Robust Regression with an Application to Nigeria's Micro-Economic Data

David Peter Ilesanmi, Oluwarotimi Williams Samuel*, Mojisola Grace Asogbon
*Corresponding author (O.W. Samuel; timitex92@gmail.com)

Abstracts
In this research work, variable selection method is used to select the useful variables from Nigeria's macro- economic data. Robust regression method is then used to analyse them. Robust regression analysis provides an alternative to a least squares regression model when fundamental assumptions are unfulfilled by the nature of the data. Sometimes, the analyst can transform his variables to agree to those assumptions. However, a transformation will not eliminate the leverage of influential outliers that bias the prediction and distort the significance of parameter estimates. Under these circumstances, robust regression that is resistant to the influence of outliers may be a potential recourse. The weaknesses of the Least Squares estimator are highlighted, and the idea of error in data refined. Properties such as breakdown point, efficiency and leverage are discussed with respect to M and MM-estimators in relation to these properties. The study reveals that Fixed Asset, Foreign Exchange, Consumption, Net Savings, and External Reserve contributed to the improvement of the Nigerian economy. However, one of the recommendations of the study is that, the method of robust regression should be used in situations where the presence of outliers is suspected.
Keywords: Gross Domestic Products (GDP), Outliers, Multicollinearity, Biweight, Correlation matrix, Cook's Distance, Residual, Influence.

IJSER

1 INTRODUCTION

Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable. There is a vast arsenal of regression methods that can be used to determine this.Nearly all regression analysis relies on the method of least squares for estimation of parameters in the model. Out of many possible regression techniques, the OrdinaryLeast Squares (OLS) is the most commonly used method. The assumption is that

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i \quad \ldots\ldots \quad (1)$$

Where $\beta_0, \ldots, \beta_k$unknown parametersi=1,….n.Thereϵ_i are independent random variables, which are$E(\epsilon_i) = 0, VAR(\epsilon_i) = \sigma^2, and\ \epsilon_i$ independent of x_{ik} each k, where k=1,….n. This model implies that the conditional mean of y_i given $(x_{i1}, \ldots, x_{ik})$, is $\beta_0 + \Sigma\beta_k x_{ik}$, a linear combination of the predictors.

Equation 1 is a *homoscedastic* model, meaning that theyϵ_i have a common variance. If the error termϵ_i has a variance ${\sigma_i}^2$ and${\sigma_i}^2 \neq {\sigma_j}^2$for some,$i \neq j$, the model is said to be *heteroscedastic*.

- *David Peter Ilesanmi, MSc Statistics, University of Lagos, Nigeria, PH-08064864045. E-mail: davidilesanmi@yahoo.com*
- *Oluwarotimi Williams Samuel, PhD student,Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, PH-8615814491870. E-mail: timitex92@gmail.com*
- *Mojisola Grace Asogbon, Prospective PhD candidate,Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, E-mail: mojiasho@yahoo.com*

Even when ϵ has a normal distribution, heteroscedasticity can result in relatively low efficiency when using the Ordinary Least Square estimator which means that the estimator of β can have a relatively large standard error. The probability coverage can be poor when computing confidence intervals. Thus, Ordinary Least Squares is said to be not robust to violations of its assumptions.

Another problem that we often encounter in the application of regression is the presence of an outlier or outliers in the data. As defined by Barnett and Lewis (1994), outliers are observations that appear inconsistent with the rest of the data. Outliers can be generated from a simple operational mistake or naturally. They have serious effects on statistical inference. Even one outlying observation can destroy least square estimation, resulting in parameter estimates that do not provide useful information to the majority of the data. Robust regression analyses have been developed as an improvement of least square estimation in the presence of outliers and to provide us information about what valid observation is and whether this should be thrown out. The primary purpose of robust regression analysis is to fit a model which represents the information in the majority of the data.

The property of efficiency, breakdown point, and bounded influence are used to define the measure of robust technique performance in a theoretical sense. Efficiency can tell us how high efficiency is mostly desired on estimation. The breakdown point is a measure for stability of the estimator when the sample contains a large fraction of outliers (Hampel, 1975). It gives the minimum fraction of outliers which may produce an infinite bias. It is referred to as the measure of global robustness in this sense. For example, least square has a breakdown point of $\frac{1}{n}$. This indicates that only one outlier can make the estimate useless. In

contrast, some robust regression estimates attach approximately 50% breakdown point and it is called a high breakdown point in this case. Lastly, bounded influence is designed to encounter the tendency of least squares to allow exterior X-space or high leverage points to exhibit greater influence which can be important especially if these points are outliers. Robust regression estimators were first introduced by Huber (1973, 1981), and it is well-known as M regression estimators. Rousseeuw (1984) introduced the least median of squares (LMS) and the least trimmed squares (LTS) estimators. These estimators minimise the median and the trimmed mean of the squared residuals respectively. They are very high breakdown point estimators. However, the high breakdown point estimation has some drawbacks.

