to calculate *p* values, we must know the distribution of the error term. If the error term is normally distributed, the regression estimates will also be normal over repeated samples and the *t* statistic, defined as the ratio of regression estimate to its standard error, will follow *Student's t* distribution with $n - k - 1$ degrees of freedom when the null hypothesis of no effects

holds (Wooldridge, 2009, Chapter 4), where n is the sample size and k is the number of predictors in the equation. Even then, the normality assumption is not particularly important because the regression estimates are asymptotically normal regardless of the distribution of the error term (Wooldridge, 2009, Chapter 5). That is, the distribution of the regression estimates approaches normality as sample size increases and consequently the non-normality of the error term is rarely an important issue for applied research (Cohen et al., 2003, p. 120; Wooldridge, 2009, pp. 174–175).

Impact of Transformations on Standard Errors

While normality is largely a non-issue for statistical inference based on OLS estimates (and we note that OLS regression is prevalent in AIS research; also Templeton and Burney (2017) focused on OLS regression), the covariance matrix of the transformed data will almost certainly not have the same sampling distribution as the covariance matrix of the original data. Therefore, it is unclear whether the commonly used test statistics will follow their theoretical distributions after the data have been transformed (Yuan et al., 2000), thus compromising statistical inference. We demonstrate this issue with a Monte Carlo simulation of 10 000 samples of 1000 from the model with a normally distributed independent variable x and a chi-squared distributed error u, again applying OLS to each sample first using the original data, then after applying the two-step transformation on the dependent variable, and finally on both variables. Table 5 displays the variances of the regression estimates and the means of estimated variances of the estimates

----- Insert Table 5 about here -----

The results show that applying the two-step transformation increases the variance of the estimates (i.e., decreases the precision) and produces variance estimates (squares of standard errors) that are severely negatively biased. The variance estimates are on average just half of the

true variance of the estimates, which translates to a standard error bias of about -30 percent. In this scenario, the estimates are positively biased and thus applying the two-step leads to overstating both the effects and the precision of the estimates.

Transformations and Interpretation of Non-linear Effects

While transformations are largely unnecessary for achieving normality, there are scenarios where transformations are useful because, in many cases, effects are non-linear and thus using a linear regression model would not be appropriate (e.g., Wooldridge, 2009, sec. 6.2). However, the use of transformations can complicate the interpretation of regression results. As highlighted by Lin et al. (2013) it is not sufficient to just asses the significance and sign of a regression coefficient, but we must also assess the size of the effects, particularly when using large samples where trivially small effects become statistically significant. Lin et al. (2013) illustrate effect size assessment with the following two examples "each additional apple consumed per day reduces the chances of going to the doctor on average by 33%" and "including an apple a day in your diet is likely to reduce your risk of becoming ill from 3% to 2%". Making such interpretations is straightforward in regression analysis with untransformed data, because the coefficients can be interpreted directly. However, applying transformations may alter the meaning of the coefficients (Lin et al., 2013, Table 2), which we illustrate in our first empirical example in the next section.

EMPIRICAL EXAMPLES

Empirical Example 1: Econometrics Textbook

Our first example is an analysis presented in a highly regarded introductory econometrics book (Wooldridge, 2009). We chose this example because the data are freely available and thus allow for easy replication of the analysis, and because the availability of a textbook format treatment of the example greatly facilitates learning over what can be presented in the tight space

of a journal article. The example from Wooldridge (2009) is from a chapter that explains that transformations should be driven by theoretical considerations over empirical ones. One particularly relevant scenario is when the effect of change in x is expected to be proportional to the current level of x. For example, raises in salary are often proportional to current salary levels. Indeed, in many countries, labor unions negotiate salary increases on a percentage rather than absolute unit basis. The logarithm transformation (Wooldridge, 2009, secs. 2.4, 6.2) is common choice for linearization in this scenario. To showcase these issues, we used a cross-sectional dataset on the wages of 526 working individuals from 1976 used by Wooldridge (2009, Example 2.10) and analyzed the non-linear effect of years of education on wages.

----- Insert Table 6 about here -----

The regression results in Table 6 show that years of education has a statistically significant and positive effect on wages regardless of which transformation is applied. This result in and of itself is not particularly interesting because trivially small effect sizes become significant in large samples (Lin et al., 2013); instead, we should focus on the magnitude of the estimates and what they mean. The interpretations of the three regression models are "One additional year of education increases the logarithm of hourly wages by 0.08", "One additional year of education increases the two-step transformed hourly wages by 0.59", and "One additional unit of two-step transformed education increases the two-step transformed hourly wages by 0.62". However, we are rarely interested in the transformed units but on the effects expressed in the original units instead. The question that we must therefore ask is how large the effect of one additional year of education on the expected hourly wages is. For the model with logarithm transformation, the interpretation is straightforward because the logarithm transformation itself has a natural interpretation as relative change. The regression coefficient of 0.08 means that an additional year

of education increases the expected wages by 8 percent compared to the current wage level. Templeton and Burney (2017) state that the two-step transformed regression results can be interpreted in the original metric: "One additional year of education increases the hourly wages by 0.59/0.62". This interpretation would be problematic for two reasons: a) it masks the fact that relative effects have a better fit with how wages behave both theoretically and empirically, and b) because the interpretation is linear (i.e. constant, absolute effect), it means that the two-step transformation is essentially an alternative way (estimator) to draw the regression line. However, it is not clear what kind of statistical properties a line drawn this way would have.

In the case of logarithm transformation, the transformed results have a natural interpretation. However, in some cases a natural interpretation may not be available or interpreting the effects may be difficult for some other reason. In this case marginal prediction plots, available in many commonly used statistical packages (Fox, 2003; Williams, 2012), provide an invaluable tool for interpretation of the size and nature of the effects (Lin et al., 2013, p. 909). Marginal prediction plots are prepared by calculating fitted values for the regression model using several combinations of the independent variables, transforming these values back to the original metric, and plotting the data. To do this, we need an inverse transformation that allows interpreting the results. We demonstrate the use of a marginal prediction plot in Figure 3 below using the wage data example. The figure makes it clear on how the expected hourly wage depends on the years of education.

----- Insert Figure 3 about here ------

Because the two-step transformation does not have an inverse transformation, interpreting the results by back transforming the predictions to the original metric as done above is impossible. It is for this reason that introductory texts on statistical analysis only explain

transformations that have inverse transformations (e.g., Kline, 2011, pp. 63–64; Schumacker, 2010, p. 28).

Empirical Example 2: Actual Accounting Data

For our second empirical example, we replicated parts of the analyses presented by Templeton and Burney (2017). We started by obtaining data for all companies in the Compustat database from 1990 to 2014, resulting in 286,236 observations of 29,866 unique companies[5]. Thereafter, we typed the 1992 Computerworld Premier 100 table ("Premier 100 tables," 1992) into a data file. Out of the 100 companies in the list, we were able to match 94 with the Compustat data, out of which 86 observations had complete data for replicating the regression models presented by Templeton and Burney (2017).

Instead of presenting a full replication of all analyses done by Templeton and Burney (2017), we focus on one particular example. However, before presenting the example, there is one general issue that requires addressing. Many of the accounting variables used by Templeton and Burney (2017) are ratios, which can cause problems in analysis regardless of how the data are analyzed. The two key problems in ratios are that a) if the numerator and denominator have different interpretations on their own, just by observing a change in the ratio does not allow us to infer which of the two components changed, thus leading to confounding of the effects and b) if the range of the denominator includes zero, the ratio may become infinitely large or even indeterminate. While these issues have been noted in other disciplines (see Certo, Busenbark, Kalm, & LePine, 2018 for a review) and textbooks (e.g., Cohen et al., 2003, pp. 60–61) they

[5] These numbers are about 15% and 20% larger than the numbers reported by Templeton and Burney (2017). The difference may be due to database versions. We used the Compustat Monthly Updates - Fundamentals Annual dataset.

have been largely ignored in AIS research. Because of these issues, we chose to focus on the non-ratio variable net income (NI, in millions of dollars), to demonstrate effects of the two-step transformation without confounding its effects with the challenges caused by using ratio variables.

We will first consider the relationship between SIZE, defined as the natural logarithm of total revenue in dollars, which was one of the control variables used by Templeton and Burney (2017), and NI. Figure 4 shows the relationship between the original variables and the same relationships after the two-step transformation. The first panel shows the full data, the second panel the part of the original data that fall into the range of the two-step transformed NI, and the third panel shows the two-step transformed NI. The first panel shows that the variance of NI depends on SIZE and that there are a few potential outliers. Therefore, we followed the guidelines presented by Aguinis, Gottfredson, and Joo (2013) and first inspected what the extreme observations are. These included Fannie Mae, Freddie Mac, American International Group, and General Motors, each of which produced an outlier observation due to government intervention during the 2007-2008 financial crisis, Time-Warner and JDS Uniphase, which are extreme observations due to write-offs of acquisitions done during the dot-com bubble, and Vodafone's 2013 sale of its stake of Verizon Wireless. Because these extreme observations clearly have nothing to do with how IT investments influence company performance, the most appropriate course of action would be elimination of these observations as outliers.

