Visual Assessment of Residual Plots in Multiple Linear Regression: A Model-Based Simulation Perspective

Hongwei Yang
University of Kentucky

This article follows a recommendation from the regression literature to help regression learners become more experienced with residual plots for identifying assumption violations in linear regression. The article goes beyond the usual approach to residual displays in standard regression texts by taking a model-based simulation perspective: simulating the data from a generating model and using them to estimate an analytical model. The analytical model is a first order linear regression model; whereas the generating model violates the assumptions of the analytical model. The residuals from the analytical model are plotted to demonstrate assumption violations to provide experience for regression learners with characterized residual patterns. The article also briefly discusses remedial measures.

Multiple regression models answer questions about the relationship between a response variable and one or more predictors. Such models start with a collection of potential predictors. Some of these predictors may be continuous measurements, like miles traveled or salary. Some may be discrete, but ordered, such as the number of employees in a company. Other potential predictors can be categorical, like gender or hair color. All these types of potential predictors can be useful in a multiple regression model.

From the pool of potential predictors, terms, or model effects, can be created. The terms are the x variables that appear in the multiple regression model in Equation 1

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad (1)$$

where $\beta_1, \beta_2, \ldots, \beta_k$ are model parameters, $x_1, x_2, \ldots, x_k$ are terms or model effects, and $\varepsilon_i \sim N(0,\sigma^2)$ are conditional on all predictors with σ^2 being an unknown, but with common variance $i = 1, 2, \ldots, n$. The relationship is stochastic in the sense that the model is not exact, but subject to random variation, as expressed in the error term ε. In a stochastic relationship, the values of one set of variables (predictors) determine the (conditional) probability distribution of another variable (response) (Goldberger, 1968). The part of the model without the random error is its deterministic component, which is also known as the mean function, as it describes the mean of y conditional on values of all predictors.

In Equation 1, each term x_i, $i = 1, 2, \ldots, k$, could be one of the followings: 1) a quantitative predictor in its original or transformed metric, 2) an interaction of two or more predictors, 3) a higher order term of a predictor, and 4) a dummy indicator for a qualitative predictor. A regression with k predictors may combine to give fewer than k terms or expand to require more than k terms (Weisberg, 2005).

Despite the fact that the number of terms in Equation 1 could be either greater than or smaller than the number of predictors, a commonly used type of model is one where the former equals the latter and each predictor is a term and vice versa. This type of model usually features the absence of such terms as interactions or polynomials because the principle of hierarchy requires that a polynomial term of a given order and/or an interaction term not be retained in the model unless all related terms also stay in the model; thus, leading to more terms than predictors. Because the functional form of Equation 1 is already linear in parameters, the absence of both interactive and polynomial terms causes the model to be also linear in terms and their effects to be additive. Thus, making this type of model a first-order regression model where the response is linearly related to each parameter and each term (predictor) and the effects of all terms (predictors) are additive.

A first order regression model as described above can be estimated using the method of ordinary least squares (OLS), given that the underlying model assumptions are satisfied to an acceptable level. The OLS method is a model fitting mechanism that is based on minimizing the residual sum of squares (RSS), and it is capable of estimating the (linear-in-parameter) model as described by Equation 1 regardless of whether or not the model is first order. That is, even when the model by Equation 1 stops being first order by containing polynomial (nonlinear) and/or interaction (non-additive) terms like in the case of general linear models, the OLS method can also serve as its estimator.

How well the ordinary least squares method performs in estimating an applicable linear regression

model largely counts on the extent to which the assumptions of the model are satisfied. However, although linear regression is widely used, research has indicated it is also probably the most abused in terms of ignoring the underlying assumptions of the model (Berry, 1993). Usually without being certain in advance that a regression model is appropriate for an application, the model is considered for that application anyway. This is usually at the cost of one, or several, of the assumptions of the model such as the normality assumption for the error term being violated or being inappropriate for the particular data at hand. Then, after the model is estimated using the data from the application, people fail to test whether the underlying assumptions are reasonably satisfied, which renders the final results questionable. After that, despite likely assumption violations, results from the regression analysis are still interpreted, reported, and published. This is one of many types of common scenario regression abuse that is rooted in failure to attend to regression assumptions.

Therefore, a correct understanding of regression assumptions is critical for regression learners to appreciate weaknesses and strengths of his or her estimates from a regression model. These assumptions need to be checked using regression diagnostics. Diagnostic techniques can be graphical, which are more flexible, but harder to definitely interpret or numerical (statistical tests), which are narrower in scope, but require no intuition (Faraway, 2004). Both types of diagnostic techniques are usually based on an analysis of residuals because their analysis is an effective way to discover several types of model inadequacies including assumption violations (Montgomery, Peck, & Vining, 2001). In the literature, some argue the final judgment on model adequacy should be based on statistical tests of residuals, instead of their graphical displays (Mendenhall & Sincich, 2003); whereas others insist residual plots are much more useful than formal testing procedures when assessing model adequacy (Chatterjee & Hadi, 1988). Despite the difference in opinions, both approaches remain the most commonly used regression diagnostic tools. This article primarily examines basic residual plotting for checking model assumptions.

With that said, it is important to train regression learners to effectively judge plots of residuals to properly evaluate the validity of assumptions for a fitted regression model. To that end, the article resorts to computer simulation. This is because simulation is widely recognized as a useful tool for teaching statistics and an aid to learning statistical methods (Burton, Altman, Royston, & Holder, 2006; Hodgson & Burke, 2000; Mooney, 1995). Particularly, simulating data can help people visualize abstract statistical concepts, and to see dynamic processes, rather than statistic figures and illustrations (Dupuis & Garfield, 2010). On the other hand, it has long been recognized that there can be a lot of slippage between the model assumed and the right/true model (Berk, 2004). With real world applications, the right model underlying the data is almost never known. So, given a mismatch between the assumed model and the true model that is evidenced by characterized residual patterns, even experienced regression experts may find it difficult to properly bridge the gap between the pattern of residuals and the remedial measures to take, not to mention inexperienced regression learners.

Therefore, Faraway (2004) recommends that artificial plots, where the true relationship is known, be generated using computer simulation to help regression learners gain some prior experience in judging residual plots. That way, when a characterized shape in a residual plot indicates a problem (i.e., assumption violation), the cause of that problem can be easily spotted; thus, providing experience for regression learners. Unfortunately, although artificial generation of plots is a good way to help people become experienced in judging residual plots, a review of literature indicates that standard regression texts seldom bother to cover this topic (Belsley, Kuh, & Welsch, 1980; Berk, 2004; Chatterjee & Hadi, 1988; Chatterjee & Price, 1991; Cook, 1998; Draper & Smith, 1998; Fox, 2008; Gelman & Hill, 2007; Goldberger, 1968; Kutner, Nachtsheim, Neter, & Li, 2005; Mendenhall & Sincich, 2003; Montgomery et al., 2001; Rao & Toutenburg, 1999; Weisberg, 2005; Yan & Su, 2009). Almost always, standard regression texts just use residuals from real world applications where the true model is unknown and/or present prototype residual plots with no information provided regarding the structure of the data (the true model) and the assumed model to which the data are fitted. In the absence of the true relationship between the two models, such plots are limited in providing regression learners with the type of experience that they will need for properly calibrating residual patterns. Fortunately, there are exceptions. Faraway (2004) and Grob (2003) go beyond those usual approaches to displaying residuals. They computer-simulate the data to make available the information about both the assumed model and the true model to help people obtain a better feel for characterized residual behaviors. Even so, their simulation is limited mainly in two aspects: 1) restricted to only one or two assumptions and 2) without presenting the characterized pattern

in a dynamic process or as a function of other critical factors (e.g., level of autocorrelation, etc.). Such limitations are disappointing, so this article aims to extend their works by improving on those aspects.