In a study that involves several *X* variables, selection method becomes so important so that the best predictors can be achieved. Therefore, variable selection which is also known as feature selection is the process of selecting a subset of relevant features used in model construction. The central assumption when using a variable selection technique is that the data contains many redundant or irrelevant variables. Redundant variables are those which provide no more information than the currently selected variables and it provides no useful information in any context. Variable selection techniques are a subset of more general field of variable extraction.
Selecting the appropriate variables for an analytical model is an important issue in this research. This may involve finding the right combination of confounders to adjust for when estimating the association between the respond variable and the explanatory variables.

In this era, variable selection procedures are an integral part of virtually all widely used statistics package, and their uses will only increase as the information revolution brings us larger data sets with more and more variables. The demand for variable selection will be strong and it will continue to be a basic strategy for data analysis.

1.2 STATEMENT OF THE PROBLEM

Let *Y* be a respond variable of interest and*let*$X_1, \ldots, X_p$ *be* a set of potential explanatory variables or predictor variables, whichare vectors of *n* observations.
Consider the multiple regression model:
$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$ *(2)*

Selecting these several variables will be so tedious. The problem of variable selection arises when one wants to model the relationship between *Y* and a subset of$X_1, \ldots, X_p$, but there is uncertainty about which subset to use in such a situation.It is particularly of interest when p is large and $X_1, \ldots, X_p$is considered to contain many redundant or noise variables.
Suppose we have p independent variables $X_1, X_2, \ldots, X_p$. Then there are 2^p subsets of variables. The purpose of all-possible – regressions approach is to identify a small group of regression models that are good according to a specific criterion so that a detailed examination can be made of these models.It would be impossible for an analyst to make a detailed examination of all possible regression models. For example, in this research, the number of *X* variables is thirteen; there would be $2^{13} = 8192$ possible regression models.

1.3 OBJECTIVES

The objectives of this study are:

- To select the best independent variables;
- To develop the best model for predicting the economic growth (GDP) of Nigeria when other independent variables are known: to fit a model which represents the information in the majority of the data;
- To increase the model predictive ability;
- To check the relationship among the variables selected and
- To identify outliers and influential cases.

1.4SIGNIFICANCE OF THE STUDY

As a result of the fact that many researchers who had studied the Nigerian macro - economic data had failed to take cognizance of the fact that the macro – economic data of Nigeria have outliers imbedded in them, it has led to a misleading interpretation of Nigeria's economic growth. We observed that an investigation on the influence of outliers on our data set will fill the gap and contribute immensely to scholarship. So, the purpose of this study is to use robust regression to analyse the macro – economic data of Nigeria, since our data contains outliers which have significant influence on our data set and the biweights regression is robust to these outliers.

2 RESEARCH METHODOLOGY

The key stepwise procedures of selection are discussed as follows.

2.1Backward Elimination

A variable selection procedure in which all variables are entered into the equation and then sequentially removed. This is the simplest of all variable selection procedures and can be easily implemented without special software. In situations where there is a complex hierarchy, backward elimination can be run manually while taking account of what variables are eligible for removal. It follows the procedures below:

- Start with all the predictors in the model
- Remove the predictor with highest p-value greater thanα_{crit}
- Refit the model and go to 2
- Stop when all p-values are less than α_{crit}.

2.1.1 Forward Selection

A stepwise variable selection procedure in which variables are sequentially entered into the model.
This just reverses the backward method. The procedures;

- Start with no variables in the model.
- For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than α_{crit}.
- Continue until no new predictors can be added.

2.2Model Specification and Interpretation of the Coefficients of the Model

Dependent variable: our dependent variable for this model is Gross Domestic Products (GDP).

Independent variables are:

X_1(Public Expenditure), X_2 (Fixed Asset), X_3 (Private Investment), X_4 (Foreign Exchange), X_5 (Inflation Rate), X_6 (Public Debt), X_7 (Interest Rate), X_8 (Net Export), X_9

(Consumption), X10 (Net Savings), X_{11} (External reserve), X_{12} (Net Import) and X_{13} (Balance of Payment).
Therefore, the model for our data set is:

$$Gdp = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8 + \beta_9X_9 + \beta_{10}X_{10} + \beta_{11}X_{11} + \beta_{12}X_{12} + \beta_{13}X_{13} + \in_i \quad (3)$$

where $\in_i$ is the random error?

