HHS Public Access

Author manuscript

Am J Clin Nutr. Author manuscript; available in PMC 2016 September 01.

Published in final edited form as:

Am J Clin Nutr. 2015 September ; 102(3): 533–539. doi:10.3945/ajcn.115.113498.

Best (but of t-forgotten) practices: checking assumptions concerning regression residuals[1,2]

Lawrence E Barker[*] and **Kate M Shaw**

Centers for Disease Control and Prevention, Chamblee, GA

Abstract

The residuals of a least squares regression model are defined as the observations minus the modeled values. For least squares regression to produce valid CIs and *P* values, the residuals must be independent, be normally distributed, and have a constant variance. If these assumptions are not satisfied, estimates can be biased and power can be reduced. However, there are ways to assess these assumptions and steps one can take if the assumptions are violated. Here, we discuss both assessment and appropriate responses to violation of assumptions.

Keywords

assumptions; regression; statistics; residuals; variance; normality

Introduction

One wants to conduct a linear regression, a model in which a dependent variable (assumed, to limit scope, continuous and uncensored) is predicted from one or more independent variables. Alternately, one might consider a 1- or 2-sample *t* test, an ANOVA, or an ANCOVA, because all are special cases of regression; for simplicity, we will continue to say “regression,” unless one of the special cases is specifically intended.

One gathers observations of both the dependent variable and the independent variable(s). One uses software that implements regression. Suppose the slope is significantly different from zero. Does that mean that one is done? Actually, additional steps are needed to make sure the conclusion is valid. We will discuss some (but not all) of them.

Linear regression (the example below assumes one independent variable, but there can be any number) involves a model of the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

[1]The authors reported no funding received for this study.

[2]The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC.

[*]To whom correspondence should be addressed. lsb8@cdc.gov.

The authors' responsibilities were as follows—LEB and KMS: responsible for design, writing, and final content. Neither author declared a conflict of interest related to this study.

where, for the ith observation, Y_i is the dependent variable, X_i is the independent variable, β_0 and β_1 are fixed but unknown constants, and ε_i is a random variable accounting for measurement error/lack of fit of model. However, for least squares, the most common form of regression, to work, certain assumptions concerning $\{\varepsilon_i\}$ must be satisfied:

1. $\{\varepsilon_i\}$ must be uncorrelated.
2. $\{\varepsilon_i\}$ must be normally distributed.
3. Var$\{\varepsilon_i\}$ must be constant.

For a more complete explanation, see Berry (1).

To be clear, these are assumptions about $\{\varepsilon_i\}$ and do not apply to the dependent or independent variables. In particular, neither the dependent nor independent variables need to be normally distributed.

Violation of these assumptions can cause coverage of CIs (the probability that the CI contains the true parameter value) to be very different from the nominal value (the value that was calculated under the assumptions listed above). Similarly, the actual and nominal values of the probabilities of type I error and type II error can be far apart. The more severe the violation, the more severe the impact. In practical terms, although a minor violation might or might not have much practical consequence, a severe violation often leads to poor statistical performance.

Of course, $\{\varepsilon_i\}$ cannot be directly observed. We must, therefore, work with the estimated residuals, $\{\hat{\varepsilon_i}s\}$, defined as the observed ith value of the dependent variable minus the modeled ith value. How can we decide, based on the estimated residuals, if the assumptions about the unobserved $\{\varepsilon_i\}$ are satisfied? There are both formal tests and less formal graphical methods, both of which have advantages. Tests are objective and can, if necessary, be automated. However, with large sample sizes, tests can flag trivial deviations from the assumptions. Tests only reject the null hypothesis if the evidence is strong; with small sample sizes, tests might fail to detect violations that, although not statistically significant, might still be problematic if real. Graphical methods are subjective. However, graphical methods often allow one to judge the severity of the departures from the assumptions. The degree of severity determines how badly remedial measures are needed—and, indeed, if they are likely to make much difference. Admittedly, this might be difficult for those without extensive statistical experience, so the reader should assess his or her own ability to interpret graphics before implementing them. Here, we discuss both tests and graphical methods for assessing assumptions and what to do if the assumptions are violated.