In sum, this article revisits the topic of residual analysis in assessing regression model assumptions through a model-based simulation perspective. This simulation approach supplements the existing literature that primarily uses real world applications and/or prototype plots for demonstrating residual pattern as a function of assumption violation. A primary goal of this article is to help regression learners become more experienced in judging the shape of a residual plot. To that end, the article proposes that the data be simulated with a known structure from a generating model (GM) and then fitted to an analytical model (AM) that differs from the generating model in various aspects: distribution of the error, the relationship between the response and each predictor, and the relationship between predictors. With the residuals obtained from the analytical model, which is mis-specified relative to the generating model; thus, making the former model inappropriate for the data produced by the latter model, it becomes possible to pin-point the effect of the violation of one or more assumptions (of the analytical model) on the tendency exhibited in the residuals of the analytical model.

The article is organized as follows. The section that immediately follows provides an overview of statistical assumptions that are considered to be critical in the literature for the model described by Equation 1 when it is first-order and is estimated using ordinary least squares. Next, this section also presents the generating model for simulating the data with a given structure along with the analytical model that produces the residuals. The residuals are plotted accordingly in certain basic, but common ways (i.e., against predicted values, each predictor) to produce characterized patterns of assumption violations and a discussion of the patterns as a function of assumption violations is provided. Given this for each assumption, the article proceeds to another section that discusses more issues pertaining to residual analysis including remedial measures. The article concludes with a summary of the study.

Critical Assumptions and Assessment Under Simulations

In this section, the article examines several assumptions considered to be critical for the model described by Equation 1 when it is first-order, where there is a one-to-one correspondence between terms and predictors (i.e., each term is a predictor and vice versa), and when it is the ordinary least squares method that is used as its estimator. This article discusses each assumption before providing a generating model for simulating the data with a known structure and fitting the data to a first-order, analytical model from Equation 1 to produce residuals. The residuals are next presented graphically to demonstrate characterized patterns from each assumption violation. This process is performed separately for each critical assumption to be examined.

In Gelman and Hill (2007), several of the most critical assumptions are presented in order of importance for the model described by Equation 1 when it is first order: 1) validity, 2) additivity and linearity (i.e., two assumptions counted as one), 3) independence of errors, 4) equal variance of errors, and 5) normality of errors. The article examines all these statistical assumptions using Monte Carlo simulations with the exception of the first one that focuses on validity of the data. The validity assumption usually includes three aspects: 1) the response should accurately reflect the phenomenon of interest, 2) the model should include all relevant predictors, and 3) the model should generalize to the case to which it will be applied. Clearly, this is a very important assumption. However, it is so broad that it cannot be practically demonstrated and tested using the available data alone. Therefore, the validity assumption is omitted from this article due to its great breadth, but not because it is unimportant. Then, this article is left with the remaining four critical assumptions. That they are critical is also endorsed by other works in the literature (Kutner et al., 2005; Tamhane & Dunlop, 1999).

The entire analysis is coded in R. The R code is available upon request. In all simulations, the sample size is fixed at 2000, or $n = 2000$. The random seed that the simulation process begins with is randomly set at 1024. During the process, the random seed for assessing the independence of errors assumption is set at 12345. Finally, coefficients for all generating models are all randomly picked up.

Assumptions of Linearity and Additivity

Before proceeding to the main argument, it is worthwhile to clarify a few mathematical concepts (Goldberger, 1968). Consider first the function of one variable: $y = f(x)$. We say that it is linear in x if and only if $d(y)/dx$ does not involve x. That is, if and only if $d(d(y)/dx)/x = 0$, and if and only if the effect of a

given change in x does not depend on the level of x itself. Consider next the function of two independent variables: $y = f(x_1,x_2)$. We say that it is linear in x_1 if and only if $\partial(y)/\partial x_1$ does not involve x_1. That is, if and only if $\partial(\partial(y)/\partial x_1)/\partial x_1 = 0$, and if and only if the effect of a given change in x_1 does not depend on the level of x_1 itself. Similarly, we say that it is linear in x_2 if and only if $\partial(y)/\partial x_2$ does not involve x_2. That is, if and only if $\partial(\partial(y)/\partial x_2)/\partial x_2 = 0$, and if and only if the effect of a given change in x_2 does not depend on the level of x_2 itself. Thus, essentially the same concept of linearity applies whether there are one or more independent variables. But, a new concept also applies. We can say that $y = f(x_1,x_2)$ is additive in x_1 and x_2 if and only if $\partial(y)/\partial x_1$ does not involve x_2 and $\partial(y)/\partial x_2$ does not involve x_1. That is, if and only if $\partial(\partial(y)/\partial x_1)/\partial x_2 = 0 = \partial(\partial(y)/\partial x_2)/\partial x_1$, and if and only if the effect of a given change in each of the independent variables does not depend upon the level of the other. Additivity is an appropriate name for this feature since it means that the combined effect of given changes in both variables can be obtained by adding together the separately computed effects of the given changes in each of them. Also, the extension to the case of a function of many variables is straightforward. The concepts can be formulated in terms of finite changes to cover the case where derivatives are not defined.

Based on the above description regarding linearity and additivity, it is clear that a first order model described by Equation 1 is both linear and additive when it comes to the relationship between the response and each predictor. In such a model, two things are true, respectively for linearity and additivity: 1) the expected change in the response associated with a one-unit increase in a predictor when holding constant the values of all other predictors, is the same regardless of the value of that predictor (i.e., a property of linearity) and 2) the expected change in the response associated with a one-unit increase in a predictor when holding constant all other predictors, remains the same regardless of the values of the other predictors that are being held constant (i.e., a property of additivity). That the first order model is linear in each predictor is due to the fact that each and every predictor is included in the model as a first-order term. However, linearity between the response and a predictor x_i, $i = 1, 2, \ldots, k$, may be violated if at least one term in the model involves a transformation of the predictor, such as x_i^2, $1/x_i$, $\log(x_i)$, etc. By contrast, that the first order model is additive in the effects of the k predictors is due to the fact that no term in the model is in the form of an interaction between two or more of those predictors. However, the effects of two or more predictors may no longer be additive if at least one term in the model is the product (or some other functions like ratio) of two or more of those predictors.

Since a first order model in Equation 1 is assumed to be both linear and additive, after it is fitted to a data set with a certain structure using ordinary least squares, what should be done to evaluate the extent to which the two assumptions are satisfied? An assessment of the two assumptions involves multiple scatter plots of residuals in two categories: 1) residuals against each predictor and 2) residuals against the fitted values.

To assess the linearity assumption, the residuals are to be plotted against each predictor to determine whether or not the corresponding relationship is linear. In each plot, randomness of residuals over the range of the observed predictor values is needed to support a linear relationship between the response and the corresponding predictor. Any systematic pattern of the residuals, particularly a parabolic-shape or bow-shape pattern, is a clear indicator of nonlinearity. Usually, given a violation of the linearity assumption, the residual-versus-predicted plot tends to exhibit similar patterns as each residual-versus-predictor plot. This is due in part to the fact that the predicted values are themselves a linear combination of all involved predictors when the data are fitted to the linear-in-parameter model by Equation 1.

