International Journal of Statistics and Probability; Vol. 11, No. 4; July 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Using Residual Plots to Distinguish Cases of Predictor Omission in Linear Models

Emily Nystrom[1], Julia L. Sharp[2] & William C. Bridges, Jr.[1]

[1] School of Mathematical & Statistical Sciences, Clemson University, Clemson, SC, USA

[2] Department of Statistics, Colorado State University, Fort Collins, CO, USA

Correspondence: Emily Nystrom, Independent Researcher, USA

Received: April 23, 2022 Accepted: June 13, 2022 Online Published: June 29, 2022

doi:10.5539/ijsp.v11n4p25 URL: https://doi.org/10.5539/ijsp.v11n4p25

Abstract

Residual plots are commonly used to diagnose possible model misspecification, including predictor omission. In this paper, we present a systematic workflow for using residual plots and partial residual plots to detect and distinguish several types of model misspecification in linear models. Our workflow uses a set of four Yes/No questions and is accessible to statisticians and practitioners of all experience levels.

Types of model misspecification considered by our workflow include four cases of predictor omission and two types of nonconstant variance. In particular, these cases of predictor omission are defined by the correlation and interaction status between the omitted predictor and the predictor included in the fitted model. Distinguishing cases of predictor omission is important because the impact of predictor omission can vary among cases. The interpretation of the parameter estimates in the statistical model can change depending on the approach.

Keyword: model misspecification, model diagnostics, nonconstant variance, residual analysis

1. Introduction

Linear models are commonly used to understand relationships between predictors and a response. Residual plots can be used to check model assumptions [Gray, 1989, Tsai, Cai, & Wu, 1998, Faraway, 2005]. For example, the linear model assumptions of linearity and homogeneity of variance can be checked using residual versus fitted plots, and the normality assumption can be checked using quantile-quantile plots of the residuals.

A partial residual plot (or added variable plot [Kutner, Nachtsheim, Neter, & Li, 2005] is a special type of residual plot that displays a residual on the vertical axis and a partial residual on the horizontal axis. The partial residual plots introduced by [Ezekiel, 1924] were initially used to visualize "curvilinear correlation" among dependent variables. In practice, a commonly noted feature of the partial residual plot is its ability to account for predictors already in the fitted model when considering the additional explanatory value of candidate predictors [Mansfield & Conerly, 1987, Faraway, 2005, Kutner et al., 2005].

There are several ways in which models can be mis-specified, including predictor omission [Xu & Sinha, 2021]. Predictors that are omitted from a linear model could be uncorrelated or correlated and may or may not interact with a predictor already included in the model. Non-constant variance could also arise due to model misspecification. Our work extends the literature regarding consideration of omitted predictors to diagnose model misspecification of these types. Distinguishing cases of predictor omission is important because the impact of predictor omission can impact the interpretation of results [Greene, 2003, Woolridge, 2012, Nystrom, Sharp, & Bridges, 2019, Yoon & Welsh, 2020]. In particular, the interpretation of the parameter estimates in the statistical model can change.

In this paper, we present a systematic, accessible workflow for using residual plots and partial residual plots to detect and distinguish several types of model misspecification in linear models. In Section 2, residuals and partial residuals are defined. In Section 3, traditional uses of residual and partial residual plots are described; then, we explain how to compare residual and partial residual plots to detect correlation between a predictor already included in the fitted model and a candidate predictor that has not been included in the fitted model but is included in the dataset being analyzed. In Section 4, several types of model misspecification are described. In Section 5, our workflow is presented, and examples for using our workflow are provided, illustrating how residual and partial residual plots can be used to detect and distinguish the types of model misspecification discussed in Section 5. Examples include both simulated datasets (Sect. 5.2) and a real SAT dataset (Sect. 5.3).

2. Linear Models and Residuals

In Section 2, we provide brief definitions and notation for linear models and their corresponding residuals. For a detailed motivation and development of linear models, see [Kutner et al., 2005] and [Ott & Longnecker, 2011]. See [Kutner et al., 2005] for more information about residuals from linear models.

2.1 Linear Models

A simple linear model (LM) (Eq. 1) equates a response (Y) to a linear function of a single predictor (X_1) and a random error (ϵ):

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{X}_1 + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_\epsilon^2\mathrm{I}\right). \tag{1}$$

As a matter of notation, we denote the response vector, the j^{th} predictor vector, and the error vector as follows: $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^T$, $\mathbf{X}_j = \left(X_{j,1}, X_{j,2}, ..., X_{j,n}\right)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$, respectively, where n denotes the number of observations (sample size).

A two-predictor LM is given by

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_\epsilon^2\mathrm{I}\right),$$

and a two-predictor LM with an interaction term (X_1X_2) is given by

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \beta_3\mathbf{X}_1 \circ \mathbf{X}_2 + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_\epsilon^2\mathrm{I}\right),$$

and $\circ$ denotes elementwise vector multiplication. In general, a k-predictor LM (Eq. 2) is given by

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{X}_1 + ... + \beta_k\mathbf{X}_k + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_\epsilon^2\mathrm{I}\right). \tag{2}$$

The linear predictor (η) is defined as the linear combination of predictors in the model (e.g., $\eta = \beta_0 + \sum_{j=1}^{k}\beta_jX_k$ in Eq. 2) and represents the explained portion of the response.

For a k-predictor LM (Eq. 2), the corresponding estimated linear equation is given by

$$\widehat{\mathbf{Y}} = \widehat{\beta}_0\mathbf{1} + \widehat{\beta}_1\mathbf{X}_1 + ... + \widehat{\beta}_k\mathbf{X}_k,$$

where $\widehat{Y}$ denotes the predicted response, and $\widehat{\beta}_0, \widehat{\beta}_1, ...,$ and $\widehat{\beta}_k$ denote the estimated coefficients. For our purposes, these estimated coefficient values are determined through least squares estimation.

In each of the LMs discussed thus far (e.g., Eq.s 1 and 2), errors are assumed to be independent and normally distributed with constant variance (i.e., $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_\epsilon^2\mathrm{I}\right)$), and predictors are assumed to be linearly related to the response; these assumptions are common for traditional LMs. Because these errors are unobserved, related LM assumptions in the population model are often assessed by analyzing *estimated* errors (e.g., residuals) that are produced when fitting the model to observed sample data. The intuition behind residual analysis is grounded in considering whether the estimated errors could reasonably be considered samples from the populations from which they were assumed to be generated.