$= x_i'\beta + \in_i$

Where $\beta_0$is the intercept β_1 is (it in) the coefficient of Public Expenditure, *β_2 is (it in) the coefficient of* Fixed Asset, *β_3 is (it in) the coefficient of* Private Investment,*$\beta_4$is (it in) the coefficient of Foreign Exchange, β_5 is (it in) the coefficient of Inflation Rate,β_6 is (it in) the coefficient of Public Debt, β_7 is the coefficient of Interest Rate, β_8 is (it in) the coefficient of Net Export, $\beta_9$is (it in) the coefficient of Consumption,β_{10} is (it in) the coefficient of Net Savings, β_{11} is (it in) the coefficient of External Reserve, $\beta_{12}$is (it in) the coefficient of Net Import and $\beta_{13}$is (it in) the coefficient of Balance Payment?*

2.3 Assessing the Regression as a Whole

We Want to assess the performance of the model as a whole

H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

The model has no worth

H_1: At least one regression coefficient is not equal to zero

The model has worth

If all the β's are close to zero, then the SSR will approach zero

$$\frac{SSR}{SSE} \Rightarrow 0$$

2.4 Multiple Regression

Multiple regression refers to regression with two or more X (independent) variables. Multiple regression analysis is widely used for predicting and forecasting. It is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.
The general problem of fitting the model:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots. + \beta_kx_k + \in \quad (4)$$

is called multiple linear regression problem.
The unknown parameters [β_i] are called regression coefficients. The model describes a hyperplane to the k-dimensional space of the independent variables[x_i].

2.5 Robust Regression

It is an alternative to Least Square method when errors are non-normal. All estimation methods rely on assumptions for their validity. We say that an estimator or statistical procedure is robust if it provides useful information even if some of the assumptions used to justify the estimation method are not applicable. It uses iterative methods to assign different weights to residuals until the estimation process converges. More complex to evaluate the precision of the regression coefficients, compared to ordinary model.

2.5.1 Uses of Robust Regression

- Useful to detect outliers by finding cases whose final weights are relatively small.
- Can be used to confirm the appropriateness of the ordinary least square model.
- Primarily helpful in finding cases that are outlying with respect to their y values (long-tailed errors). They can't overcome problems due to variance structure.

2.5.2 Breakdown and Robustness

The finite sample breakdown of an estimator/procedure is the smallest fraction of data points such that if [nα] points $\rightarrow\infty$ then the estimator/procedure also becomes infinite.
The sample mean of$x,\ldots,x_n$is

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{1}{n}[\sum_{i=1}^{n-1} x_i + x_n]$$

$$= \frac{n-1}{n}\bar{x}_{n-1} + \frac{1}{n}x_n \quad (3)$$

And so if x_n is large enough then $\bar{x}_n$can be made as large as desired regardless of the other $n-1$ values.
Unlike the mean, the sample median, as an estimate of a population median, can tolerate up to 50% bad values. In general, breakdown cannot exceed 50%.

Linear least-squares estimates can behave badly when the error distribution is not normal, particularly when the errors are heavy-tailed. One remedy is to remove in influential observations from the least-squares fit. Another approach, termed robust regression, is to use a fitting criterion that is not as vulnerable as least squares to unusual data.

The most common general method of robust regression is M-estimation. This class of estimators can be regarded as a generalization of maximum-likelihood estimation, hence the term "M"-estimation.
We consider only the linear model

$$y_i = \propto + \beta ixi + e_i\, i = 1,2,3,\ldots.n$$

$$y_i = \propto + \beta_1x_{i1} + \beta_2x_{i2} + \ldots. + \beta_kx_{ik} + \in_i$$

$$= x'\beta + \in_i$$

for the ith of n observations.

Where $\propto$=the constant of the equation and, β_i = the coefficients of the predictor variables. Given an estimator b for β, the fitted model is

2.6 MM-estimator

MM-estimator, introduced by Yohai (1987) and Yohai et al. (1991), combines high breakdown value estimation and M estimation. It has the highest breakdown point, 0.5, and a high efficiency under normality. The first stage of the three stages procedure calculates an S-estimate with influence function.

Yohai (1987: 644) describes the three stages that define an MM-estimator:

Stage 1 A high breakdown estimator is used to find an initial estimate, which we denote $\tilde{\beta}$. The estimator need not be efficient. Using this estimate the residuals, $r_i(\tilde{\beta}) = y_i - x_i{}^T\tilde{\beta}$, are computed.

Stage2 Using these residuals from the robust fit and $\frac{1}{n}\sum_{i=1}^{n} p\left(\frac{r_i}{s}\right) = k$ *where k is constant,*
an M-estimate of scale with 50% BDP is computed. This $\left(r_1(\tilde{\beta}), \dots\dots, r_n(\tilde{\beta})\right)$ is denoted s_n. The objective function used in this stage is labelled ρ_0.