Linearity Assumption

To show how a violation of the linearity assumption renders the residual plots nonrandom/systematic, the article proposes the generating model in Equation 2 to simulate data with a known nonlinear relationship between y and x_1.

$$y = 0.5179 + [-0.0040(x_1 - \bar{x}_1)] \pm 0.4184(x_1 - \bar{x}_1)^2 + \varepsilon\,, \qquad (2)$$

where $x_1 \sim N(2,3^2)$ and $\varepsilon \sim N(0,3^2)$. In Equation 2, the response is simulated from a model establishing a nonlinear relationship between y and x_1 under two cases: 1) a positive x_1^2 coefficient, or $(+)x_1^2$, and 2) a negative x_1^2 coefficient or $(-)x_1^2$. That the relationship is nonlinear is because Equation 2 is a 2nd order polynomial allowing curvature in the change of y as a function of x_1. The x_1 data are mean-centered to avoid potentially strong correlation between the 1st order and the 2nd order model effects.

Next, the data are fitted to an analytical model that assumes a linear relationship between y and x_1 as described by Equation 3:

$$y = \beta_0 + \beta_1(x_1 - \bar{x}_1) + \varepsilon\,, \tag{3}$$

where ε follows a conditional normal distribution with the mean being zero and the variance being unknown, but constant over all settings of the only predictor: x_1. Then, the residuals from the fitted analytical model are plotted against the centered predictor (x_1 - $\bar{x}_1$). Additionally, they are also plotted against the predicted values from the model. The resulting graphical outputs are found in Figure 1.

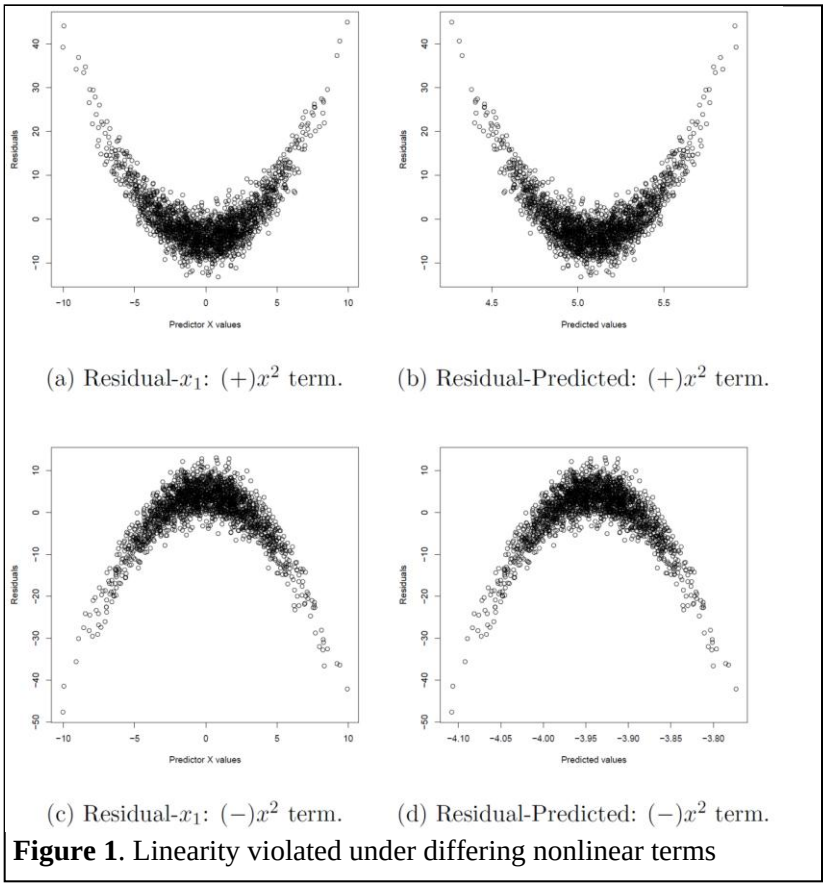


Figure 1. Linearity violated under differing nonlinear terms

As is indicated in Figure 1, fitting the data with a nonlinear relationship between y and x_1 to a model assuming a linear relationship between the two variables, causes both the residual-versus-predictor and the residual-versus-predicted plots to show systematic patterns in the shape of bows. Given $(+)x_1^2$, both plots open upward whereas with $(-)x_1^2$, they open downward. Regardless of the sign of the x_1^2 term, such bow-shape patterns in both plots indicate there is additional variability in the response that cannot be sufficiently explained by the fitted analytical model that assumes a linear relationship between y and x_1. This is not surprising because the data are simulated with curvature in it whereas the analytical model assumes no curvature at all. Therefore, fitting such curvature-present data to this curvature-absent model violates the linearity assumption of the model, thus leading to systematic patterns in both residual plots.

Additivity Assumption

To show how a violation of the additivity assumption causes the Residual plots to be of systematic patterns, the article proposes the generating model in Equation 4 to simulate data with known non-additive (interactive) effects of predictors:

$$y = -0.0814 + (-0.4927)x_1 + (0.4946)x_2 + (-0.0776)x_1x_2 + \varepsilon\,, \tag{4}$$

where $x_1 \sim N(2,3^2)$, $x_2 \sim N(0,0.3^2)$ and $\varepsilon \sim N(0,\sigma^2)$ with $\sigma^2 = 1^2$ for large error variability, $\sigma^2 = 0.08^2$ for medium error variability, and $\sigma^2 = 0.01^2$ for small error variability. In Equation 4, the response data are simulated from an interactive (non-additive) model where the effect of $x_1(x_2)$ on the response depends on the value of $x_2(x_1)$; hence, the effects of the two predictors are not additive, but are interactive. The interactive effect is presented in the form of a product of x_1 and x_2. This simulation is performed for the three different values of the error variance; hence, leading to three sets of simulated data.