2.2 Residuals

In the context of linear models, a residual (e) is defined as the difference between an observed response value and a predicted response value and is used to estimate the error (ϵ) in the LM. The residual vector is given by

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathrm{I} - \mathrm{H})\mathbf{Y},$$

where $\mathrm{H} = \mathrm{X}\left[\mathrm{X}^T\mathrm{X}\right]^{-1}\mathrm{X}^T$ is a projection matrix for the design matrix (X), which contains a column for the intercept and for each predictor in the model (e.g., $\mathrm{X} = [\mathbf{1}, \mathbf{X}_1, ..., \mathbf{X}_k]$ for Eq. 2). If model errors ($\boldsymbol{\epsilon}$) are assumed to be independent and normally distributed with zero mean and constant error variance (i.e., $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_\epsilon^2\mathrm{I}\right)$), then the distribution of the residual vector is given by

$$\mathbf{e} \sim N\left(\mathbf{0}, \sigma_\epsilon^2\left(\mathrm{I} - \mathrm{H}\right)\right). \tag{3}$$

A standardized residual ($e^{(s)}$) is derived by dividing the residual by the estimated standard deviation. The vector of standardized residuals is given by

$$\mathbf{e}^{(s)} = \frac{\mathbf{e}}{\sqrt{\widehat{\sigma_\epsilon^2}\ \mathrm{Diag}(\mathrm{I} - \mathrm{H})}} = \frac{\mathbf{e}}{\sqrt{\frac{\mathbf{e}^T\mathbf{e}}{n-(k+1)}\ \mathrm{Diag}(\mathrm{I} - \mathrm{H})}},$$

where k denotes the number of predictors in the model. Instead of dividing by the sample standard deviation of the residuals ($\sqrt{\frac{\mathbf{e}^T\mathbf{e}}{n-1}}$), the denominator of the standardardized residual uses an estimate of the standard deviation based on the residual vector's assumed distribution (from Eq. 3). Our motivation for using standarized residuals in our workflow is essentially to remove the issue of scale when plotting residuals (further discussed in Section 5).

A partial residual ($e_{(X_m|X_1,...,X_k)}$) is a residual from an estimated linear model that uses a candidate predictor (X_m) as the response regressed on all other predictors (X_1, ..., X_k) that are included in the fitted model (i.e., $\widehat{\mathbf{X}}_m = \widehat{\gamma}_0\mathbf{1} + \sum_{j=1,j\neq m}^{k} \widehat{\gamma}_j\mathbf{X}_j$) [Kutner et al., 2005]. In particular, when a simple LM (Eq. 1) is estimated using predictor X_1, and X_2 is a candidate predictor, the vector of partial residuals is given by

$$\mathbf{e}_{(X_2|X_1)} = \mathbf{X}_2 - \widehat{\mathbf{X}}_2 = \mathbf{X}_2 - (\widehat{\gamma}_0\mathbf{1} + \widehat{\gamma}_1\mathbf{X}_1)\,. \tag{4}$$

The use of partial residuals will be discussed in Section 3.

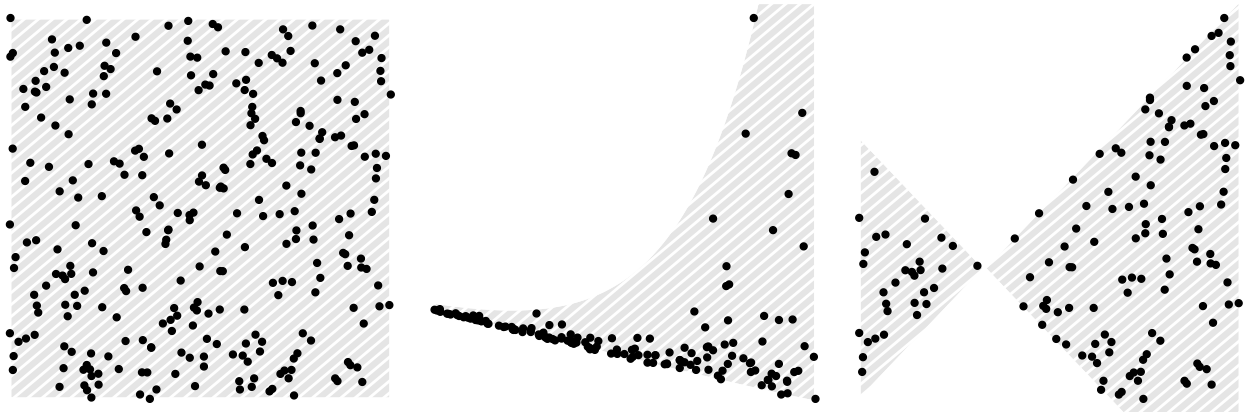
3. Using Residual Plots to Check Model Assumptions

When independence and homoscedasticity (constant variance) assumptions are satisfied, corresponding residual plots should have random scatter (Fig. 1a). Conversely, the appearance of nonrandom patterns in residual plots may indicate that an error assumption has been violated and/or that the linear predictor in the fitted model is misspecified.

3.1 Checking for Nonconstant Error Variance

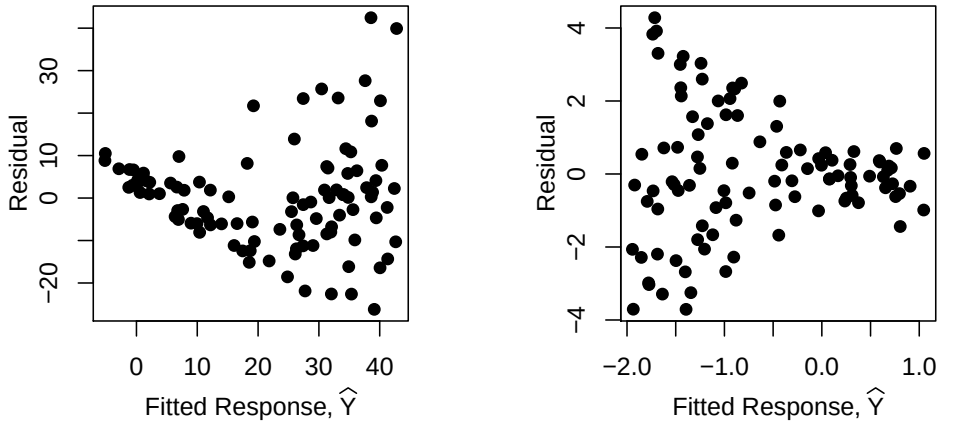
The residual versus fitted response ($\widehat{Y}$) scatterplot may be used to check the constant variance assumption (homoscedasticity). A pattern in a residual plot may be indicative of nonconstant error variance [Weisberg, 1985], especially if the residual scatter displays a fanning or funnel pattern (Fig. 1b). If the response variable needs to be log transformed to produce a linear predictor (often called "intrinsically or transformably linear"), a funnel pattern may also occur in the residual plot (Fig.s 1b and 2a). The appearance of nonconstant error variance may also be caused by omission of an interaction term from the fitted model [Fox,1991], and a bowtie (Fig. 1c) or partial bowtie pattern may occur. And points in the residual plot are evenly-spaced (vertically) within the pattern when an interaction term has been omitted from the fitted model (Fig.s 1c and 2b), which is in contrast to the uneven spacing inside the funnel pattern when the response needs to be transformed (Fig.s 1b and 2a).