Stage 3 The MM-estimator is now defined as an M-estimator of βusing a redescending score function, $\psi_1(u) = \frac{\partial_{p1}(u)}{\partial u}$, and the scale estimate $s_n$obtained from Stage 2. So an MM-estimator $\hat{\beta}$is defined as a solution to

$$\sum_{i=1}^{n} x_{ij}\,\psi_1\left(\frac{y_i - x_i{}^T\beta}{s_n}\right) = 0 \qquad j = 1, \dots p.$$

The objective function ρ_1 associated with this score function does not have to be the same as ρ_0 but it must satisfy:

i) $\rho$is symmetric and continuously differentiable, and $\rho(0) = 0$.
ii) There exists *a* >0 such that ρ is strictly increasing on [0, *a*] and constant on [*a* ∞).
iii) $\rho_1(u) \le \rho_0(u)$

A final condition that must be satisfied by the solution to 3.12.1 is that

$$\sum_{i=1}^{n} \rho_1\left(\frac{y_i - x_i{}^T\hat{\beta}}{s_n}\right) \le \sum_{i=1}^{n} \rho_1\left(\frac{y_i - x_i{}^T\tilde{\beta}}{s_n}\right).$$

Often S-estimators are used in stage one (Andersen 2008: 57).

The MM-estimation option in the R MASS package, using the function rlm, uses an S-estimator with a Tukey bisquare objective function, having set *a* = 1.548 in the first two stages. Such an S-estimator has a BDP very close to 50%. In the final stage, the rlm function also uses the Tukey bisquare objective function with *a* = 4.685 is used to obtain 95% asymptotic relative efficiency.

The MM-estimator generally performs well except in areas with high leverage (Simpson and Montgomery, 1998b).

2.7 Inference about Robust Regression Parameters

Hypotheses

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ (All the slope parameters are equal to zero.)

H1: $\beta i \neq 0$ (At least one slope parameter is not equal to zero.)

The test statistic is: $t_{n-K-1} = \frac{\beta_K - 0}{S_{\beta_K}}$

where K = number of independent variables

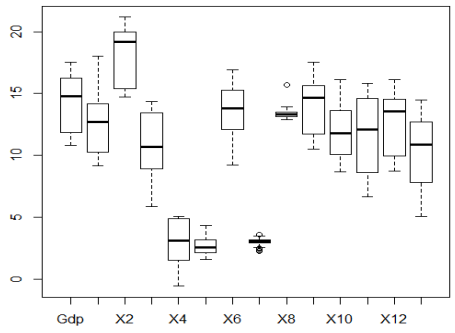
The denominator, S_{β_K} , is the standard error of the regression coefficient, β_k

Take the standard errors of the regression coefficients from the computer output

Note: *In this research work, all these will be tested with p-value at 5%*

3 RESULTS AND DISCUSSIONS

Figure 1



Boxplot figure above shows that only X10 (Net Savings) is symmetric. Our Gdp, X_1 (Public Expenditure), X_2 (Fixed Asset), X_6 (Public Debt), X_7 (Interest Rate), X_8 (Net Export), X_9 (Consumption), X_{11} (External reserve), X_{12} (Net Import) and X_{13} (Balance of Payment) are skewed to the left while X_3 (Private Investment), X_4 (Foreign Exchange) and X_5 (Inflation Rate) are skewed to the right. X_7 (Interest Rate) and X_8(Net Export) contain outliers.

3.1 Multicolinearity

Variance Inflation factor for macroeconomic aggregates with thirteen variables.

Multicollinearity.

Variables	coefficients	VIF
X_1	-.015	56.939
X_2	.246	209.267
X_3	.035	23.689
X_4	.065	78.509
X_5	.012	2.100
X_6	.012	63.299
X_7	-.009	5.748
X_8	-.001	1.830
X_9	.369	320.697
X_{10}	.108	53.985
X_{11}	.125	51.411
X_{12}	.050	160.415
X_{13}	.025	7.134

The maximum Variance Inflation Factor (*VIF)* value is 320.697 and X_1, X_2, X_3, X_4 X_6, X_9, X_{10}, X_{11}, and X_{12} *VIF* values greatly exceed 10, which indicates that serious multicollinearity problem exist among these variables. But X_5, X_7, X_8, and X_{13} *VIF* values do not show serious multicollinearity, though there is also multicollinearity among these variables.

3.2 Backward stepwise regression method of variable selection in linear regression

In this stage, we carry out the selection process using the method of backward stepwise regression.