Assessing Assumptions

Residuals are uncorrelated

Knowledge of how the data were gathered often makes independence (which implies no correlation) plausible. However, correlation of population residuals can be assessed. Note that one must distinguish $\{\varepsilon_i\}$, the population residual, from the estimated residuals, $\{\hat{\varepsilon_i}\}$;

some correlation is always present in the $\{\hat{\varepsilon_i}\}$, because it is a property of linear regression that $\Sigma\{\hat{\varepsilon_i}\}$ must always be 0.

There are infinitely many forms of correlation. For example, observations from similar individuals might be correlated. Clustering (e.g., if the data were collected in a few locations, observations collected at a single location) can induce correlation. We limit our discussion to what is perhaps the most common form of correlation, serial correlation—if observations are gathered sequentially, residuals that occur near one another might be correlated (autocorrelation).

Test—The most common test for assessing serial dependence is based on the Durbin-Watson statistic (2, 3). Values of the Durbin-Watson statistic that are larger than the upper tail critical value suggest positive correlation, whereas values that are smaller than the lower tail critical value suggest negative correlation. The null hypothesis is lack of correlation and so is rejected only if evidence of correlation is strong.

Graphics—The Durbin-Watson statistic can be interpreted by noting that it approximately equals $2(1 - r)$, where r is the sample correlation between estimated residuals and lag-one estimated residuals (hereafter, simple "residuals") (4). This can be graphically exploited by plotting a scatterplot of residuals vs. lag-one residuals. If the assumption is satisfied, one should see a patternless blob (in particular, it does not resemble a straight line); a pattern (particularly scatter around a line) suggests there might be an issue. In Figure 1, we illustrate what a scatterplot in which residuals are truly independent and what a scatterplot in which the residuals are positively correlated might look like. (All figures were generated by using simulated data—details of the simulation are not given here, for the sake of brevity. Because the residuals are the result of a simulation, we know, unlike the situation that would arise with real data, for certain that they are independent/positively correlated.)

Normality

The Gauss-Markov theorem shows us that regression provides the linear unbiased estimate with the smallest possible variance, even if residuals are not normally distributed. This is sometimes misinterpreted to mean that normality is not important. The Gauss-Markov theorem only concerns point estimates, not tests or CIs. Regression estimates can be especially sensitive to heavytailed distributions (5).

Test—Many tests for normality of residuals have been proposed. The D'Agostino test (6) is based on sample skewness (a measure of symmetry) and kurtosis (a measure of how heavy the distribution's tails are). In some cases, the heaviness of the tails is the most important feature.

The Shapiro-Wilk test (7) is a formalization of the quantile-quantile plot (8), a comparison of theoretical and empirical quantiles. Lilliefor's test (9) is a normality test related to the one-sample Kolmogorov-Smirnov test (Lilliefor's test is for a family of distributions, whereas the Kolmogorov-Smirnov test is for a single specified distribution). In fact, Lilliefor's test is sometimes incorrectly identified as the Kolmogorov-Smirnov test.

In all of these tests, the null hypothesis is normality. The null is rejected only if evidence of nonnormality is strong. Therefore, one should be cautious in implementing these tests with small sample sizes. In such cases, departures from normality that can have substantial consequences are sometimes not detected.

Graphics—The simplest way to graphically evaluate normality of residuals is to plot a histogram and examine it for departures from normality. This can often reveal skewed residuals. It can be problematic for heavy-tailed residuals, because some heavy-tailed symmetric distributions can look quite normal. A more sophisticated method, and one that often reveals deviations from normality that are difficult to see in a histogram, is to plot the empirical quantiles of the residuals against the theoretical quantiles of a normal distribution (a quantile-quantile plot). A straight line suggests normality, whereas a curved suggests a departure from normality. In Figure 2, we illustrate what normal and nonnormal quantile-quantile plots of residuals might look like. The residuals are the result of a simulation and therefore very clear-cut. In particular, the residuals in Figure 2A were truly normally distributed, instead of being "approximately normal," as one usually encounters in practice. Actual data are likely to be more ambiguous, requiring judgment. For example, a small departure from linearity might have very small practical consequences.