Note that there are other causes of nonconstant variance and patterns in the residual plots (such as non-normal response variables), but this article focuses on scenarios that can be appropriately modeled as traditional linear models.



(a) Rectangle (b) Funnel (c) Bowtie

Figure 1. Patterns of scatter within residual plots when considering the homogeneity of variance assumption: (a) rectangle (random scatter); (b) funnel; (c) bowtie. Spacing of points within each shape is emphasized by background shading. In these examples, points are relatively evenly spaced (vertically within each shape) for the rectangle and bowtie patterns, whereas vertical spacing of points within the funnel is nonconstant. We note that scatterplots in this figure are stylized to illustrate the patterns discussed. Scatter shown in Figures 1a and 1c is conceptual, and actual residuals were not used for these subfigures. Residual plots with actual residuals are included in later figures (e.g., Fig.s 1b, 2-4, 7, 8)



(a) Nonconstant variance from incorrect response transformation in fitted model

(b) Nonconstant variance from omission of interaction term in fitted model

Figure 2. Example residual plots for two sources of nonconstant variance: (a) incorrect response transformation; (b) omission of interaction term

3.2 Checking for Predictor Omission

A residual versus candidate predictor scatterplot may be used to check whether a candidate predictor should be added to the fitted model. Random scatter in a residual versus candidate predictor scatterplot supports leaving out the candidate (X_m), whereas a nonrandom pattern supports adding X_m (or a function of X_m, such as X_m^2) to the current model.

Comparing the residual versus candidate predictor scatterplot to the corresponding partial residual plot may help to identify cases in which a candidate predictor is correlated with a predictor that is already included in the fitted model. When X_1 and X_2 are uncorrelated, the candidate residual scatterplot (e vs. X_2) is similar to the residual scatterplot (e vs. $e_{(X_2|X_1)}$) (Fig. 3a). On the other hand, when the patterns of the two scatterplots are noticeably different, correlation between X_1 and X_2 is suspected (Fig. 3b).

The impact of correlation on the partial residual plot can be explained mathematically by considering the difference between X_2 and $e_{X_2|X_1}$ (the horizontal components of the residual versus candidate plot and the corresponding partial residual plot, respectively). Recall (Eq. 4) that $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ are estimated coefficients for an LM with response X_2 and predictor X_1, and the estimated slope of the simple LM between X_2 and X_1 is given by $\widehat{\gamma}_1 = r_{1,2}\frac{s_2}{s_1}$. That is,

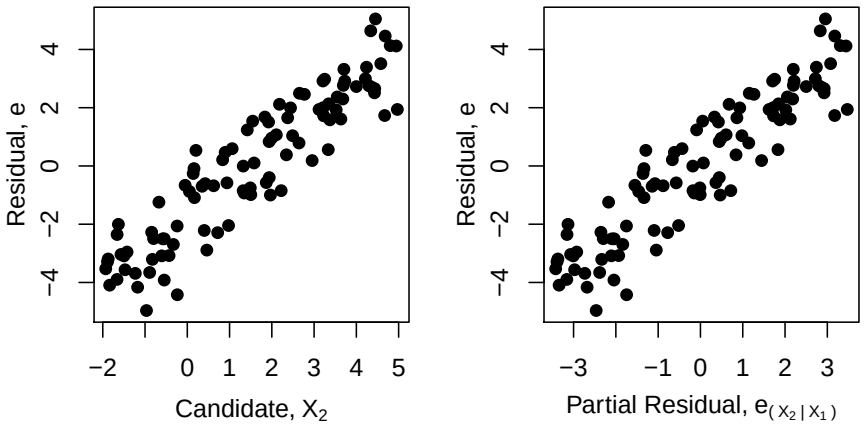
$$X_2 - e_{X_2|X_1} = X_2 - \left(X_2 - \widehat{X}_2\right) = \widehat{X}_2 = \widehat{\gamma}_0 + \widehat{\gamma}_1 X_1 = \widehat{\gamma}_0 + r_{1,2}\frac{s_2}{s_1}X_1,$$

where $r_{1,2}$ denotes the sample correlation between X_1 and X_2, and s_1 and s_2 denote sample standard deviations of X_1 and X_2, respectively. Thus when other parameters, such as s_1 and s_2, are fixed, $|\widehat{\gamma}_1|$ decreases as $|r_{1,2}|$ decreases. When the correlation between X_1 and X_2 is negligible, the difference between X_2 and $e_{X_2|X_1}$ is approximated by $\widehat{\gamma}_0$ (i.e., $X_2 - e_{X_2|X_1} \approx \widehat{\gamma}_0$ if $r_{1,2} \approx 0$), in which case the partial residual plot is a horizontally-shifted version of the candidate residual plot.

