ISSN 2581-3463

AJMS
Asian Journal of Mathematical Sciences

RESEARCH ARTICLE

Regression Diagnostic Checking for Survey Data on Mothers' Weights and Ages as Factors Responsible for Babies' Weight at Birth

Habiba Danjuma

Department of Statistics, Federal Polytechnic Bali, PMB 05 Bali, Taraba State, Nigeria

Received: 22-07-2022; Revised: 25-08-2022; Accepted: 25-10-2022

ABSTRACT

When a regression model is considered for an application, we can usually not be certain in advance that the model is appropriate for that application, any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, the diagnostic checking techniques on the regression model are essential. This study therefore is on model diagnostic checking techniques with application to linear regression analysis. In this study, a method useful for diagnosing violation of basic regression assumptions are presented and tested using a secondary data on babies' weight at birth which serves as dependent variable and mothers' weight and ages as independent variables. All the assumptions tested from the objectives (Normality of residual, collinearity between the independent variable, outlier/leverage, and linearity of the model) are met and no one deviated from the assumptions of multiple linear regression fitted on the data.

Key words: Diagnostic checking, Regression assumption, Collinearity, Outlier, Normality, linearity

INTRODUCTION

Regression analysis is a statistical technique for investigating and modeling the relationship between two or more variables. More specifically, it is an attempt to explain movement in one variable by reference to movements in one or more other variables. To make this definition more concrete, we can denote the variable whose movements the regression seeks to explain by y and the variables which are used to explain those variation by x_1, x_2,… x_3. Hence, in this relative set up, it would be said that the variations in k variables (the x's) cause changes in some variable, y's.[1]

Applications of regression analysis exist in almost every field of human endeavor. In econometrics, the dependent variable might be a family' s consumption expenditure and the independent variables might be the family's income, number of children in the family, and other factors that would affect the family's consumption patterns. In political science, the dependent variable might be a state's level of welfare spending and the independent variables measures of public opinion and institutional variables that would cause the state to have higher or lower levels of welfare spending. In sociology, the dependent variable might be a measure of the social status of various occupations and the independent variables characteristics of the occupations (pay, qualifications, etc.). In psychology, the dependent variable might be individual's racial tolerance as measured on a standard scale and with indicators of social background as independent variables. In education, the dependent variable might be a student's score on an achievement test and the independent variables characteristics of the student's family, quality of teachers, or school. One major way to judge whether a variable adds to the explanatory power of a model is by looking at the impact its inclusion has on the value of R-Squared. If the value of R-Squared increases significantly when a variable is added to the model, then the extra information provided by this variable increases the model's ability to explain the variation in the response variable Imam.[2]

Address for correspondence:
Habiba Danjuma
E-mail: yumarhabiba@gmail.com

Diagnostic techniques were gradually developed to find problems in model-fitting and to assess the quality and reliability of regression estimates. These concerns turned into an important area in regression theory intended to explore the characteristics of a fitted regression model for a given data set. Discussion of diagnostics for linear regression models is often indispensable chapters or sections in most of the statistical textbooks on linear models. One of the most influential books on the topic was regression diagnostics: Identifying influential data and sources of collinearity by Belsley *et al.*[3]

Survey literature has not given much attention to diagnostics for linear regression models. Deville and Särndal,[4] and Potter[5] discuss some possibilities for locating or trimming extreme survey weights when the goal is to estimate population totals and other simple descriptive statistics. Hulliger[6] and Moreno-Rebollo *et al.*[7] addresses the effect of outliers on the Horvitz-Thompson estimator of a population total. Smith (1987) demonstrates diagnostics based on case deletion and a form of the influence function. Chambers and Skinner,[8] Gwet and Rivest,[9] Welsh and Ronchetti,[10] and Duchesne (1999) conduct research on outlier robust estimation techniques for totals. Elliott[11] and Korn and Graubard (1995)[12] are two of the few references which introduce techniques for the evaluation of the quality of regression on complex survey data.