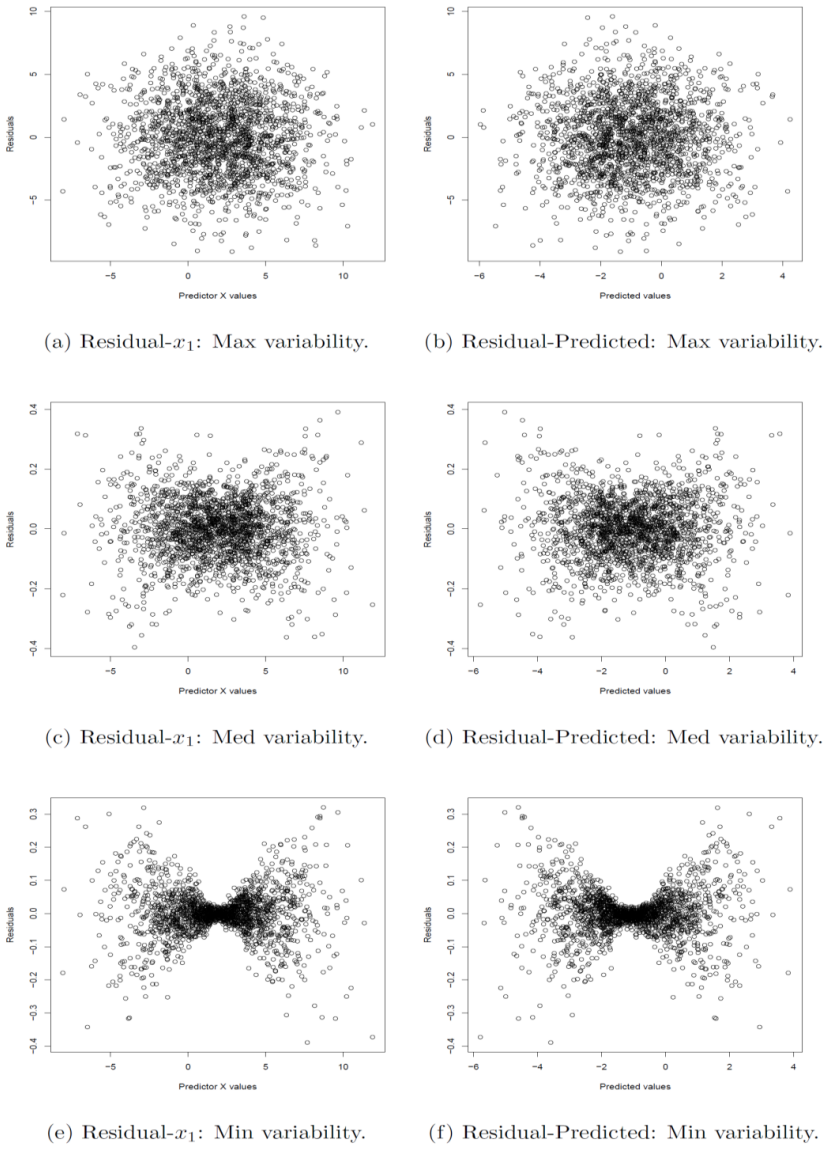
Each set of data simulated from the non-additive model is fitted to an analytical model in Equation 5 where the effects of the two predictors on the response are additive:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon\,, \tag{5}$$

where ε follows a conditional normal distribution with the mean being zero and the variance being unknown, but constant over all settings of the two predictors: x_1 and x_2. Next, the residuals from the fitted

model are plotted against each of the two predictors, x_1 and x_2. They are also plotted against the predicted values from the fitted model. This is done for all three sets of simulated data. The resulting graphical outputs are in Figure 2. The residual plots against x_2 are omitted due to their high similarity to those against x_1.

As is indicated in Figure 2, fitting the data simulated from a non-additive (interactive) model to an additive model, results in systematic patterns in both the residual-versus-predictor and the residual-versus-predicted plots. Such patterns are more pronounced under small variability of the error than under large variability of the error. Under Figures 2a and 2b, the patterns in the two residual plots are hard to detect, despite a clear violation of the additive assumption for Equation 5. This indicates that residual plots are not always effective in identifying interaction effects. As the error variance decreases from 1^2 to 0.08^2, and further down to 0.01^2, the transition from less-marked to more-marked systematic patterns is evidently observed in both residual plots.



(a) Residual-x_1: Max variability. (b) Residual-Predicted: Max variability.

(c) Residual-x_1: Med variability. (d) Residual-Predicted: Med variability.

(e) Residual-x_1: Min variability. (f) Residual-Predicted: Min variability.

Figure 2. Additivity violated under differing amount of error variability

Assumption of Independence of Errors

The independence of observations assumption for Equation 1 requires that the error terms for any two observations be uncorrelated. A typical scenario where this assumption is likely to be violated is when data are obtained at successive time points for the same sampling unit(s) or time-series data (Tamhane & Dunlop, 1999). According to Mendenhall and Sincich (2003), because time series data tend to follow economic trends and seasonal cycles, the value of a time series at time t is usually indicative of its value at time (t+1) and values at time points that are even farther away. Therefore, the value of a time series at time t is correlated with its value at time (t+1) and values at other further apart time points. If such a series is used to estimate the model in Equation 1 assuming independence of errors under ordinary least squares, the result is that the residuals tend to exhibit long term trends over extended periods of time.

Error terms correlated over time are said to be autocorrelated, and usually measured by ρ, with the autocorrelation parameter ranging from (−1) to (+1). That the residuals are correlated over time is just one of a number of problems derived from fitting time series data set to the model in Equation 1 that assumes independence of errors. Among other problematic issues are misleading statistical test results that lead to overoptimistic evaluations of a model's predictive ability and underestimates of standard errors of

regression coefficients. To more appropriately model time series data and avoid the stated problems, Equation 1 has to be extended to account for a different, usually more complex, structure of the random errors. This extension of the random error structure is typically in the form of a separate equation for errors that is known as the autoregressive error equation involving the use of the autocorrelation parameter ρ to allow for correlated errors over time.

In view of the problems from ignoring autocorrelated errors, it is important that their presence be detected so that an extension of the random error term of Equation 1 for properly modeling a time series is justified and; thus, implemented. To that end, a plot of residuals against time is an effective, though subjective, way of detecting autocorrelated errors. This plot is known as a run chart. Because a run chart of residuals takes into account the time sequence, it is able to make visible time-related patterns if any.

Next, the article proposes a generating model that simulates a time-series data set under first-order autocorrelation measured by ρ. The simulated data are fitted under Equation 1 without extending the random error term to account for such autocorrelation. Residuals from the fitted model are then plotted against time to examine the pattern of residuals as a function of time under a given level of autocorrelation. This process is repeated for six different values of ρ to assess the effect of ρ on the pattern of residuals as a function of time. The six selected values for ρ are 0.00, 0.30, 0.50, 0.90, 0.99, and 1.00, which represent a wide range of levels of autocorrelation from non-existent, to weak, to moderate, to strong, to very strong, and to perfect, respectively. Next, the generating model for simulating the time series data is given by Equations 6 and 7:

$$y_t = 0.2345 + 0.5911x_t + \varepsilon_t, \tag{6}$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \tag{7}$$

where $x_t \sim$ Uniform(0,10), $u_t \sim N(0,1)$, and ρ is an autocorrelation parameter. For each of the six selected values of ρ, a set of 2000 observations is simulated from 2000 time points. Next, each simulated data set is fitted to an analytical model from Equation 1 that assumes error independence. Specifically, the analytical model used here is in Equation 8:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \tag{8}$$

where ε_t follows a conditional normal distribution with the mean being zero and the variance being unknown, but constant over all settings of the only predictor: x_t.

The graphical results are in Figure 3. Each subplot in Figure 3 represents the pattern of residuals from using ordinary least squares to fit the data simulated given a certain level of autocorrelation to Equation 1 that does not assume any autocorrelation. When $\rho = 0$, or no autocorrelation between consecutive time points is indicated, the residuals do not show any systematic patterns. That is, adjacent error terms do not tend to be of the same sign or of the same magnitude causing the pattern plot to be heavily dense around the zero line, which is the mean of the residuals from the ordinary least squares method. In other words, the residuals are not consistently above or below the zero line for extended periods; rather, they fluctuate very rapidly around that line during the entire 2000 time points. Such a non-systematic pattern of residuals becomes less and less obvious as the autocorrelation parameter ρ keeps increasing. Stated differently, the higher the absolute value of the ρ parameter, the more systematic the pattern of residuals will become: Residuals are more and more likely to stay consistently above or below the zero line for longer periods of time. Although such a change from $\rho = 0$ to 0.3 is not quite noticeable, the change from $\rho = 0.3$ to 0.5 is much more phenomenal; where residuals become more and more clustered consistently above/below the mean and the pattern plot becomes gradually less dense around the zero line. However, at this point, it is still difficult to distinguish the periods when the residuals are consistently above the zero line from those when the residuals are consistently below the zero line. Then, when the ρ parameter jumps from 0.50 to 0.90, such a distinction between those two types of patterns (above/below the zero line) could already be made; though, with some efforts. When ρ increases to as high as 0.99, it is easy to tell that from the very beginning up to the 900^{th} time point, the residuals are almost always consistently above the zero line with some exceptions at around the 500^{th} time point; whereas most of the remaining residuals during the remaining time periods are consistently below the zero line. For the case of $\rho = 1$, the residuals remain consistently negative from the beginning up to about the 250^{th} time point. Then, from the 250^{th} time point up to the 1250^{th} time point, almost all residuals are consistently positive and during the remaining period, all residuals are again consistently negative.