The general equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + e_i$$

The final result that yields the best predictor variables is:

lm(formula = Gdp ~ X_2 + X_4 + X_9 + X_{10} + X_{11})

Residuals:

Min	1Q	Median	3Q	Max
-0.15281	-0.06111	-0.00523	0.04158	0.31778

Coefficients:

Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.89143	0.41792	2.133	0.042166 *
X2	0.28933	0.07371	3.926	0.000539 ***
X4	0.11434	0.03170	3.607	0.001239 **
X9	0.38724	0.09311	4.159	0.000290 ***
X10	0.12026	0.04224	2.847	0.008329 **
X11	0.07127	0.03113	2.289	0.030105 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1005 on 27 degrees of freedom

Multiple R-squared: 0.9984, Adjusted R-squared: 0.9982

F-statistic: 3459 on 5 and 27 DF, p-value: < 2.2e-16

Interpretation

As with any regression, the positive coefficients indicate a positive relationship with the dependent variables. Multiple R-squared: 0.9984, this shows that 99.8% of the changes in Gdp are explained jointly by variation of the explanatory variables (Fixed Asset, Foreign Exchange, Consumption, Net Savings and External Reserve). They are all significant and have meaningful effects on the economy performance. The f – statistics is 3459 and is significant at 5% level, indicating that the model is adequate, from the regression results obtained.

We have the final subsets *(X_2, X_4, X_9, X_{10}, X_{11},)* that satisfy the criterion for stopping during selection process: that is, all the p-values of the variables selected are less than 0.05.

The model: $Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + e$

3.3 $\mathbf{R_p^2}$ Criterion

In this case, we use R^2 to select the best model

$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$

Consideration of the subsets X_2, X_4, X_9, X_{10} *and* X_{11} in the regression model appear to be more reasonable according to the R_p^2 criterion.

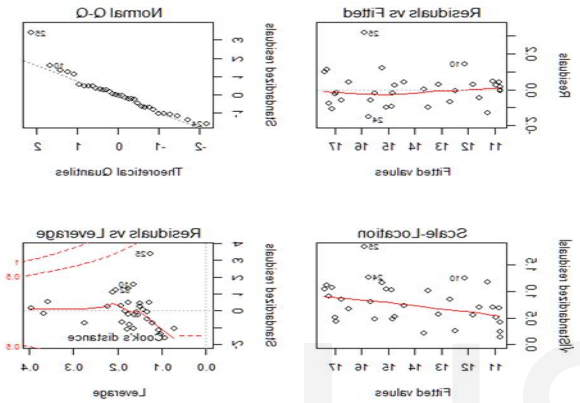
From the result of (*what*?)$\boldsymbol{R_p^2 = 0.9984}$. Then we have the best model:

$$Gdp = 0.89143 + 0.28933\,X_2 + 0.11434\,X_4 + 0.38724X_9 + 0.12026X_{10} + 0.07127X_{11}$$

3.4 Relationship between Selected Variables

	Gdp	X2	X4	X9	X10	X11
Gdp	1.0000000	0.9944099	0.9591168	0.9971924	0.9790777	0.9822965
X2		1.0000000	0.9473365	0.9932582	0.9628573	0.9725396
X4			1.0000000	0.9477366	0.9213973	0.9543428
X9				1.0000000	0.9784477	0.9737696
X10					1.0000000	0.9657851
X11						1.0000000

A very strong relationship exists among all the selected variables.
Figure 2



From these plots, we can identify observation 25 as being possibly problematic to our model.
The Cook's distance shows the impact of individual cases on the regression line.B[c(25),2:1]
GDP Year 25 16.25 2004

This shows that the year 2004 has either high leverage or large residuals.

3.5Multiple Regression Output Without observation 25.

Call:
lm(formula = Gdp ~ X2 + X4 + X9 + X10 + X11, data = B[-c(25),])

Residuals:

Min	1Q	Median	3Q	Max
-0.134795	-0.047990	0.003096	0.031323	0.175969

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.92910	0.32293	2.877	0.007914	**
X2	0.25098	0.05760	4.357	0.000183	***
X4	0.10509	0.02458	4.276	0.000227	***
X9	0.44437	0.07309	6.080	2e-06	***
X10	0.12267	0.03263	3.759	0.000873	***
X11	0.05799	0.02424	2.393	0.024241	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07762 on 26 degrees of freedom
Multiple R-squared: 0.9991, Adjusted R-squared: 0.9989
F-statistic: 5650 on 5 and 26 DF, p-value: < 2.2e-16

Since at least one regression coefficient is not equal to zero. The model has worth.

The multiple regression output when observation 25 was dropped. The effects of the outlier have been taken care of. The Multiple R-squared is now higher and residual standard error is also lower than what we had before removing the outlier.