When a regression model is considered for an application, we can usually not be certain in advance that the model is appropriate for that application, any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand.[13] Hence, it is important to examine the suitability of the model for the data before inferences based on that model are undertaken. This achieved using model diagnostic checking techniques on the regression model. In this section, we discuss some simple graphic methods for studying the appropriateness of a model, as well as some remedial measures that can be helpful when the data are not in accordance with the conditions of the regression model. Therefore, the aim of this study is to examine departures from the assumption of linear regression model with normal errors through model diagnostic checking techniques. We shall consider following six important types of departures from linear regression model with normal errors: non-linearity, non-constant of variances of error term, multicollinearity, outlier in observations, and non-normality of error term.

METHODOLOGY

This section involves the full procedure for the diagnosis testing on deviations of regression model from some its assumptions that are considered in this paper. The data used are secondary which comprises a dependent/response and two independent/predictor variables. In regression analysis, tests based on the square of the residuals may be used to detect non-linearity. This study considers the following five important types of departures from linear regression model. These are as follows: the regression function is not linear, the error terms do not have constant variance, the error terms are not independent, the model fits all but one or a few outlier observations, and the error terms are not normally distributed. The procedures to carry out these departures are stated as follows:

Linear Regression Models

We consider a basic linear model where there is only one predictor variable and the regression function is linear. Model with more than one predictor variable is straight forward. The model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

Where Y_i is the value of the response variable in the ith trial β_0 and β_1 are parameters, X_i is a known constant, namely, the value of the predictor variable in the ith trial, ε_i is a random error term with mean zero and variance σ_2 and ε_i and ε_j are uncorrelated so that their covariance is zero.

Regression model (1) is said to be simple, linear in the parameters, and linear in the predictor variable. It is "simple" in that there is only one predictor variable, "linear in the parameters," because no parameters appears as an exponent or its multiplied or divided by another parameter, and "linear in predictor variable" because this variable appears only in the first power. A model that is linear in the parameters and in the predictor variable is also called first order model.

Diagnostics and Remedial Measures

When a regression model is considered for an application, we can usually not be certain in advance that the model is appropriate for that application, any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model for the data before inferences based on that model are undertaken. In this section, we discuss some simple graphic methods for studying the appropriateness of a model, as well as some remedial measures that can be helpful when the data are not in accordance with the conditions of the regression model.

Non-linearity of Regression Model

Whether a linear regression function is appropriate for the data being analyzed can be studied from a residual plot against the predictor variable or equivalently from a residual plot against the fitted values.

Figure 1a shows a prototype situation of the residual plot against X when a linear regression model is appropriate. The residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative. Figure 1b shows a prototype situation of a departure from the linear regression model that indicates the need for a curvilinear regression function. Here, the residuals tend to vary in a systematic fashion between being positive and negative. However, Figure 2a displays a prototype situation of a departure from the linear regression model that indicates the need a nonlinear relationship while Figure 2b shows a linear relationship.

Non-constancy of Error Variance

Plots of residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant. The prototype plot in Figure 1a exemplifies residual plots when error term variance is constant. Figure 2a shows a prototype picture of residual plot when the error variance increases with X.

Presence of Outliers

Outliers are extreme observations. Residual outliers can be identified from residual plots

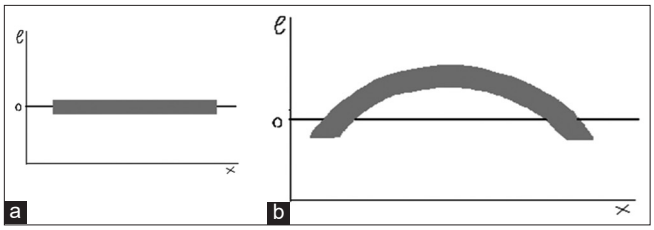

Figure 1: (a and b) Prototype situation of the residual plot against x when a linear regression model is appropriate

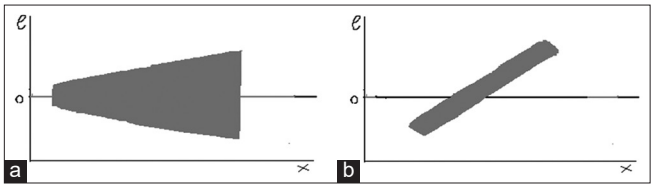

Figure 2: (a and b) A prototype situation of a departure from the linear regression model

against X or $\hat{Y}$. Outliers can create great difficulty. When we encounter one, our first suspicion is that the observation resulted from a mistake or other extraneous effect. On the other hand, outliers may convey significant information, as when an outlier occurs due to an interaction with another predictor omitted from the model. A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents in error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstances.