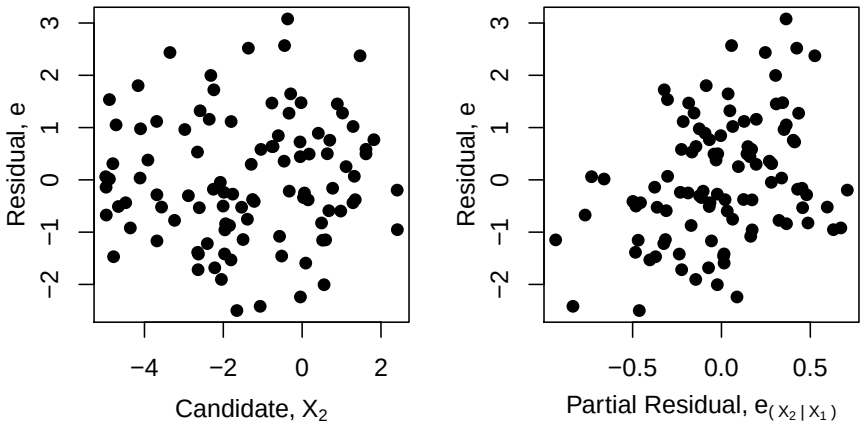
While a candidate residual plot can be compared to a partial residual plot to detect correlation, the partial residual plot also provides insight into the usefulness of adding the candidate predictor into a fitted model with its current predictors. A linear trend in a partial residual plot gives evidence that a candidate predictor is linearly related to the response and that the candidate predictor has *additional* explanatory value beyond that of the predictor(s) already included in the model.

Substitute vertical and horizontal components (e and $e_{X_2|X_1}$, respectively) of the partial residual plot into a slope-intercept equation of a line (i.e., $e = me_{X_2|X_1} + b$). Suppose $e = me_{X_2|X_1} + b$. Then the residual $(Y - \widehat{\beta}_0 - \widehat{\beta}_1 X_1 = m\left(X_2 - \widehat{\gamma}_0 - \widehat{\gamma}_1 X_1\right) + b)$ can be rearranged as $Y = \left(\widehat{\beta}_0 - m\widehat{\gamma}_0 + b\right) + \left(\widehat{\beta}_1 - m\widehat{\gamma}_1\right)X_1 + mX_2$. Thus, the slope of the partial residual plot (m) is related to the additional contribution of candidate X_2 to a fitted model that already includes X_1.

A nonzero slope in the partial residual plot suggests that the current model may benefit from the addition of a particular candidate predictor (e.g., X_2 in Fig. 4a), whereas random scatter in the partial residual plot indicates that a particular candidate predictor (e.g., X_3 in Fig. 4b) will not add much explanatory value to a fitted model that already contains X_1.

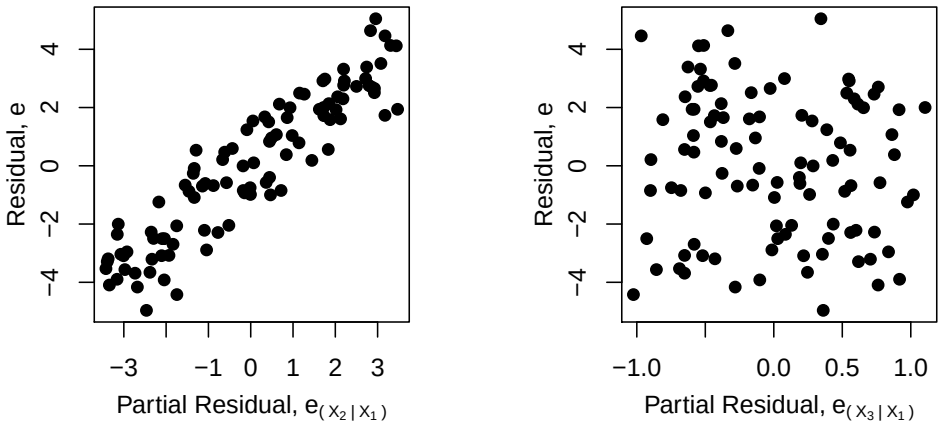


(a) The partial residual plot and the residual plot are nearly identical because X_1 and X_2 are uncorrelated ($r = -0.003$).



(b) The partial residual plot and the residual plot are visibly different because X_1 and X_2 are highly correlated ($r = 0.985$).

Figure 3. Comparison of candidate residual plots and partial residual plots to detect correlation between a candidate predictor (e.g., X_2) and a predictor already included in the fitted model (e.g., X_1)



(a) Linear trend occurs when a candidate predictor (e.g., X_2 shown here) should be added to the fitted LM.

(b) Random scatter occurs when a candidate predictor (e.g., X_3 shown here) should not be added to the fitted LM.

Figure 4. Evaluating patterns in partial residual plots: (a) linear trend; (b) random scatter

4. Cases of Misspecification Considered

Residual plots may be used to check for model misspecification. In this paper, we consider six LMs (Models 0 to 5), each of which has one predictor (X_1) or two predictors (X_1 and X_2), and possibly an interaction term (X_1X_2). Each of these "true" models (Table 1) relates the response ($T(Y)$) to a linear combination of predictor(s) (denoted by the linear predictor, η) and a random error (ϵ):

$$\textbf{True Model: } T(\mathbf{Y}) = \boldsymbol{\eta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma_\epsilon^2 \mathrm{I}\right). \tag{5}$$

Function $T(Y)$ represents the appropriate transformation for a response variable – possibly the identity ($T(Y) = Y$).

Table 1. Response, linear predictor, and X_1-X_2 correlation status, for each of Models 0 to 5 (Eq. 5). Model 0 is a simple LM with response Y and predictor X_1. Model 1 includes two predictors that are uncorrelated with one another. Model 2 includes two predictors that are correlated with one another. Model 3 includes two uncorrelated predictors and their interaction term. Model 4 includes two correlated predictors and their interaction term. Model 5 is a simple LM with response $ln(Y)$ and predictor X_1. As a matter of notation, $\rho_{1,2}$ represents the population correlation between X_1 and X_2. When X_1 and X_2 are uncorrelated, $\rho_{1,2}$ is zero; when X_1 and X_2 are correlated, $\rho_{1,2}$ is some nonzero value in $[-1, 1]$. For the linear predictors shown in this table, β_0 can have any real value, whereas β_1, β_2, and β_3 are nonzero so that models are distinct (e.g., so that Model 1 is not a special case of Model 3)

Model	**Response**, $T(Y)$	**Linear Predictor**, η	**Correlation**
0	Y	$\beta_0 + \beta_1 X_1$	$\rho_{1,2} = 0$
1	Y	$\beta_0 + \beta_1 X_1 + \beta_2 X_2$	$\rho_{1,2} = 0$
2	Y	$\beta_0 + \beta_1 X_1 + \beta_2 X_2$	$\rho_{1,2} \neq 0$
3	Y	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$	$\rho_{1,2} = 0$
4	Y	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$	$\rho_{1,2} \neq 0$
5	$ln(Y)$	$\beta_0 + \beta_1 X_1$	$\rho_{1,2} = 0$

Suppose a simple LM (Model 0) is fit to data generated from Models 0 to 5. Residual plots and partial residual plots can be used to identify whether or not the fitted model (Model 0) is misspecified and to distinguish types of misspecification (e.g., transformed response, missing predictor, missing interaction term), which we classify into cases (Fig. 5). We defined six cases of model misspecification as follows. Case 0 is the null case (no misspecification) when the fitted model has the appropriate form (i.e., is correctly specified). Otherwise, Case c ($c = 1, ..., 5$) denotes misspecification that occurs when Model 0 is fit for data that are properly described by another model (i.e., Model c).