Therefore, the autocorrelation parameter controls the pattern of residuals in the residual-against-time plot. The higher the absolute value of the ρ parameter, the more systematic the residual plot will become either consistently positive or consistently negative or a combination of both over extended periods of time. Such observed time trends of residuals that come from the fitted analytical model in Equation 1, assuming independence of errors, indicate that this assumption is violated. So, this analytical model does not provide a valid representation of the autocorrelated data in terms of the problems previously described such as misleading statistical test results of regression coefficients. Therefore, an extension of the random error term of Equation 1 is needed to better account for the correlation of errors over time.

In the end, it should be noted that time order is not the only factor likely to cause the assumption of independence of observations to be violated. According to Tamhane and Dunlop (1999), even cross-sectional data (i.e., data collected on different units at about the same time) may be correlated because of spatial proximity, family relationships, or other reasons. Such autocorrelation, due to spatial clustering of sampling units, needs also to be taken into account appropriately when building the model.

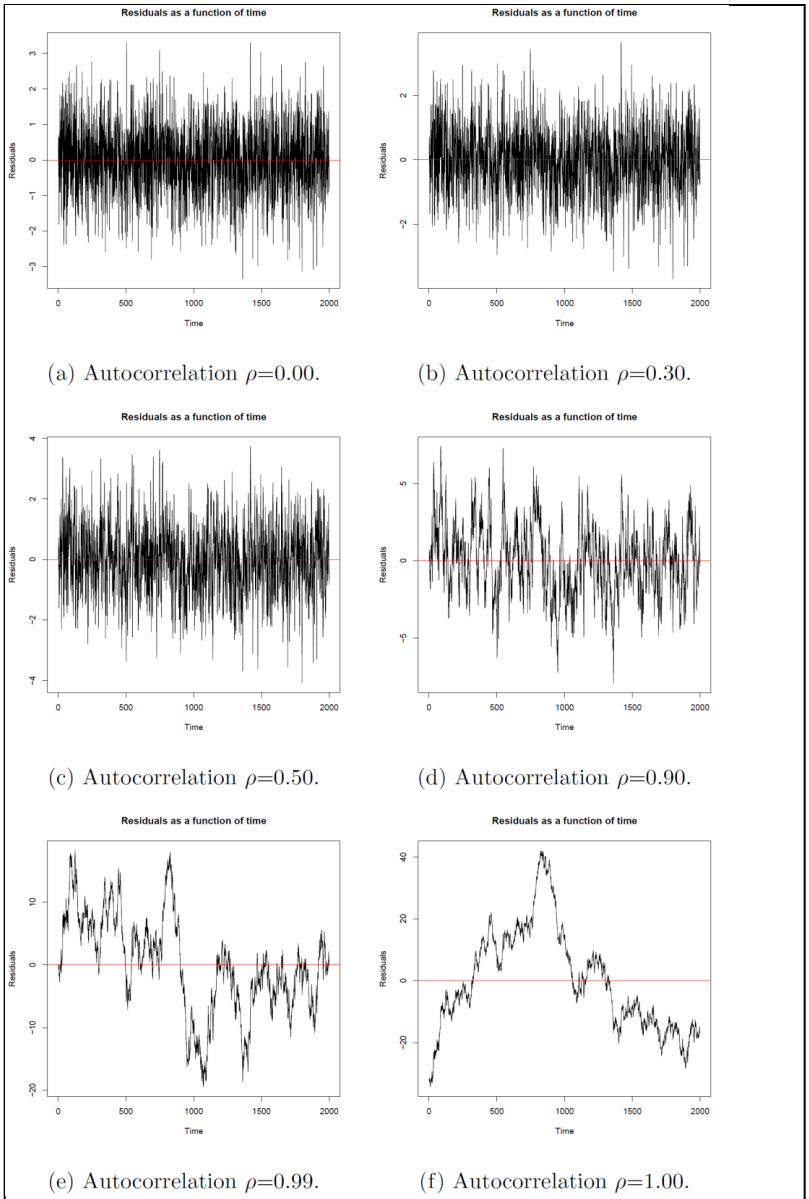


Figure 3. Independence of errors violated under differing degree of autocorrelation

Assumption of Constant Variance of Errors

The constant variance assumption is also called the assumption of homoscedasticity. Under Equation 1, it is related to the normality assumption because it deals with the unknown variance parameter, σ^2, parameter for ε in Equation 1, shared by the conditional normal distribution of the response and that of the error. The assumption states that the conditional variance of the error term in Equation 1 remains constant over the range of all predictor values. In other words, the constant variance assumption requires that the unknown σ^2 parameter for ε remain invariant under any set of predictor characteristics.

The opposite of homoscedasticity is called heteroscedasticity or unequal variances for different settings of predictors. Among the many underlying reasons for heteroscedasticity is one where the response variable follows a distribution in which the variance is functionally-related to the mean. In most such cases, significant non-normality in the response is observed as well. A typical scenario here is when the response is a count/frequency variable that follows a Poisson distribution where the variance of the response equals its mean. Under such a distribution, the mean of the response, $E(y)$, changes with a change in one or more predictors. Therefore, the conditional distribution of the response cannot have

constant variance at all levels of the predictors because the variance of a Poisson distribution equals its mean, which is changing with changing values of the predictors.

An assessment of the homoscedasticity assumption could be based on a plot of residuals against the model-predicted values for the response, which are estimates of its means conditional on levels of the predictors. If the dispersion of the residuals is approximately constant around zero with respect to the predicted values of the response, the assumption of homoscedasticity is supported. Otherwise, the assumption is questionable. Finally, residuals are also plotted against each predictor to assess if the constant variance assumption holds reasonably.