----- Insert Figure 4 about here -----

The main thing to observe in Figure 4[6] is that it demonstrates why the correlations after the two-step transformation can be larger than when using the original data. In this sample, the correlation between SIZE and NI was originally 0.21, but increased to 0.50 when the two-step transformation was applied to NI. This difference in the correlations can be attributed to how the two-step transformation treats outliers. Outliers increase the variance of the data and produce a distribution with longer tails. The transformations that are specifically designed to address outliers, such as winsorizing, address this issue by pulling the tails containing the outliers toward the mean of the data, thus decreasing the variance. While the two-step transformation does the same, to maintain the original variance, the technique also pulls other observations towards the outliers, as comparing the second and third panel in Figure 4 clearly shows. We are not aware of any statistical principle that would justify transforming non-outlier observations to be more similar to outliers. To summarize, in this example, the two-step transformation does not eliminate the effect of outliers but, on the contrary, amplifies their effect by pulling other observations toward the outliers, which is almost certainly not something that an AIS researcher would want to do.

The larger correlations produced by the two-step transformation lead to larger R^2 values in regression analysis, as shown in Table 7 below. To demonstrate this feature, we first regressed NI on SIZE, LEVERAGE, and the four Computerworld variables used by Templeton and Burney (2017, pp. 152–153). The R^2 values for the two models were 0.42 and 0.48. We then

[6] As a technical side note, Figure 4 also demonstrates that while the two-step transformation can make the data closer to univariate normality, the data are not multivariate normal because the two-variable plot lacks the elliptical shape of a multivariate normal distribution. That univariate normality does not imply multivariate normality and the implications of this fact for research methods literature is beyond the article, by we refer interested readers to e.g., the work by Mair, Satorra, and Bentler (2012).

regressed NI on just SIZE producing R^2 values of 0.36 and 0.46 showing that the difference in R^2 between the transformed and non-transformed data can be traced back to how the two-step transformation increased the correlation between the two focal variables in our example. To demonstrate this effect graphically, Figure 5 shows the added variable or partial regression plots for SIZE for the original NI (first plot) and two-step transformed NI (second plot). The plots show that, also in the multiple regression context, the two-step transformation pulls outliers toward the mean of the data while also pulling other non-outlier observations away from the mean and toward the outliers as shown before, producing a steeper regression line.

----- Insert Table 7 and Figure 5 about here -----

Figure 5 also demonstrates clearly that with the original data (first plot), the relationship between SIZE and NI is nonlinear, but that a linear relationship is not an appropriate explanation for the data goes unnoticed when the two-step transformation is used (second plot). In this rare case, we know that the relationship between SIZE and NI is non-linear by definition: SIZE is defined as a natural logarithm of total revenues, but total revenues has a direct linear relationship with NI: holding all other things constant, one dollar increase in total revenues will lead to one dollar increase in net income. If two variables are originally linearly associated, they cannot be associated in the nonlinear fashion that the logarithm transformation presents. With the original data, this misspecification of the model would have been detected using standard regression diagnostics, but this misspecification is masked by the two-step transformation.

DISCUSSION AND CONCLUSIONS

Our simple simulation examples show that the two-step transformation produces regression coefficients that are biased and less precise than would be the case using the original untransformed data, and which have biased standard errors. Moreover, the two-step

transformation does not have an inverse transformation, making interpretation of the size of the effects nearly impossible. Our research is not the only study demonstrating these problems (Beasley et al., 2009). That our assessment of the two-step technique differs radically from that by Templeton and Burney (2017) naturally raises the question of why this is the case.

While our assessment of the two-step transformation presented by Templeton and Burney (2017) is negative, we do agree that severe non-normality may require some adjustments to the statistical techniques used in AIS research. However, non-normality should not be viewed as a problem per-se, but as a symptom of one or more different problems that each require different remedies. In the following two sections, we will address the methodological implications and practical implications for AIS research in turn.

Methodological Implications

The underlying reason for the radically different conclusions between our work here and Templeton and Burney's (2017) article is that most of the results by Templeton and Burney (2017) are based on correlations, which may be inflated (positively biased) by the two-step procedure, as shown by our simple simulation study above. These results highlight the fact that, in the same way we cannot judge a weather forecaster solely on how frequently she predicts sunny weather, we cannot assess results of statistical procedures based on just whether the results produced from a real samples of data are desirable from the point of view of researchers, but must instead focus on whether the results can be expected to be correct in some sense.

Because we can generally only know the population value of a statistic in simulated datasets, but not in empirical ones, whether a technique produces correct results can only be judged based on a simulation study. This is in stark contrast to the claim by Templeton and Burney (2017): "Thus, our findings suggest that real data may often be preferred in studies

comparing research methods, since inferences from our comparisons using real data normalized by the Two-Step procedure are not possible with randomly generated data" (p. 159). The logic of this sentence is difficult to understand. It is certainly possible to analyze the performance of the two-step using simulated data, which we prove by doing so. To assess the claim, we need to understand its context, which starts by that stating the results from original data were more similar to results obtained from analyzing uncorrelated random variables than they were to results after the two-step transformation. This does not invalidate using simulated datasets, for two reasons. First, such argument can only be made by demonstrating that analyzing simulated datasets using the proposed approach leads to erroneous results. However, Templeton and Burney (2017) actually never apply the two-step procedure to simulated datasets in their article. Instead, they compare their results to a regression model where the dependent variable is a random variable generated independently of the other data, which is hardly representative of an informative simulation study. Second, we cannot infer that because the results from two-step were quite different from the original results and results from random data, they would be more appropriate. An equally plausible explanation, and one that is supported by the evidence that we present, is that the results based on the two-step transformation were simply substantially biased. Because the true value of a relationship is rarely, if ever, known with real data, we simply cannot say which of the results using alternative techniques produced the most accurate results. Therefore, simulated datasets from known populations must be used. This is also exclusively the approach followed in methodological investigations of existing or proposed new analytical approaches.

There are also purely statistical reasons, which go beyond the use of simulated datasets, which suggest that the use of the two-step transformation is likely not an optimal analysis

technique. A variable gives us information about the location of each observation as well as the ranks of the observations in the sample. By applying the two-step transformation, we first convert the data to ranks, thus discarding the location information. Mapping the ranks on to the normal distribution does not add new information to the data, but simply expresses the rank information in a different metric. This raises two important concerns. First, the two-step transformation is information destroying, which is never a desirable feature. Second, because the transformed data contain only information about ranks, it would be more appropriate to use statistical tools specifically designed for rank data, rather than assuming – incorrectly – that the transformed data also contain location information. Indeed, rank-based approaches have been used in the past in the AIS literature (e.g., Stratopoulos & Dehning, 2000).

The information destroying nature of the transformation also has a direct implication on the claim that the two-step transformation would have a positive effect on reliability. The concept of reliability originates from classical test theory (Markus & Borsboom, 2013, sec. 3.1.1) and is about the degree to which the data are contaminated with random noise. If a variable has low reliability, then the ratio of random noise to useful information in the variable is low. It is mathematically impossible that a transformation would improve reliability because statistical transformation can only destroy but not create new information in the data (MacKay, 2003). The reason why Templeton and Burney (2017) reached a different conclusion is because they did not actually study reliability, but rather the behavior of test-retest statistic over a five-year lag. This presents two problems: First, the test-retest statistic is a valid estimate of reliability only if the underlying trait remains unchanged between the test and the retest. This is hardly true for companies and financial ratios, particularly over such a long period of time. Second, this statistic

is based on a correlation, and it is simply possible that the two-step transformation produced a positively biased correlation estimate, which lead to overstating test-retest reliability.

Implications for AIS Research

Finally, we address the implications of non-normality for the productivity paradox and for the AIS research more generally. Templeton and Burney (2017) begin their article by stating that "non-normality is barely mentioned as a potential reason for confounding statistical results in AIS" (p. 149) thus suggesting that non-normality of the data may be one of the causes of the productivity paradox. Dating back to seminal article by Brynjolfsson (1993), who attributed the lack of a positive correlation between productivity statistics and IT investments to mismeasurement of inputs and outputs, lags due to learning, redistribution of profits, and mismanagement, various explanations – both methodological as well as substantial – have been put forward for the inconsistent finding between studies (Schryen, 2013). That non-normality is not mentioned as a potential cause of these inconsistent finding in AIS research is a natural outcome considering that in regression analysis, which is one of the most commonly used statistical tools in AIS research, no normality assumptions are made about the explanatory variables and sample sizes are typically large enough that non-normality of the error term is hardly an issue.