Non-independence of Error Terms

Whenever data are obtained in a time sequence or some other type of sequence, such as for adjacent geographical areas, it is good idea to prepare a sequence plot of the residuals. The purpose of plotting the residuals against time or some other type of sequence is to see if there is any correlation between error terms that are near each other in the sequence. A prototype residual plot showing a time related trend effect is presented in Figure 2b, which portrays a linear time related trend effect. When the error terms are independent, we expect the residuals in a sequence plot to fluctuate in a more or less random pattern around the base line 0.

Non-normality of Error Terms

Small departures from normality do not create any serious problems. Major departures, on the other hand, should be of concern. The normality of the error terms can be studied informally by examining the residuals in a variety of graphic ways.

Comparison of frequencies when the number of cases is reasonably large is to compare actual frequencies of the residuals against expected frequencies under normality. For example, one can determine whether, say, about 90% of the residuals fall between ±1.645 $\sqrt{MSE}$. Normal probability plot: Still another possibility is to prepare a normal probability plot of the residuals. Here, each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal.

Omission of Important Predictor Variables

Residuals should also be plotted against variables omitted from the model that might have important effects on the response. The purpose of this additional analysis is to determine whether there are any key variables that could provide important additional descriptive and predictive power to the model. The residuals are plotted against the additional predictor variable to see whether or not the residuals tend to vary systematically with the level of the additional predictor variable.

Overview of Tests

Graphical analysis of residuals is inherently subjective. Nevertheless, subjective analysis of a variety of interrelated residuals plots will frequently reveal difficulties with the model more clearly than particular formal tests.

TESTS FOR RANDOMNESS

A run test is frequently used to test for lack of randomness in the residuals arranged in time order. Another test, specially designed for lack of randomness in least squares residuals is the Durbin-Watson test.

Durbin–Watson test

The Durbin–Watson test assumes the first order autoregressive error models. The test consists of determining whether or not the autocorrelation coefficient (ρ, say) is zero. The usual test alternatives considered are:
H_0: $\rho = 0$
H_0: $\rho > 0$
The Durbin–Watson test statistic D is obtained using ordinary least squares to fit the regression function, calculating the ordinary residuals: $e_t = Y_t - \hat{Y}_t$, and then calculating the statistic:

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{n\sum_{t=1} e_t^2}$$

Exact critical values are difficult to obtain, but Durbin–Watson has obtained lower and upper bound d_L and d_U such that a value of D outside these bounds leads to a definite decision. The decision rule for testing between the alternatives is:
if $D > d_U$, conclude H_0
if $D < d_L$, conclude H_1
if $d_L \le D \le d_U$, test is inconclusive.

Small value of D leads to the conclusion that $\rho > 0$.

Correlation Test for Normality

In addition to visually assessing the appropriate linearity of the points plotted in a normal probability plot, a formal test for normality of the error terms can be conducted by calculating the coefficient of correlation between residuals ei and their expected values under normality. A high value of the correlation coefficient is indicative of normality.

TESTS FOR CONSTANCY OF ERROR VARIANCE

Modified Levene Test

The test is based on the variability of the residuals. Let e_{i1} denotes the i^{th} residual for group 1 and e_{i2} denotes the i^{th} residual for group 2. Furthermore, we denote n_1 and n_2 to denote the sample sizes of the two groups, where: $n_1 + n_2 = n$.
Further, we shall use $\tilde{e}_1$ and $\tilde{e}_2$ to denote the medians of the residuals in the two groups. The modified Levene test uses the absolute deviations of the residuals around their median, to be denoted by d_{i1} and d_{i2}:

$$\bar{d}_{i1} = \left|e_{i1} - \tilde{e}_1\right|, \quad \bar{d}_{i2} = \left|e_{i2} - \tilde{e}_2\right|$$

With this notation, the two-sample *t*-test statistic becomes:

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2}, respectively, and the pooled variance s^2 is:

$$s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i2} - \bar{d}_2)^2}{n-2}$$

If the error terms have constant variance and n_1 and n_2 are not too small, t_L^* follows approximately the t distribution with n-2° of freedom. Large absolute values of t_L^* indicate that the error terms do not have constant variance.

