Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.37301

Statistical Diagnostics and Reliability Evaluation of Linear Regression Models

Haoran Ni

LuHe International Academy, Beijing, China
nihaoran618@163.com

Abstract. Linear regression is often used to explain, infer statistically and predict, but a good fit within the training data does not guarantee the reliability of the model. The main diagnostic and validation tools for assessing linear regression models are introduced in this paper, including residual analysis, multicollinearity, influential observations and out-of-sample evaluation. First, it presents the assumptions of ordinary least squares and explains how residual plots, heteroscedasticity tests, autocorrelation diagnostics, variance inflation factors, leverage measures, Cook's distance and other statistics can reveal different sources of model instability. Then it will present recent studies on improved multicollinearity diagnosis, robust estimation, influence measures and cross-validation. Based on the above comparisons, the diagnostic conclusions may differ depending on the sample size, model selection, data quality and decision-making threshold. Other reasons are that multiple assumptions have been violated simultaneously, there is sensitivity to influential observations, instability in high-dimensional spaces, and discrepancies among validation methods. In the future, good estimation, uncertainty-aware diagnosis, high-dimensional methods and transparent out-of-sample validation will be integrated to enhance the reliability and reproducibility of linear regression analysis.

Keywords: linear regression, statistical diagnostics, residual analysis, multicollinearity, cross-validation

1. Introduction

Linear regression is a general-purpose statistical method that can be used in many places, such as economics, engineering, natural sciences, social sciences, etc. [1]. The model can quantify the relationship between the response variable and one or more explanatory variables, estimate parameters, conduct hypothesis tests and make out-of-sample predictions [1]. A large coefficient of determination (R^2) does not guarantee real-world applicability. When the assumptions of linearity, exogeneity, homoscedasticity or error independence are violated, the ordinary least squares (OLS) estimates and their standard errors may be unreliable; residual normality is additionally required for exact small-sample inference [2-4]. In addition to assumption violations, multicollinearity among the explanatory variables, as well as the presence of outliers and influential observations, can destabilize regression coefficients, reduce the significance of tests, and harm prediction [5-8]. A high R^2 value under these circumstances may be due to model misspecification or overfitting, and thus

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.37301

does not necessarily indicate a real relationship. Therefore, statistical diagnosis and systematic model evaluation are required prerequisites for building a reliable linear regression model [8, 9]. In recent years, many developments have occurred in the field of diagnostics and validation for linear regression models. Residual analysis is used to check the assumptions of the model, and standardised residuals, studentised residuals and residual plots can help identify non-linear trends, heteroscedasticity, influential observations and deviations from normality [2, 8, 9]. Recent methods also address heteroscedasticity and clustered dependence in regression estimation and inference [3, 4]. Diagnostics for multicollinearity have moved beyond fixed cutoffs of Variance Inflation Factor (VIF); recently, data-dependent decision rules based on redefined VIF values have been proposed [5]. Outlier and influential point identification is based on measures such as Cook's Distance, standardised differences in coefficient estimates (DFBETAS), and leverage values [8]. Recently, more studies have refined the distribution and critical values of Cook's Distance [10]. Besides in-sample diagnostic indicators, gradually more researchers have employed out-of-sample testing. The data can be split into a training set and a test set to evaluate the model's generalisation capability, and cross-validation can be used for multiple estimates of prediction error [11]. The estimand and uncertainty of cross-validation nevertheless require careful interpretation [11]. Although an all-encompassing diagnostic and verification system has been built, many empirical studies still focus too much on fitting indicators and neglect full-process tests, thus yielding statistically invalid conclusions. Based on the above background, this paper reviews statistical diagnostics and reliability assessment for linear regression, and examines how multiple diagnostic indicators and cross-validation can be used to evaluate model robustness.

2. Statistical foundations and basic assumptions of linear regression

The two general types of linear regression, and thus the selection of a single-predictor or multi-predictor model, depend on the specific analysis circumstances of the researcher. A single-predictor model relates the outcome to a single independent variable, whereas multiple linear regression can be employed to consider multiple predictors simultaneously.

The end result of the study or prediction in this analytical system is the main object of exploration, and the explanatory variables are factors associated with it. The intercept is the expected value of the response when all predictor variables are zero, and partial regression coefficients show the association between a predictor and the response after controlling for other predictors. The random error term includes all unobserved factors, measurement errors and model misspecification that are not included in the predictors.

Parameter estimation is carried out using OLS to obtain fitted values that approximate the actual observations, and a residual is defined as the difference between the recorded outcome and the model's prediction; this gap shows how well each individual sample fits the overall trend captured by the model [1, 8].