Constant variance

Variance of residuals can, in a departure from the assumptions that underlie linear regression, change as independent variables do. It is particularly common that the variance of the residuals increases with values of an independent variable. Even with nonconstant variance, linear regression provides unbiased point estimators (Gauss-Markov theorem). However, with nonconstant residual variance, the nominal and actual probabilities of type I and type II errors can be very different. Similarly, coverage of CIs can be far from their nominal values.

This is of particular concern with ANOVA. If a subpopulation has both a larger variance and a larger sample size, the resulting test becomes conservative (and of low power). Conversely, if a subpopulation has both a smaller sample size and a larger variance, the resulting test can be anticonservative (10).

Test—The Breusch-Pagan test (11) is a χ^2 test based on regressing the squared residuals on the independent variables. The Breusch-Pagan test can be sensitive to violations of normality. The White test (12) might be less sensitive to nonnormality. Like the Breusch-Pagan test, it depends on an auxiliary regression. Levene's test (13) estimates the variance under the assumption that the null hypothesis is true and again under more general assumptions and then calculates the ratio of those estimates. Values statistically significantly larger than 1.00 indicate nonconstant variance. However, whichever test is used, it is probably best to look for evidence of nonnormality before testing for constant variance. The null hypothesis is constant variance and so is rejected only if evidence of nonconstant variance is strong. In cases with small sample size, nonconstant variances that can have substantial consequences are sometimes not detected.

Graphics—The simplest form of graphical evaluation is to plot residuals vs. each of the independent variables (or, alternately, vs. the modeled values of the dependent variable). If the assumption is satisfied, one should see a patternless blob. A pattern suggests there might be an issue. The “v-shaped” pattern (or, if one prefers, “fan shaped” or “pie-wedge shaped”), where absolute values of residuals tend to increase as an independent variable increases, is reasonably common in real data. Therefore, one should be very wary of it. In Figure 3, we illustrate what a patterned and v-shaped scatterplot might look like.

What to Do When Assumptions Are Violated

The discussions of this section are, of necessity, somewhat sketchy. However, we provide references where the details of each method can be found by someone who needs to implement the method.

Residuals are correlated

How one handles correlated residuals depends on how much one knows about the correlation structure of the residuals. For example, if one knows that residuals are likely to be auto-correlated, this can be accounted for in modeling. Feasible generalized least squares is a method that has broad applicability. An explanation appears in Baltagi (14).

Residuals are not normally distributed

If a parametric family can be identified, then one can often achieve greatest power by explicit modeling. This can be done through generalized linear models (15). If no parametric family is found, a transformation of either the independent or dependent variables might help. The Box-Cox transform (16) provides a model-based transformation of the dependent variable. Tukey's ladder of transformations (17) provides a graphics-based method of choosing transformation of both the dependent and independent variables.

Robust regression is a variant of regression for which outliers in the residuals (but not necessarily in the independent variables) have little impact on the estimates. There are too many forms of robust regression to discuss here, although most involve down-weighting, in some manner, “extreme” residuals. For example, least trimmed squares (18) minimizes the sum of the squared “middle” residuals, deleting the extremes. An overview of robust regression appears in Rousseeuw and Leroy (19).

In the case of residuals of totally unknown parametric form, one can use resampling methods, in which one samples from the sample, to obtain estimates of standard errors that do not depend on parametric assumptions (although other assumptions, not specified here, must be made). For example, one can bootstrap residuals (a classic resampling method) or use jackknifing (another resampling method) on the entire set of observations. At one time, objections to resampling due to computational intensiveness were common. In today's world of cheap and easy computing, these objections are no longer valid. Indeed, many common statistical software packages (e.g., SAS, Stata) implement bootstrapping and/or the jackknife with a single command. An overview of resampling, as applied to regression, appears in Wu (20).

Finally, some authors [e.g., Valdar et al. (21)] advocate using the inverse normal transformation to make the dependent variable normally distributed. However, the assumption of normality concerns the residuals, not the dependent variable. Although the inverse normal transformation sometimes makes residuals normally distributed, it can also fail to do so (22). Thus, if one uses this method, one should still check residuals for normality.