TESTS FOR OUTLYING OBSERVATIONS

(i) Elements of Hat Matrix: The Hat matrix is defined as $H=X(X'X)^{-1}X'$, X is the matrix for explanatory variables. The larger values reflect data points are outliers.

(ii) $WSSD_i$: $WSSD_i$ is an important statistic to locate points that are remote in x-space. $WSSD_i$ measures the weighted sum of squared distance of the i[th] point from the center of the data. In general, if the $WSSD_i$ values progress smoothly from small to large, there are probably no extremely remote points. However, if there is a sudden jump in the magnitude of $WSSD_i$, this often indicates that one or more extreme points are present.

(iii) Cook's D_i: Cook's D_i is designed to measure the shift in $\hat{y}$ when i[th] observation is not used in the estimation of parameters. D_i follows approximately $F_{(p,n-p-1)}(1-\alpha)$. Lower 10% point of this distribution is taken as a reasonable cutoff (more conservative users suggest the 50% point). The cutoff for D_i can be taken as $\frac{4}{n}$.

(iv) $DFFITS_i$: DFFIT is used to measure difference in i[th] component of $\left(\hat{y} - \hat{y}_{(i)}\right)$. It is suggested that $DFFITS_i \geq 2\left(\frac{p+1}{n}\right)^{1/2}$ may be used to flag off influential observations.

(v) $DFBETAS_{j(i)}$: Cook's D_i reveals the impact of i[th] observation on the entire vector of the estimated regression coefficients. The influential observations for individual regression coefficient are identified by $DFBETAS_{j(i)}, j = 1,2,...,p+1$, where each $DFBETAS_{j(i)}$ is the standardized change in b_j when the i[th] observation is deleted.

(vi) $COVRATIO_i$: The impact of the i[th] observation on variance-covariance matrix of the estimated regression coefficients is measured by the ratio of the determinants of the two variance-covariance matrices. Thus, COVRATIO reflects the impact of the i[th] observation on the precision of the estimates of the regression coefficients. Values near 1 indicate that the i[th] observation has little effect on the precision of the estimates. A value of COVRATIO >1 indicates that the deletion of the i[th] observation decreases the precision of the estimates; a ratio <1 indicates that the deletion of the observation increases the precision of the estimates. Influential points are indicated by,

$$\left|COVRATIO_i - 1\right| > \frac{3(p+1)}{n}$$

TEST OF LINEARITY IN THE DATA

Testing whether a linear regression function is appropriate for the data being analyzed can be studied from a residual plot against the predictor variable or equivalently from a residual plot against the fitted values.

This is displayed as follows:

Figure 3 shows a residual plot against the predictor variable. Ideally, the residual plot will show no fitted pattern. That is, the red line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspect of the linear model. In this Figure 3 above, there is no pattern in the residual plot and the red line is approximately horizontal. This suggests that we can assume linear relationship between the predictors and the outcome variables.

Test of Normality of Residuals

Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of *squared*

error, a few extreme observations can exert a disproportionate influence on parameter estimates. The best test for normally distributed errors is a normal probability plot or normal quantile plot of the residuals. These are plots of the fractiles of error distribution versus the fractiles of a normal distribution having the same mean and variance. If the distribution is normal, the points on such a plot should fall close to the diagonal reference line. A bow-shaped pattern of deviations from the diagonal indicates that the residuals have excessive skewness (i.e., they are not symmetrically distributed, with too many large errors in one direction). An S-shaped pattern of deviations indicates that the residuals have excessive kurtosis, that is, there are either too many or two few large errors in both directions. Sometimes the problem is revealed to be that there are a few data points on one or both ends that deviate significantly from the reference line ("outliers"), in which case they should get close attention. The QQ plot of residuals is used here to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In our example, all the points fall approximately along this reference line, so we can assume normality.

Figure 4 displays the normal probability plot of residuals. The normal probability plot of residuals should approximately follow a straight line. In our Figure 4 above, all the points fall approximately along this reference line, so we can assume normality.