5. Workflow for Distinguishing Cases of Misspecification

In Section 5, we describe our workflow for distinguishing cases of misspecification (Sect. 5.1) and provide examples for using our workflow (Sect.s 5.2, 5.3).

5.1 Description of the Workflow

To distinguish among the particular cases of misspecification described in Section 4 (Cases 0 to 5), our workflow uses residual and partial residual plots to answer four questions. Figure 5 lists these questions, shows how they can be mapped to cases, and describes which residual plots can be used to address each question. Questions are numbered for convenience in referencing, and it is not necessary to answer these questions in any particular order. A template that combines scatterplots and questions used in our workflow is shown in Figure 6.

5.2 Simulated Examples Using the Workflow

We will demonstrate how our workflow (described in Sect. 5.1) can be applied to four example datasets (Fig. 7). These datasets were simulated using models described in Table 1. Thus, the "true" model is known for each dataset. For Examples 1 to 4, β_0 was fixed; its value does not impact the pattern of scatter of residual plots. The sample size (n) was 100 observations per dataset. Values of $\beta_1, \beta_2, \beta_3$, and $\rho_{1,2}$ (for these examples) are shown in Table 2.

Standardized residuals were used in Figure 7 so that the vertical scale would not itself be a distinction between examples with transformation needed (unlike Fig. 2) since scale would vary across cases in realistic scenarios.

Example 1 (Fig. 7a): The standardized residual vs. predictor (left) scatterplot has random scatter and does not show evidence of nonconstant variance, which indicates that transformation and interaction can be eliminated as possible types of misspecifications. Candidate X_2 can be eliminated as a predictor based on random scatter in the candidate versus residual plot (middle) and the partial residual plot (right). The candidate residual plot (middle) and the partial residual plot (right) are nearly identical, which indicates that X_2 and X_1 are uncorrelated. Thus, this is an example of Case 0 (fitted model is correct) since we answered "No" to all four misspecification questions.

Residual plots can be used to answer model misspecification questions, such as the questions shown below.

Figure 5. Tree diagram for cases of misspecification (Cases 0 to 5). This tree diagram relates Cases 0 to 5 through binary questions about types of model misspecification that can occur when a simple LM is fit to data generated from Models 0 to 5.

Question	Use of Residual Plots
1. Should a Y transformation be considered?	Look for funnel pattern in residual vs. $\widehat{Y}$.
2. Should X_2 be added to the fitted model?	Look for linear trend in residual vs. X_2.
3. Is X_2 correlated with X_1?	Compare partial residual plot to residual vs. X_2.
4. Should X_1X_2 be added to the fitted model?	Look for bowtie pattern in residual vs. $\widehat{Y}$.