While non-normality is hardly an issue itself, it can be a symptom of a number of other underlying concerns. As our results indicate, there is no panacea for non-normal data. Rather than applying transformations, researchers should first learn from their data and apply their understanding of the studied phenomenon to identify the sources of non-normality, because not all sources of non-normality are problematic and even when non-normality may be an indication of a problem, the appropriate remedies depend on what exactly the problem is. We provide four

recommendations for dealing with non-normality. First, as Templeton and Burney (2017) note, non-normality can be due to outliers. An outlier can be simply a data entry mistake, in which case the data can often be corrected. Another alternative is that the outlier falls outside the study population and hence should be removed. Regardless of the type of outlier, these should be studied and researchers should make informed decisions rather than automatically trimming or transforming outliers (Aguinis et al., 2013), as we did in our second empirical example.

Second, the source of non-normality can be a nonlinear effect in the model. In the case of a regression model, residual diagnostics plots (e.g., Cohen et al., 2003, Chapter 4; Draper & Smith, 1998, Chapter 2; Fox, 2015, Chapter 6) provide an invaluable set of tools for detecting these problems and models can be adjusted accordingly. While there is no excuse for not performing these diagnostics, reporting that models were diagnosed is unfortunately rare in AIS research. A typical remedy for non-linearity is to linearize the model by transforming the variables, most commonly by taking a natural logarithm of a variable. In these cases, the interpretation of the effect sizes should be carried out with the help of marginal prediction plots (Fox, 2003; Lin et al., 2013; Williams, 2012), as demonstrated in our first empirical example. If the model cannot be linearized with a transformation, we suggest that AIS researchers consider the use of non-linear models, such as fractional polynomials (Nikolaeva, Bhatnagar, & Ghose, 2015) or other similar techniques (Greene, 2012, Chapter 6; Tan, Shiyko, Li, Li, & Dierker, 2012).

Third, while it is beyond the scope of our article to address the issue in full detail, future AIS research should pay more attention to the difficulties in analyzing ratio data. Consider return on assets (ROA), often used as a dependent variable in AIS research (Lim, Dehning, Richardson, & Smith, 2011) and defined as the ratio of net income to average total assets. When a firm invests in IT, the investment is added to the firm's assets where it is depreciated over time. If the

IT investment influences firm productivity, it can also be expected to influence net income. Because the IT investment can thus be expected to influence both the numerator and the denominator of the ratio, these effects, which have clearly different interpretations, are easily confounded. Wiseman (2009) and Certo et al. (2018) provide an accessible introduction to the problems of using financial ratios as dependent variables and discusses some of the solutions to these problems. While we are not advocating that AIS researchers completely abandon financial ratios as dependent variables, this is clearly an issue that future AIS research should investigate.

Finally, if all else fails, we have a few robust and non-parametric approaches that could be used when an approach that does not assume normality in any way is truly needed. One alternative that has seen some use in AIS research are non-parametric techniques, particularly in the form of rank comparisons (Stratopoulos & Dehning, 2000). However, these techniques, while effective for comparing groups, have a major weakness in that they do not allow for assessing what is the (statistical) impact of IT spending on various performance measures. That is, they do not readily answer the question of how large productivity impact an IT investment is expected to have. Nevertheless, based on the analysis presented in this article, it is difficult to see a scenario where the proposed two-step transformation technique would be preferable over these alternative techniques.

REFERENCES

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Bandalos, D., Hancock, G. R., & Mueller, R. O. (2006). The use of Monte Carlo studies in structural equation modeling research. In *Structural equation modeling: A second course* (pp. 385–426). Greenwich, Conn: Information Age Publishing Inc.

Beasley, T. M., Erickson, S., & Allison, D. B. (2009). Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behavior Genetics*, *39*(5), 580–595. https://doi.org/10.1007/s10519-009-9281-0

Bliss, C. I. (1967). *Statistics in biology. Statistical methods for research in the natural sciences.* New York: McGraw-Hill Book Company.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variables*. New York: Wiley.

Boomsma, A. (2013). Reporting monte carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 518–540. https://doi.org/10.1080/10705511.2013.797839

Brynjolfsson, E. (1993). The Productivity Paradox of Information Technology. *Communications of the ACM*, *36*(12), 67–77.

Certo, S. T., Busenbark, J. R., Kalm, M., & LePine, J. A. (2018). Divided we fall: How ratios undermine research in strategic management. *Organizational Research Methods*, 1094428118773455. https://doi.org/10.1177/1094428118773455

Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (Fifth edition). Hoboken, New Jersey: Wiley.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. London: Lawrence Erlbaum Associates.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed). New York: Wiley.

Fisher, R. A., & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver and Boyd·Ltd. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1614539/

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, *8*(15), 1–27.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications.

Goodhue, D. L., Lewis, W., & Thompson, R. (2012). Comparing PLS to regression and Lisrel: A Response to Marcoulides, Chin, and Saunders. *MIS Quarterly*, *36*(3), 703-A10.

Greene, W. H. (2012). *Econometric analysis*. Boston: Prentice Hall.

Henseler, J., Hubona, G., & Ray, P. A. (2016). Using PLS path modeling in new technology research: Updated guidelines. *Industrial Management & Data Systems*, *116*(1), 2–20. https://doi.org/10.1108/IMDS-09-2015-0382

Kennedy, P. (2008). *A guide to econometrics* (6th ed). Malden, MA: Blackwell Pub.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Lim, J.-H., Dehning, B., Richardson, V. J., & Smith, R. E. (2011). A Meta-Analysis of the Effects of IT Investment on Firm Financial Performance. *Journal of Information Systems*, *25*(2), 145–169. https://doi.org/10.2308/isys-10125

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research commentary - Too big to fail: Large samples and the p-value problem. *Information Systems Research*, *24*(4), 906–917. https://doi.org/10.1287/isre.2013.0480

Liu, P., Chen, J., Lu, Z., & Song, X. (2015). Transformation Structural Equation Models With Highly Nonnormal and Incomplete Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 401–415. https://doi.org/10.1080/10705511.2014.937320

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Mair, P., Satorra, A., & Bentler, P. M. (2012). Generating nonnormal multivariate data using copulas: Applications to SEM. *Multivariate Behavioral Research*, *47*(4), 547–565. https://doi.org/10.1080/00273171.2012.692629

Markus, K. A., & Borsboom, D. (2013). *Frontiers in test validity theory: measurement, causation and meaning*. New York, N.Y.: Psychology Press.

Montfort, K. van, Mooijaart, A., & Meijerink, F. (2009). Estimating structural equation models with non-normal variables by using transformations. *Statistica Neerlandica*, *63*(2), 213–226.

Mooijaart, A. (1993). Structural equation models with transformed variables. In Haagen, Klaus, D. J. Bartholomew, & M. Deistler (Eds.), *Statistical modeling and latent variables* (pp. 249–258). Amsterdam; New York: North-Holland.

Nikolaeva, R., Bhatnagar, A., & Ghose, S. (2015). Exploring Curvilinearity Through Fractional Polynomials in Management Research. *Organizational Research Methods*, *18*(4), 738–760. https://doi.org/10.1177/1094428115584006

Premier 100 tables. (1992, September 14). *Computerworld*, pp. 54–61.

R Core Team. (2016). R: a language and environment for statistical computing (Version 3.3.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rigdon, E. E. (2016). Choosing PLS path modeling as analytical method in European management research: A realist perspective. *European Management Journal*, *34*(6), 598–605. https://doi.org/10.1016/j.emj.2016.05.006

Rönkkö, M., McIntosh, C. N., & Antonakis, J. (2015). On the adoption of partial least squares in psychological research: Caveat emptor. *Personality and Individual Differences*, *87*, 76–84. https://doi.org/10.1016/j.paid.2015.07.019

Rönkkö, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management*, *47–48*, 9–27. https://doi.org/10.1016/j.jom.2016.05.002

Schryen, G. (2013). Revisiting IS business value research: what we already know, what we still need to know, and how we can get there. *European Journal of Information Systems*, *22*(2), 139–169.

Schumacker, R. E. (2010). *A beginner's guide to structural equation modeling* (3rd ed). New York: Routledge.

Stratopoulos, T., & Dehning, B. (2000). Does successful investment in information technology solve the productivity paradox? *Information & Management*, *38*(2), 103–117. https://doi.org/10.1016/S0378-7206(00)00058-6

Tan, X., Shiyko, M., Li, R., Li, Y., & Dierker, L. (2012). Intensive longitudinal data and model with varying effects. *Psychological Methods*, *17*, 61–77.