Test of Normality Shapiro–Wilk Normality Test

Shapiro–Wilk normality test is used in this study. Calculation of confidence intervals and various significance tests for coefficients is all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow. The normality analysis in the Table 1 shows that the residual values is normal because the p-value is >5% level of significance and therefore accept H_o.

Detection of Outliers and High Leverage Points

The presence of outliers may affect the interpretation of the model, because it increases the RSE. Outliers can be identified by examining the standardized residual (or studentized residual),

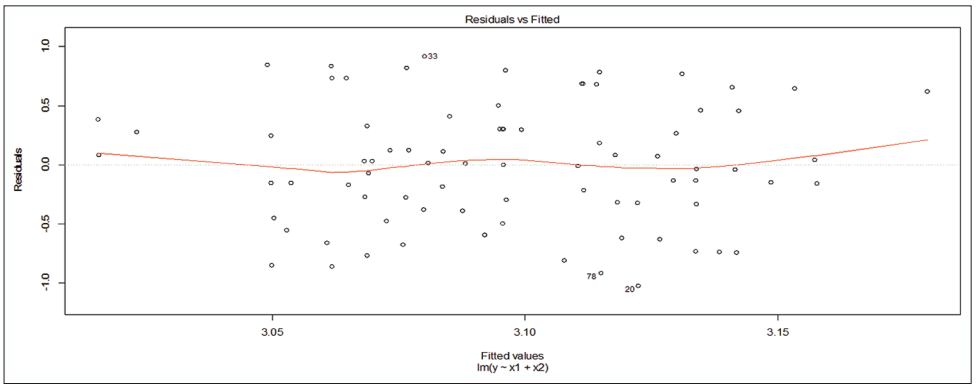

Figure 3: Checking Linearity in the Data

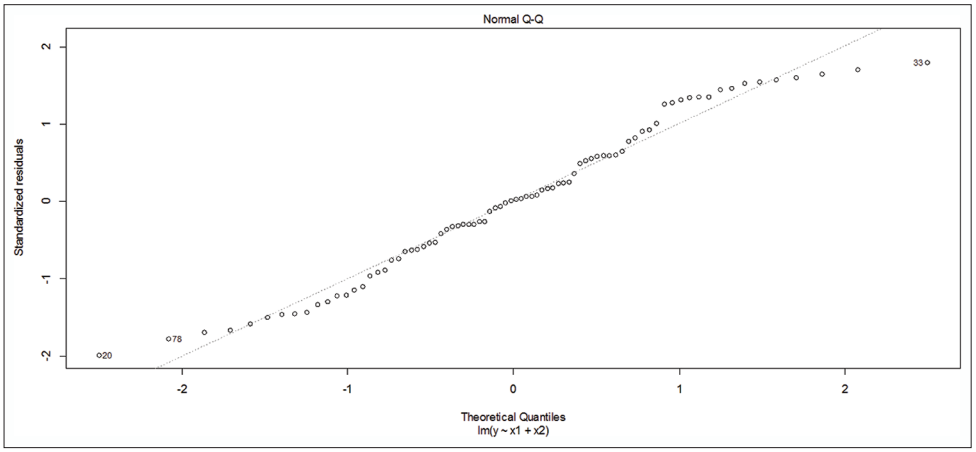

Figure 4: Normality of Residual Plot

which is the residual divided by its estimated standard error. Standardized residuals can be interpreted as the number of standard errors away from the regression line. Observations whose standardized residuals are >3 in absolute value are possible outliers. A data point has high leverage, if it has extreme predictor x values. This can be detected by examining the leverage statistic or the hat-value. A value of this statistic above 2(p + 1)/n indicates an observation with high leverage;[1] where, *P* is the number of predictors and n is the number of observations.

The plot above in Figure 5 on outlier checking highlights the top 3 most extreme points (#40, #26 and #16), with a standardized residuals higher 3. However, there are outliers that exceed 3 standard deviations.

In addition, in Figure 6 on leverage checking, there is high leverage point in the data. That is, some data points have a leverage statistic higher than 2 (p + 1)/n = 6/80 = 0.08.