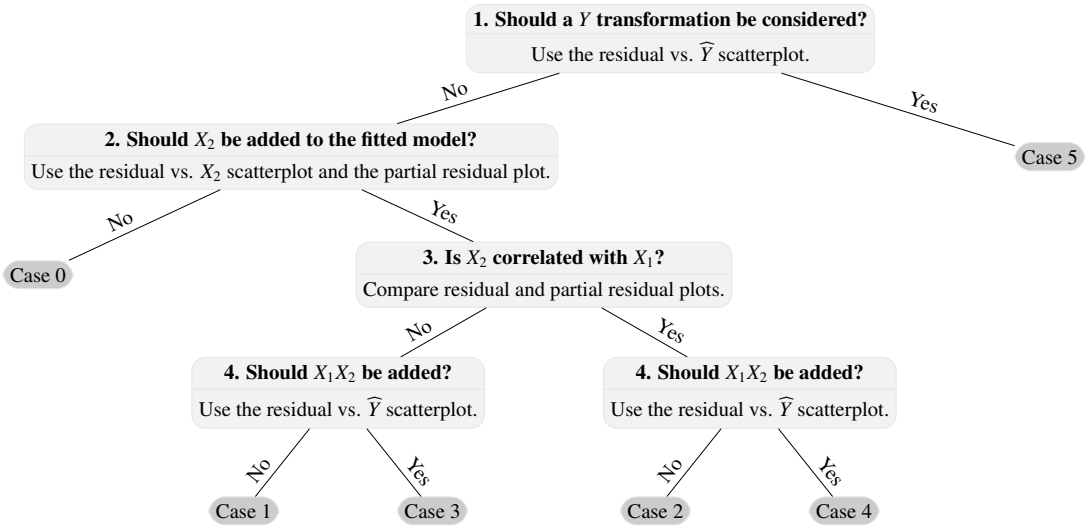


Figure 6. A template with the scatterplots and questions used in our workflow

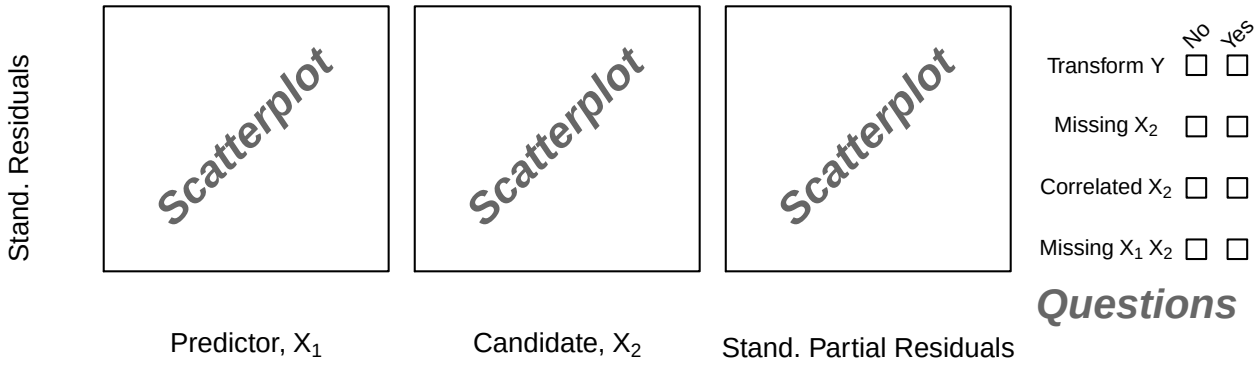
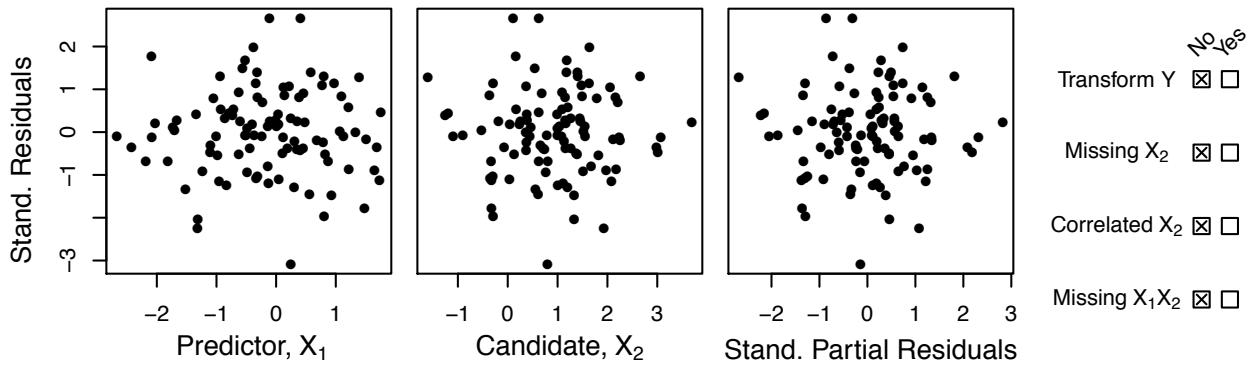
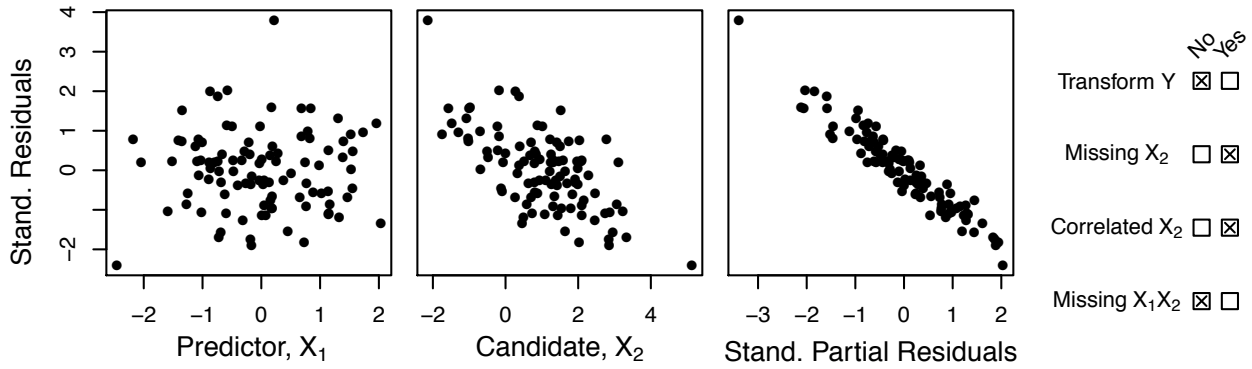


Table 2. Parameters for simulated examples in Section 5.2. Regression coefficients β_1, β_2, and β_3 correspond to coefficients for the linear model from which response data was generated (Eq. 5; Table 1). The correlation coefficient, $\rho_{1,2}$, denotes the correlation between the populations from which X_1 and X_2 data were generated. Example numbers in this table corresponds to simulated examples in Section 5.2

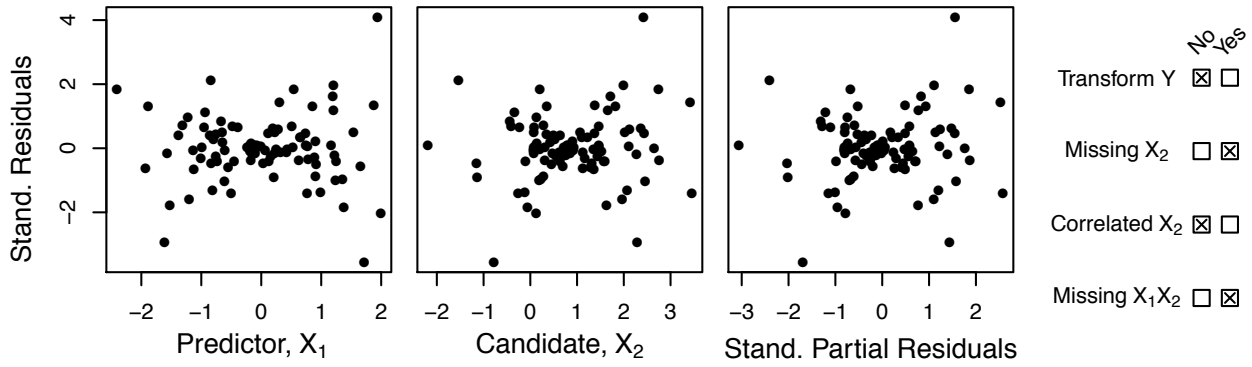
Example	β_1	β_2	β_3	$\rho_{1,2}$
1	-3	0	0	0.00
2	10	-3	0	-0.62
3	-1	1	10	0.00
4	-1	0	0	0.00



(a) Example 1: This Case 0 example exhibits random scatter in all residual plots.



(b) Example 2: This Case 2 example has nonrandom scatter in the candidate residual plot and the partial residual plot. And the candidate residual plot is noticeably different than the partial residual plot.