Variance is not constant

Nonconstant variance can be extremely problematic, because data points with greater variance can have a disproportionate impact on the estimates. Fortunately, methods already mentioned can often help. For example, the Box-Cox transformation or Tukey's ladder of transformations can sometimes make variances approximately constant. Robust regression methods can be less sensitive to nonconstant variance than traditional methods. Weighted least squares [for an explanation, see Strutz (23)], feasible generalized least squares regression, and generalized linear models are all ways that one might deal with nonconstant variance. Huber-White standard errors (12, 24), readily available in many software packages, provide estimates that converge, as the sample size increases, to the true population value. This is yet another valid approach.

In conclusion, if the residuals of a least squares regression model do not satisfy the assumptions, *P* values and CIs might not perform as one expects. However, there are ways to assess these assumptions and steps one can take if the assumptions are violated. Many of these are easily implemented by using common statistical packages. With some effort, one can produce more defensible regressions.

Historically, some authors check assumptions and some do not. For example, Hirose et al. (25) did an ANOVA and *t* tests (both special cases of regression). They clearly reported checking assumptions. Similarly, so did Hussein et al. (26). Thus, we need not worry about the issue of validity of assumptions when assessing their research. On the other hand, Vors et al. (27) conducted an analysis that included both analysis of variance and *t* tests. The authors did not indicate that assumptions were tested. That does not necessarily mean that the assumptions were violated. It does not even mean that they were not tested. However, it does mean that the readers do not know. Thus, one does not know whether to worry about violation of assumptions in this work.

Finally, this report has not covered all important issues. We have not addressed outliers in either the independent variables (high-leverage data points) or the dependent variables (all but a small number of residuals are approximately normally distributed, but those few suggest very heavy tails). We have not addressed collinearity (some linear combination of independent variables is approximately constant). We have not discussed linearity of relation between dependent and independent variable (s). These issues are important but beyond the scope of this report.

Acknowledgments

The authors thank Pamela Sedgwick-Barker and Srila Sen for their editorial contributions.

References

1. Berry, WD. Understanding regression assumptions. Newbury Park (CA): Sage; 1993.
2. Durbin J, Watson GS. Testing for serial correlation in least squares regression, I. Biometrika. 1950; 37:409–28. [PubMed: 14801065]
3. Durbin J, Watson GS. Testing for serial correlation in least squares regression, II. Biometrika. 1951; 38:159–78. [PubMed: 14848121]
4. Gujarati, DN.; Porter, DC. Basic econometrics. 5th. Boston: McGraw-Hill Irwin; 2009.
5. Hogg, RV. An introduction to robust estimation. In: Launer, RL.; Wilkinson, GN., editors. Robustness in statistics. New York: Academic Press; 1979. p. 1-17.
6. D'Agostino RB, Pearson ES. Tests for departure from normality: empirical results for the distributions of b_2 and $\sqrt{b_2}$. Biometrika. 1973; 58:341–8.
7. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965; 52:591–611.
8. Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis of data. Biometrika. 1968; 55:1–17. [PubMed: 5661047]
9. Lilliefors H. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J Am Stat Assoc. 1967; 62:399–402.
10. Glass, GV.; Hopkins, KD. Statistical methods in education and psychology. 3rd. Pearson; New York: 2008.
11. Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient variation. Econometrica. 1979; 47:1287–94.
12. White HA. Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica. 1980; 48:817–38.
13. Levene, H. Robust tests for equality of variance. In: Olkin, I., editor. Contributions to probability and statistics: essays in honor of Harold Hotelling. Stanford (CA): Stanford University Press; 1960. p. 278-92.
14. Baltagi, BH. Econometrics. 4th. New York: Springer; 2008.
15. Lindsey, JK. Applying generalized linear models. New York: Springer; 1997.
16. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc B. 1964; 26:211–52.
17. Tukey, JW. Exploratory data analysis. Reading (MA): Addison-Wesley; 1977.
18. Atkinson AC, Cheng TC. Computing least trimmed squares regression with the forward search. Stat Comput. 1999; 9:251–63.
19. Rousseeuw, PJ.; Leroy, AM. Robust regression and outlier detection. New York: John Wiley; 2003.
20. Wu CFJ. Jackknife, bootstrap, and other resampling methods in regression. Ann Stat. 1986; 14:1261–95.
21. Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JN, Mott R, Flint J. Genetic and environmental effects on complex traits in mice. Genetics. 2006; 174:959–84. [PubMed: 16888333]
22. Beasley TM, Ericson S. Rank based inverse normal transformations are increasingly used, but are they merited? Behav Genet. 2009; 39:580–95. [PubMed: 19526352]
23. Strutz, T. Data fitting and uncertainty: A practical introduction to weighted least squares and beyond. Springer Vieweg Vieweg+Teubne; Berlin: 2011.
24. Huber, PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In: Le Cam, LM.; Nehman, J., editors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. City of publication; Berkley CA: 1967. p. 221-33.
25. Hirose A, Terauchi M, Tamura M, Akiyoshi M, Owa Y, Kato K, Kubot T. Tomato juice intake increases resting energy expenditure and improves hypertriglyceridemia in middle-aged women: an open-label, single-arm study. Nutr J. 2015; 14:34. [PubMed: 25880734]
26. Hussein MO, Hoad CL, Wright J, Singh G, Stephenson MC, Cox EF, Placidi E, Pritchard SE, Costigan C, Ribeiro H, et al. Fat emulsion intragastric stability and droplet size modulate