Templeton, G. F. (2011). A Two-Step Approach for Transforming Continuous Variables to Normal: Implications and Recommendations for IS Research. *Communications of the Association for Information Systems*, *28*(1). Retrieved from http://aisel.aisnet.org/cais/vol28/iss1/4

Templeton, G. F., & Burney, L. L. (2017). Using a Two-Step Transformation to Address Non-Normality from a Business Value of Information Technology Perspective. *Journal of Information Systems*, *31*(2), 149–164. https://doi.org/10.2308/isys-51510

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

Van der Waerden, B. L. (1952). Order tests for the two-sample problem and their power. In *Indagationes Mathematicae (Proceedings)* (Vol. 55, pp. 453–458). Elsevier. https://doi.org/10.1016/S1385-7258(53)50012-5

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *Stata Journal*, *12*(2), 308–331.

Wiseman, R. M. (2009). On the use and misuse of ratios in strategic management research. In *Research Methodology in Strategy and Management* (Vol. 5, pp. 75–110). Emerald Group Publishing Limited. https://doi.org/10.1108/S1479-8387(2009)0000005004

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th ed). Mason, OH: South Western, Cengage Learning.

Yuan, K.-H., Chan, W., & Bentler, P. M. (2000). Robust transformation with applications to structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *53*(1), 31–50.

TABLES AND FIGURES

Table 1 Regression of y on non-normal x

	Original data	Two-step transformation applied to y	Two-step transformation applied to x and y	Population value
Intercept	-0.04	0.14**	0.20***	0
	(0.04)	(0.05)	(0.05)	
x	1.02***	0.85***	0.80***	1
	(0.02)	(0.03)	(0.03)	
R^2	0.69	0.48	0.43	0.67
Adj. R^2	0.69	0.48	0.43	

N = 1000. Standard errors in parentheses.
*** p < 0.001, ** p < 0.01, * p < 0.05, two-tailed tests

Table 2 Regression of y on normal x with chi-squared u

	Original data	Two-step transformation applied to y	Two-step transformation applied to x and y	Population value
Intercept	0.08	0.08	0.08	0
	(0.05)	(0.04)	(0.04)	
x	0.99***	1.18***	1.18***	1
	(0.05)	(0.04)	(0.04)	
R^2	0.31	0.45	0.45	0.33
Adj. R^2	0.31	0.45	0.45	

N = 1000. Standard errors in parentheses.
*** p < 0.001, ** p < 0.01, * p < 0.05, two-tailed tests

Table 3 Covariance matrix of original variables, shift, and two-step transformed y

	x	u	y	$shift$	$y_{two-step}$
x	1.06				
u	-0.01	2.29			
y	1.05	2.28	3.33		
$shift$	1.26	1.91	3.17	3.33	
$y_{two-step}$	0.21	-0.37	-0.16	0.16	0.32

Table 4 Regression of y on normal x and x^2

	Misspecified model			Correctly specified model			
	Original data	Two-step transformation applied to y	Two-step transformation applied to x and y	Original data	Two-step transformation applied to y	Two-step transformation applied to x and y	Population value
Intercept	0.96***	0.96***	0.96***	-0.05	0.11*	0.12*	0
	(0.05)	(0.06)	(0.06)	(0.04)	(0.05)	(0.05)	
x	0.87***	0.78***	0.79***	0.97***	0.86***	0.83***	1
	(0.05)	(0.06)	(0.06)	(0.06)	(0.04)	(0.04)	
x^2				1.04***	0.87***	0.86***	1
				(0.02)	(0.03)	(0.03)	
R^2	0.20	0.16	0.17	0.74	0.54	0.53	0.75
Adj. R^2	0.20	0.16	0.17	0.74	0.54	0.53	

N = 1000. Standard errors in parentheses.
*** p < 0.001, ** p < 0.01, * p < 0.05, two-tailed tests

Table 5 Results of Monte Carlo study of the two-step transformation

	Variance of regression estimate ($var(\widehat{\beta_1})$)	Mean of estimated variance of regression estimate ($\mathbb{E}\left[(\widehat{var(\widehat{\beta_1})}\right]$)
Original data	0.0020	0.0020
Two-step applied to y	0.0034	0.0016
Two-step applied to both y and x	0.0034	0.0016

Monte Carlo simulation with 10 000 replications. N = 1000.

Table 6 Regression of logarithm of hourly wages on years of education

	Logarithm transformation applied to wages	Two-step transformation applied to wages	Two-step transformation applied to wages and education
Intercept	0.58***	-1.48*	-1.82**
	(0.10)	(0.67)	(0.67)
Years of education	0.08***	0.59***	0.62***
	(0.01)	(0.05)	(0.05)
R^2	0.19	0.20	0.21
Adj. R^2	0.18	0.19	0.21

N = 526. Standard errors in parentheses.
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, two-tailed tests

Table 7 Regression of NI (net income) as modeled by Templeton and Burney (2017)

	Original NI		Two-step transformation applied to NI	
Intercept	-2820.01***	-3157.18***	-3199.71***	-3628.57***
	(653.03)	(557.38)	(615.08)	(512.15)
IS budget as % revenue	-14.97		-17.92	
	(16.44)		(15.48)	
% of IS budget for staff	-3.42		-8.62	
	(8.18)		(7.71)	
% of IS budget for training	8.94		4.98	
	(25.02)		(23.56)	
% of employess with PCs/terminals	-3.68		-0.67	
	(2.74)		(2.58)	
LEVERAGE	98.35*		20.81	
	(47.50)		(44.74)	
SIZE	427.24***	436.25***	492.44***	493.26***
	(63.44)	(63.65)	(59.76)	(58.49)
R2	0.42	0.36	0.48	0.46
Adj. R2	0.37	0.35	0.44	0.45

N = 86. Standard errors in parentheses.
*** p < 0.001, ** p < 0.01, * p < 0.05, two-tailed tests

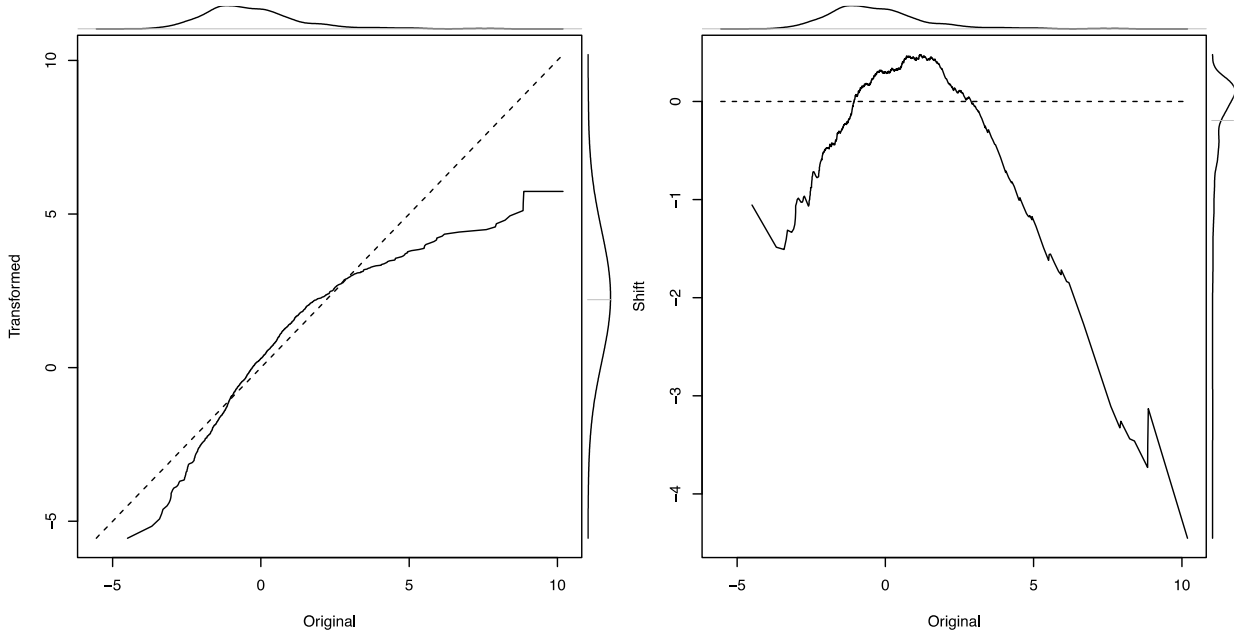


Figure 1 Non-parametric transformation curve and shift as a function of original value