Table 1: Normality test

Test statistic (W)	*P*-value (p)	Null hypothesis (H_0)	Decision
0.97052	0.06126	There is normality of residual	Accept H_0 (Residual is normal)

Testing the Homoscedasticity Assumption

This assumption was checked by examining the scale-location plot, also known as the spread-location plot.

Figure 7 shows if residuals are spread equally along the ranges of predictors. It is good if you see a horizontal line with equally spread points. In our Figure 7, this is the case. We can assume Homogeneity of variance.

Non-Constant Error Variance (NCV) Test

A further test was carried out using non-constant variance (NCV) test to determine whether the assumption of homoscedacity holds in the data. Table 2 below shows the result/output of the test.

The analysis in Table 2 is testing of homoscedasticity assumption. It shows that, there homoscedacity because *P*-value is >5% level of significance and therefore accept H_o.

Testing the Multicollinearity Assumption

Variance inflation factor for multicollinearity was tested using Farrar-Glauber test for multicollinearity. This is to ascertain the weather there is a collinearity between the two independent

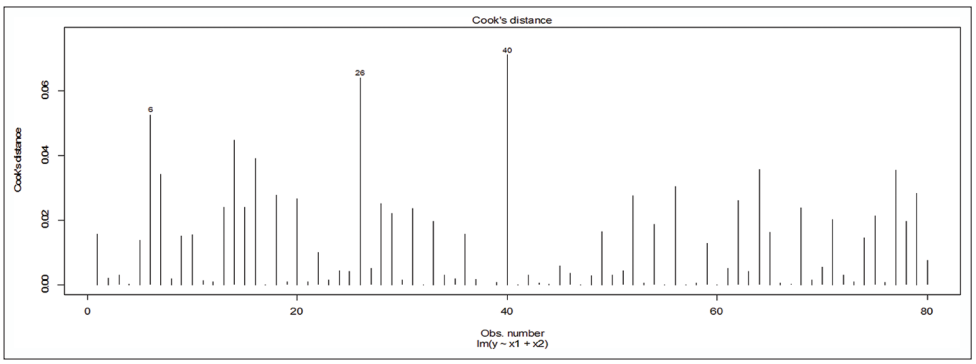

Figure 5: Outlier Checking

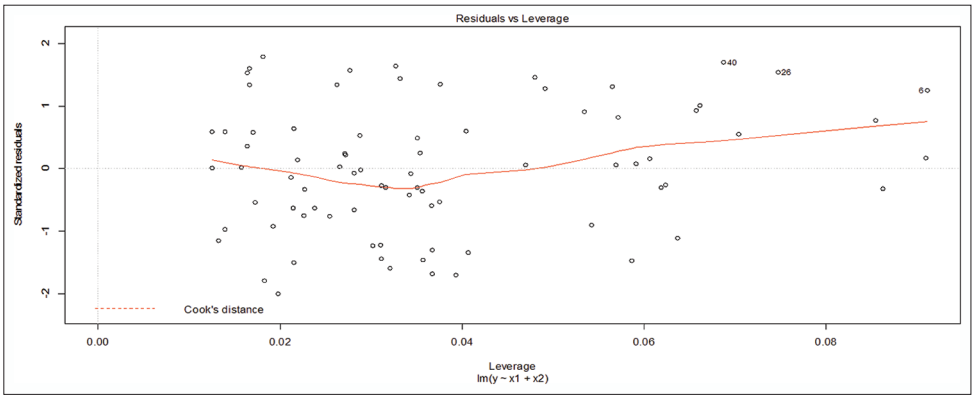

Figure 6: Leverage Checking

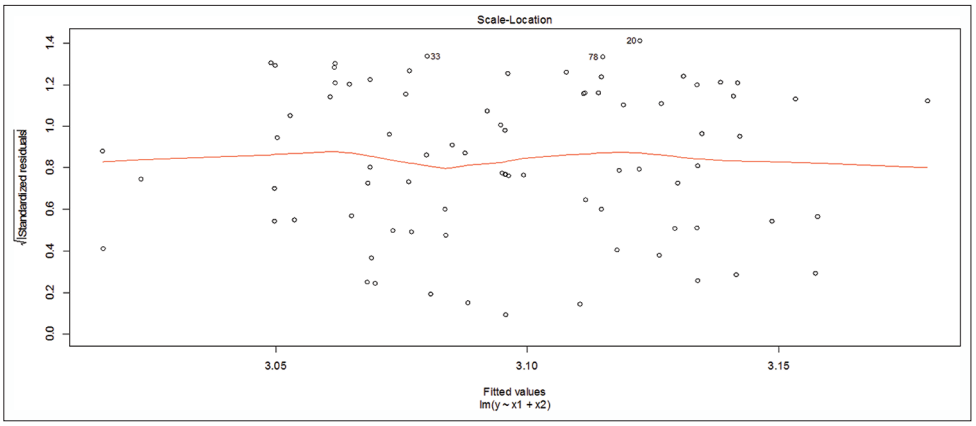

Figure 7: Homoscedasticity Assumption Checking

Table 2: Homoscedasticity assumption test using NCV statistic

Test statistic (X^2)	P-value (p)	Null Hypothesis (H_0)	Decision
0.0196848	0.8884	There is homoscedacity	Accept H_0 (homoscedacity exist)

Table 3: Collinearity test

VIF	1.008809	<3	No collinearity is detected by the test
Farrar Test	2.05	<3	No collinearity is detected by the test
Durbin WatsonTest	0.0473677	1.882538	0.634

Table 4: Durbin–Watson Test

Correlation value	Test statistic (D-W)	P-value (p)	Null hypothesis (H_0)	Decision
0.0473677	1.882538	0.634	There is no correlation	Accept H_0 (no correlation)

variables (Mothers' weights and mothers' ages) or not. The results are given as follows.

Tables 3 and 4 above show the collinearity and Durbin–Watson Test. The result in Table 3 indicates in that there is no collinearity between the two independent variables since the values of both tests are <3. Furthermore, in Table 4, there is no correlation between the two independent variables due to its *P*-value higher than 5% level of significance.