Next, the article proposes a generating model that simulates a data set where the variance of the error is decided by the square of a linear function of three predictors. In this model, any change in one or more of the three predictors usually causes this linear function to change as well, which in turn changes the variance of the response. In other words, the variance of the response cannot remain constant over the range of levels of all predictors due to the mathematical expression that determines how the former is related to the latter. With that said, the generating model used here is:

$$y = -0.0764 + (-0.2124)x_1 + (-0.4998)x_2 + (-0.4463)x_3 + \varepsilon\,, \tag{9}$$

where $x_1 \sim N(2,3^2)$, $x_2 \sim N(0,0.3^2)$, $x_3 \sim N(5,2^2)$ and $\varepsilon \sim N(0,[1+2x_1+3x_2+4x_3]^2)$. In Equation 9, the variance of the error equals the square of a linear function of the three predictors: ,$[1+2x_1+3x_2+4x_3]^2$. So, the error variance changes as a function of one or more predictors. The simulated data are fitted using ordinary least squares to an analytical model from Equation 1 that assumes constant variance over levels of all predictors. So, the analytical model used here is Equation 10:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon\,, \tag{10}$$

where ε follows a conditional normal distribution with the mean being zero and the variance being unknown, but constant over all settings of the three predictors: x_1, x_2, and x_3. The residuals from the fitted analytical model are plotted against the model predicted values of the response and each predictor. The resulting graphical outputs are in Figure 4. As is indicated in Figure 4, fitting the data with an un-constant variance structure using ordinary least squares to a model that assumes constant variance, results in systematic patterns in both the residual-versus-predicted and the residual-versus-predictor plots. Under Figure 4a, when plotted against the predicted values, the residuals are in the shape of a right cone with the circular base pointing to the lower end of the predicted values and the apex to the higher end of the predicted values, which indicates that the variability of the residuals tends to decrease with increasing predicted values. Figures 4b and 4d also show a cone shape of residuals, but in the opposite direction. Variability of the residuals is smaller at the lower end of x_1 and x_3, but larger at the higher end of the two predictors. By contrast, the residuals in Figure 4c do not tend to exhibit any systematic patterns despite a violation of the

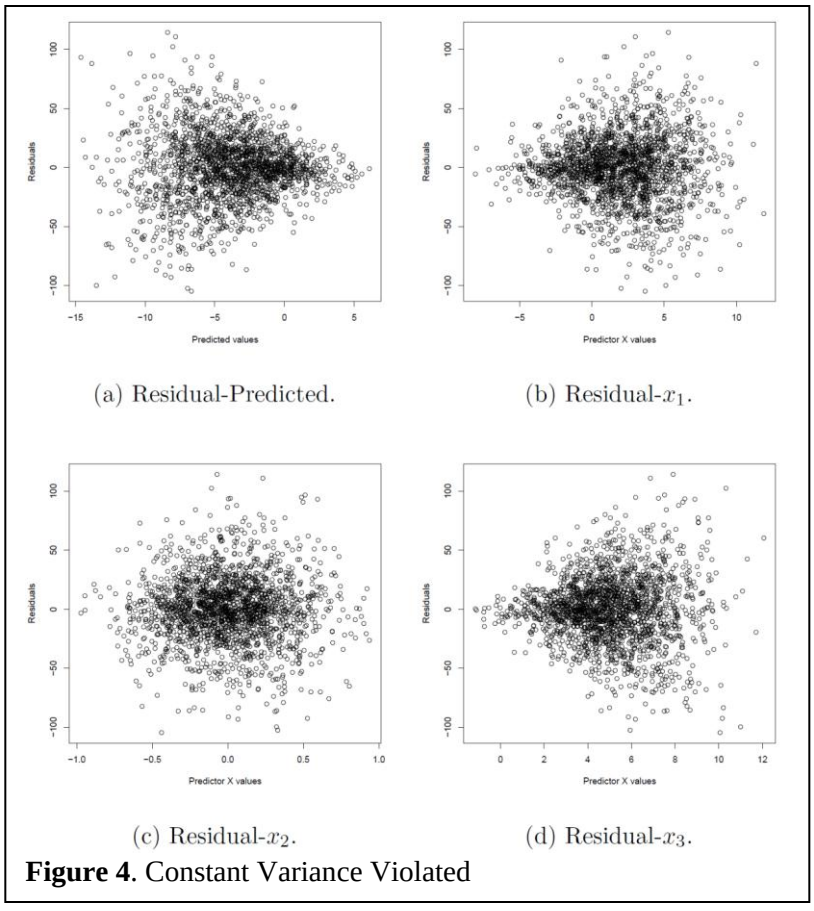


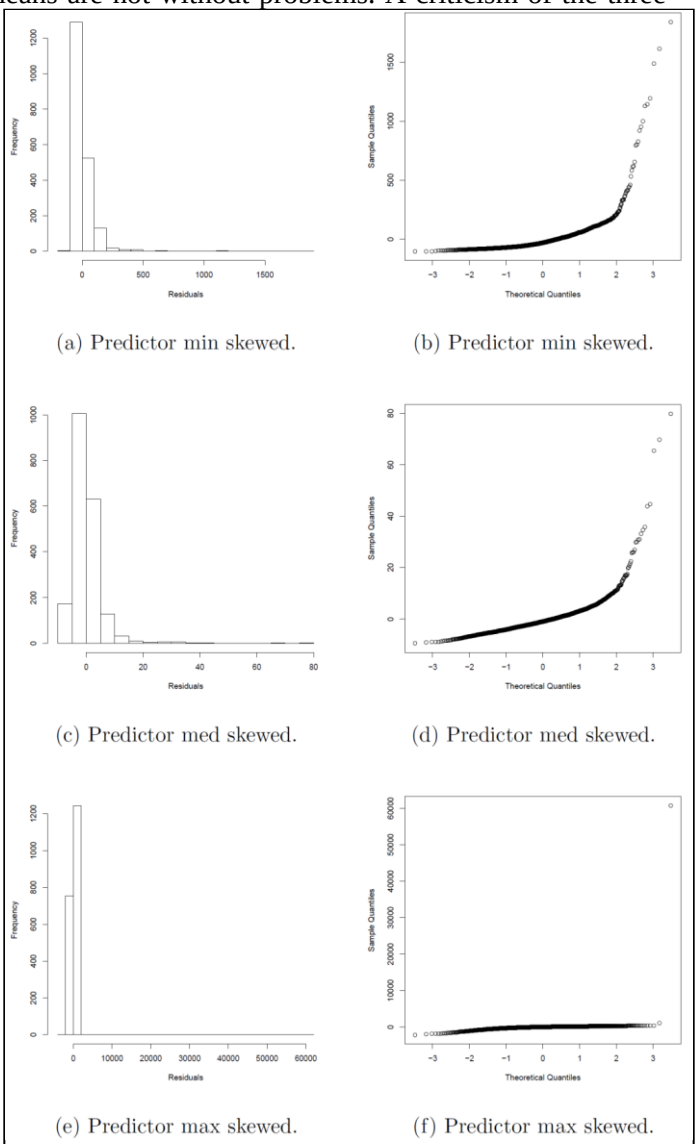
Figure 4. Constant Variance Violated

constant variance assumption. This observation indicates that residual plots are not always effective in identifying assumption violations. Often, their use may have to be supplemented by statistical tests.

Assumption of Normality of Errors

The normality assumption states that given any characteristics of all predictors, the random error is assumed to be normally distributed with the mean being zero and variance being a constant σ^2, which is usually unknown. Of the four critical assumptions examined in this article, the normality assumption is the least critical one. Many times, regression analysis is robust with respect to non-normal errors. By robust, this means that inferences from regression analysis tend to remain valid when the assumption of normal errors is not quite satisfied. It is only when the errors come from a heavy-tailed, or highly skewed, distribution that a violation of the normality assumption of errors becomes a concern.