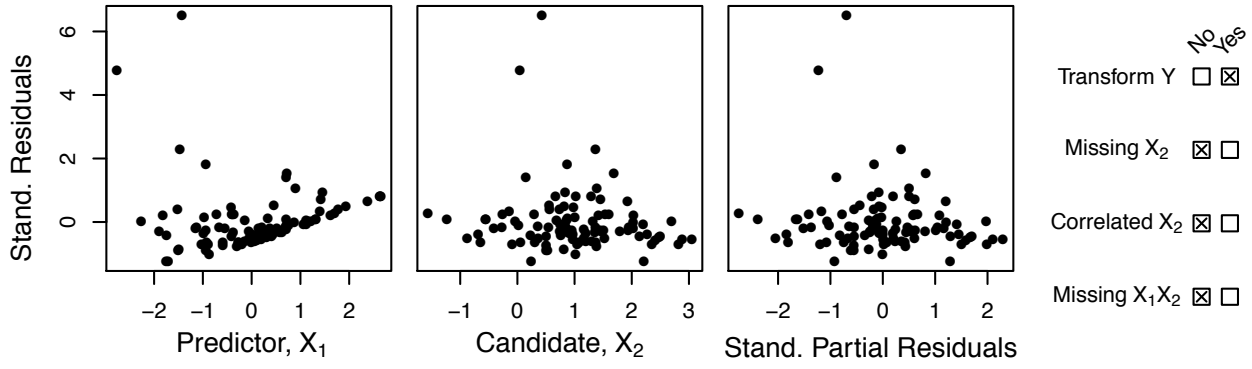
(c) Example 3: This Case 3 example shows evidence of a missing, uncorrelated candidate predictor X_2 and a missing X_1X_2 interaction term. That is, X_2 and X_1X_2 should be added to the fitted model.

Figure 7. Example residual and partial residual plots sets to illustrate our workflow

Example 2 (Fig. 7b): Transformation and interaction can be eliminated based on random scatter in the left plot. Based on the linear trend in the middle and right plots, candidate X_2 should be added to the fitted model. Finally, the difference in shape between the middle and right plots indicates that X_2 is correlated with X_1. Thus, this is an example of Case 2, which means X_2 is linearly related to Y and correlated with X_1.

Example 3 (Fig. 7c): The spacing of residuals within the "bowtie" pattern in the left plot indicates that an interaction term should be added to the fitted model and that no response transformation is needed. The shape of the scatter in the candidate and partial residual plots (middle and right plots, respectively) are nearly identical, which indicates that X_2 is uncorrelated with X_1. Unlike Example 1, the scatter in Example 3's candidate residual plot and the partial residual plot appear to be nonrandom. And even though Example 3's X_2 data are linearly related to Y, neither the candidate residual plot nor the partial residual plot show a linear trend because of the missing interaction term. Thus, this is an example of Case 3, where X_2 and X_1X_2 should be added to the fitted model and X_2 is uncorrelated with X_1.

Example 4 (Fig. 7d): The funnel pattern and uneven vertical spacing of residuals in the left plot indicates that a transformation is needed to correct this nonconstant error variance. By process of elimination, we can determine that Example 4 belongs to Case 5 because, of the six cases (Cases 0 to 5) discussed in this article, Case 5 is the only case with a transformation needed. Thus, within our simulation, Case 5 can be distinguished from Cases 0 to 4 by answering a single question. In general, additional cases could be defined such that some individual cases have both predictor omission problems as well as an incorrect response transformation. Considering other misspecification questions, the middle and left plots appear to be nearly identical, which indicates that X_1 and X_2 are uncorrelated. After accounting for uneven vertical spacing of residuals spacing due to untransformed response, the pattern in the candidate and partial residual plots does not appear to provide support for adding X_2 to the model. If $log(Y)$ had been linearly related to X_2 and X_1 and X_2 were uncorrelated, we would expect to see evidence in the middle and right plots of an exponential relationship between residuals and X_2, which we do not see in Example 4.



(d) Example 4: This Case 5 example could be corrected using the log-transformed response.

Figure 7. Workflow Examples (continued)

5.3 Application to SAT Dataset

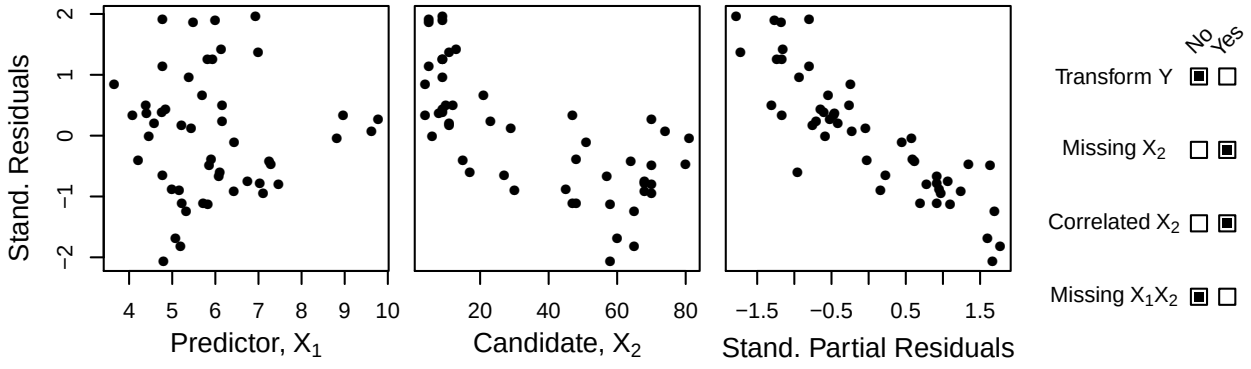
We will illustrate our workflow with a real dataset. From R's UsingR package, we used the SAT dataset [Verzani, 2015], a well-known example of the impact of predictor omission introduced by [Guber, 1999]. The SAT dataset includes statistics about schools and SAT performance and participation by state for the 1994-1995 school year. The following fields are included for each state: average expenditure per student, average student-teacher ratio, average teacher salary, student SAT participation rate, and average SAT scores (total, verbal, and math). Additional information about the dataset is provided at http://jse.amstat.org/datasets/sat.txt.

We first note that unlike the previous, simulated examples, the "true" model is unknown for this real dataset. A typical approach in this situation is to fit a simple linear model to a dataset (based on the principal of parsimony) and then use residual plots to determine if a more complicated model (e.g., models 1 through 5) may be more appropriate.

We fit a simple linear model to the SAT dataset to predict average SAT math score (Y) using expenditure (X_1). SAT participation (X_2; the percentage of eligible students who took the SAT) is considered as a candidate predictor that could be added to the fitted model. Then we applied Steps 1 to 4 of our workflow (shown in Fig. 5) to assess this fitted model based on the residual and partial residual plots in Figure 8.