gastrointestinal responses and subsequent food intake in young adults. J Nutr. 2015; 145:1170–7. [PubMed: 25926408]

27. Vors C, Pineau GI, Gabert L, Drai J, Louche-Pe'lissier C, Defoort C, Lairon D, De'sage M, Danthine S, Lambert-Porcheron S, et al. Modulating absorption and postprandial handling of dietary fatty acids by structuring fat in the meal: a randomized crossover clinical trial. Am J Clin Nutr. 2013; 97:23–36. [PubMed: 23235199]

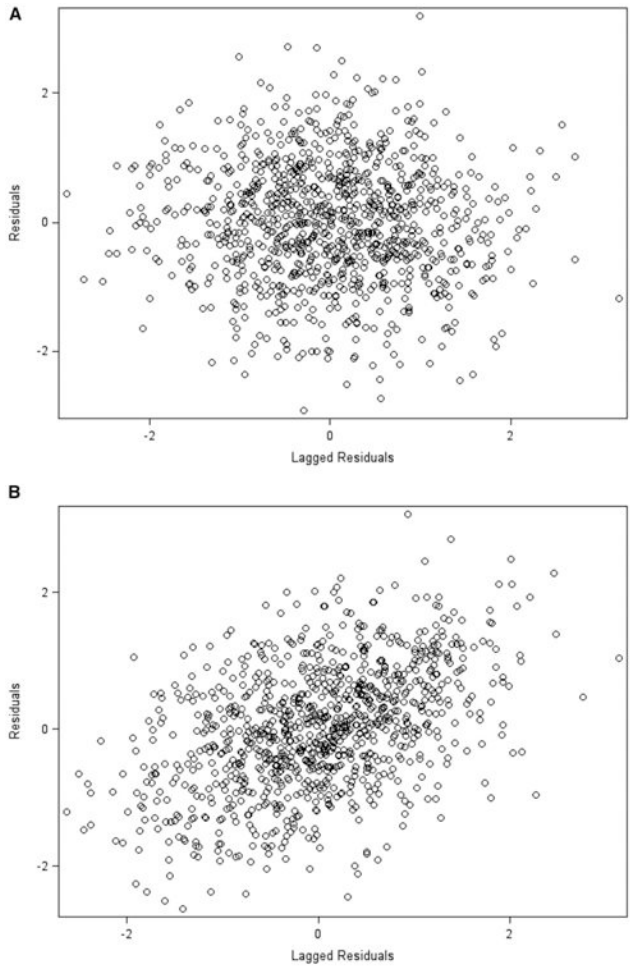

Figure 1.
Uncorrelated residuals, Pearson correlation = −0.06 (A). Correlated residuals, Pearson correlation = 0.45 (B).