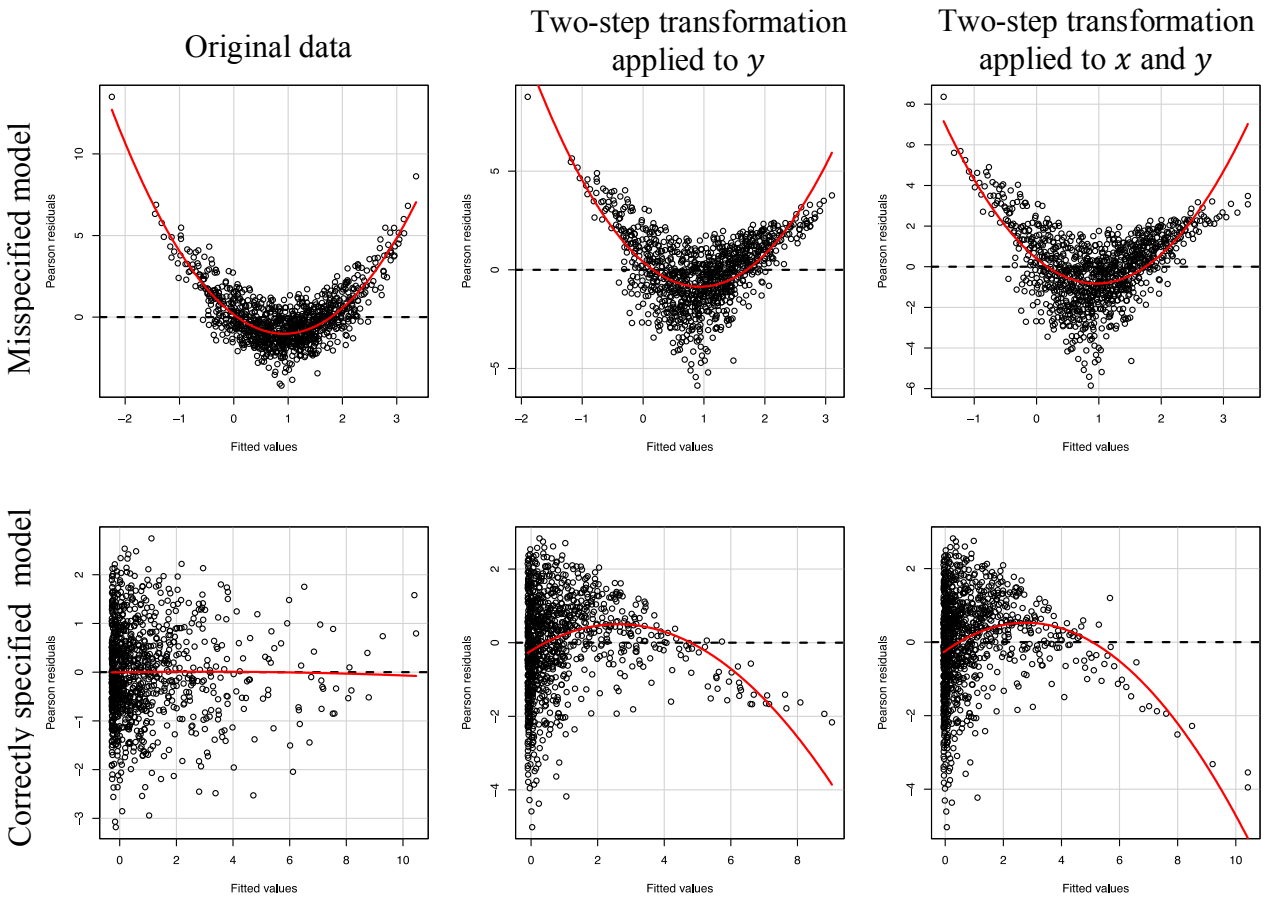


Figure 2 Residual vs. fitted regression diagnostic plots for misspecified linear and correctly specified quadratic models

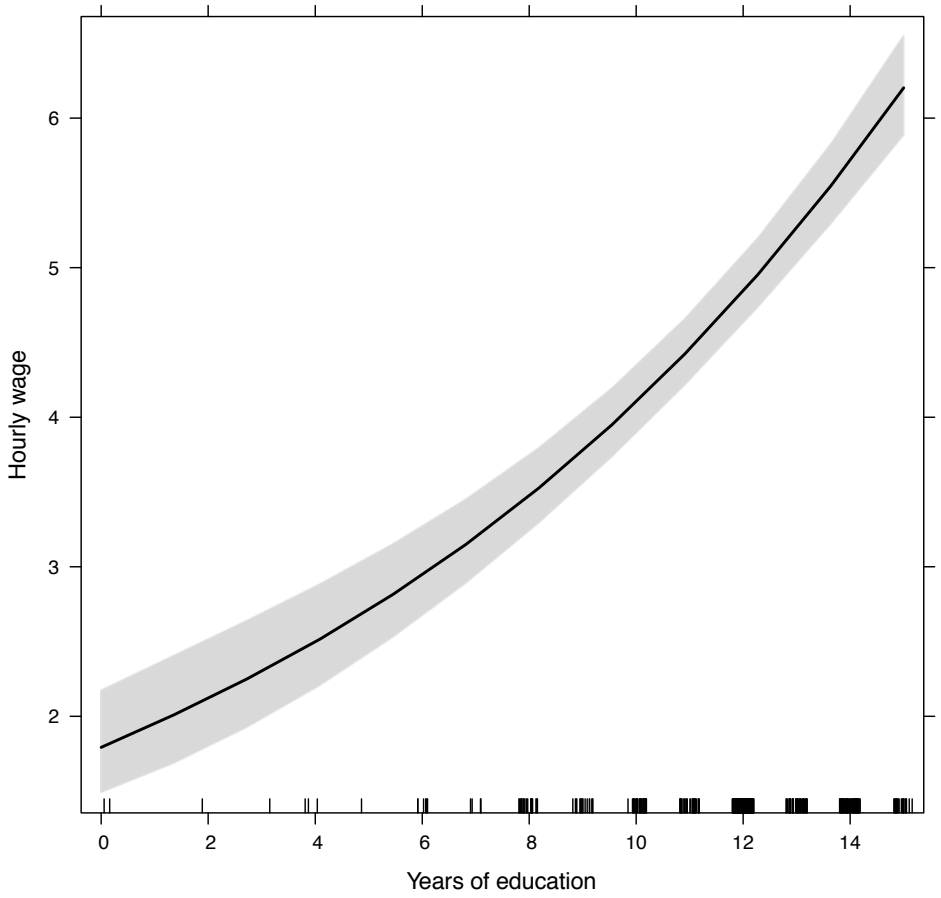


Figure 3 Marginal prediction plot of hourly wages on years of education

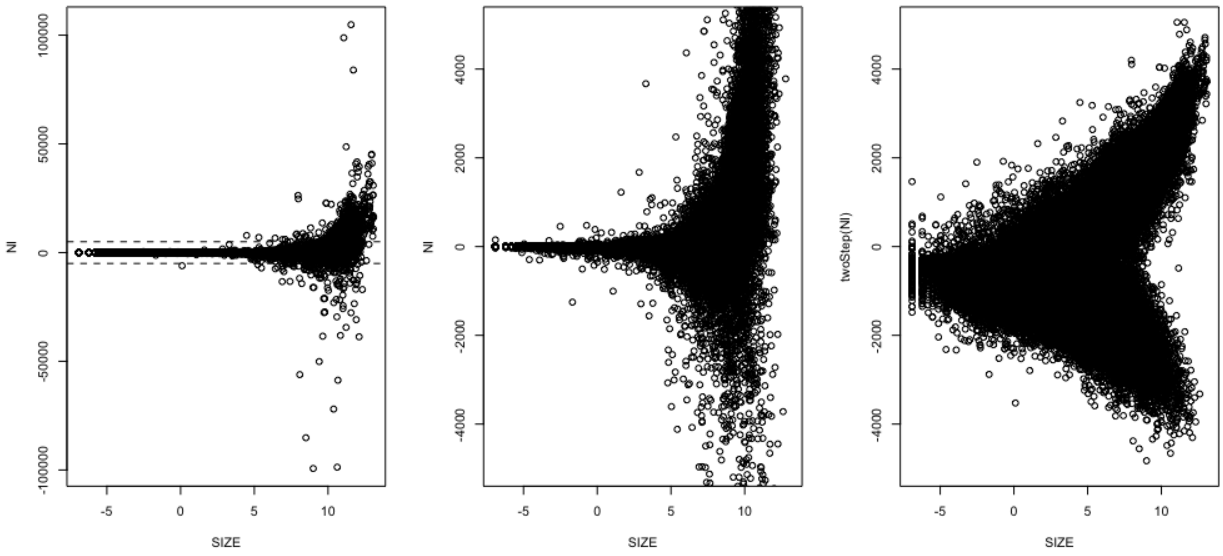


Figure 4 Relationship between NI and SIZE and two-step(NI) and SIZE.