- Step 1: Vertical spacing of residuals in the residual versus predictor scatterplot does not show evidence that a transformation of the response variable (math score) is needed.
- Step 2: Based on the linearity in the partial residual plot, we determined that the candidate predictor (X_2, participation) should be added to the fitted model.
- Step 3: Differences in scatter between the candidate versus residual plot and the partial residual plot indicate that the candidate predictor (participation) is correlated with the predictor already included in the model (expenditure).
- Step 4: For this small example dataset, the consideration of including the interaction term between expenditure and participation in Step 4 is more challenging in that patterns may be less distinguishable when there are fewer data points. There does not appear to be strong evidence for adding the expenditure-participation interaction term to this model because there is not a general bow-tie shape to the residual pattern.

Figure 8. Residual and partial residual plots from a simple LM for SAT math score regressed on Education Expenditure (X_1). Participation (X_2), the candidate predictor, is the percentage of eligible students who took the SAT. Black filled in boxes in Figure 8 indicate the authors' answers to workflow questions, whereas "x"s in previous figures indicate correct answers (based on knowledge of the true model)



5.4 Discussion: Using Simulation for Training Purposes

We advocate for the use of simulation when training students to identify potential problems with model assumptions. Use of simulated datasets allows students to learn to recognize patterns, such as those indicative of misspecification types discussed in this paper, while knowing the "true model." Some benefits of using simulated data for initial training include sample size control, more examples, and continued exploration. First, simulation allows for sample size to be chosen, which, as discussed in Section 5.3, is important because the sample size can impact the recognizeability of some patterns. Simulation can be used to illustrate what happens under different sample size scenarios. Simulation also allows for increasing the number of examples to provide students with more practice, in general, and more practice distinguishing differences between random variation in data generated by the same true model and systematic differences in the true model. Third, in addition to the ability to apply lessons learned in a deductive way (as proposed in our workflow), some types of interactive simulation allow for continued exploration through inductive reasoning. Students can come up with their own scenarios of the true underlying models and then "see" how the residual patterns from these models appear under different fitted models. Such hands-on, student-guided exploration has the potential to encourage and strengthen students' statistical intuition, while still within a known simulation framework.

6. Conclusion

A workflow for using residual plots to distinguish among cases of correlated and/or interacting predictor omission has not been previously outlined in the literature. Rather than suggesting a new type of residual plot, we suggest using a combination of traditional residual and partial residual plots to distinguish among five types of misspecification, include two types of nonconstant error variance and four cases of predictor omission. Our suggested methods are accessible to researchers of all experience levels with the goal of enabling a wide audience to distinguish types of misspecification encountered in practice. In this paper, we described our workflow and illustrated its use for simulated examples and a real example with SAT data.

We have identified four avenues of future work that build on the workflow described in this paper. First, future research could be conducted to consider generalizing our workflow to include nonnormal distribution misspecification issues (i.e.,

the generalized linear model (GLM) context). Specific questions for consideration include: *Does combining residual plots and partial residual plots identify correlation among predictors in the GLM context?* and *Can traditional residual plots be used to distinguish sources of nonconstant error variance (e.g., missing interaction term versus transformed response) in the GLM context?* Second, future research could be conducted to evaluate the effectiveness of our workflow in handling cases of model misspecification that simultaneously involve both predictor omission problems and incorrect response transformations. Third, our workflow could be implemented in a web application to enable interactive practice in using residual plots to distinguish among types of model misspecification. Fourth, future research could generalize the workflow for contexts where there are more than two predictors considered (e.g., one predictor in the fitted model and two predictors omitted from the model).

Conflict of Interest and Funding Statements

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of interest: none

Data Availability Statement

The SAT dataset from R's UsingR package was used in our article. This dataset is publicly available through R; additional information about the dataset itself is available at http://jse.amstat.org/datasets/sat.txt.

References

Ezekiel, M. (1924). A method for handling curvilinear correlation for any number of variables, *JASA*, *19,* 431-453.

Faraway, J. (2005). *Extending the Linear model with R: generalized linear mixed effects and nonparametric regression models.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Fox, J. (1991). *Regression diagnostics: An introduction*. Sage University Paper Series on Quantitative Applications in Social Sciences. 07-79. Newbury Park, CA: Sage.

Gray, J. (1989). On the use of regression diagnostics, *Journal of the Royal Statistical Society. Series D (The Statistician), 38*(2), 97-105. https://doi.org/10.2307/2348307

Greene, W. (2003). Specification analysis and model specification. In *Econometric analysis*, 5th ed. (pp. 148-161). Upper Saddle River: Prentice Hall.

Guber, D. (1999). Getting what you pay for. *Journal of Statistics Education, 7*(2).

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models*, 5th ed. Boston: McGraw Hill.

Mansfield, R., & Conerly, M. (1987). Diagnostic value of residual plots. *American Statistician, 41*(2), 107-116. https://doi.org/10.1080/00031305.1987.10475457

Nystrom, E., Sharp, J. L., & Bridges, W. (2019). The impact of correlated and/or interacting predictor omission on estimated regression coefficients in linear regression. *Journal of Statistical Theory and Practice, 13*(56). https://doi.org/10.1007/s42519-019-0056-5

Ott, R., & Longnecker, M. (2001). Linear regression and correlation. In *An introduction to statistical methods and data analysis*. (6th ed.). (pp.572-663). Belmont, CA: Brooks/Cole Cengage Learning.

Tsai, C., Cai, Z., & Wu, X. (1998). The examination of residual plots. *Statistica Sinaca, 8*(2), 445-465.

Verzani, J. (2015). *SAT data*. In *UsingR* R package. Retrieved from https://CRAN.R-project.org/package=UsingR

Weisberg, S. (1985). *Applied linear regression*. (2nd ed.). New York: Wiley.

Woolridge, J. (2012). Multiple regression analysis: estimation. In *Introductory econometrics: A modern approach*, 5th ed. (pp.68-117). Mason, OH: South-Western Cengage Learning.

Xu, X., & Sinha, S. (2021). Robust designs for generalized linear mixed models with possible model misspecification, *Journal of Statistical Planning and Inference, 210*, 20-41. https://doi.org/10.1016/j.jspi.2020.04.006

Yoon, H., & Welsh, A. (2020). On the effects of ignoring correlation in the covariates when fitting linear mixed models, *Journal of Statistical Planning and Inference, 204*, 18-34. https://doi.org/10.1016/j.jspi.2019.04.001