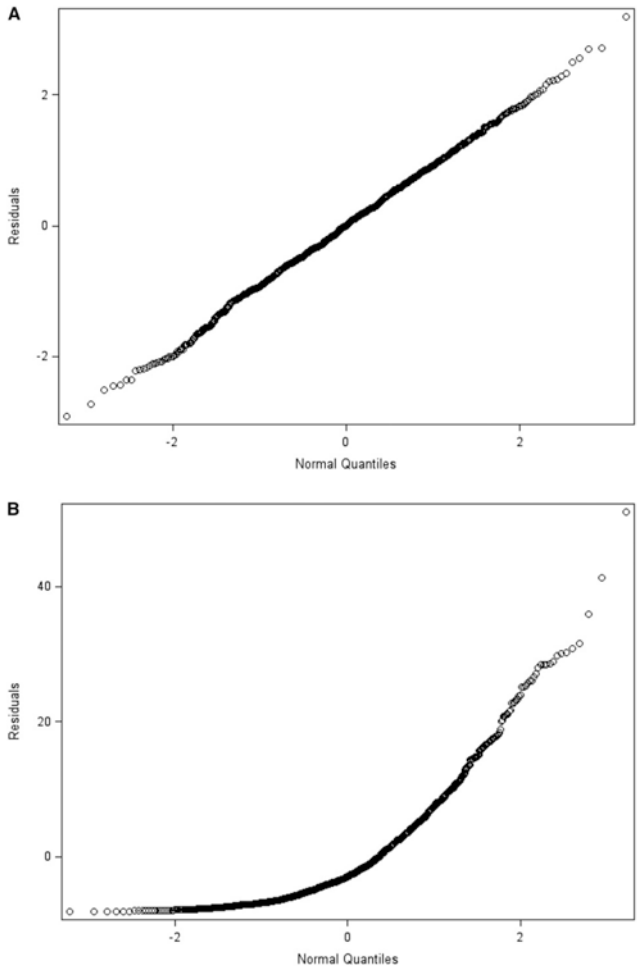

Figure 2.
Residuals normally distributed, quantile-quantile plot (A). Residuals not normally distributed, quantile-quantile plot (B).

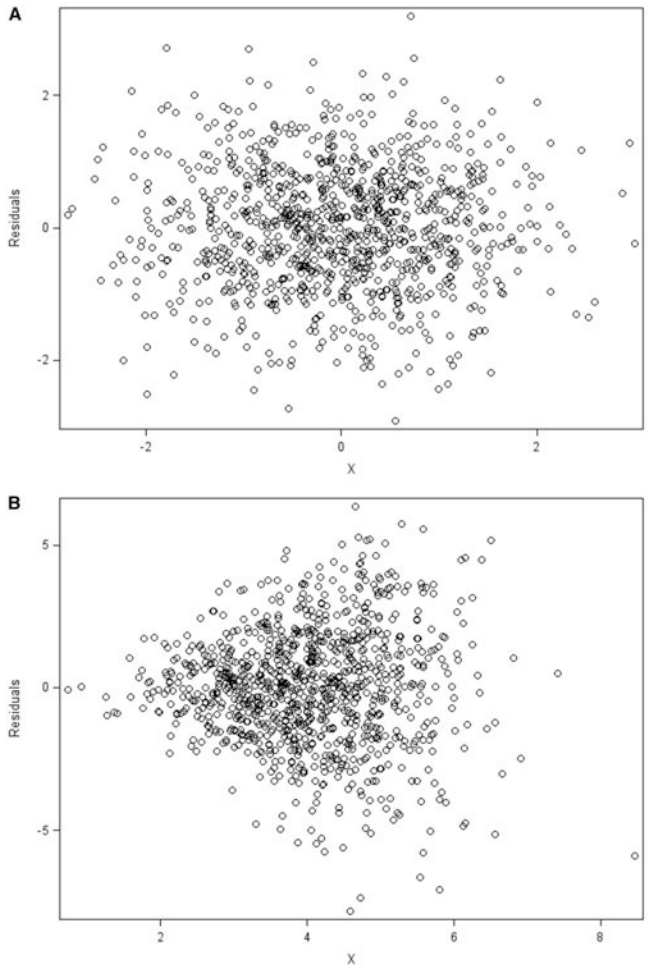

Figure 3.
Constant variance of residuals, Pearson correlation = 0.00 (A). Increasing variance of residuals, Pearson correlation = 0.00 (B).