3.6 Huber Regression Output

Call: rlm(formula = Gdp ~ X2 + X4 + X9 + X10 + X11, data = BB)

Residuals:

Min	1Q	Median	3Q	Max
-0.128862	-0.038165	0.002854	0.033827	0.366593

Coefficients:

	Value	Std. Error	t value
(Intercept)	0.7783	0.3649	2.1330
X2	0.2685	0.0644	4.1718
X4	0.0990	0.0277	3.5781
X9	0.4341	0.08135	3401
X10	0.1264	0.0369	3.4282
X11	0.0537	0.0272	1.9739

Residual standard error: 0.05374 on 27 degrees of freedom

Since at least one regression coefficient is not equal to zero. The model has worth.

We reject the null hypothesis and conclude that Fixed Asset, Foreign Exchange, Consumption and External Reserve are significantly and positively related to Real Gross Domestic Product. As we have with the multiple regression, all explanatory

variables are significant.

Year	resid	weight
25 2004	0.366592811	0.1971847
21 2000	0.180587427	0.4002956
10 1989	0.162607561	0.4445425
7 1986	-0.128862280	0.5609527
24 2003	-0.116723145	0.6192908
30 2009	-0.100844897	0.7168707
32 2011	0.100204617	0.7212670
19 1998	-0.091818286	0.7873178
33 2012	0.087416552	0.8267580
31 2010	-0.086173399	0.8389907
14 1993	-0.072725332	0.9939332
1 1980	0.030890534	1.0000000
2 1981	-0.004334863	1.0000000
3 1982	-0.021179372	1.0000000
4 1983	0.015686886	1.0000000

We can see roughly, as the absolute residual goes down, the weight goes up. In order words, cases with large residuals tend to be down-weighted. This output shows us that the observation for the year 1986 will be down-weighted the most. Observation for 2004 will also be substantially down-weighted. All observations not shown above have a weight of 1. In OLS regression, all cases have a weight of 1. Hence, the more cases in the robust regression that have a weight close to one (1), the closer the results of the OLS and robust regressions. Huber weights have difficulties with several outliers.

3.7 Tukey Biweight Regression Output

Call: rlm(formula = Gdp ~ X2 + X4 + X9 + X10 + X11, psi = psi.bisquare)

Residuals:

Min	1Q	Median	3Q	Max
-0.128141	-0.024255	0.005514	0.041738	0.392528

Coefficients:

	Value	Std. Error	t value
(Intercept)	0.7732	0.3283	2.3550
X2	0.2462	0.0579	4.2523
X4	0.0913	0.0249	3.6665
X9	0.4701	0.0731	6.4272
X10	0.1260	0.0332	3.7977
X11	0.0473	0.0245	1.9351

Residual standard error: 0.05937 on 27 degrees of freedom

Since at least one regression coefficient is not equal to zero. The model has worth.

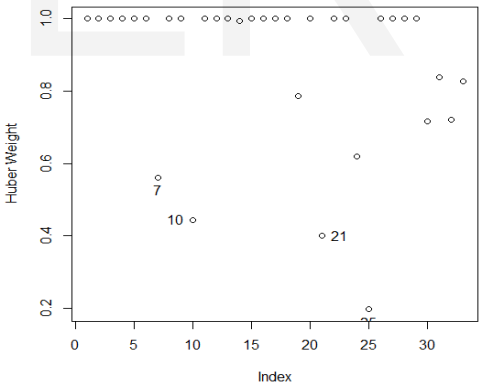
We reject the null hypothesis and conclude that Fixed Asset, Foreign Exchange, Consumption and External Reserve are significantly and positively related to Real Gross Domestic Product. As we have with the multiple regression, all explanatory variables are significant. The large differences in the intercept of our multiple regressions and the Tukey biweight regression suggest that the model parameters are being influenced by outliers.The standard errors are also lower than what we obtained with OLS. Compared to the Huber estimates, the (Tukey) biweight estimate of Consumption is larger.

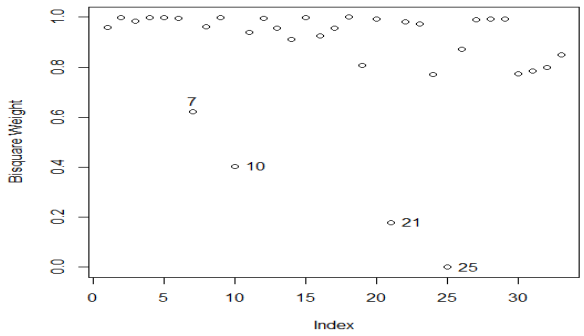
	Year	resid	weight
25	2004	0.39252819	0.0000000
21	2000	0.21127491	0.1789580
10	1989	0.16794735	0.4036986
7	1986	-0.12814095	0.6205402
24	2003	-0.09726981	0.7703221
30	2009	-0.09675716	0.7726238
31	2010	-0.09389097	0.7851176
32	2011	0.09052999	0.7992827
19	1998	-0.08844069	0.8080077
33	2012	0.07801955	0.8487713
26	2005	0.07167701	0.8715799
14	1993	-0.05942490	0.9107766
16	1995	0.054266770	0.9253244
11	1990	0.04901711	0.9388590
17	1996	0.04173811	0.9554705

We can see that the weight given to the year 2004 is dramatically lower using the Tukey (bisquare) weighting function than the Huber weighting function and the parameter estimates from these two different weighting methods differ. However, bisquare weights have difficulties converging or may yield multiple solutions.