CONCLUSION

The findings from the above preliminary investigation and diagnostics of assumptions/ deviation from the assumption give indications that data on weight of child at birth are dependent on his/ her mother weight or age during the birth. All the assumptions tested from the objectives (normality of residual, collinearity between the independent variable, outlier/leverage, and linearity of the model) are met and no one deviated from the assumptions of multiple linear regression fitted on the data.

REFERENCES

1. Oyeyemi GM, Bukoye A, Akeyede I. Comparison of outlier detection procedures in multiple linear regressions. Am J Math Stat 2015;5:37-41.
2. Imam A. Investigation of parameter behaviors in stationarity of autoregressive and moving average models through simulations. Asian J Math Sci 2020;4:30-7.
3. Belsley DA, Kuh E, Welsch R. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley; 1980.
4. Deville JC, Särndal CE. Calibration estimators in survey sampling. J Am Stat Assoc 1992;87:376-82.
5. Potter FJ. A study of procedures to identify and trim extreme sampling weights. In: Proceedings Section on Survey Research Methods Association. New Jersey: American Statistical Association; 1990. p. 225-30.
6. Hulliger B. Outlier robust Horvitz-Thompson estimators. Surv Methodol 1995;21:79-87.
7. Moreno-Rebollo JL, Muñoz-Reyes A, Muñoz-Pichardo J. Influence diagnostic in survey sampling: Conditional bias. Biometrika 1999;86:923-8.
8. Chambers RL, Skinner CJ. Analysis of Survey Data. New York: John Wiley; 2003.
9. Gwet J, Rivest L. Outlier resistant alternatives to the ratio estimator. J Am Stat Assoc 1992;87:1174-82.
10. Welsh AH, Ronchetti E. Bias-calibrated estimation from sample surveys containing outliers. J R Statist Soc Ser B Methodol 1998;60:413-28.
11. Elliott MR. Bayesian weight trimming for generalized linear regression models. Surv Methodol 2017;43:23-34.
12. Korn EL, Graubard BI. Examples of differing weighted and un-weighted estimates from a sample survey. Am Stat 1995;49:291-5.
13. Warha AA, Yusuf AM, Akeyede I. A comparative analysis on some estimators of parameters of linear regression models in presence of multicollinearity. Asian J Probab Stat 2018;2:1-8.