As the standard parameter estimation method for linear regression, OLS aims to minimise the sum of squared residuals between the observed and fitted values. In terms of geometry, this way finds the hyperplane that is closest to all the data points vertically. If the Gauss-Markov conditions are met, such as a correct linear relationship, exogenous variables, no perfect multicollinearity, homoscedasticity, and uncorrelated random errors, then OLS estimators are the best linear unbiased estimators [1, 8].

Consistent and stable OLS results are based on the basic theoretical assumptions of valid regression inference. The first requires a linear additive link between predictors and outcomes [1, 8]. Second, the random errors of different samples should be uncorrelated to avoid serial correlation [8].

DOI: 10.54254/2753-8818/2026.37301

Third, for OLS to be efficient and for the conventional standard errors to be valid, the error variance should be constant across predictor values [2, 3]. Fourth, the predictors must be exogenous; thus, under any given set of covariates, the expected value of the random errors is zero. Normality of the errors is not required for BLUE properties, but it enables exact small-sample inference [2]. Fifth, explanatory variables cannot form exact linear combinations; otherwise, coefficient estimation would be infeasible [5, 8].

The three applications of linear regression in applied research are: fitting to the existing data, statistical hypothesis testing, and forecasting of unseen out-of-sample observations. Clarifying the connection among these functions will form an all-encompassing analytical system for residual inspection, multicollinearity screening, outlier identification and cross-validation testing [8, 11].

3. Statistical diagnostic framework for linear regression models

Regression diagnostic techniques can be organised into interconnected modules that work together to assess the stability of a model: residual-based assumption checks, multicollinearity screening, and detection of outliers and influential observations. No single indicator can show all the problems of a model, so multiple indicators need to be used together for empirical analysis [8, 9].

3.1. Model assumption testing based on residuals

Residual analysis is a type of diagnosis for regression, and charts of the residuals can be used to identify violations of the classical assumptions. Analysts generally use residual-versus-fitted plots, quantile-quantile plots and residual time-series plots to identify problems in the model. Curvature can indicate non-linearity; changing spread may suggest heteroscedasticity; serial patterns indicate autocorrelation; and large deviations from the quantile-quantile reference line are associated with non-normal residuals [2, 8, 9]. However, visual residual interpretation is subjective; therefore, it should be used in conjunction with effect sizes and other diagnostic indicators [9].

3.2. Multicollinearity diagnostic indicators

Severe multicollinearity reduces the stability of regression results by altering coefficient estimates, expanding standard errors, and thus decreasing the reliability of significance tests for individual predictors [5]. A number of supporting instruments are employed by researchers to identify collinearity: pairwise correlation coefficients of predictors, tolerance statistics, the Variance Inflation Factor (VIF), and condition indices derived from the design matrix. VIF > 10 should be considered a relatively lenient screening criterion for the threshold, and it is not necessarily a general statistical cutoff for all datasets [5, 8].

3.3. Outlier and influential point diagnostics

Several supporting statistical indicators are used to examine the abnormal samples. Standardised and studentised residuals flag observations with unusually large residuals compared to the fitted model, and absolute values exceeding 2 or 3 are often used as heuristic screening thresholds [8]. Leverage values measure the extent to which an observation's predictor values deviate from the mean, and a high-leverage point can become influential when it also has a large residual [8, 10]. Cook's distance measures the total change in fitted regression results after a single sample is removed and is often used to identify influential observations [10]. Difference in fits (DFFITS) is another measure that indicates how much each data point changes the fitted value of the model [8]. In addition to the

DOI: 10.54254/2753-8818/2026.37301

traditional parametric linear regression model, some special outlier detection methods have also been proposed for nonparametric and circular regression data to address the problem of directional response variables with wrapped residual distributions [12].

4. Recent research review on statistical diagnostics and regression model evaluation

Recently, research on diagnostics for linear regression has begun to address the traditional problems of high-dimensionality, ill-conditioned matrices, severe outliers and estimation of generalisation error.

Salmerón-Gómez and García-García (2025) proposed a redefined Variance Inflation Factor and a statistical decision rule for identifying troubling multicollinearity [5]. They compare the VIF values with a reference based on orthonormal predictors and define a region of non-rejection according to a significance level, rather than using a single universal cutoff [5].

Lin and others (2026) have derived the distribution of Cook's Distance and the corresponding critical values for the detection of influential observations [10]. The distribution is shown as a function of sample size and model dimension; relevant quantiles are estimated numerically, and these are compared with the heuristic threshold [10].