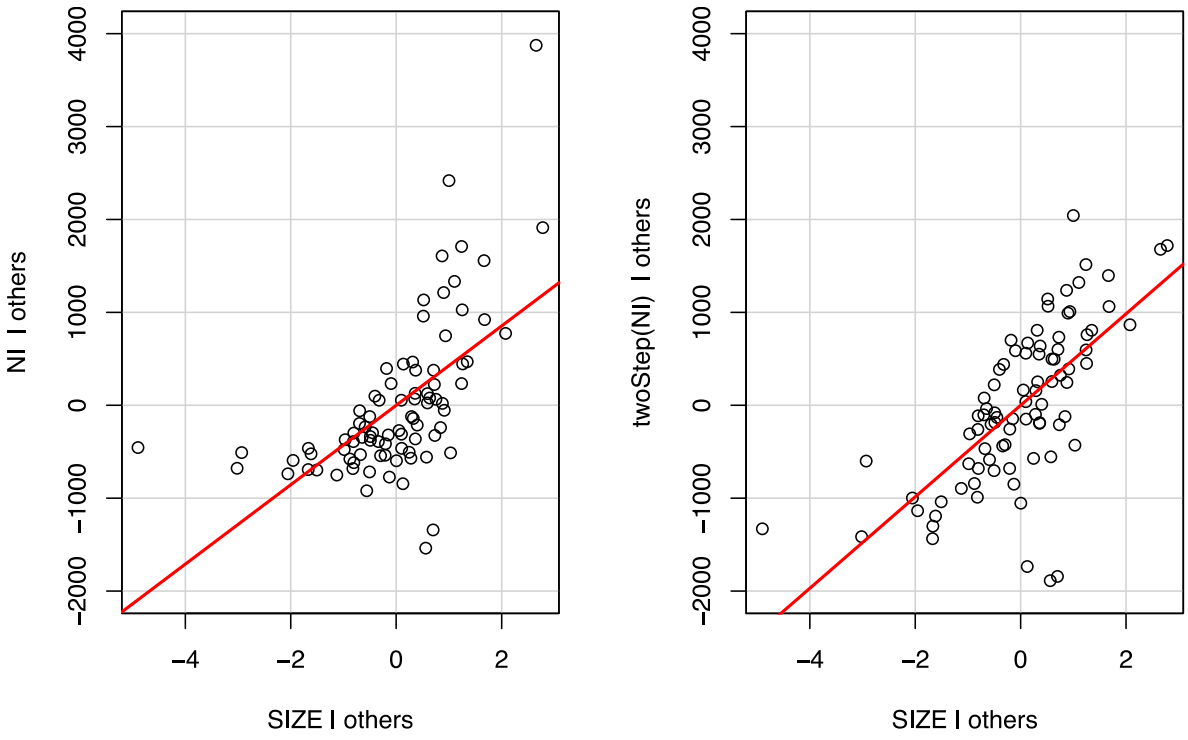


Figure 5 Added variable plots of NI on SIZE with and without two-step transformation

Online supplement 1: R code for the study

```
# Packages for producing effects plots and doing regression diagnostics
library(texreg)
```

```
## Version:  1.36.23
## Date:     2017-03-03
## Author:   Philip Leifeld (University of Glasgow)
##
## Please cite the JSS article in your publications -- see citation("texreg").
```

```
library(car)

# Two-step transformation, as presented by Templeton et al.

twoStep <- function(x){
  prank <- rank(x)/length(x)
  prank[prank == 1] <- 1-1/length(x)  # A workaround for percentile rank = 1.
  r <- qnorm(prank)
  r/sd(r)*sd(x)+mean(x)               # Scale to have original mean and variance
}

# Set the seed and sample size
set.seed(1); N <- 1000
```

Example 1: Non-normal x

```
x <- rnorm(N)^2
y <- x + rnorm(N)

ty <- twoStep(y)
tx <- twoStep(x)

m1 <- lm(y ~ x)
m2 <- lm(ty ~ x)
m3 <- lm(ty ~ tx)

screenreg(list(m1, m2, m3))
```

```
##
## ================================================
##               Model 1      Model 2      Model 3
## ------------------------------------------------
## (Intercept)    -0.04         0.14 **      0.20 ***
##                (0.04)       (0.05)       (0.05)
## x               1.02 ***     0.86 ***
##                (0.02)       (0.03)
## tx                                        0.80 ***
##                                          (0.03)
## ------------------------------------------------
## R^2             0.69         0.48         0.43
## Adj. R^2        0.69         0.48         0.43
```

```
## Num. obs.    1000         1000         1000
## RMSE            1.04         1.34         1.41
## ================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Example 2: Non-normal error

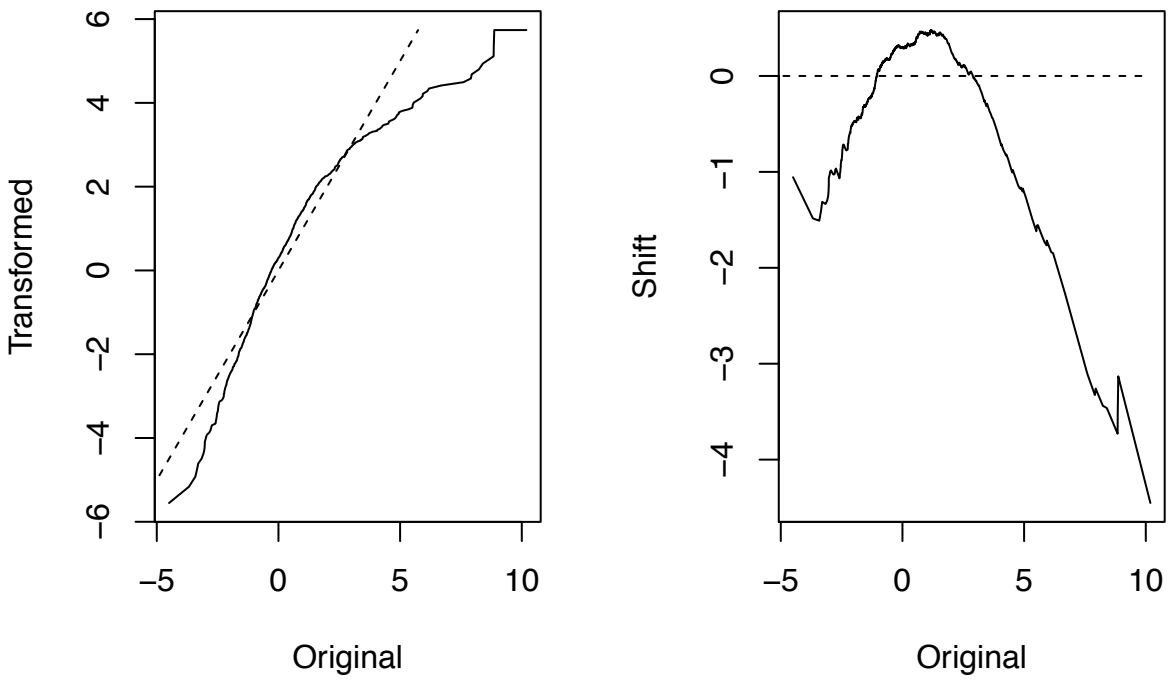
```
x <- rnorm(N)
# Subtract 1 from the error term to center at zero
u <- rnorm(N)^2 - 1
y <- x + u

ty <- twoStep(y)
tx <- twoStep(x)

m1 <- lm(y ~ x)
m2 <- lm(ty ~ x)
m3 <- lm(ty ~ tx)

screenreg(list(m1, m2, m3))
```

```
##
## ================================================
##              Model 1      Model 2      Model 3
## ------------------------------------------------
## (Intercept)     0.08         0.08         0.08
##                (0.05)       (0.04)       (0.04)
## x               0.99 ***     1.18 ***
##                (0.05)       (0.04)
## tx                                        1.18 ***
##                                          (0.04)
## ------------------------------------------------
## R^2             0.31         0.45         0.45
## Adj. R^2        0.31         0.45         0.45
## Num. obs.    1000         1000         1000
## RMSE            1.51         1.36         1.36
## ================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Analysis of the source and direction of the bias

```
# First construct the transformation function

o <- order(y)

par(mfrow = c(1,2))
plot(y[o], ty[o], type ="l", ylab="Transformed", xlab = "Original")
lines(x = c(min(ty),max(ty)), y = c(min(ty),max(ty)), lty = 2)

shift <- ty - y
```

```
plot(y[o], shift[o], type ="l", ylab="Shift", xlab = "Original")
lines(x = c(min(ty),max(y)), y = c(0,0), lty = 2)
```