An examination of the normality assumption under residual analysis is usually based on one or more of the three types of plots: 1) a histogram of residuals, 2) a normal QQ plot of residuals, and 3) a stem-and-leaf plot of residuals. But, these graphical means are not without problems. A criticism of the three plots is that each one lumps the residuals together from multiple settings of the predictors for a test of normality, when regression theory states that it is conditional on each possible setting of all predictors that the errors are normally distributed, which indicates that a check of the normality assumption should be done separately conditional on each possible setting of all predictors; a task practically impossible to accomplish. Lumping together residuals from multiple settings of all predictors, and performing a single check of the normality assumption using all residuals, could cause some non-normal patterns to be hidden or become invisible. This is particularly true when the distribution of residuals is skewed to one direction for some values of all predictors, but skewed to the other direction for other values of all predictors. In such a case, combining these residuals and examining them in a single plot could produce a distribution that is relatively symmetric (Mendenhall & Sincich, 2003). Despite such criticism, the above plots are still commonly used for checking normality.



(a) Predictor min skewed. (b) Predictor min skewed.

(c) Predictor med skewed. (d) Predictor med skewed.

(e) Predictor max skewed. (f) Predictor max skewed.

Figure 5. Normality violated under differing skewness in predictor

With that description concerning one problem associated with the three graphical means, this article implements the first two types of plots to demonstrate departures from normality for the residuals from a fitted regression model; namely 1) a histogram and 2) a normal QQ plot of residuals. In order to support normality, the histogram needs to be reasonably symmetric and the QQ plot approximately linear with data points falling almost along a straight line.

Next, the article presents a generating model that simulates data with a non-normal distribution for the response. The data are then fitted to an analytical model from Equation 1 assuming normality of errors. The choice of the non-normal distribution here is a Poisson distribution for modeling count data. The rationale for such a choice is that count data, typically treated as interval/ratio variables, are analyzed

frequently under Equation 1 assuming normal errors using ordinary least squares (Nussbaum, Elsadat, & Khago, 2007). Such a practice is questionable because count data are not interval/ratio and the correct choice for a count response should be either the Poisson or the negative binomial model. With that said, the generating model here is Equation 11, a Poisson regression model. Therefore, the response follows a Poisson distribution with the mean being $\exp[(0.2)+(0.3)x_1+(0.9)x_2+(0.5)x_3]$ conditional on the values of the predictors:

$$y = \text{Poisson}(\exp[(0.2)+(0.3)x_1+(0.9)x_2+(0.5)x_3]) , \quad (11)$$

where $x_1 \sim N(2,3^2)$ and $x_2 \sim N(0,0.3^2)$. x_3 is simulated separately under three cases: 1) $N(5,2^2)$, 2) $F(5,100)$, and 3) $F(1,20)$. Clearly, from case 1 to case 3, the skewness of x_3 data is anticipated to keep increasing: For case 1, skewness of x_3 data is computed to be 0.0811; for case 2, 1.2374; and for case 3, 3.4359.

The simulated data containing y (count data), x_1, x_2, and x_3 from the Poisson model are fitted to an analytical model from Equation 1 assuming normality of the random errors. Therefore, the analytical model used here is Equation 12:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon , \quad (12)$$

where ε follows a conditional normal distribution with the mean being zero and the variance being unknown, but constant over all settings of the three predictors: x_1, x_2, and x_3. In such a case, the normality assumption of the analytical model is violated. To demonstrate the assumption violation, residuals from the fitted analytical model are plotted in a histogram and in a normal QQ plot, respectively. The results are in Figure 5.

As is indicated in Figure 5, the normality assumption of Equation 1 is not even close to being reasonably satisfied. In all three residual histograms, they are strongly positively skewed, which contradicts normality. Such an observation is further supported by all three normal QQ plots where the data points deviate substantially from being linear. Therefore, both plots indicate the common practice of treating a count response as interval/ratio and analyzing it using a linear model as Equation 1 under ordinary least squares could lead to a serious violation of the normality assumption of the model. Finally, from case 1 to case 3, the skewness of x_3 keeps increasing with everything else held constant. However, there does not seem to be a clear, monotonic pattern for change in residual skewness as a function of monotonic increase in predictor skewness: From case 1 to case 2, the predictor skewness increases substantially by about 15 times (from 0.0811 to 1.2374) and the residual skewness by contrast drops from 6.9626 to 5.0638. From case 2 to case 3, the former skewness increases by only about 2 times (from 1.2374 to 3.4359) and the latter skewness sharply increases from 5.0638 to as high as 40.9812.

Discussion

So far, this article has discussed four assumptions for the regression model in Equation 1 when it is first order that is considered to be critical in the literature. For each assumption, this article first presents a generating model simulating the data violating one or more assumptions of Equation 1. Then, the simulated data are fitted to the corresponding analytical model from Equation 1 to obtain the residuals. Finally, several basic residual plots are utilized to graphically present the effect of assumption violations on the pattern of residuals to help regression learners gain experience with residual plots.

With that said, the article proceeds to provide some additional discussion regarding the issue of regression assumption violations and that of residual analysis. Three topics are briefly discussed here: 1) likely inclusiveness of assumption violations, 2) alternatives to basic residual plots such as statistical tests and partial residual plots, and 3) remedial measures given one or more assumption violations. Although each of the three topics is broad enough to deserve an individual article on its own merits, the scope of this study limits their discussion to a brief overview.

- Many times, one regression assumption violation is usually coupled by one or more other assumption violations. That is, assumption violations are usually not exclusive of each other; rather, they are inclusive of each other. This indicates that when one type of residual plot indicates a violation of one assumption, another type of residual plot may likely indicate a violation of another assumption at the same time. For example, a violation of the normality assumption is often accompanied by a violation of the constant variance assumption. This could happen when the variance of the response is a function of its mean conditional on the settings of

all predictors; such as in the case of a count response that has a Poisson distribution. Many times, though, when appropriate remedial measures are taken to address one assumption violation, another assumption violation is likely to be addressed simultaneously. Back to the case when the response y is a count variable following a Poisson distribution where non-normality is coupled by heteroscedasticity, both assumption violations frequently can be rectified by applying certain variance-stabilizing transformations to the response (Mendenhall & Sincich, 2003) via the square root transformation $\sqrt{y}$.