3.8 Graphical Display of Weights

The plots show the weight given to each case. Observations 7, 10, 21, and 25 received the least weight of all the observations.





3.9 MM-Estimators

Call: rlm(formula = Gdp ~ X2 + X4 + X9 + X10 + X11, data = B, method = "MM")
Residuals:

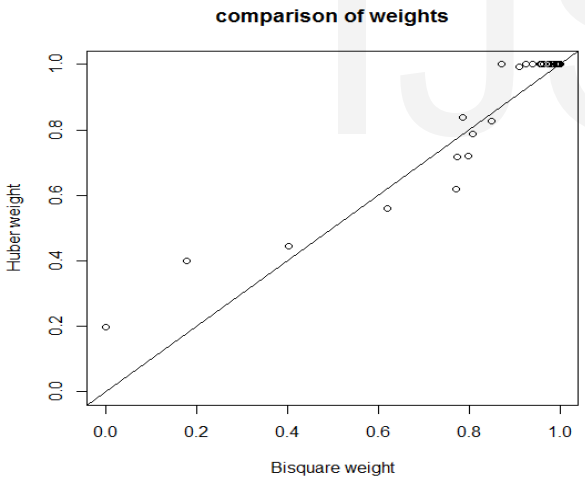
Min	1Q	Median	3Q	Max
-0.131151	-0.029757	0.004867	0.038803	0.382408

Coefficients:

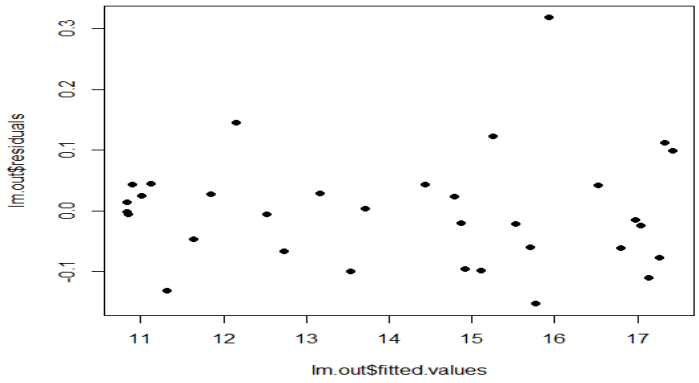
	Value	Std. Error	t value
(Intercept)	0.8321	0.3267	2.5472
X2	0.2486	0.0576	4.3147
X4	0.0966	0.0248	3.8991
X9	0.4595	0.0728	6.3139
X10	0.1252	0.0330	3.7908
X11	0.0511	0.0243	2.0991

Residual standard error: 0.07071 on 27 degrees of freedom
Since at least one regression coefficient is not equal to zero. The model has worth.

We reject the null hypothesis and conclude that Fixed Asset, Foreign Exchange, Consumption and External Reserve are significantly and positively related to Gross Domestic Product. As we have with the multiple regression, all explanatory variables are significant.



We notice that, although the relative weights are similar, the Huber method gives weight of 1 to far more cases than does the bisquare weight.



The graph above is showing the fitted values of our data set against their residuals

3.10 Comparing the models

Regression coeffs.	MULTIPLEREG (all)	MULTIPLEREG (~~25~~)	M-Est. (Biweight)	M-Est. (Huber)	MM-Est.
INTERCEPT	0.89143	0.9291	0.7732	0.7783	0.8321
X2	0.28933	0.25098	0.2462	0.2685	0.2486
X4	0.11434	0.10509	0.0913	0.099	0.0966
X9	0.38724	0.44437	0.4701	0.4341	0.4595
X10	0.12026	0.12267	0.126	0.1264	0.1252
X11	0.07127	0.05799	0.0473	0.0537	0.0511
SE	MULTIPLEREG (all)	MULTIPLEREG (~~25~~)	M-Est. (Biweight)	M-Est. (Huber)	MM-Est.
INTERCEPT	0.41792	0.32293	0.3283	0.3649	0.3267
X2	0.07371	0.0576	0.0579	0.0644	0.0576
X4	0.0317	0.02458	0.0249	0.0277	0.0248
X9	0.09311	0.07309	0.0731	0.0813	0.0728
X10	0.04224	0.03263	0.0332	0.0369	0.033
X11	0.03113	0.02424	0.0245	0.0272	0.0243