To address multicollinearity and outliers simultaneously, Wasim et al. (2024) proposed weighted penalised M-estimators for robust ridge regression [7]. Through simulation and an application to gasoline-consumption data, they compared the proposed estimators with ordinary least squares, ridge regression, and existing penalised M-estimators and reported a lower mean squared error under contaminated conditions [7].

Bates et al. (2024) have investigated what cross-validation estimates and the uncertainty of common cross-validation procedures in terms of model generalisation performance [11]. They showed that standard cross-validation aims to find the average prediction error of models trained on hypothetical unseen training sets, rather than the prediction error of a single model fitted to the observed training set, and that conventional confidence intervals can be misleading because fold errors are correlated [11]. Therefore, the estimand and uncertainty reported by cross-validation [11] need to be interpreted carefully.

5. Research limitations and future research directions

5.1. Core limitations of existing regression diagnostic systems

Although many modern diagnostic tools have been optimised numerous times and a large number of improved algorithms have been released, the entire regression evaluation system still has structural defects that affect the authenticity and external validity of the empirical results. Many of these defects are due to the inherent logic of the main diagnostic pipeline and interact with each other in actual data analysis, rather than being isolated technical flaws.

A practical problem is that many of the commonly used diagnostic cutoffs are heuristic and may not be suitable for different sample sizes, model structures, and data distributions [5, 8, 10]. Although analysts often use fixed benchmarks, such as a VIF value exceeding 10 or a Cook's Distance greater than 4/n, these rules should not be considered general statistical thresholds before reaching a final conclusion [5, 10]; rather, they are screening heuristics. Fixed heuristic boundaries are used to make diagnostic judgments in the absence of context-aware interpretation.

In addition, a real-world dataset may contain multiple diagnostic problems simultaneously [7, 8]. When multicollinearity and outliers occur together, ordinary least squares and ridge estimators can

DOI: 10.54254/2753-8818/2026.37301

be adversely affected, motivating robust penalised estimators designed for joint contamination [7]. Therefore, the diagnostic results should be interpreted in combination rather than assuming that each model defect occurs independently [8].

Sample size further reduces diagnostic reliability [9, 11]. Assumption tests may have low power in small samples and flag trivial deviations in very large samples, so graphical evidence and effect size should be provided alongside formal tests [9]. Cross-validation uncertainty also requires caution because fold errors are correlated and conventional confidence intervals may undercover [11].

Although the influence diagnostic method has identified an outlier, it cannot determine whether this observation is due to a data error or an actual rare event in the data [6]. Outliers may reflect measurement errors, idiosyncratic events, unpredictable shocks, or structural differences in the data-generating mechanism [6]. Domain knowledge is thus required to determine whether the abnormal observation should be kept, corrected or excluded.

The design of cross-validation may also reduce the comparability of the results from different regression models, as they are based on different data-splitting schemes [11]. Fold errors are correlated, so uncertainty estimates based on the assumption of naive independence are unreliable [11]. Thus, the split strategy should be consistent with the data structure and prediction objectives.

High-dimensional analytical scenarios are a serious problem for regression diagnosis when the number of predictors is close to or exceeds the sample size [4, 6]. The above settings will make the traditional estimation and variance calculation unstable, so methods have been developed to address many covariates or high-dimensional contamination specifically [4, 6].

5.2. Targeted future research directions

In response to the six core bottlenecks listed above, future research on the statistical diagnosis and reliability assessment of linear regression can be expanded in six specific directions to address the inherent defects of the current system [5, 7, 10, 11]. First, future research should develop interpretable integrated diagnostic pipelines for the simultaneous assessment of collinearity, residual problems, influential points and generalisation error, and provide outputs that identify the likely defect source and appropriate corrective measures. Secondly, in the future, research can be conducted to further explore the application of robust statistical estimation, such as M-estimation and least absolute deviation (LAD) regression, for outlier and influential observation reduction in parameter estimation. Third, in the future, data-dependent calibration methods will be developed to quantify the uncertainty of diagnostic indicators and, if necessary, replace fixed empirical critical values. Fourth, future research should design high-dimensional diagnostic algorithms that are stable in the presence of many covariates and can distinguish between pairwise and multiple collinearity. Fifth, future research will carry out cross-dataset stability verification using multiple open datasets and appropriate stratified sampling to assess the generalisability of diagnostic conclusions and reduce the dependence on results from a single dataset. Sixth, future research should incorporate causal-inference principles when causal interpretation is required, distinguish between model fit in association and causal effects, and avoid equating diagnostic adequacy with evidence of causation. Research in the six directions above can help address the problems of subjective empirical thresholds, interacting model violations and high-dimensional diagnostic failure, thus improving the transparency and statistical credibility of linear regression models in economics, engineering and the social sciences.