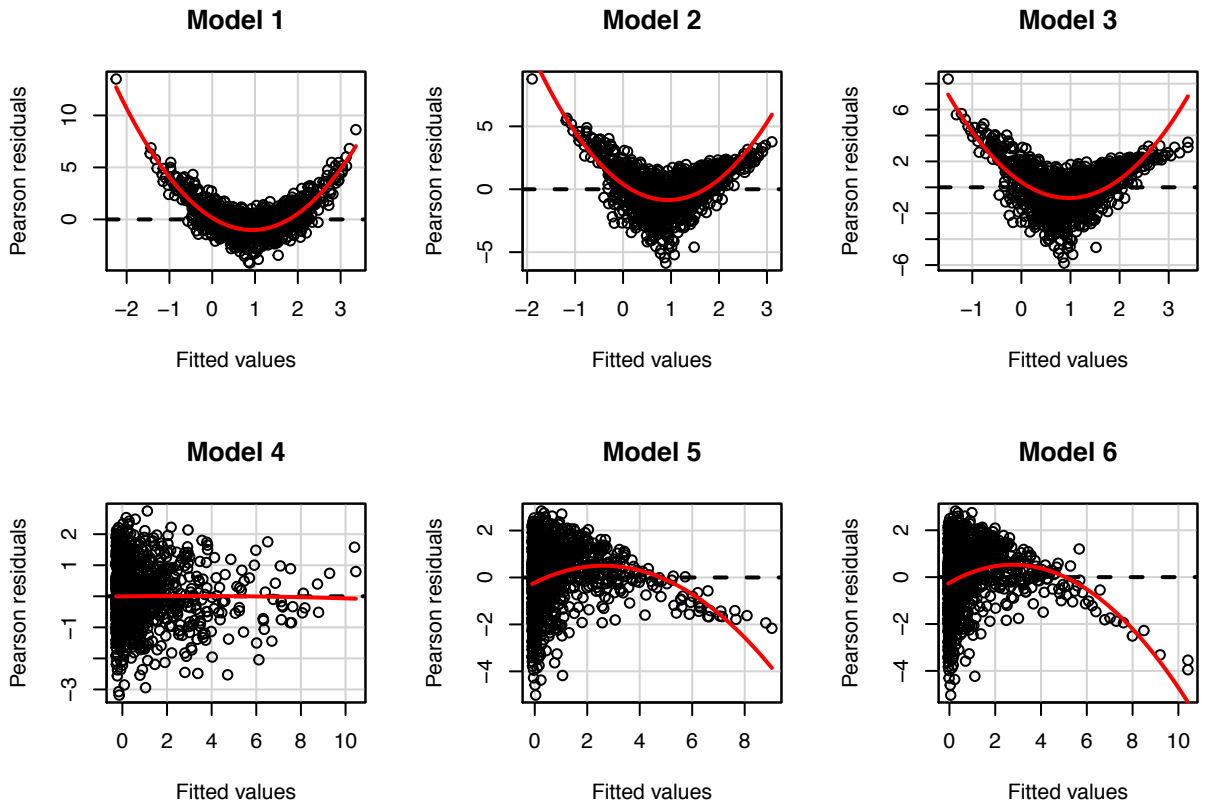

```
# Print covariances
round(cov(cbind(x, u ,y, ty, shift)), digits =2)
```

```
##           x     u     y   ty shift
## x      1.06 -0.01  1.05 1.26  0.21
## u     -0.01  2.29  2.28 1.92 -0.36
## y      1.05  2.28  3.33 3.18 -0.15
## ty     1.26  1.92  3.18 3.33  0.15
## shift  0.21 -0.36 -0.15 0.15  0.31
```

Example 3: Nonlinear relationship

```
x <- rnorm(N)
y <-x+x^2+ rnorm(N)

ty <- twoStep(y)
tx <- twoStep(x)

m1 <- lm(y ~ x)
m2 <- lm(ty ~ x)
m3 <- lm(ty ~ tx)
m4 <- lm(y ~ x + I(x^2))
m5 <- lm(ty ~ x + I(x^2))
m6 <- lm(ty ~ tx + I(tx^2))

# Diagnostics plots

par(mfrow=c(2,3))
```

```r
residualPlot(m1, main = "Model 1")
residualPlot(m2, main = "Model 2")
residualPlot(m3, main = "Model 3")
residualPlot(m4, main = "Model 4")
residualPlot(m5, main = "Model 5")
residualPlot(m6, main = "Model 6")
```




```r
screenreg(list(m1, m2, m3, m4, m5, m6))
```

```
##
## ============================================================================================
##                Model 1      Model 2      Model 3      Model 4      Model 5      Model 6
## --------------------------------------------------------------------------------------------
## (Intercept)     0.96 ***     0.97 ***     0.97 ***    -0.05         0.11 *       0.14 **
##                (0.05)       (0.06)       (0.06)       (0.04)       (0.05)       (0.05)
## x               0.87 ***     0.78 ***                  0.97 ***     0.86 ***
##                (0.05)       (0.06)                    (0.03)       (0.04)
## tx                                        0.80 ***                                0.81 ***
##                                          (0.06)                                  (0.04)
## I(x^2)                                                 1.04 ***     0.88 ***
##                                                       (0.02)       (0.03)
## I(tx^2)                                                                           0.85 ***
##                                                                                  (0.03)
## --------------------------------------------------------------------------------------------
## R^2             0.20         0.16         0.17         0.74         0.54         0.53
## Adj. R^2        0.20         0.16         0.17         0.74         0.54         0.53
## Num. obs.    1000         1000         1000         1000         1000         1000
## RMSE            1.71         1.75         1.75         0.98         1.30         1.31
```

```
## ================================================================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

Inference

```
r <- replicate(10000,{
  x <- rnorm(N)
  y <- x + rnorm(N)^2
  ty <- twoStep(y)

  m1 <- lm(y ~ x)
  m2 <- lm(ty ~ x)
  x <- twoStep(x)
  m3 <- lm(ty ~ x)

  # Return estimates and their estimated variances
  c(coef(m1)[2], vcov(m1)[2,2],
    coef(m2)[2], vcov(m2)[2,2],
    coef(m3)[2], vcov(m3)[2,2])

})

round(matrix(c(var(r[1,]),
               var(r[3,]),
               var(r[5,]),
               mean(r[2,]),
               mean(r[4,]),
               mean(r[6,])),3,2), digits = 4)
```

```
##        [,1]   [,2]
## [1,] 0.0020 0.0020
## [2,] 0.0035 0.0016
## [3,] 0.0035 0.0016
```

Empirical example of transformation from Wooldridge.

```
#The data are from the crs package
library(crs)
```

```
## Categorical Regression Splines (version 0.15-27)
## [vignette("crs_faq") provides answers to frequently asked questions]
```

```
library(effects)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'carData'
```

```
## The following objects are masked from 'package:car':
##
##     Guyer, UN, Vocab
```

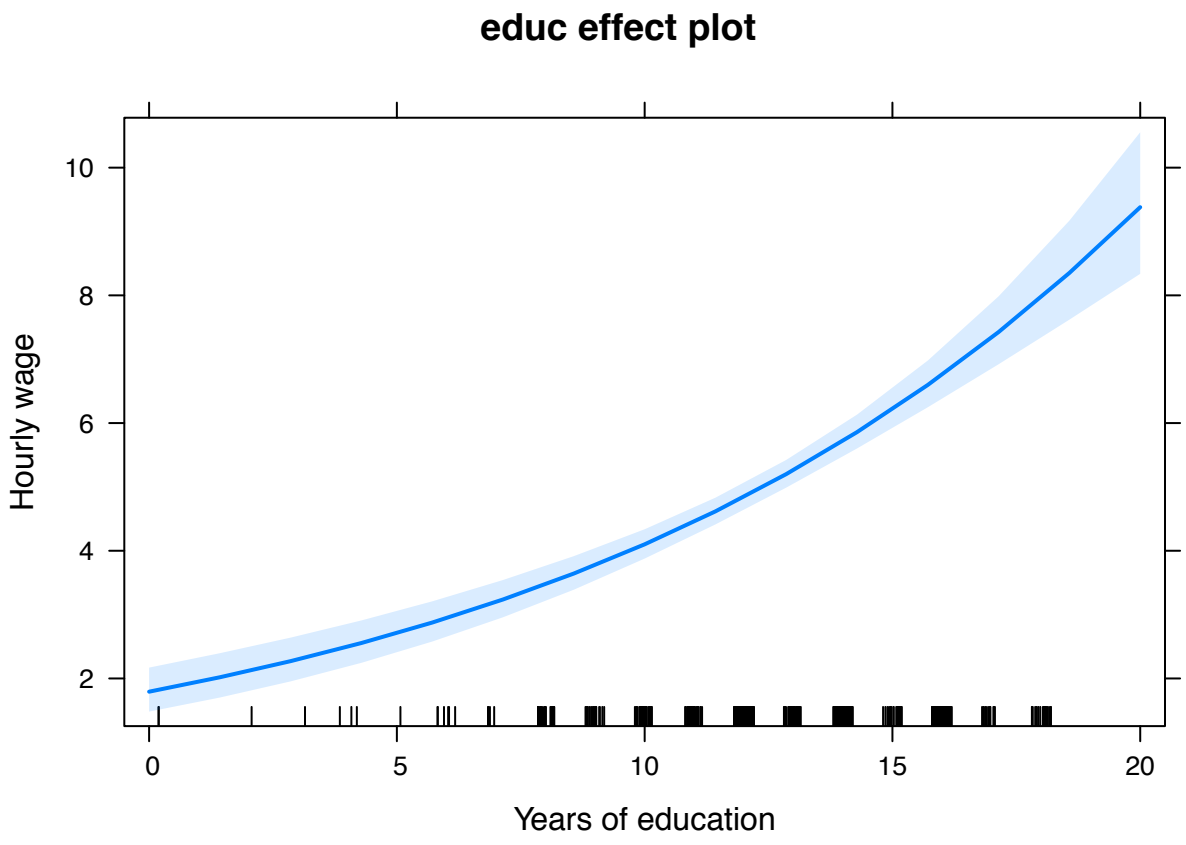
```
## lattice theme set by effectsTheme()
```