The Huber M-Estimation technique gives similar estimates to the multiple regression with the three influential cases deleted. Biweight M-estimates and MM-estimates were slightly less affected by the influential cases than the Huber estimates. We then use the MM-estimates because it is the best.

$$Gdp = 0.8321 + 0.2486\,X_2 + 0.0966X_4 + 0.4595X_9 + 0.1252X_{10} + 0.0511X_{11}$$

4 CONCLUSION

The method of backward stepwise variable selection process was used and 8192 possible regression models were generated. In the selection, the variable with the highest p-value was dropped until all the p-values were less than 0.05. The best subset variables that met the stopping criterion were achieved at the ninth selection. The variables are: Fixed Asset (X_2), Foreign Exchange(X_4), Consumption (X_9), Net Savings (X_{10}) and External Reserve (X_{11}). We applied R_p^2 Criterion to select the best regression model: $Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + e$ was selected as the best regression model.

Robust regression was carried out because of the violation of the assumptions of Ordinary Least Squares (OLS) model especially, the presence of outliers in our data set. The methods used assigned different weights to our data set to reduce the influence of these outliers that were detected. It was observed that observation 25 which was an influential outlier received the least weights of all the observations. From our analyses; Fixed Asset, Foreign Exchange, Consumption, Net Savings and External Reserve play significant roles in the growth of the economy (Gdp). Furthermore, the variable selection method should be adoptedwhen large numbers of explanatory variables are to be considered. The robust regression model would perform better when outliers are bound to occur. Finally, for proper economic growth, careful attention should be paid to Fixed Asset, Foreign Exchange, Consumption, Net Savings, and External Reserve.

REFERENCES

[1] Acha. J "The Role of oil in Nigerian Economy". http://Ezine Articles. Com/? Expert= AchaJoy.

[2] Adedipe. B. (2004), "The Impact of Oil on Nigeria's Economic Policy Formulation".

[3] Akanni, O.P (2004) "Oil Wealth and Economic Growth in Oil Exporting African countries". AERC Research paper 170.

[4] Ali, A., Qadir, M. F. (2005). A Modified M-estimator for the detection of outliers. *Pakistan journal of Statistics and Operations Research*, 1, 49-64.

[5] Benjamini Y. and Gavrilov Y. A simple forward selection procedure based on false discovery rate control. eprint arXiv: 0905.2819, 2009.

[6] Breiman L. and D. Freedman. How many variables should be entered in a regression equation?

[7] Breiman L. and Spector P.; Sub-model selection and evaluation in regression. The X random than n. Annals of Statistics, 35(6):2313–2351, 2007.

[8] Bill Jacoby, "Regression III: Advanced Methods", *Michigan State University:* http://polisci.msu.edu/icpsr/regress3

[9] Bullion Publication of CBN, vol.32, No.2, April-June, 2008.

[10] Central Bank of Nigeria Statistical Bulletin, December 2012.

[11] Chukwu and et al (2010) "Oil Price Distortions and their short and Long-Run impacts on the Nigerian Economy". MPRA paper No. 24434, August.

[12] Draper, N. R. and Smith, H. (1966). Applied Regression Analysis, John Wiley & Sons, New York, 342-344.

[13] Ehanmo.J. N. (2002), "Is sustainable Development Compatible with Economic Growth in An Oil Dominated developing economy?" A case study of Nigeria.

[14] Genova, A; Toyin Falola (2003) Oil in Nigeria: A Bibliographical Reconnaissance History in Africa, Vol. 30, pp. 133-156.

[15] Huber, P. (1981). Robust Statistics. John Wiley & Sons: New York, 153-199.

[16] Jaidah A.M. (1982), "Perspective on the Oil Market", OPEC Review (autumn) 254-260.

[17] Longman Dictionary of Contemporary English, 5th Edition.

[18] Jureckova, J. (1980). Asymptotic representation of M-estimators of location. Math. Operationsforsch. Und Statistic, *Ser. Statistics*, 11:1, 61-73.

[19] Jureckova J. and Picek, J. (2006) Robust Statistical Methods with R, Chapman and Hall.

[20] Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006) Robust Statistics: Theory and Methods, Wiley.

[21] Odularu, .G.O (2008), "Crude oil and Nigerian Economic Performance".

[22] Odularu, .G.O., C. Okonkwo, (2009), "Does Energy Consumption Contribute to the Economic performance?" Empirical Evidence from Nigeria.

[23] Rousseeuw, P.J. and Leroy, A.M. (1987). Robust Regression and Outlier Detection. New York: Wiley.

[24] Wilcox, R. (1997). Introduction to the Robust Estimation and Hypothesis Testing. Academic Press: San Diego, 14.