DOI: 10.54254/2753-8818/2026.37301

6. Conclusion

This paper presents the statistical diagnostics and reliability tests of linear regression models based on the assumptions of the model, behaviour of residuals, multicollinearity, influential observations and out-of-sample validation. Ordinary least squares is still convenient because its coefficients are easy to interpret and it has good theoretical properties; however, these properties rely on certain assumptions that need to be verified rather than assumed. Residual plots and formal tests can reveal nonlinearity, heteroscedasticity, autocorrelation and other distributional problems; variance inflation factors and condition-based measures can identify unstable coefficient estimation caused by multicollinearity. Leverage, studentised residuals, Cook's distance and other influence measures provide complementary information on unusual observations, but none of these statistics should be used as an automatic rule for deleting data. Recently, some research has also introduced robust estimation and data-dependent diagnostic thresholds to address violations of the classical assumptions. Cross-validation can also provide information on generalisation that is not available in the in-sample measure of R^2. Therefore, the main problem is not to find a single diagnostic index, but rather to gather multiple sources of evidence and interpret them according to the sample size, model objectives, data quality and domain knowledge. The diagnostic decisions should also be recorded so that another researcher can reproduce the analysis and learn why a model was chosen, modified or not chosen. Future work will cover high-dimensional diagnosis, uncertainty quantification, robust estimation and reproducible validation on multiple datasets. A good regression analysis should show both the fit and diagnostic results of the model, distinguish statistical warning signs from problems in the real data, and assess the predictive ability of the model for new data. The above workflow can be used to interpret the regression results more cautiously and reduce the risk of drawing strong conclusions from models that have not met the assumptions or generalisation requirements.

References

[1] Huang, M. H., Xiao, S. T., Ouyang, Y. G., & Li, Z. Q. (2025). Statistical inference of linear regression models for small sample skewed data. *Journal of Beijing University of Chemical Technology (Natural Science), 52*(3), 132–138. https://doi.org/10.13543/j.bhxbzr.2025.03.014

[2] Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods, 53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

[3] Dar, I. S., & Chand, S. (2024). New heteroscedasticity-adjusted ridge estimators in linear regression model. *Communications in Statistics - Theory and Methods, 53*(19), 7087–7101. https://doi.org/10.1080/03610926.2023.2258427

[4] Anatolyev, S., & Ng, C. F. (2026). Many covariate and cluster robust estimation and inference. *Econometric Reviews, 45*(5), 703–732. https://doi.org/10.1080/07474938.2025.2608657

[5] Salmerón-Gómez, R., & García-García, C. B. (2025). rvif: A decision rule to detect troubling statistical multicollinearity based on redefined VIF. *The R Journal, 17*(4), 193–216. https://doi.org/10.32614/RJ-2025-040

[6] Farnè, M., & Vouldis, A. (2024). ROBOUT: A conditional outlier detection methodology for high-dimensional data. *Statistical Papers, 65*, 2489–2525. https://doi.org/10.1007/s00362-023-01492-3

[7] Wasim, D., Suhail, M., Albalawi, O., & Shabbir, M. (2024). Weighted penalized M-estimators in robust ridge regression: An application to gasoline consumption data. *Journal of Statistical Computation and Simulation, 94*(15), 3427–3456. https://doi.org/10.1080/00949655.2024.2386391

[8] Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons. https://doi.org/10.1002/0471725750

[9] Shatz, I. (2024). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. *Behavior Research Methods, 56*(2), 826–845. https://doi.org/10.3758/s13428-023-02072-x

[10] Lin, Z., Cousineau, D., & Chartier, S. (2026). The distribution of the Cook's distance with critical values. *Communications in Statistics - Theory and Methods, 55*(18), 6289–6309.

DOI: 10.54254/2753-8818/2026.37301

https://doi.org/10.1080/03610926.2026.2666188
[11] Bates, S., Hastie, T., & Tibshirani, R. (2024). Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association, 119*(546), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686
[12] Sert, S., & Kardiyen, F. (2023). A method for detecting outliers in linear-circular nonparametric regression. *PLOS ONE, 18*(6), e0286448. https://doi.org/10.1371/journal.pone.0286448