```
## See ?effectsTheme for details.
```

```
data("wage1")
attach(wage1)
m1 <- lm(log(wage) ~ educ)
screenreg(m1)
```

```
##
## ========================
##               Model 1
## ------------------------
## (Intercept)     0.58 ***
##                (0.10)
## educ            0.08 ***
##                (0.01)
## ------------------------
## R^2             0.19
## Adj. R^2        0.18
## Num. obs.     526
## RMSE            0.48
## ========================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

```
par(mfrow=c(1,1))
# Marginal effects plot
plot(effect("educ",m1, transformation = list(link = log, inverse = exp)), type ="response",
     ylab ="Hourly wage", xlab="Years of education")
```



Online supplement 2: Monte Carlo simulations of the examples

```
# Two-step transformation, as presented by Templeton et al.

twoStep <- function(x){
  prank <- rank(x)/length(x)
  prank[prank == 1] <- 1-1/length(x)  # A workaround for percentile rank = 1.
  r <- qnorm(prank)
  r/sd(r)*sd(x)+mean(x)               # Scale to have original mean and variance
}

# Set the seed, sample size, and number of replications
set.seed(1); N <- 1000; rep <- 10000
```

Example 1: Non-normal x

```
r <- replicate(rep,{
  x <- rnorm(N)^2
  y <- x + rnorm(N)
  ty <- twoStep(y)

  m1 <- lm(y ~ x)
  m2 <- lm(ty ~ x)
  x <- twoStep(x)
  m3 <- lm(ty ~ x)

  # Return estimates and their estimated variances
  c(coef(m1)[2], vcov(m1)[2,2],
    coef(m2)[2], vcov(m2)[2,2],
    coef(m3)[2], vcov(m3)[2,2])

})

m <- matrix(c(mean(r[1,]), mean(r[3,]), mean(r[5,]),
              var(r[1,]), var(r[3,]), var(r[5,]),
              mean(r[2,]), mean(r[4,]), mean(r[6,]),
              var(r[2,]), var(r[4,]), var(r[6,])),3,4)

# Print mean and variance of estimates and their estimated variances
colnames(m) <- c("mean est x", "var est x", "mean est x var", "var est x var")
rownames(m) <- c("Original", "Two-step y", "Two-step x and y")

m
```

```
##                  mean est x    var est x mean est x var var est x var
## Original          1.0000942 0.0004912537   0.0005073543  4.244487e-09
## Two-step y        0.8418287 0.0008196411   0.0007991223  2.721078e-09
## Two-step x and y  0.7901424 0.0008151196   0.0008836465  4.210087e-09
```

Example 2: Non-normal error

```
r <- replicate(rep,{
  x <- rnorm(N)
  y <- x + rnorm(N)^2
  ty <- twoStep(y)

  m1 <- lm(y ~ x)
  m2 <- lm(ty ~ x)
  x <- twoStep(x)
  m3 <- lm(ty ~ x)

  # Return estimates and their estimated variances
  c(coef(m1)[2], vcov(m1)[2,2],
    coef(m2)[2], vcov(m2)[2,2],
    coef(m3)[2], vcov(m3)[2,2])

})

m <- matrix(c(mean(r[1,]), mean(r[3,]), mean(r[5,]),
              var(r[1,]), var(r[3,]), var(r[5,]),
              mean(r[2,]), mean(r[4,]), mean(r[6,]),
              var(r[2,]), var(r[4,]), var(r[6,])),3,4)

# Print mean and variance of estimates and their estimated variances
colnames(m) <- c("mean est x", "var est x", "mean est x var", "var est x var")
rownames(m) <- c("Original", "Two-step y", "Two-step x and y")

m
```

```
##                  mean est x   var est x mean est x var var est x var
## Original          0.9997521 0.002032579    0.002008972  6.315940e-08
## Two-step y        1.1836703 0.003402467    0.001605224  3.205844e-08
## Two-step x and y  1.1828521 0.003384242    0.001607182  3.217682e-08
```

Example 3: Nonlinear relationship

```
r <- replicate(rep,{
  x <- rnorm(N)
  y <-x+x^2+ rnorm(N)

  ty <- twoStep(y)

  m1 <- lm(y ~ x)
  m2 <- lm(ty ~ x)
  m4 <- lm(y ~ x + I(x^2))
  m5 <- lm(ty ~ x + I(x^2))

  x <- twoStep(x)

  m3 <- lm(ty ~ x)
```

```r
  m6 <- lm(ty ~ x + I(x^2))

  # Return estimates and their estimated variances
  c(coef(m1)[2], vcov(m1)[2,2],
    coef(m2)[2], vcov(m2)[2,2],
    coef(m3)[2], vcov(m3)[2,2],
    coef(m4)[2], vcov(m4)[2,2], coef(m4)[3], vcov(m4)[3,3],
    coef(m5)[2], vcov(m5)[2,2], coef(m5)[3], vcov(m5)[3,3],
    coef(m6)[2], vcov(m6)[2,2], coef(m6)[3], vcov(m6)[3,3])

})

# Print mean and variance of estimates and their estimated variances
m <- matrix(c(mean(r[1,]), mean(r[3,]), mean(r[5,]),
              var(r[1,]), var(r[3,]), var(r[5,]),
              mean(r[2,]), mean(r[4,]), mean(r[6,]),
              var(r[2,]), var(r[4,]), var(r[6,]),
              mean(r[7,]), mean(r[11,]), mean(r[15,]),
              var(r[7,]), var(r[11,]), var(r[15,]),
              mean(r[8,]), mean(r[12,]), mean(r[16,]),
              var(r[8,]), var(r[12,]), var(r[16,]),
              mean(r[9,]), mean(r[13,]), mean(r[17,]),
              var(r[9,]), var(r[13,]), var(r[17,]),
              mean(r[10,]), mean(r[14,]), mean(r[18,]),
              var(r[10,]), var(r[14,]), var(r[18,])),3,12)

colnames(m) <- c("mean est x", "var est x", "mean est x var", "var est x var",
                 "mean est x", "var est x", "mean est x var", "var est x var",
                 "mean est x2", "var est x2", "mean est x2 var", "var est x2 var")
rownames(m) <- c("Original", "Two-step y", "Two-step x and y")

# Misspecified models
m[,1:4]
```

```
##                 mean est x   var est x mean est x var var est x var
## Original          1.0005600 0.010836474    0.002993815  3.553639e-08
## Two-step y        0.8801684 0.007737571    0.003223797  4.159906e-08
## Two-step x and y  0.8863784 0.005856969    0.003214689  4.419034e-08
```

```r
# Correctly specified models that include interactions
m[,5:12]
```

```
##                 mean est x   var est x mean est x var var est x var
## Original          1.0000497 0.001026762    0.001007681  4.200643e-09
## Two-step y        0.8810495 0.001340722    0.001904941  2.987884e-08
## Two-step x and y  0.8642437 0.002378002    0.001899727  1.634617e-08
##                 mean est x2   var est x2 mean est x2 var var est x2 var
## Original          0.9998288 0.0005090145    0.0005094466   4.155089e-09
## Two-step y        0.8179674 0.0010222612    0.0009570562   6.684633e-09
## Two-step x and y  0.8242887 0.0009728987    0.0009797492   6.604082e-09
```