- Two alternatives to basic residual plots exist as aids to residual interpretation and as indicators of potential model improvements: 1) formal statistical tests and 2) partial residual plots. These are in addition to any sensible way that residuals may be plotted for the particular problem under consideration (Draper & Smith, 1998).
- For the alternative of statistical tests, many of them are conducted to assess the significance of model terms. For example, given a quadratic relationship between *y* and x_1, such as the one in Equation 2, a plot of residuals obtained from the first order model by Equation 3 may clearly indicate a lack of fit of the model to the data. But, equivalently, a partial *t*-test of whether the second order term x_1^2 contributes to the model may serve the same purpose. Besides, decisions based on residual plots are subjective, and possibly even more subjective than statistical tests, so it is recommended that statistical tests be used to help decide if residual plots are really suggesting any assumption violations (Weisberg, 2005). However, these tests are often subject to spurious statistical significance due to large sample sizes. Therefore, it would not always be wise to make decisions solely on the basis of such tests.
- For the alternative of partial residual plots, they are used when the model contains more than one predictor. A partial residual plot measures the influence of one predictor on the response after the effects of all other predictors have been removed or considered. Often, by indicating more precisely how to modify the model, a plot of the partial residuals against a predictor reveals more information about the relationship between the response and the predictor than do those basic residual plots.
- Given one or more assumption violations for the regression model by Equation 1 when it is first order, it is recommended that remedial measures be taken to rectify model deficiencies. To that end, there are two basic choices (Kutner et al., 2005): 1) abandon the regression model in Equation 1 and search for a more appropriate model (e.g., using a Poisson (generalized linear) model rather than a linear model for modeling count data) and 2) employ some transformation on the data so that the model in Equation 1 is appropriate for the transformed data. Each approach has advantages and disadvantages. The first approach may entail a complex model that could nevertheless yield better insights; whereas the second approach usually leads to a simple model, but at the likely cost of obscuring the fundamental interconnections between the variables through transformations. Because it is linear regression that this article primarily discusses, the focus is next on how to conduct transformations on the data so that the linear model becomes appropriate in the transformed scale. Here, transformation refers to changing the form of the data using a mathematical function applied to each point in the data set. Or, equivalently, transformation means the application of a mathematical function to the model hypothesized to be a good approximation to the underlying one that generates the data (i.e., the generating model).
- Transformations could be done on the response *y* and/or one or more of the predictors' *x*'s. A transformation of *x*'s is sometimes preferred over that of *y*. This is because the predictors are assumed to be measured perfectly (i.e., they are not subject to measurement error or an additional assumption of many regression models). Thus, there is little problem in transforming them. However, for transformations of *y*, which is assumed to be subject to measurement error, extra care should be taken. This is because a transformation of the response is likely to affect the distribution of errors. Therefore, it is important to examine the residuals for the model finally fitted (i.e., the transformed model) to see if relevant assumptions appear to be violated. On the other hand, transformations of *y* can also lead to difficulties with interpretation. For example, given heteroscedasticity, a transformation of *y* (e.g., logarithm or square root) may lead to a constant variance model. The difficulty here is that while the original scale *y* may be meaningful

or make easy and intuitive sense, log(y) or $\sqrt{y}$ may not; thus, making it difficult to give a simple interpretation of the regression coefficients with the transformed y Therefore, transforming y should generally be applied with care and should even be avoided whenever a suitable transformation of the x's is available (Draper & Smith, 1998; Faraway, 2006).

- A critical principle shared by most transformations is to reduce a complicated (i.e., usually nonlinear and/or higher-order) model to a simple, lower-order model whose mean function is linear in the transformed scale; thus, satisfying the assumptions studied in this article for the linear model in Equation 1. For example, it is very common that a linear regression model, as described by Equation 1, is inappropriate for the collected data in their raw form because additivity and/or linearity are not reasonable assumptions for them such as data with the structure $y = x_1x_2x_3$ in the original scale. In this case, nonlinear transformations of the data can sometimes remedy the situation by making the data in the transformed scale more closely satisfy those assumptions (Gelman & Hill, 2007) such as data with the structure log(y) = log(x_1)+log(x_2)+log(x_3) in the transformed scale. However, note that not all nonlinear models can be transformed to be a linear model, and those that can be transformed to be linear are called intrinsically linear; whereas those that cannot be transformed are called intrinsically nonlinear.
- As far as the type of transformations is concerned, the family of power transformation is commonly used that creates a rank-preserving transformation of the original data using power functions. A power transformation is usually conducted on the response variable y and is of the form:

$$\begin{cases} y' = y^{\lambda}, \text{ when } \lambda \neq 0, \\ y' = log(y), \text{ when } \lambda = 0, \end{cases}$$

where λ is a transformation parameter to be determined from the data. Some commonly used power transformations include y^2 when $\lambda = 2$, $\sqrt{y}$ when $\lambda = 0.5$, and log(y) when $\lambda = 0$ (by definition). The power transformation can be effective in correcting several assumption violations discussed in this article (Kutner et al., 2005): 1) skewed (non-normal) distribution of error terms, 2) unequal error variances, and 3) nonlinearity of the regression function. With the aid of the Box-Cox procedure (Box & Cox, 1964), many computer programs are able to automatically identify a maximum likelihood estimate of the transformation parameter λ; thus, making it easier for people to take advantage of the power transformation to remedy assumption violations. However, despite the availability of such automatic transformation search algorithms, it may still be difficult to identify an appropriate transformation on the response and/or predictors. Many times, several alternative transformations on the response may be tried, as well as some simultaneous transformations on one or more of the predictors. In the end, it is important to note that the effectiveness of a transformation is best assessed by trying it on the data and then checking the fit of the finally fitted model and the pattern of the resultant residuals (Draper & Smith, 1998). Therefore, residual plots and other graphical/non-graphical means should be prepared to determine the most effective transformations.

Conclusion

This article supplements the existing regression literature by taking a model-based simulation perspective to help regression learners gain some experience in judging residual plots for identifying regression assumption violations. Four critical assumptions for a multiple linear regression model are examined in this study: 1) linearity and additivity (i.e., two assumptions counted as one), 2) independence of errors, 3) constant variance of errors, and 4) normality of errors. The approach that the article follows is to generate and graphically demonstrate the pattern of residuals from estimating a first-order linear regression model using a data set simulated with one or more assumptions of the said model violated. In the end, the article also discusses other related issues and possible remedial measures for assumption violations with a focus on data transformations.

References

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Hohn Wiley & Sons.

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage Publications, Inc.

Berry, W. D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage Publications Inc.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of Royal Statistical Society, 26,* 211–246.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25,* 4279–4292.

Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity analysis: Linear regression*. New York: John Wiley & Sons, Inc.

Chatterjee, S., & Price, B. (1991). *Regression analysis by example* (2nd ed.). New York: John Wiley & Sons, Inc.

Cook, R. D. (1998). *Regression graphics: Ideas for studying regressions through graphics*. New York: Wiley & Sons, Inc.

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

Dupuis, D., & Garfield, J. (2010). *Teaching with data simulations*. Retrieved from http://serc.carleton.edu/sp/cause/datasim/index.html.

Faraway, J. (2004). *Linear models with R*. Boca Raton, FL: Chapman & Hall.

Faraway, J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: Chapman & Hall.

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage Publications, Inc.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Goldberger, A. S. (1968). *Topics in regression analysis*. New York: The Macmillan Company.

Grob, J. (2003). *Linear regression*. New York: Springer-Verlag.

Hodgson, T., & Burke, M. (2000). On simulation and the teaching of statistics. *Teaching Statistics, 22,* 91–96.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York: McGraw-Hill/Irwin.

Mendenhall, W., & Sincich, T. (2003). *A second course in statistics: Regression analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis* (3rd ed.). New York: John Wiley & Sons, Inc.

Mooney, C. (1995). Conveying truth with the artificial: Using simulated data to teach statistics in the social sciences. *SocInfo Journal, 1,* 1–5.

Nussbaum, E. M., Elsadat, S., & Khago, A. H. (2007). Best practices in analyzing count data: Poisson regression. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 306–323). Thousand Oaks, CA: Sage Publications, Inc.

Rao, C. R., & Toutenburg, H. (1999). *Linear models: Least squares and alternatives* (2nd ed.). New York: Springer-Verlag.

Tamhane, A. C., & Dunlop, D. D. (1999). *Statistics and data analysis: From elementary to intermediate*. Upper Saddle River, NJ: Prentice Hall.

Weisberg, S. (2005). *Applied linear regression* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Yan, X., & Su, X. G. (2009). *Linear regression analysis: Theory and computing*. Singapore: World Scientific.

Send correspondence to:
Hongwei Yang
University of Kentucky
Email: patrick.yang@uky.edu